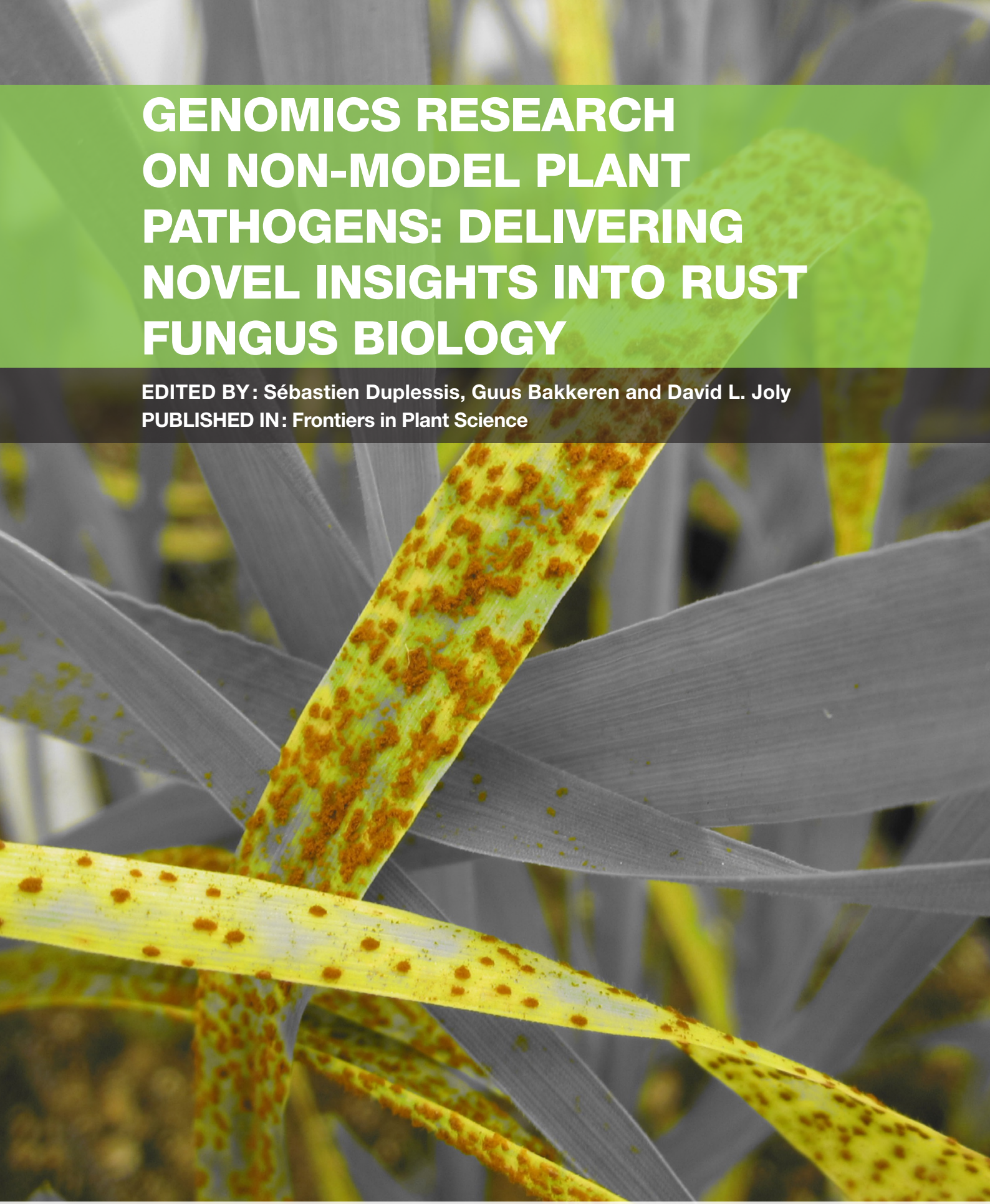
GENOMICS RESEARCH ON NON-MODEL PLANT PATHOGENS: DELIVERING NOVEL INSIGHTS INTO RUST FUNGUS BIOLOGY

EDITED BY: Sébastien Duplessis, Guus Bakkeren and David L. Joly

PUBLISHED IN: Frontiers in Plant Science



frontiers Research Topics

Frontiers Copyright Statement

© Copyright 2007-2016 Frontiers Media SA. All rights reserved.

All content included on this site, such as text, graphics, logos, button icons, images, video/audio clips, downloads, data compilations and software, is the property of or is licensed to Frontiers Media SA ("Frontiers") or its licensees and/or subcontractors. The copyright in the text of individual articles is the property of their respective authors, subject to a license granted to Frontiers.

The compilation of articles constituting this e-book, wherever published, as well as the compilation of all other content on this site, is the exclusive property of Frontiers. For the conditions for downloading and copying of e-books from Frontiers' website, please see the Terms for Website Use. If purchasing Frontiers e-books from other websites or sources, the conditions of the website concerned apply.

Images and graphics not forming part of user-contributed materials may not be downloaded or copied without permission.

Individual articles may be downloaded and reproduced in accordance with the principles of the CC-BY licence subject to any copyright or other notices. They may not be re-sold as an e-book.

As author or other contributor you grant a CC-BY licence to others to reproduce your articles, including any graphics and third-party materials supplied by you, in accordance with the Conditions for Website Use and subject to any copyright notices which you include in connection with your articles and materials.

All copyright, and all rights therein, are protected by national and international copyright laws.

The above represents a summary only. For the full conditions see the Conditions for Authors and the Conditions for Website Use.

ISSN 1664-8714
ISBN 978-2-88919-814-6
DOI 10.3389/978-2-88919-814-6

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view.
By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: **researchtopics@frontiersin.org**

GENOMICS RESEARCH ON NON-MODEL PLANT PATHOGENS: DELIVERING NOVEL INSIGHTS INTO RUST FUNGUS BIOLOGY

Topic Editors:
Sébastien Duplessis, INRA, France
Guus Bakkeren, Agriculture & Agri-Food Canada, Canada
David L. Joly, Université de Moncton, Canada

Uredinia of *Puccinia triticina* (causal agent of wheat leaf rust), 10 days after inoculation on *Triticum aestivum* "Thatcher".
Image by David L. Joly.

Fungi of the order Pucciniales cause rust diseases on many plants including important crops and trees widely used in agriculture, forestry and bioenergy programs; these encompass gymnosperms and angiosperms, monocots and dicots, perennial and annual plant species. These fungi are obligate biotrophs and -except for a few cases- cannot be cultivated outside their hosts in a laboratory. For this reason, standard functional and molecular genetic approaches to study these pathogens are very challenging and the means to study their biology, i.e. how they infect, develop and reproduce on plant hosts, are rather limited, even though they rank among the most devastating pathogens. Among fungal plant pathogens, rust fungi display the most complex lifecycles with up to five different spore forms and for many rust fungi, unrelated alternate hosts on which sexual and clonal reproduction

are achieved. The genomics revolution and particularly the application of new generation sequencing technologies have greatly changed the way we now address biological studies and has in particular accelerated and made feasible, molecular studies on non-model species, such as rust fungi.

The goal of this research topic is to gather articles that present recent advances in the understanding of rust fungi biology, their complex lifecycles and obligate biotrophic interactions with their hosts, through the means of genomics. This includes genome sequencing and/or resequencing of isolates, RNA-Seq or large-scale transcriptome analyses, genome-scale detailed annotation of gene families, and comparative analyses among the various rust fungi and, where feasible, with other obligate biotrophs or fungi displaying distinct trophic modes.

This Research Topic provides a great opportunity to provide an up-to-date account of rust fungus biology through the lens of genomics, including state-of-the-art technologies developed to achieve this knowledge.

Citation: Duplessis, S., Bakkeren, G., Joly, D. L., eds. (2016). Genomics Research on Non-model Plant Pathogens: Delivering Novel Insights into Rust Fungus Biology. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-814-6

Table of Contents

frontiers in Plant Science

EDITORIAL
published: 23 February 2016
doi: 10.3389/fpls.2016.00216

Editorial: Genomics Research on Non-model Plant Pathogens: Delivering Novel Insights into Rust Fungus Biology

Guus Bakkeren[1]*, *David. L. Joly*[2]* and *Sébastien Duplessis*[3,4]*

[1] Agriculture and Agri-Food Canada, Summerland Research and Development Centre, Summerland, BC, Canada, [2] Department of Biology, Université de Moncton, Moncton, NB, Canada, [3] INRA, UMR 1136 Interactions Arbres/Microorganismes INRA/Université de Lorraine, Centre INRA Nancy Lorraine, Champenoux, France, [4] Faculté des Sciences et Technologies, Université de Lorraine, UMR 1136 Interactions Arbres/Microorganismes Université de Lorraine/INRA, Vandoeuvre-lès-Nancy, France

Keywords: rust fungi, basidiomycota, obligate biotrophy, genomics, resequencing, genetic variation, transcriptomics, effectors

OPEN ACCESS

Edited and reviewed by:
Joshua L. Heazlewood, The University of Melbourne, Australia

Reviewed by:
Peter Dodds, Commonwealth Scientific and Industrial Research Organisation, Australia

***Correspondence:**
Guus Bakkeren
guus.bakkeren@agr.gc.ca;
David L. Joly
david.joly@umoncton.ca;
Sébastien Duplessis
duplessi@nancy.inra.fr

Specialty section:
This article was submitted to Plant Biotic Interactions, a section of the journal Frontiers in Plant Science

Received: *08 December 2015*
Accepted: *08 February 2016*
Published: *23 February 2016*

Citation:
Bakkeren G, Joly DL and Duplessis S (2016) Editorial: Genomics Research on Non-model Plant Pathogens: Delivering Novel Insights into Rust Fungus Biology. Front. Plant Sci. 7:216. doi: 10.3389/fpls.2016.00216

The Editorial on the Research Topic

Genomics Research on Non-model Plant Pathogens: Delivering Novel Insights into Rust Fungus Biology

The diversity among rust fungi is simply astounding: over 7000 species that evolved to colonize niches all over the plant kingdom. This diversification likely involved major host jumps, especially considering that the life cycles of heteroecious rusts, such as the cereal rusts, involve sexual and asexual stages that take place on completely unrelated host plants (Savile, 1976; McTaggart et al., 2015), but also co-evolution in diverse natural settings, establishing equilibrium (e.g., Thrall et al., 2012). These interactions resulted in complex life styles including for many rusts, the production of up to five different spore types, each requiring very different developmental programs and hence gene expression.

Unfortunately, despite the intriguing biology of these fascinating pathogens, many rust fungi have gained notoriety because of the fact that some of their hosts were selected as food sources by humans leading to their extensive cultivation, and more recently their monocultures over large areas. Upsetting the balance, the rust fungi took the occasion, having a ball.

Because of their importance, plenty of research has been done on rust fungi describing life cycles, host range, and infection processes. Driven by the need for resistant crop cultivars, the genetics of race-cultivar interactions and fungal race identification pioneered in the 1940's, became the staple for breeding programs worldwide. However, because of their strict biotrophic life styles and recalcitrance to genetic transformation, molecular genetic research on rust fungi remains difficult. The advance of genomics has really impacted research on rust fungi, demonstrated by the expansion of labs embarking on and receiving funding for molecular work over the last 5 years. With this in mind, we invited submissions for this *Frontiers in Plant Science Research Topic* and present here 14 papers.

Recent studies have revealed that genome sizes vary widely among the rust fungi but are on average much larger than other basidiomycete genomes: 89 Mbp for *Puccinia graminis* f. sp. *tritici* and 101 Mbp for *Melampsora larici-populina* and harboring at least 45% repeat and transposable elements (Duplessis et al., 2011). Here we present papers indicating that this may generally be true. A high quality assembly of the *Melampsora lini* genome indicated a genome size of 189

Mbp with at least 45% harboring repetitive elements, primarily retrotransposons (Nemri et al.). A larger size is expected from the initial assembly of a draft genome of *Uromyces fabae* which here is estimated at 329 Mbp (Link et al.). Several groups have attempted to sequence the genome of the destructive Asian soybean pathogen *Phakopsora pachyrhizi*. It became clear early on that its genome was massive, possibly above 850 Mbp but here, a new report estimated it to be in the 500 Mbp-range (Loehrer et al.). To better prepare for genome sequencing projects, genome size estimates thus become important. Here, Tavares et al. employ nuclear fluorescence and flow cytometry to present rough estimates of the genome sizes of a wide variety of rust fungi, contrasting the sizes of the previously reported genomes. It is widely believed that active transposable elements contribute to and are responsible for these expanded, over-sized genomes since the gene space, and number of gene models predicted (and confirmed by RNAseq data in many cases) is roughly similar for all of them (Spanu, 2012). Indeed, it is believed that polyploidy and/or the activity of transposable elements could be important diversity-creating factors in such mostly asexual species, being the main mechanisms driving genome expansions (Ramos et al., 2015). The coffee rust fungus *Hemileia vastatrix* is among those exhibiting a particularly large genome and a striking richness in repeated elements, making it difficult to assemble genome reads. A draft genome of 333 Mb was generated through a hybrid assembly of eight different isolates and helped in the identification of more than 14,000 putative genes (Cristancho et al.). Performing genetic studies in the laboratory with heteroecious rust fungi is complex as it requires two hosts to generate the various spore stages for crossing. A major breakthrough is reported here on a successful self-cross of the reference genome *M. larici-populina* isolate and the benefit of resequencing offspring in order to identify recombination breaks in its genome (Pernaci et al.). This also allowed corrections to the initial (dikaryotic) genome assembly, generating a framework for the construction of a future genetic map.

Two transcriptomic studies of specific stages of plant-rust interactions (early infection and telia development) illustrate the importance of expression profiling in gaining an understanding of rust biology. Whereas most recent rust transcriptomic studies focus on time-course infections of the telial host or on isolated haustoria collected from such infected host tissues (see Duplessis et al., 2014), other stages have been less investigated. Here transcript profiles of the coffee rust fungus at early stages of the infection (i.e., germinating urediniospores and appressoria) were compared to previously reported expression profiles *in planta* highlighting candidate effector genes expressed only during biotrophic growth in the plant and also signaling pathways and metabolic shifts that most likely condition the success of infection (Talhinhas et al.). Teliospores are survival structures produced on decaying telial host tissues and molecular mechanisms responsible for this process are mostly unknown. Here, Hacquard et al. report on transcriptome profiling in early *M. larici-populina* telia sampled in autumn, revealing adaptation to the drastic change in environmental conditions and the tight coordination of karyogamy and meiotic processes. Comparing these profiles with poplar infection identified genes encoding small secreted proteins that most likely do not play a role in the biotrophic phase.

Available genome sequences have accelerated the discovery of candidate effectors, which can support the design of disease management and resistance breeding strategies (Vleeshouwers and Oliver, 2014). Because of their importance, a mini review on effector proteins of rust fungi has been included in this issue (Petre et al.). Predicting candidate effectors in fungal pathogen genomes has been difficult due to a lack of signature sequence motifs, resulting in too many candidates for functional validation. Here Nemri et al. established a bioinformatic pipeline that prioritizes effector candidates by integrating multiple lines of evidence, taking advantage of the knowledge acquired from flax rust Avr proteins and other known rust effectors. Pendleton et al. investigated gene gain and loss in rust fungi and found gene family losses and contractions, and numerous lineage-specific duplications, probably accounting for their large proteome size. As described above, their highly plastic genomes probably played a major role in this process. As found in *M. lini* (Nemri et al.), such lineage-specific families could be enriched in Avr proteins targeted by host intracellular immune receptors (i.e., R proteins). Working on *P. graminis* f. sp. *tritici,* Sperschneider et al. employed another similar pipeline based on taxonomic information, expression data and evolutionary signatures of diversifying selection. Among the 42 effector candidates identified as part of pathogen associated gene families up-regulated during infection and rapidly evolving, was one that was previously shown to trigger genotype specific defense responses in wheat (Upadhyaya et al., 2014). Surprisingly, most of those candidates lacked features frequently associated with fungal effectors such as small size and high cysteine content, reinforcing the need for unbiased predictors.

A lack of identifiable homologs or clustering in families could result in unidentified effector candidates. Comparative genomic and transcriptomic approaches can circumvent this bias by correlating mutations within candidate effector genes with virulence phenotypes. By sequencing the transcriptome of six wheat leaf rust races and focusing on genes encoding secreted proteins during infection, Bruce et al. found amino acid changes in 15 *P. triticina* effector genes that correlated with virulence shifts. Similarly, in *M. larici-populina*, correlations between non-synonymous mutations and avirulence phenotypes were uncovered using genomic resequencing data from multiple isolates (Persoons et al.). A similar study in *P. graminis* f. sp. *tritici* (Upadhyaya et al.) compared amino acid changes between a given isolate and field-derived mutants that have gained virulence toward certain *R* genes. Because of the close genetic relationship between these isolates, a relatively low number of SNPs were identified, reducing the number of candidate genes requiring experimental validation as potential molecular determinants of the observed recognition specificities.

Rust fungi are among the most devastating plant pathogens and a serious threat to major domesticated crops (Dean et al., 2012). Though their obligate biotrophic growth habit and complex lifestyles make these fungi difficult to study with most standard genetic and molecular tools and techniques, we show here in this research topic that the genomics revolution that we

are currently experiencing leads to new opportunities to dissect the biology of these fascinating fungi. These studies represent a few examples illustrating that these novel genomic approaches can deliver on promises made, including anticipated progress toward crop protection.

AUTHOR CONTRIBUTIONS

All authors listed, have made substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Dean, R., Van Kan, J. A. L., Pretorius, Z. A., Hammond-Kosack, K. E., Di Pietro, A., Spanu, P. D., et al. (2012). The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.* 13, 414–430. doi: 10.1111/j.1364-3703.2011.00783.x

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). "Advancing knowledge on biology of rust fungi through genomics," in *Advances Botanical Research,* ed F. M. Francis (London: Academic Press), 173–209. doi: 10.1016/b978-0-12-397940-7.00006-9

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

McTaggart, A. R., Shivas, R. G., van der Nest, M. A., Roux, J., Wingfield, B. D., and Wingfield, M. J. (2015). Host jumps shaped the diversity of extant rust fungi (Pucciniales). *New Phytol.* 209, 1149–1158. doi: 10.1111/nph.13686

Ramos, A. P., Tavares, S., Tavares, D., Do Céu Silva, M., Loureiro, J., and Talhinhas, P. (2015). Flow cytometry reveals that the rust fungus *Uromyces bidentis* (Pucciniales) possesses the largest fungal genome reported, 2,489 Mbp. *Mol. Plant Pathol.* 16, 1006–1010. doi: 10.1111/mpp.12255

Savile, D. B. O. (1976). "Evolution of the rust fungi (Uredinales) as reflected by their ecological problems," in *Evolutionary Biology,* Vol. 9, eds M. K. Hecht, W. C. Steere, and B. Wallace (New York, NY: Plenum), 137–207. doi: 10.1007/978-1-4615-6950-3_4

Spanu, P. D. (2012). The genomics of obligate (and nonobligate) biotrophs. *Annu. Rev. Phytopathol.* 50, 91–109. doi: 10.1146/annurev-phyto-081211-173024

Thrall, P. H., Laine, A. L., Ravensdale, M., Nemri, A., Dodds, P. N., Barrett, L. G., et al. (2012). Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. *Ecol. Lett.* 15, 425–435. doi: 10.1111/j.1461-0248.2012.01749.x

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M. A., Ellis, J. G., and Dodds, P. N. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol. Plant Microbe Interact.* 27, 255–264. doi: 10.1094/MPMI-07-13-0187-FI

Vleeshouwers, V. G. A., and Oliver, R. P. (2014). Effectors as tools in disease resistance breeding against biotrophic, hemi-biotrophic and necrotrophic plant pathogens. *Mol. Plant Microbe Interact.* 27, 196–206. doi: 10.1094/MPMI-10-13-0313-IA

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Bakkeren, Joly and Duplessis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 24 March 2014
doi: 10.3389/fpls.2014.00098

The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*

Adnane Nemri[1], Diane G. O. Saunders[2], Claire Anderson[3], Narayana M. Upadhyaya[1], Joe Win[2], Gregory J. Lawrence[1], David A. Jones[3], Sophien Kamoun[2], Jeffrey G. Ellis[1] and Peter N. Dodds[1]*

[1] CSIRO Plant Industry, Canberra, ACT, Australia
[2] The Sainsbury Laboratory, Norwich Research Park, Norwich, UK
[3] Research School of Biological Sciences, College of Medicine, Biology and Environment, Australian National University, Canberra, ACT, Australia

Edited by:
Sébastien Duplessis, INRA, France

Reviewed by:
Pietro Daniele Spanu, Imperial College London, UK
David L. Joly, Université de Moncton, Canada

***Correspondence:**
Peter N. Dodds, CSIRO Plant Industry, GPO Box 1600, Clunies Ross Street, Canberra, ACT 2601, Australia
e-mail: peter.dodds@csiro.au

Rust fungi cause serious yield reductions on crops, including wheat, barley, soybean, coffee, and represent real threats to global food security. Of these fungi, the flax rust pathogen *Melampsora lini* has been developed most extensively over the past 80 years as a model to understand the molecular mechanisms that underpin pathogenesis. During infection, *M. lini* secretes virulence effectors to promote disease. The number of these effectors, their function and their degree of conservation across rust fungal species is unknown. To assess this, we sequenced and assembled *de novo* the genome of *M. lini* isolate CH5 into 21,130 scaffolds spanning 189 Mbp (scaffold N50 of 31 kbp). Global analysis of the DNA sequence revealed that repetitive elements, primarily retrotransposons, make up at least 45% of the genome. Using *ab initio* predictions, transcriptome data and homology searches, we identified 16,271 putative protein-coding genes. An analysis pipeline was then implemented to predict the effector complement of *M. lini* and compare it to that of the poplar rust, wheat stem rust and wheat stripe rust pathogens to identify conserved and species-specific effector candidates. Previous knowledge of four cloned *M. lini* avirulence effector proteins and two basidiomycete effectors was used to optimize parameters of the effector prediction pipeline. Markov clustering based on sequence similarity was performed to group effector candidates from all four rust pathogens. Clusters containing at least one member from *M. lini* were further analyzed and prioritized based on features including expression in isolated haustoria and infected leaf tissue and conservation across rust species. Herein, we describe 200 of 940 clusters that ranked highest on our priority list, representing 725 flax rust candidate effectors. Our findings on this important model rust species provide insight into how effectors of rust fungi are conserved across species and how they may act to promote infection on their hosts.

Keywords: rust, flax, *Melampsora*, effector, virulence, avirulence

INTRODUCTION

Ever since the work of Flor (1955), the interaction between the flax plant (linseed; *Linum usitatissimum*) and the flax rust fungus *Melampsora lini* (*Mli*) has served as a model pathosystem to study the genetics underlying host-pathogen interactions in plants. Flor's work led to the formulation of the gene-for-gene model describing the interaction between host resistance (R) genes and pathogen avirulence (Avr) genes (Flor, 1955). This gene-for-gene relationship was later found to apply to most interactions between plants and their adapted pathogens and pests, both in natural, or agricultural systems (Jones and Dangl, 2006). In the 1990s, rust resistance genes from flax were among the first R genes to be cloned in plants, followed by the identification of flax rust Avr genes as encoding secreted proteins that activate R gene-encoded intracellular immune receptors containing Toll-like nucleotide-binding leucine-rich repeats (TIR-NB-LRR) domains (reviewed in Ellis et al., 2007). To date the flax rust pathosystem continues to play an important role in the genetic dissection of plant-pathogen interactions (Lawrence et al., 2007; Ravensdale et al., 2010). For example, the feasibility of genetic transformation and artificial gene silencing in flax rust (Lawrence et al., 2010) makes it an important pathosystem for the study of virulence in biotrophic fungi. However, to date genomic resources have been lacking. In this study, we remedy this by describing the genome sequence of the flax rust fungus.

Rust fungi (Phylum Basidiomycota, order Pucciniales) constitute the largest group of fungal pathogens of plants, with more than 7000 species described (Cummins and Hiratsuka, 2003). They are responsible for considerable yield losses in cultivated crops such as wheat or barley, with wheat stem rust alone having the potential to cause global losses of up to USD 54 billion per annum (Pardey et al., 2013). They can also significantly impact biodiversity, e.g., myrtle rust (*Uredo rangelii*) is a recently introduced rust species in Australia currently spreading on Myrtaceae on a continental scale threatening many native tree species and ecosystems (Australian Nursery and Garden Industry, 2012).

Consequently, understanding mechanisms of virulence in rust fungi and devising innovative ways to protect crops against them is essential.

Rust fungi are obligate biotrophs, that is they need one or more living hosts to grow and complete their complex reproductive cycle. During infection they form haustoria, specialized structures surrounded by invaginated host cell membrane, with a role in nutrient uptake from the host and delivery of secreted effector proteins into host cells (Mendgen and Nass, 1988; Kemen et al., 2005; Rafiqi et al., 2010). These effectors are proposed to promote pathogen reproductive fitness by mediating the suppression of host immunity, creating a suitable environment for the pathogen and promoting nutrient uptake (see review by Duplessis et al., 2011c). For example, the RTP1p effector, originally identified in the bean rust pathogen, is delivered into host cells during infection and may act as an inhibitor of host proteases to promote disease (Kemen et al., 2005; Pretsch et al., 2013). Importantly, a subset of these effectors elicit host resistance, including four Avr genes cloned from *M. lini* (*AvrL567*, *AvrM*, *AvrP123*, and *AvrP4*; Dodds et al., 2004; Catanzariti et al., 2006; Barrett et al., 2009), and one recently identified Avr candidate from the wheat stem rust fungus (Upadhyaya et al., 2013). Their function in pathogenicity is unknown. A further eight Avr loci from *M. lini* have been genetically characterized, as well as one inhibitor gene that specifically suppresses host resistance against normally avirulent isolates (Lawrence et al., 1981; Jones, 1988; Lawrence, 1995). Consistent with their role in mediating adaptation of rust fungi to their hosts, rust genes encoding effectors can exhibit high levels of polymorphism and signatures of positive selection (Dodds et al., 2006; Van Der Merwe et al., 2009; Joly et al., 2010). Thus, identifying the effectors possessed by rust fungi to infer their function and the evolutionary processes acting on them is key to understanding mechanisms of pathogenicity in rust fungi and the evolution of their often narrow range of host species. The elucidation and comparisons of the genome sequences of rust pathogen species is an important step toward achieving this goal (McDowell, 2011).

Over the past few years, a number of rust fungi genomes have been sequenced, including the poplar leaf rust pathogen *Melampsora larici-populina* (*Mlp*, ∼101 Mbp), a close relative of *Mli*, as well as the wheat and barley stem rust pathogen *Puccinia graminis* f.sp. *tritici* (*Pgt*, ∼88 Mbp; Duplessis et al., 2011a) and the wheat yellow (stripe) rust pathogen *P. striiformis* f.sp. *tritici* (*Pst,* between 65 and 130 Mbp; Cantu et al., 2011; Zheng et al., 2013). In comparison, the dikaryotic genome of *M. lini* uredospores (2n = 36; Boehm and Bushnell, 1992) was estimated from nuclear fluorescence studies to be ∼2.5 times larger than that of *Pgt*, giving a predicted size of ∼220 Mbp (Eilam et al., 1992). In addition, transcriptome analyses have identified rust fungi genes expressed during infection, including effectors, in flax rust (Catanzariti et al., 2006), poplar rust (Duplessis et al., 2011b; Hacquard et al., 2012), wheat stripe rust (Yin et al., 2009; Cantu et al., 2013; Garnica et al., 2013), faba bean, common bean and soybean rusts (*Uromyces viciae-fabae*, *U. appendiculatus* and *Phakopsora pachyrhizi*, respectively; Jakupović et al., 2006; Link and Voegele, 2008; Link et al., 2013) and coffee rust pathogens (Fernandez et al., 2011). The availability of genome and transcriptome sequences from multiple rust species and isolates allows interspecies comparisons to identify shared rust fungal effectors and determinants of host specificity both among and within species (Duplessis et al., 2011a; Saunders et al., 2012; Cantu et al., 2013).

The total number of effectors in the flax rust fungus and how many are unique to this species is unknown. To gain insights, we sequenced and annotated the genome of *Mli* isolate CH5. Taking advantage of previous knowledge of flax rust avirulence genes, we then characterized its predicted effector complement in relation to those of three other rust species. The availability of a sequenced genome and a compilation of candidate effectors from the flax rust fungus together with the available genetic tools, will help in future studies to identify determinants of host specificity in the flax-flax rust interaction as well as better understanding the mechanisms of rust pathogen infection.

RESULTS

DE NOVO GENOME ASSEMBLY AND ANNOTATION

We selected the flax rust pathogen isolate CH5, the F1 parent of a well-characterized F2 family segregating for 10 Avr and one inhibitor loci (Lawrence et al., 1981) to build the *Mli* reference genome sequence. Illumina sequencing data were obtained using paired-end and mate-paired libraries of four sizes (∼300, 2000, 3000, and 5000 bp; **Table S1**). The genome assembly and initial scaffolding of contigs were performed using *SOAPdenovo* with a k-mer value of 41, followed by multiple rounds of gap-closing and scaffolding. The final 189 Mbp genome assembly (including 14.1% of *N*'s) consisted of 21,310 scaffolds and represented 86.4% of the predicted 220 Mbp genome size (**Table 1**). A *de novo* search

Table 1 | Summary statistics of assembly and annotation of the genome of flax rust pathogen *Melampsora lini* isolate CH5.

Cumulative size of scaffolds	189.5 Mbp (86.4% of expected size)
No. scaffolds	21,310
Fraction of *N*'s in assembly	14.1%
Longest scaffold	239.7 kbp
N50 scaffold length	31.5 kbp
L50 scaffold count	1799
GC content	41%
Gene space completeness (CEGMA)	95%
Protein-coding genes	16,271
Mean scaffold size	8.9 kbp
Median scaffold size	1.1 kbp
No. scaffolds >1 Mbp	0
No. scaffolds >100 kbp	81 (0.4% of scaffolds)
No. scaffolds >10 kbp	5339 (25.1% of scaffolds)
No. scaffolds >1 kbp	10,798 (50.7% of scaffolds)

Only scaffolds larger than 200 bp were retained in the final assembly. The N50 of scaffold length indicates that 50% of the total assembled sequence is on scaffolds larger than that size. The L50 scaffold count indicates the number of scaffolds larger than the N50 length. Gene space completeness indicates the fraction of 248 conserved eukaryotic genes (CEGMA) present with >70% length in the assembly.

for repetitive elements identified ∼45% of the genome sequence as interspersed repeats (**Table 2**).

To assess the gene space coverage in the genome assembly we used three different sources of evidence. First, an analysis searching for the CEGMA set of 248 conserved eukaryotic genes (Parra et al., 2009) in the assembly found 95% of them present "in full," indicating a high level of completeness for the genome assembly. In a second test, we used EST sequences from an haustorial cDNA library from Catanzariti et al. (2006). In addition to the 856 ESTs previously described we sequenced an additional 1937 cDNA clones. After filtering out ESTs coming from flax, flax rust ribosomal RNA or retrotransposons, of the 1399 remaining ESTs, only 3 (0.2%) did not match the assembled *Mli* genome but did match genes of other fungal species including *Mlp*, *Melampsora magnusiana*, and *Magnaporthe oryzae*, again supporting that most of the gene space is covered in the assembled genome sequence. Finally, the assembly was checked against a total of 79 kbp of genomic sequences from *Mli* previously derived by Sanger sequencing of cloned DNA. The sequenced regions serving as positive controls included loci carrying *AvrP123*, *AvrP4*, *AvrL567-C*, *AvrM-A*, *AvrM-B*, *AvrM-C*, β-*tubulin*, *transcription elongation factor 1*α and a gene fragment from 25S ribosomal RNA. All tested regions were present in full at least once in the assembly, with the exception of regions containing genes from the *AvrM* family (**Figure 1**). In that case, the five previously sequenced paralogs *AvrM-A*, *-B -C, -D*, and *-E* from the avirulence haplotype and *avrM* from the virulence allele in isolate CH5 were assembled as a single gene sequence corresponding to the coding region of *AvrM*, whereas their repeat-rich flanking regions were assembled as separate contigs. Such "collapse" of paralogs into a single assembled sequence was not seen in all cases, e.g., two paralogs of the β-*tubulin* (TUB1) gene were found on the same scaffold, consistent with expectations (Ayliffe et al., 2001), and four copies of 25S rRNA fragment were assembled, including two on the same scaffold (**Figure 1**). Altogether, this suggests that the *Mli* gene space, including that of complex gene families and effectors, is mostly present in the assembly. Also, we have found that, on the limited number of genomic regions tested, the sequence contiguous to genes has been assembled mostly correctly.

Table 2 | *De novo* identification of sequence repeats in the genome of the flax rust pathogen *Melampsora lini* isolate CH5.

Type of repeat	Number of elements	length (kbp)	Percentage occupied of sequence
INTERSPERSED REPEATS			
Retrotransposons			
LTR elements	84,252	42,907	22.64
LINEs	3641	2834	1.5
incl. LINE1	179	133	0.07
incl. LINE2	320	188	0.1
incl. L3/CR1	155	31	0.02
SINEs of type MIRs	70	45	0.02
DNA transposons	37,834	14,273	7.53
Unclassified	81,255	25,603	13.51
OTHER			
Simple repeats	19,637	1114	0.59
Low complexity	4212	237	0.13
Satellites	190	63	0.03
Small RNA	129	58	0.03
Total detected		87,008	45.91%

Retrotransposons include LTRs, long tandem repeats; LINEs, long interspersed elements; and SINEs, short interspersed elements of sub-class; MIRs, mammalian interspersed repeats.

To aid gene annotation, we generated a transcript assembly based on RNAseq data (∼58 million reads, 75 bp single-end) from an RNA sample collected from rust infected flax leaves 6 days post infection. The transcriptome assembly was performed using both assembly-by-alignment in *Cufflinks* and genome-guided *de novo* assembly in *Trinity*. To identify protein-coding genes in the assembled genome scaffolds, several types of evidence were weighted and aggregated to derive consensus gene calls (**Figure 2**). In order of decreasing weights (ω in **Figure 2**), this evidence included: (1) assembled transcripts from the RNAseq library and haustorial-specific ESTs; (2) spliced protein-to-genome alignments using *Mlp* and *Pgt* proteomes; and (3) *ab initio* gene predictions. In total, 31,485 transcriptional units were identified, including 6999 derived from predicted transposable elements, 8215 pseudo-genes (predicted proteins less than 50 amino-acids long or missing a start codon) and 16,271 protein-coding genes. This is similar to the 16,399 genes identified in *Mlp* (Duplessis et al., 2011a). To validate the annotation process, we found that 98% of the conserved eukaryotic (CEGMA) genes assembled "in full" (95% of total) were correctly annotated as protein-coding genes with correct protein sequence. Therefore, the *Mli* assembly presented herein likely contains most of the gene space, with largely complete protein sequences. Version 1 of the genome and proteome sequence and annotation can be found in **Additional file 1** (NCBI Bioproject ID PRJNA239538). Also, the current genome sequence and annotation can be browsed at and downloaded from http://webapollo.bioinformatics.csiro.au:8080/melampsora_lini.

GENOMIC EVIDENCE FOR NUTRIENT ASSIMILATION PATHWAYS IN THE FLAX RUST FUNGUS

Based on genome comparisons with non-biotrophic basidiomycetes, it has been hypothesized that the evolution of obligate biotrophy in rust pathogens is associated with reductions in metabolic abilities via losses of whole metabolic pathways, coupled with an expansion in transporter gene families for enhanced uptake of host-derived nutrients. For example, previous studies noted the absence of some members of the nitrate assimilation cluster in several rust fungi (Duplessis et al., 2011a; Garnica et al., 2013). Within the *Mli* genome assembly, we identified a putative nitrate reductase gene (MELLI_sc3720.2) adjacent to a Major Facilitator Superfamily (MFS) transporter (MELLI_sc3720.1) of unknown function, which may correspond to the nitrate/nitrite transporter in the cluster. However, we did not identify a gene encoding a nitrite reductase and the nitrate reductase gene appeared to be expressed at an extremely low level in infected tissue, suggesting that this pathway may not be functional in *Mli*, similar to other rust fungi. On the

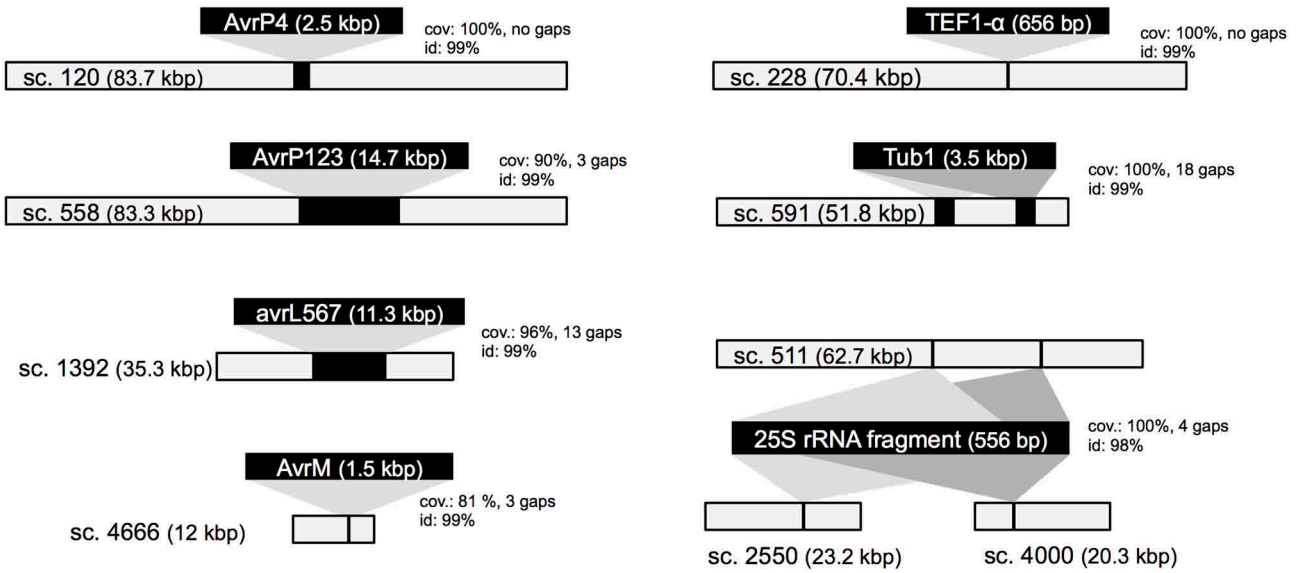


FIGURE 1 | Validation of the *M. lini* isolate CH5 genome assembly. Seven previously sequenced genomic loci were used as controls (black bars with text, not in scale). They were aligned to scaffolds from the *de novo* assembly. Scaffolds showing the best match are shown (light gray bars, in scale), with alignment scores including portion of query covered in match (cov.), sequence identity (id.), and alignment gaps (in bp).

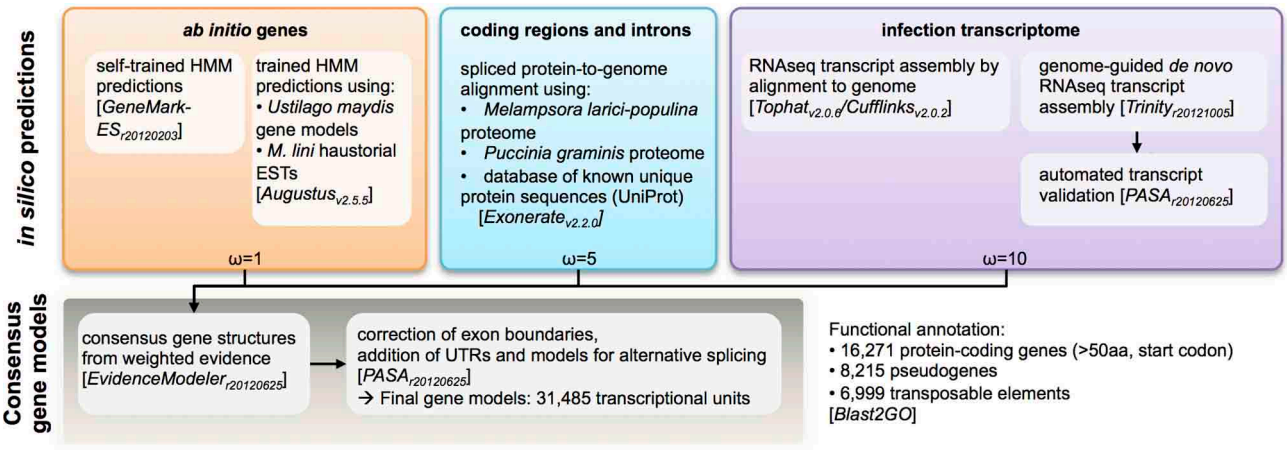


FIGURE 2 | Analysis pipeline for annotation of protein-coding genes in the genome of the flax rust pathogen *Melampsora lini*. Multiple types of evidence were assigned weights (ω) and aggregated to generate a consensus call for a transcriptional unit.

other hand, homologs of all components of the ammonium assimilation pathway were identified in the *Mli* genome, including four ammonium transporters (MELLI_ sc457.12 MELLI_sc152.7 MELLI_sc152.8), the key enzymes glutamate synthase (MELLI_sc11.10_sc11.11) and glutamine synthetase (MELLI_sc3079.2), NAD-specific glutamate dehydrogenase (MELLI_sc1197.2), aspartate aminotransferases (MELLI_sc30.24 and MELLI_sc1978.2), asparagine synthase (MELLI_sc1344.1 and MELLI_sc1683.3) and asparaginase (MELLI_sc2460.3). Thus, as proposed for *Mlp*, the major uptake of host-derived nitrogen is likely in the form of ammonium. In addition, *Mli* may also acquire amino acids and carbon via a relatively large number of amino acid and peptide transporters. The genome of *Mli* contained 16 amino acid permeases/transporters including homologs of *Uromyces vicie-fabae* AAT1, AAT2 and AAT3 (MELLI_sc114.5, MELLI_sc1561.1, and MELLI_sc1251.10, respectively) and 27 putative oligopeptide transporters, which is slightly above the 22 detected in *Mlp*. Several components of the sulfate assimilation pathway were identified in *Mli*, including four sulfate transporter genes (MELLI_sc1698.3, MELLI_sc2898.2, MELLI_sc487.5, and MELLI_sc610.2), sulfite reductase α and β subunits (MELLI_sc3167.3 and MELLI_sc1053.1, respectively) and phosphoadenosine phosphosulfate reductase (MELLI_sc358.2), although we did not identify an ATP

sulfurylase. Similar observations were made for *Mlp* (Duplessis et al., 2011a). In *Mli*, expression levels in infected tissues varied greatly among the putative members of the pathway, with one putative transporter (MELLI_sc487.5) and the sulfite reductase α subunit showing no expression and the phosphoadenosine phosphosulfate expressed at very low level. Other genes identified as putative components of the pathway were all highly expressed, suggesting an important role for them during infection. It is thus unclear whether the seemingly almost complete *Mli* sulfate assimilation pathway is functional in its entirety. Overall, this suggests that the two *Melampsora* species studied here have largely similar metabolic abilities.

In total, we have identified 190 putative proteins showing similarity to transporters from other species (**Additional file 2**, "Transporters"). There were at least 65 members of the Major Facilitator Superfamily (MFS), with 13 putative sugar transporters. This included homologs of the hexose transporter *Uvf*HXT1 and its associated H+/ATPase transporter *Uvf*PMA1 (MELLI_sc2238.1 and MELLI_sc2503.1, respectively). In addition, 17 ATP-binding cassette (ABC) transporters were identified. This was less than the numbers detected in *Mlp* and *Pgt* for both MFS transporters (88 and 51, respectively) and ABC transporters (50 and 38, respectively) (Duplessis et al., 2011a). We also identified two putative auxin efflux carriers (MELLI_sc2698.1 and MELLI_sc890.3), compared with 7 identified in *Mlp*. Fungal pathogens can synthesize auxin, which may be secreted into the host during infection to promote disease (see review by Wang and Fu, 2011). Overall, the 190 putative transporters identified in *Mli* is significantly less than the 356 identified in *Mlp*, although this reduction was not uniform across the different kinds of transporters, e.g., for oligopeptide transporters (see above) we found a larger number in *Mli* than *Mlp*.

COMPARATIVE STUDY OF EFFECTOR COMPLEMENTS FROM FOUR RUST FUNGI

Mining for candidate effectors in four species of rust fungi

To identify putative effectors within *Mli* and highlight those conserved across rust fungi species, we modified the prediction pipeline described in Saunders et al. (2012) to search the proteomes of four rust species, *Mli*, *Mlp*, *Pgt,* and *Pst* (**Figure 3**). Here we broadly define effectors as any fungal protein that is secreted by the fungal cell to act on host-derived substrates or targets or otherwise affect host responses. First, secretome predictions were performed by identifying proteins with a predicted signal peptide (SP), no transmembrane domain and no mitochondria-targeting motif, as described in Torto et al. (2003). To set the stringency of the selection criteria in the pipeline, we used known rust fungal effectors, including AvrL567, AvrM, AvrP123, AvrP4, RTP1p and their homologs from the four investigated rust pathogen species where relevant. The most stringent criteria that still allowed all the known rust fungal effectors to pass were used, e.g., a D-score value of 0.36 in *SignalP*$_{v4.1}$. In total, 7054 (9%) of the 81,399 proteins from all four rust fungi were predicted to be secreted, including 1085 from the flax rust fungus. Subsequently, similarity searches were undertaken between the predicted secreted proteins and the remaining proteomes, to ensure candidates with mis-annotated N-termini or missed SPs were not overlooked. This resulted in 23,516 proteins being selected (29% of the total proteomes). Similarity-based Markov clustering grouped these into 2917 "effector tribes" (**Additional file 2**, "Complete list of tribes"), of which 940 tribes (16,908 proteins) contained at least one protein from the flax rust fungus and were used for further analysis.

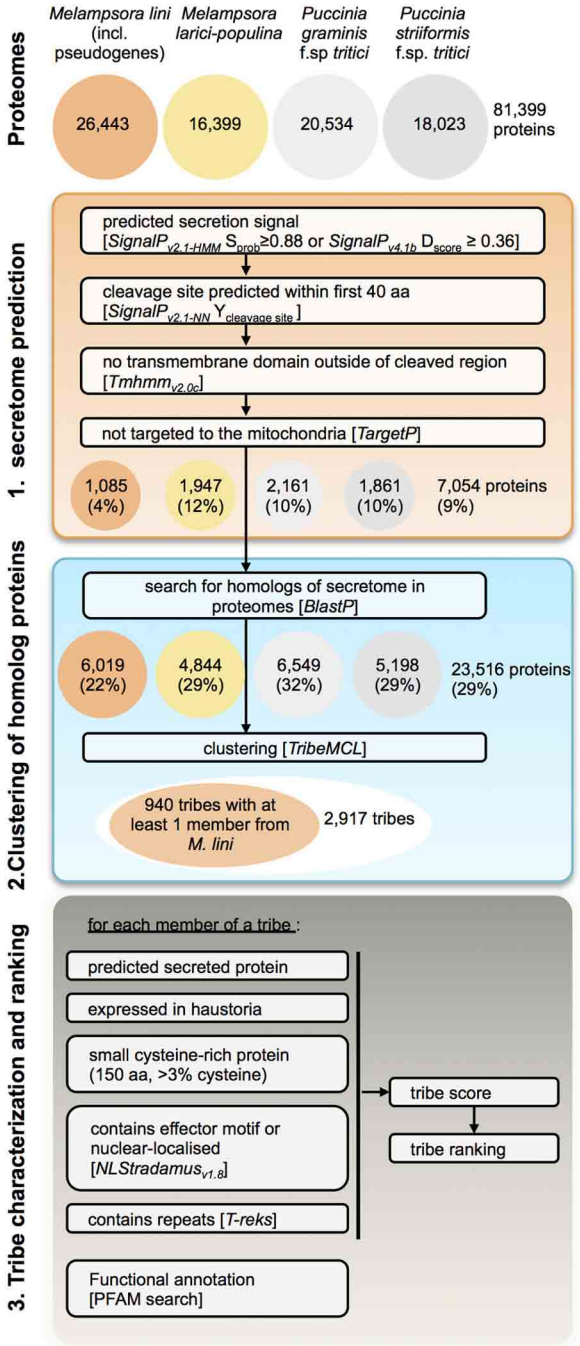


FIGURE 3 | Pipeline for prediction of candidate rust fungal effectors. Tribes of predicted secreted proteins were gathered from the proteomes of *M. lini* isolate *CH5, M. larici-populina* isolate 98AG31, *P. graminis* f.sp *tritici* isolate CDL 75–36-700–3 (race SCCL) and *P. striiformis* f.sp. *tritici* isolate 130. Tribes containing at least one member from *M. lini* were selected for characterization and ranking.

These 940 tribes were characterized *in silico* for properties associated with known effectors from *Mli* and other filamentous plant pathogens. We considered tribes with a high fraction of: (1) predicted secreted proteins; (2) proteins with similarity to *Mli* haustorial ESTs or haustorial predicted secreted proteins (HESPs; Catanzariti et al., 2006); (3) small cysteine-rich proteins; or (4) proteins with predicted effector motifs such as [Y/F/W]xC (Godfrey et al., 2010) as high priority for further analysis. In addition, the presence of a nuclear localization signal or internal repeats in members of a tribe was considered. A single tribe score was then assigned for each of the 940 tribes to order them based on their probability of containing effector proteins (Saunders et al., 2012; **Additional file 2**, "Ranked tribes incl. *Mli* members"). To validate our approach, we looked at the ranking of tribes containing AvrP4, AvrP123, AvrM, AvrL567, RTP1p and rust homologs of the corn smut pathogen (*Ustilago maydis*) effector Chorismate mutase 1 (Cmu1; Djamei et al., 2011), which all occurred among the top 232 tribes out of 940 (**Figure 4**; **Additional file 2**, "Top 200 tribes PFAM annotation"). Manual inspection and curation resulted in removal of 32 clusters that did not appear to represent true effector candidates, mostly large clusters containing only one or a few predicted secreted proteins that may have been mis-annotated. The remaining top 200 tribes were selected for further analysis (**Figure 4**). Also ranking within the top 200 were tribes containing previously identified HESPs from *Mli* and PST130 homologs of HESPs from the stripe rust pathogen isolate 79 (Garnica et al., 2013).

The majority of candidate effectors show conservation across rust fungi

In total, these selected 200 tribes contained 2642 proteins with representatives from all four rusts, including 725 proteins from the flax rust fungus. Of the 200 tribes, 105 (52.5%) had members from all four rust species, 75 (37.5%) had members from *Mli* and *Mlp* only and 16 (8%) were unique to *Mli* (**Figure 5A**). Out of 725 *Mli* predicted effectors, we found only 34 (4.6%) that were in tribes specific to *Mli*, whereas 235 (32.4%) had close homologs only in *Mlp*, and 451 (62.2%) had close homologs across the four rust species. Hence, it seems that the majority of the top-ranking candidate effectors are conserved across the four rust fungi studied here. Tribes containing members from all four rust fungi were relatively large, with an average of 18.2 proteins per tribe and a similar number of members from each species. This indicates that, across the 105 rust fungi-conserved tribes, there is no major shift toward expansion or reduction in gene numbers, although these may be observed at the individual tribe level. Indeed, 5 tribes out of 200 had one member from *Mli* and two members from either *Mlp*, *Pst*, or *Pgt*. These could potentially represent deletions occurring within one species, proteins missing from the annotation or expansion of gene families across several rust fungi genera. Overall, out of the 2642 proteins present in the 200 tribes, 1706 (64.5%) had a predicted SP. Out of the 725 *Mli* proteins in the 200 tribes, there were 395 (54.5%) with a predicted SP, representing 36% of the predicted secretome. Several of these tribes, particularly the larger ones, contain some members with a predicted SP and others lacking an SP; all are considered candidate effectors in this study. For example, tribes 54 and 55 (65 members each) consist of putative aspartate proteases and carboxipeptidases, respectively (annotation available for 63 proteins out of 65, for both tribes). However, only 27 and 29 proteins in each tribe, respectively, have an annotated SP. These tribes may include some members that act on intracellular substrates and others with extracellular activity. Such is the case for fungal chitinases which can function both intra- and extra-cellularly (Duo-Chuan, 2006). In our set, chitinases corresponded to tribe 58 with 32 out of 61 members with an annotated SP. This suggests that some effectors may have evolved from their non-secreted ancestors through the addition of an SP, with a role to perform the same or a similar biological function outside of the fungal cell, e.g., on the host cell wall or inside the host cell.

Shared candidate effectors enriched in predicted apoplastic enzymes

We found that PFAM functional annotation was primarily obtained for tribes that consisted of members from all four rust fungi (**Figure 5B**). The most reliable annotations, i.e., the lowest *E*-values in *BlastP*, were for catalytic enzymes such as glycoside hydrolases and proteases. We found that rust fungi share effectors that are likely involved in (1) degradation of the host physical barriers to infection, (2) inhibition of host immunity and (3) nutrient acquisition. Among the secreted glycoside hydrolase families (GH), we found putative cell wall degrading enzymes including cellulases (GH3, 5, 7, 10, 12, 17, 61), callases (GH16), mannanases (tribe 193, GH76), xylanases (GH10), pectinesterases (tribe 190, Hesp-412), and cutinases (tribe 94). Additionally, we identified a number of candidate effectors that may act in detoxifying the environment or inhibiting immune response signaling, including secreted superoxide dismutases (tribes 91 and 368) and thioredoxins (tribe 620). Other candidate effectors were predicted to be catalytic enzymes that target sugars and proteins either to suppress host immunity or to derive nutrients. Among these, we found sugar degrading enzymes (GH27, 31) and a number of predicted secreted proteases including subtilases (tribe 293), serine carboxipeptidases (tribe 55) and aspartate proteases (tribe 54). From their predicted functions, several of these candidate effectors would likely operate on substrates in the host apoplast.

Identification of putative translocated effectors

Previous work has shown that all four characterized Avr proteins from *Mli* are expressed in haustoria and translocated to the host cell where they are recognized by cytoplasmic TIR-NB-LRR resistance proteins. We found that all four *Mli* Avr proteins were in tribes either specific to *Mli* or in tribes shared only within the *Melampsora* genus. The tribes containing AvrP4 and AvrP123 were specific to *Mli*. Previously, it was found that AvrP4 is present broadly across the *Melampsora* genus and shows signatures of positive selection, at several coding positions, resulting in extensive diversity at the protein level (Van Der Merwe et al., 2009). Most similarity among homologs from *Mli* and *Mlp* resides in the N-terminus end of the proteins, which contains the SP domain, whereas the C-terminal end shows more polymorphism and the signature of positive selection. Here, the closest AvrP4 homologs from *Mlp* were in a separate tribe specific to *Mlp*,

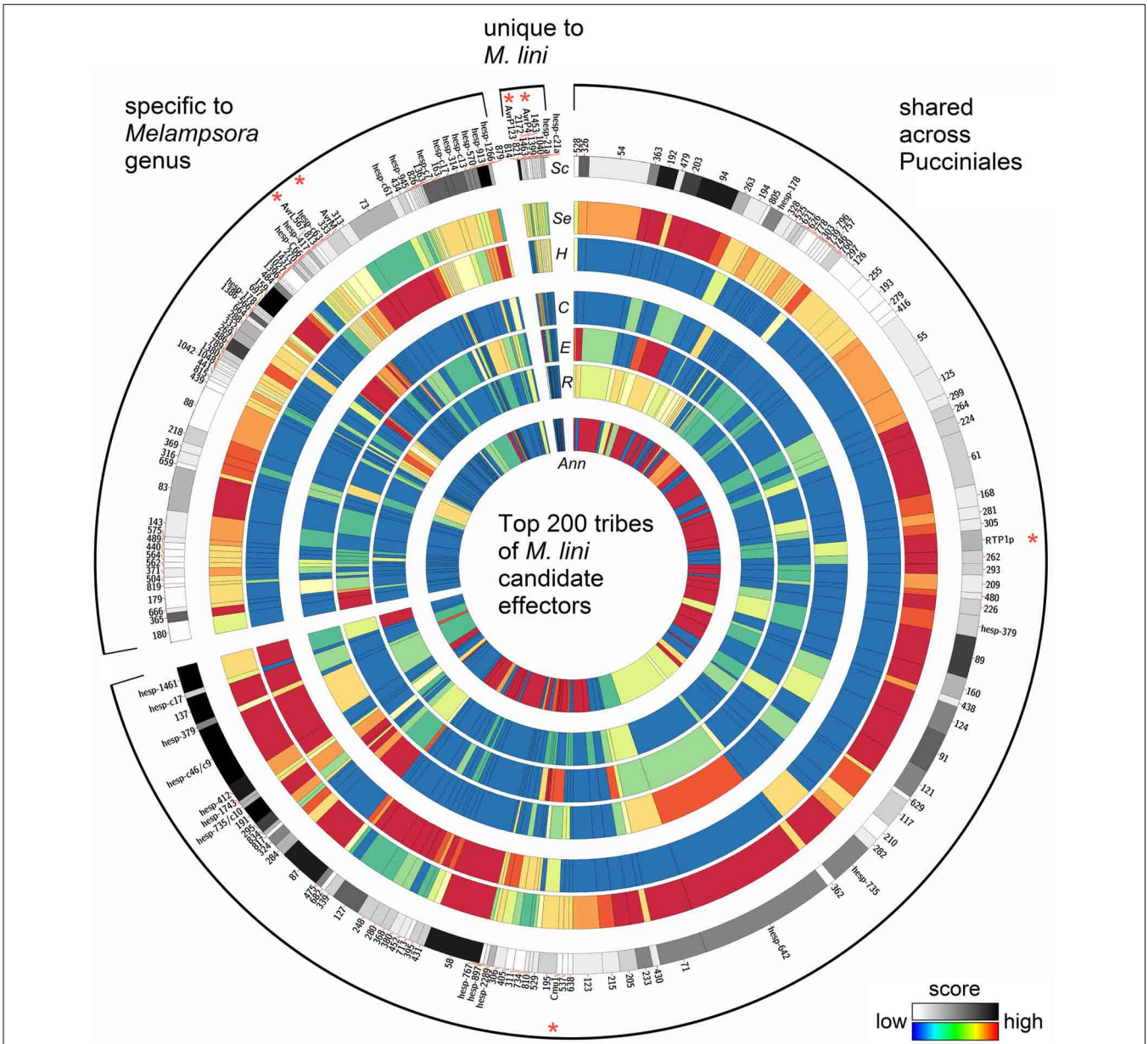


FIGURE 4 | Properties of the top 200 tribes of *M. lini* candidate effectors. Tribes were assembled with 2642 proteins coming from four rust species (*Mli*, *Mlp*, *Pgt*, and *Pst*), and all contain at least one member from *Mli*. Tribes were grouped according to whether they contain members from *Mli* only or members from the two *Melampsora* species only or members from all four rust species. On a circle, each bar represents a tribe, with the width of bar proportional to the number of members in that tribe. Previously known *Mli* avirulence proteins and fungal virulence effectors are indicated with a red asterisk. For each tribe, heatmaps indicate the scores for (*Sc*) overall tribe rank, (*Se*) prediction of secretion, (*H*) similarity to haustorial ESTs, (*C*) presence of small cysteine-rich proteins, (*E*) presence of an effector motif or nuclear localization signal, (*R*) presence of tandem repeats, and (*Ann*) fraction of proteins with a PFAM annotation. Note that the (*H*) field represents similarity across the whole tribe with *Mli* haustorial ESTs, as opposed to haustorial expression of tribe members from all four rusts. As *Mli* is the focus of this study, tribes with members expressed in the haustoria of *Mli* were highlighted. Some previously known haustorially secreted proteins either from *Mli* or other rusts may thus appear to have low score.

perhaps reflecting the fact that the clustering into tribes was performed with predicted mature proteins. In contrast, AvrL567 and AvrM were in tribes shared between *Mli* and *Mlp* and these tribes had no proteins from *Pgt* or *Pst*. These four Avr protein families may have evolved after the divergence of *Melampsora* and *Puccinia*. Also, the average sizes of tribes specific to *Mli* (2.1 proteins) or the two *Melampsora* species (8.9 proteins) were smaller than that of tribes shared across the *Melampsora* and *Puccinia* species studied here (18.2 members). This suggests that relatively small, *Melampsora*-specific tribes may be enriched in translocated effectors.

Other candidate effectors that are putatively translocated to the host cell included rust fungal homologs of the translocated chorismate mutase Cmu1 effector from *U. maydis* (tribe 558) and

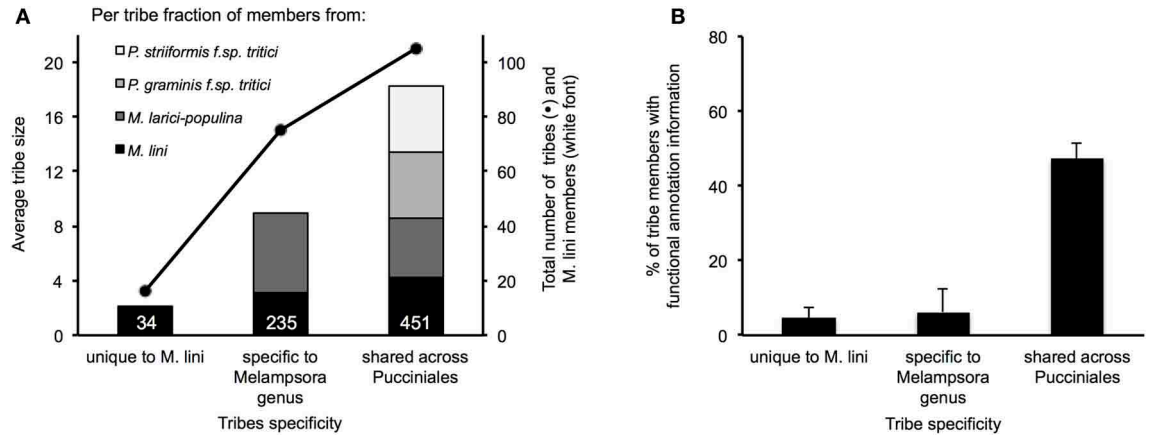


FIGURE 5 | Characteristics of the top 200 tribes of *M. lini* candidate effectors. Tribes were assembled with 2642 proteins coming from four rust species (*Mli, Mlp, Pgt,* and *Pst*), and all contain at least one member from *Mli.* Tribes were grouped according to whether they contain members from *Mli* only or members from the two *Melampsora* species only or members from all four rust species. **(A)** Average tribe size is represented in colored bars, with species of origin of tribe members shown. The total number of tribes and *Mli* members in each group is represented with dots and white font, respectively. **(B)** For each group, the average fraction per tribe of proteins with a PFAM annotation is shown with standard error bars.

of the bean rust pathogen RTP1p effector (tribe 176, hesp-327), which translocates to host cells to function as a putative protease inhibitor (Kemen et al., 2005; Pretsch et al., 2013). These two known effectors, with homologs in all four rust fungi studied here, do not have known avirulence properties. Interestingly, several additional candidate effectors expressed in haustoria (e.g., tribe 26- hesp-642, tribe 52-hesp-c46 and hesp-c9, tribe 304-hesp-1266, tribe 77-hesp-735) corresponded to predicted secreted proteins with a putative nuclear localization signal and PFAM domains similar to those found in nuclear proteins. Further work will be required to assess whether they are translocated to the host cell, as suggested by their PFAM annotation.

DISCUSSION

Here, we report the annotated genome sequence of *Melampsora lini,* providing a genomic resource on a well-established pathogen for research into rust diseases. To our knowledge, it represents the largest fungal genome sequenced so far. Interestingly, genomes from *Melampsora* species can be significantly different in size, with genomes of the poplar and flax rust pathogens estimated at ∼100 and ∼220 Mbp, respectively. It is unclear what mechanisms generate such variation, although a simple genome duplication is unlikely as we have found a comparable number of genes between *Mli* and *Mlp* (16,271 and 16,399 respectively; Duplessis et al., 2011a). The comparatively large genome size of *Mli* can be explained in part by the presence of a greater amount of interspersed repeats (mostly related to transposable elements), as repeats represent ∼45% of both *Mli* and *Mlp* genomes, but with a *Mli* genome more than twice the size of the *Mlp* genome. Similarly, there appears to be a higher absolute amount of non-repetitive sequences in *Mli* than in *Mlp*, explaining further the size difference between the two related genomes. There was significant agreement between gene models of *Mli* and *Mlp* or *Pgt*. The RNAseq data collected from infected leaf tissue 6 days post infection served to support gene predictions for *Mli* genes expressed during infection in hyphae and haustoria. Additionally, by aligning *Mlp* and *Pgt* proteomes to *Mli* genome, we were able to annotate genes that may be missing from our infection transcriptome, because they are expressed at other life-stages than captured using the RNAseq data from infected flax leaves, providing they had homologs from *Mlp* or *Pgt*. Thus, we are confident that our assembled sequence and annotation cover extensively the gene space of *M. lini*.

We have sequenced the hybrid isolate CH5, which carries the four characterized Avr proteins and eight genetically identified but not yet cloned avirulence genes. The candidate effectors reported in this study provide a starting point for screening for the corresponding avirulence phenotypes of these unknown Avr proteins. With respect to assembling a consensus haploid genome of *Mli*, using this isolate posed a number of challenges and resulted in a significant amount of genome fragmentation. The polymorphism between the alleles inherited from the parental isolates C and H generated numerous conflicting assemblies that could not be resolved by the assembly software. This may be true particularly at some effector loci where two significantly diverged alleles exist in CH5 or where the effectors occur in large multigene families surrounded by repeats. For example, at the *AvrL567* locus, the virulence allele contains only one gene (*AvrL567-C*) whereas the avirulence allele carries a tandem gene duplication (*AvrL567-A* and *-B*; Dodds et al., 2004). Reads from the three genes were assembled to form a single haploid consensus ORF. Although the predicted protein had "correct" sequence that allowed effector prediction, the resulting assembly does not reflect the complexity at this locus, i.e., unequal number of genes between the two alleles. Similarly, assembling the regions containing *AvrM* presented multiple challenges due to the presence of repeats in the flanking regions and the similarity in the ORF of the *AvrM* paralogs (Catanzariti et al., 2006). This resulted in a single scaffold being assembled for the coding space from all paralogs as well as separate scaffolds for the flanking regions. The "collapsed" scaffold

was missing the first 10 nucleotides in the ORF, resulting in an incorrect protein prediction. Overall, such issues are typical for assemblies of polymorphic diploid genomes (Yandell and Ence, 2012) and are expected to occur particularly in regions containing complex gene families. However, based on the observation that most (99.8%) filtered haustorial ESTs had strong matches in the assembly, we expect that at least one member of each complex effector gene family is present in the assembly. Further progress can still be made to reduce the fragmentation of the assembled genome of *Mli* and improve gene calls. For example, the F2 segregating population derived from CH5 can be exploited to place the genomic scaffolds on a genetic map. An initial attempt to estimate the synteny between *Mli* scaffolds and the comparatively large *Mlp* scaffolds proved computationally too difficult due to the high level of fragmentation of the *Mli* genome, although it may become feasible with improved assemblies. Here, we have used RNAseq transcripts aligned to the genome for calling genes; nonetheless, a subset of *de novo* assembled RNAseq transcripts that failed to align to the genome may be helpful to identify missing gene calls typically due to genome mis-assemblies or gaps (Haas et al., 2003). In the future, particular efforts will focus on curating manually the genome annotation in regions of interest with a specific focus on haustorially-expressed genes encoding secreted proteins, similar to the approach taken by Duplessis et al. (2011a).

We have identified a large number of candidate effectors in *Mli* that show variable degrees of evolutionary conservation, i.e., shared across the four rust species included in this comparison, shared between the two *Melamspora* species or specific to flax rust. Within the 200 selected tribes out of 940 described here, we found 37% of *Mli* candidate effectors to be specific to the *Melampsora* genus, which contrasts with the ∼74% of all small secreted proteins present in *Mlp* but without *Blast* matches in *Pgt* (Duplessis et al., 2011a). The difference may result from our clustering and tribe-ranking approaches, which emphasized larger (more shared) tribes and does not solely rely on secretome predictions. Here, we have highlighted candidate effectors that may contribute to plant pathogenicity across rust fungi, largely recapitulating previous related studies (Duplessis et al., 2011a; Saunders et al., 2012; Garnica et al., 2013), as well as potential determinants of host specificity in *Mli*. Importantly, all known *Mli* Avr proteins were in tribes that are either specific to *Mli* alone or *Mli* and *Mlp*, whereas a majority of the conserved rust fungal tribes contain enzymes with expected apoplastic activity. In flax, all resistance genes identified so far are predicted to act in the cytoplasm (Ellis et al., 2007). Taken together, this may indicate that rust fungal tribes specific to the genus level are enriched in intracellular effectors and thus may be a primary source of Avr proteins targeted by host intracellular immune receptors. It is unclear whether these genus-specific effectors determine host specificity of rust fungi species via their virulence action, and not just their potential avirulence properties. So far, most of the research on rust fungal effectors has focused on these translocated effectors (see review by Duplessis et al., 2011c). In contrast, little is known of the putative apoplastic effectors of rust fungi identified here and in previous studies (Duplessis et al., 2011a; Hacquard et al., 2012; Saunders et al., 2012; Garnica et al., 2013). The apparent wide conservation across rust fungi of some of these apoplastic effectors that perform a more general virulence function on a wide variety of hosts (e.g., cellulases) makes them particularly interesting for future studies. In the plant-pathogenic Ascomycete fungus *Cladosporium fulvum*, several apoplastic effectors have been characterized, including some that are recognized extra-cellularly by immune receptors (reviewed in Wit et al., 2009). For example, *Ecp6* and Avr4 function as chitin-binding proteins that inhibit host chitin-elicited immunity and a host-chitinase inhibitor, respectively (Van Den Burg et al., 2006; De Jonge et al., 2010). Further work is required to assess whether candidate apoplastic effectors from rust fungi have similar roles to the *C. fulvum* effectors and whether there are components of host immunity that may target them in the apoplast.

To fine-tune the search for the effector complement of the flax rust fungus, we took advantage of previous knowledge of *Mli* Avr proteins and other known rust effectors. Tribes of candidate effectors were prioritized for future studies, including functional characterization, if, similar to the previously known rust effectors, they ranked among the selected top 200. The similarity-based clustering of proteins into tribes used here is beneficial for identifying conserved gene families although just as any sequence-based clustering approach, its power decreases when dealing with related genes under accelerated rates of evolution, such as Avr gene families. Regarding the ranking approach, we elected to give weight to the presence of known effector motifs, although previous work has found that no obvious protein motif broadly characterized effectors from rust fungi species (Saunders et al., 2012). Also, effectors with no identified homologs in the flax rust fungus or another rust fungus studied here would appear in tribes of size 1, and would likely be ranked low, despite their biological relevance. Thus, it should be noted that tribes with low or intermediate ranking may still correspond to effectors, e.g., tribe 400 ranks 371 out of 940 and consists of putative extracellular invertases, probably essential to degrade sugars outside of the fungal cell. Also, a small number of effectors may still be missing from our predictions. This could result from difficulties in generating the assembly for some of them or missing gene calls, although coverage of ESTs from haustoria and CEGMA analyses suggests this is a limited occurrence (∼5%). In addition, our filtering criteria, while enriching for likely effectors could generate a number of false negatives, e.g., in the case of effectors without a conventional eukaryotic secretion signal, such as the barley powdery mildew effectors AVR_{k1} and AVR_{a10} (Ridout et al., 2006). Likewise, the 50 amino acid cutoff for gene prediction does not allow discovery of very small effectors, such as the bean rust candidate effectors PIG11 and PIG13 (24 and 31 amino acids respectively; Hahn and Mendgen, 1997). Also, a mis-annotated 5′ end or a real SP that falls just under the cut-off for prediction could cause some effectors to be missing from the effector complement. However, we limited these problems by forming tribes that contained even just one predicted secreted member and enriching those that were expressed in haustoria. Generally, however, based on our use of known Avr proteins and rust effectors to help set the parameters of the pipelines for genome assembly and annotation and effector prediction, we are confident that the bulk of *Mli* effectors are contained in our set.

Our findings agree with trends previously reported for non-rust biotrophic plant pathogens. Specifically, our results support the notion that evolution of obligate biotrophy is associated with the loss of some metabolic pathways (Kemen et al., 2011), although our results illustrate that the degree to which pathways can be affected may vary. For example, an almost complete pathway for sulfate metabolism was identified in *Mli*, and previously in *Mlp*, but appears to be absent in *Puccinia* (Duplessis et al., 2011a; Garnica et al., 2013), Ascomycetes such as the barley powdery mildew pathogen *Blumeria graminis* f.sp. *hordei* (Spanu et al., 2010) and even Oomycetes such as *Hyaloperonospora arabidopsidis* the downy mildew pathogen of *Arabidopsis thaliana* (Baxter et al., 2010). Consistent with findings on other rust fungi (Duplessis et al., 2011a; Garnica et al., 2013), in *Mli* the probable loss of the ability to import and metabolize nitrate or nitrite appears to be coupled with an expansion in the number of amino-acid and oligopeptide transporters, compared to non-biotrophic basidiomycetes, which would allow accumulation of host-derived organic nitrogen sources. During infection by *Mlp*, transporter proteins are mostly expressed after haustorial formation (∼48 h post infection; Duplessis et al., 2011b), supporting the view that they are involved in the uptake of host-derived nutrients and possibly also the efflux of virulence factors and influx of plant anti-fungal compounds for detoxification. Assigning the direction of transport and the nature of the cargo translocated by the numerous transporters described here will require significant further investigation.

Finally, studies on *M. lini* collected from wild populations infecting the native Australian wild flax (*Linum marginale*) has revealed the existence of two lineages of *Mli*, namely the AA and AB lineages, where A and B refer to the genetic constitution of the two haploid nuclei in the dikaryon (Barrett et al., 2007). The lineages exhibit substantial differences in terms of virulence and life-style, with lineage AA capable of both sexual and asexual (clonal) reproduction and lineage AB only found to reproduce clonally (Nemri et al., 2012). The complete life-cycle of *Mli* contains five different spore stages, all occurring on flax (Lawrence et al., 2007; Ravensdale et al., 2010). In this study, we were interested in genes expressed during infection with uredospores, the asexual spore stage, of isolate CH5 (lineage AA). In the future, it will be interesting to compare it with the infection transcriptome following inoculation with the four spore stages forming the sexual cycle. Also, comparing the genomes and transcriptomes of isolates of lineage AA and AB may give insight into how much within-species diversity can be found in candidate effectors or candidate genes mediating environmental adaptation and life-history differences.

In conclusion, we have identified a large number of candidate proteins potentially involved in multiple aspects of infection of flax by *Mli* uredospores. These aspects include: (1) penetration of host tissue and colonization, with cuticle and cell wall degradation enzymes; (2) detoxification and modification of host metabolism for suppression of host defenses and promotion of infection; and (3) hydrolysis and uptake of nutrients. Further work is needed to assign effector candidates and metabolic pathways to specific time-points of infection and specific fungal organs. In *Mli*, a technique for genetic transformation and gene silencing is available (Lawrence et al., 2010), creating the opportunity to dissect the role of candidate genes identified in this study, coming from *Mli* and other rust fungi. This provides a starting point for future investigations aiming to understand virulence in economically important rust fungi and developing innovative strategies to render crops resistant to them.

MATERIALS AND METHODS

SAMPLE PREPARATION

Genomic DNA was extracted from *Melampsora lini* reference isolate CH5 uredospores according to Justesen et al. (2002) with modifications. Approximately 100 mg of dried uredospores were ground with 1 g of acid washed sand using a pestle and mortar. The powder was transferred to a 15 ml polypropylene tube and resuspended in 2 ml of DNA extraction buffer (25 g/L D-sorbitol, 10 g/L sodium dodecyl sulfate; 8 g/L hexadecyltrimethylammonium bromide (CTAB), 10 g/L polyvinylpyrrolidone (PVP), 0.8 M NaCl, 20 mM EDTA pH 8.0, 0.1 M Tris HCl pH 8.0). Five microliters of 100 mg/ml RNaseA were added and the samples were incubated at 65°C for 30 min. Ten microliters of 20 mg/ml proteinase K were added and the samples were incubated at 65°C for a further 30 min before extraction using 3 ml of chloroform. DNA was precipitated by addition of 1 vol. isopropanol and DNA was recovered by centrifugation for 15 min at 16,000 g. DNA was washed with 75 % (v/v) ice-cold ethanol, air-dried and resuspended in 62.5 mM MOPS pH 7.0. DNA was then cleaned-up using a Qiagen G/20 genomic-tip according to the manufacturer's instructions. Four genomic DNAseq libraries were generated including one paired-end library of ∼300 bp and three mate-pair libraries of sizes 2000, 3000, and 5000 bp (**Table S1**). All sequencing was performed at Macrogen Inc. (Seoul, Republic of Korea) and the Australian Genome Research Facility (AGRF, Sydney, Australia) using Illumina HiSeq2000 to produce reads of 100 bp. Additionally, an RNAseq library was generated from leaf material of host line Hoshangbad infected with isolate CH5 at 6 days post infection as in Catanzariti et al. (2006). In total, ∼110 million raw 75 bp single-end reads were sequenced using Illumina Genome Analyzer II at AGRF. Finally, a haustorial-specific EST library of 2783 sequences was used, as described in Catanzariti et al. (2006), including 1961 ESTs not previously reported. The library was preprocessed with *Seqclean* (http://compbio.dfci.harvard.edu/tgi/software/) to remove polyA's and vector contamination (UniVec_Core library, http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html).

DE NOVO GENOME AND TRANSCRIPTOME ASSEMBLY

Prior to assembly, DNAseq and RNAseq reads underwent quality-based trimming using *Condetri* (Smeds and Kunstner, 2011) and trimming of Illumina adaptor sequence using *Trimmomatic* (Lohse et al., 2012). Removal of PCR duplicates from all four DNAseq libraries was done using *filterPCRdupl* (http://code.google.com/p/condetri) followed by removal of likely sequencing errors using *ErrorCorrection* (Luo et al., 2012). The genome assembly and initial scaffolding were performed using *SOAPdenovo v2r215* (Luo et al., 2012). After testing k-mer values ranging from 37 to 47, a k-mer value of 41 was found to give the best results and was used to produce the assembly reported here.

To close gaps in the scaffolded SOAPdenovo assembly, we used *GapCloser v1.12r1* (Luo et al., 2012) followed by scaffolding using *SSPACE v2.0* (Boetzer et al., 2011) for two rounds using the four paired DNAseq libraries. Only scaffolds longer than 200 bp were retained in the final genomic assembly. For the transcriptome analysis of infected leaf tissue, we obtained ∼94 million reads from the ∼110 million raw reads after quality-based filtering. We then filtered out RNAseq reads originating from flax by aligning the reads against the genome sequence of flax v1.0 (Wang et al., 2012) and a collection of flax ESTs (Fenart et al., 2010). Around 38% of the total RNAseq reads were removed as flax reads, leaving ∼58 million reads (62% of the total), mostly from rust and potentially including some contaminant. Transcript assembly was done using two strategies. First, assembly-by-alignment to the genome sequence was performed using *Tophat_v2.0.6/Cufflinks_v2.0.2* (Trapnell et al., 2010). Second, genome-guided transcript assembly was done using *Trinity r2012-10-05* (Grabherr et al., 2011), coupled with *PASA r2012-06-25* to predict terminal exons (Haas et al., 2008).

GENE PREDICTION AND ANNOTATION

Ab initio gene prediction was performed using: (1) *Augustus v2.5.5* (Stanke et al., 2006) with aligned ESTs as hints and *Ustilago maydis* as related species for training the gene finder; and (2) *GeneMark-es r2012-02-03* (Ter-Hovhannisyan et al., 2008). Spliced protein-to-genome alignment was performed using *Exonerate v2.2.0* (Slater and Birney, 2005) with Uniref90 (downloaded from www.uniprot.org) and complete proteomes of *Mlp* isolate 98AG31 (obtained from http://genome.jgi.doe.gov/) and *Pgt* isolate CDL-75-36-700-3 (race SCCL; obtained from http://www.broadinstitute.org/) (Duplessis et al., 2011a). *EvidenceModeler* (Haas et al., 2008) was used to combine ESTs, gene predictions, spliced protein alignments and transcript alignments. The annotation was then imported into *PASA* to update transcript predictions, add UTR's and alternative transcripts. *CEGMA* (Parra et al., 2009) was used to verify the quality of the assembly of the gene space in the genome and the annotation output by *EvidenceModeler*. *Blast2go* was used to filter out the predicted transposable elements in the final proteome set (Conesa et al., 2005). *Webapollo* was used for genome browsing and inspection of the annotation (Lee et al., 2013). *De novo* identification of repeats in the genome sequence was performed using *RepeatMasker* v4.0.1 (Smit et al., 1996).

RUST EFFECTOR PREDICTION

An effector prediction pipeline modified from *PexFinder* (Torto et al., 2003) was set up to search the proteomes of all four rust fungi species, *Mli*, *Mlp*, *Pgt* (isolates specified above) and *Pst* isolate 130 (*Pst*; Cantu et al., 2011), obtained from D. Saunders, Sainsbury Laboratory, Norwich, UK. Proteins were selected if (1) they exceeded the cutoffs for SP prediction of 0.88 for S-probability in *SignalP v2.1-HMM* (Nielsen and Krogh, 1998) or 0.36 for D-score in *SignalP v4.1b* (Petersen et al., 2011); (2) the predicted cleavage site occured between amino-acid 10 and 40; (3) no transmembrane domain was predicted to occur after the cleavage site using *Tmhmm v2.0c*; and (4) the protein was not predicted to be mitochondrial by *TargetP v1.1* (Emanuelsson et al., 2007). Any protein from all four rust fungi proteomes that passed the selection criteria was subsequently used as a query in BlastP against the remainder of the proteomes and effector tribes were formed, as per Saunders et al. (2012). In order to group candidate effectors with functional and/or structural similarities in the effector domain, the clustering was performed using predicted mature proteins when a SP was detected. Real effectors appearing as false negatives in SP prediction, due to a mis-annotated 5′ end (as was the case for AvrM), or a correct SP with a prediction that falls under our cut-offs, were included in the clustering if one related protein had a predicted SP. Preventing the Markov clustering from being primarily driven by the SP resulted in (a) these "recovered" effectors being assigned to their correct tribe and (b) avoiding the formation of very large tribes potentially composed of effectors with greatly divergent or unrelated functional domains but with a conserved SP. This focus on the functional domain of effectors also meant that the evolutionary information contained in the SP domain was not used to form the tribes.

Aside from the number of members in a tribe, six features contributed to ranking the tribes. For each of these features, a score was calculated as per Saunders et al. (2012). This score is based on the number of proteins within a tribe that displayed a particular feature, relative to the likelihood of a tribe of the given size containing the same number of proteins with that particular feature by chance. Features assessed for each protein from a tribe included: [1] being a predicted secreted protein, [2] having a BLAST match against HESPs (Catanzariti et al., 2006), or [3] haustorial ESTs and [4] being a small cysteine-rich protein. These four features were given high weight in the formula described below. Additionally, features included having [5] one or more effector motifs such as [L/I]xAR, [R/K]CxxCx12H, RxLR, [Y/F/W]xC, YxSL[R/K], or G[I/F/Y][A/L/S/T]R between amino acids 10–110, or a nuclear localization signal identified using *NLStradamus* (Nguyen Ba et al., 2009), or [6] one or more internal repeats identified using *T-reks* (Jorda and Kajava, 2009). These two features were assigned minor weight in the formula below. To emphasize shared properties among members of a tribe rather than particular features of one member, each feature was scored as a single 0 or 1, e.g., having two [Y/F/W]xC motifs and a nuclear localization motif was treated the same as having just one [Y/F/W]xC motif. Weights for all six features were combined to produce an overall score used to rank the tribe, calculated as $= log_2([1] + [2] + [3] + [4]) \times (1 + 0.1 \times ([5] + [6]))$. Manual curation of the ranked list was performed to remove tribes with less than 10% of secreted members. Out of the top-ranking 232 tribes, 32 tribes were removed (gray lines in **Additional file 2**, "Ranked tribes incl. *Mli* members") giving a list of the top selected 200 tribes described here. These top 200 tribes and their overall and individual feature scores were visualized using *Circos* (Krzywinski et al., 2009). The proportion of proteins with a PFAM score was also assessed, but did not contribute to tribe ranking, as many fungal and oomycete Avr proteins do not have recognizable PFAM domains. PFAM categories based uniquely on domain-recognition without associated function were removed, including cysteine-rich secretory protein. A cut-off of E-5 was used.

AUTHOR CONTRIBUTIONS

Adnane Nemri, Claire Anderson, Diane G. O. Saunders, Narayana M. Upadhyaya, and Joe Win performed computational analyses. Peter N. Dodds, Jeffrey G. Ellis, David A. Jones, and Sophien Kamoun designed and supervised the research. All authors contributed to manuscript writing.

ACKNOWLEDGMENTS

We thank Pat Moore for sequencing of the EST library. This work was funded by a grant from the CSIRO Transformational Biology Capability Platform to Adnane Nemri. Claire Anderson was supported by an ARC Discovery Grant (DP120104044) awarded to David A. Jones and Peter N. Dodds.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00098/abstract

Table S1 | Libraries used for the assembly of the genome and infection transcriptome of the flax rust pathogen *Melampsora lini* isolate CH5.

Additional file 1 | (.zip) Fasta files of *M. lini* genomic scaffolds and proteome including Blast2go annotations as sequence descriptors where available.

Additional file 2 | (.xlsx) Microsoft Excel workbook containing tables with [1] Complete list of tribes with members from four rust species, [2] *in silico* characterization of the 940 tribes of candidate effectors containing at least one member from *Melampsora lini* (gray lines correspond to tribes removed from the top 200), [3] top 200 *Melampsora lini* tribes with PFAM information where available and [4] transporters found in *Melampsora lini* genome with putative substrate where available.

REFERENCES

Australian Nursery and Garden Industry, T. (2012). *Myrtle Rust (Uredo rangelii) Management Plan*. Available online at: http://www.ngia.com.au/Section?Action=View&Section_id=527

Ayliffe, M. A., Dodds, P. N., and Lawrence, G. J. (2001). Characterisation of a β-*tubulin* gene from *Melampsora lini* and comparison of fungal β-*tubulin* genes. *Mycol. Res.* 105, 818–826. doi: 10.1017/S0953756201004245

Barrett, L. G., Thrall, P. H., and Burdon, J. J. (2007). Evolutionary diversification through hybridization in a wild host-pathogen interaction. *Evolution* 61, 1613–1621. doi: 10.1111/j.1558-5646.2007.00141.x

Barrett, L. G., Thrall, P. H., Dodds, P. N., Van Der Merwe, M., Linde, C. C., Lawrence, G. J., et al. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol. Biol. Evol.* 26, 2499–2513. doi: 10.1093/molbev/msp166

Baxter, L., Tripathy, S., Ishaque, N., Boot, N., Cabral, A., Kemen, E., et al. (2010). Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* 330, 1549–1551. doi: 10.1126/science.1195203

Boehm, E. W. A., and Bushnell, W. R. (1992). An ultrastructural pachytene karyotype for *Melampsora lini*. *Phytopathology* 82, 1212–1218. doi: 10.1094/Phyto-82-1212

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M. N., Chen, X. M., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. *sp tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:8. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., Maclean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. *sp.* tritici reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Conesa, A., Götz, S., Garcia-Gomez, J., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Cummins, G. B., and Hiratsuka, Y. (2003). *Illustrated Genera of Rust Fungi*. St Paul, MN: American phytopathological society.

De Jonge, R., Peter Van Esse, H., Kombrink, A., Shinya, T., Desaki, Y., Bours, R., et al. (2010). Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. *Science* 329, 953–955. doi: 10.1126/science.1190859

Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., et al. (2011). Metabolic priming by a secreted fungal effector. *Nature* 478, 395–398. doi: 10.1038/nature10454

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The Melampsora lini AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16, 755–768. doi: 10.1105/tpc.020040

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Teh, T., Wang, C. I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8888–8893. doi: 10.1073/pnas.0602577103

Duo-Chuan, L. (2006). Review of fungal chitinases. *Mycopathologia* 161, 345–360. doi: 10.1007/s11046-006-0024-y

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011b). *Melampsora larici-populina* transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Duplessis, S., Joly, D. L., and Dodds, P. N. (2011c). "Rust Effectors," in *Effectors in Plant–Microbe Interactions,* eds F. Martin and S. Kamoun (Oxford, UK: Wiley-Blackwell), 155–193. doi: 10.1002/9781119949138.ch7

Eilam, T., Bushnell, W. R., Anikster, Y., and McLaughlin, D. J. (1992). Nuclear DNA content of basidiospores of selected rust fungi as estimated from fluorescence of propidium iodide-stained nuclei. *Phytopathology* 82, 705–712. doi: 10.1094/Phyto-82-705

Ellis, J. G., Dodds, P. N., and Lawrence, G. J. (2007). Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. *Annu. Rev. Phytopathol.* 45, 289–306. doi: 10.1146/annurev.phyto.45.062806.094331

Emanuelsson, O., Brunak, S., Von Heijne, G., and Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP, and related tools. *Nat. Protoc.*, 953–971. doi: 10.1038/nprot.2007.131

Fenart, S., Ndong, Y.-P. A., Duarte, J., Riviere, N., Wilmer, J., Van Wuytswinkel, O., et al. (2010). Development and validation of a flax (*Linum usitatissimum* L.) gene expression oligo microarray. *BMC Genomics* 11:592. doi: 10.1186/1471-2164-11-592

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A. N. A., Petitot, A.-S., et al. (2011). 454-pyrosequencing of Coffea arabica leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant–rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Flor, H. H. (1955). Host-parasite interaction in flax rust - its genetics and other implications. *Phytopathology* 45, 680–685.

Garnica, D., Upadhyaya, N., Dodds, P., and Rathjen, J. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PLoS ONE* 8:e67150. doi: 10.1371/journal.pone.0067150

Godfrey, D., Bohlenius, H., Pedersen, C., Zhang, Z., Emmersen, J., and Thordal-Christensen, H. (2010). Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genomics* 11:317. doi: 10.1186/1471-2164-11-317

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, Jr, R. K., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

Hacquard, S., Joly, D. L., Lin, Y. C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (Poplar leaf rust). *MPMI* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hahn, M., and Mendgen, K. (1997). Characterization of in planta-induced rust genes isolated from a haustorium-specific cDNA library. *Mol. Plant Microbe Interact.* 10, 427–437. doi: 10.1094/MPMI.1997.10.4.427

Jakupović, M., Heintz, M., Reichmann, P., Mendgen, K., and Hahn, M. (2006). Microarray analysis of expressed sequence tags from haustoria of the rust fungus *Uromyces fabae*. *Fungal Genet. Biol.* 43, 8–19. doi: 10.1016/j.fgb.2005.09.001

Joly, D. L., Feau, N., Tanguay, P., and Hamelin, R. C. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora spp.*). *BMC Genomics* 11:422. doi: 10.1186/1471-2164-11-422

Jones, D. A. (1988). Genetic properties of inhibitor genes in flax rust that alter avirulence to virulence on flax. *Phytopathology*, 342–344. doi: 10.1094/Phyto-78-342

Jones, J. D., and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323–329. doi: 10.1038/nature05286

Jorda, J., and Kajava, A. (2009). T-reks: identification of tandem repeats in sequences with a k-means based algorithm. *Bioinformatics* 25, 2632–2638. doi: 10.1093/bioinformatics/btp482

Justesen, A. F., Ridout, C. J., and Hovmøller, M. S. (2002). The recent history of *Puccinia striiformis* f.sp. *tritici* in Denmark as revealed by disease incidence and AFLP markers. *Plant Pathol.* 51, 13–23. doi: 10.1046/j.0032-0862.2001.00651.x

Kemen, E., Gardiner, A., Schultz-Larsen, T., Kemen, A. C., Balmuth, A. L., Robert-Seilaniantz, A., et al. (2011). Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol.* 9:e1001094. doi: 10.1371/journal.pbio.1001094

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., et al. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res.* 19, 1639–1645. doi: 10.1101/gr.092759.109

Lawrence, G. J. (1995). Do plant pathogens produce inhibitors of the resistance reaction in plants? *Trends Microbiol.* 3, 475–476. doi: 10.1016/S0966-842X(00)89014-9

Lawrence, G. J., Dodds, P. N., and Ellis, J. G. (2007). Rust of flax and linseed caused by *Melampsora lini*. *Mol. Plant Pathol.* 8, 349–364. doi: 10.1111/j.1364-3703.2007.00405.x

Lawrence, G. J., Dodds, P. N., and Ellis, J. G. (2010). Transformation of the flax rust fungus, *Melampsora lini*: selection via silencing of an avirulence gene. *Plant J.* 61, 364–369. doi: 10.1111/j.1365-313X.2009.04052.x

Lawrence, G. J., Mayo, G. M. E., and Shepherd, K. W. (1981). Interactions between genes controlling pathogenicity in the flax rust fungus. *Phytopathology* 71, 12–19. doi: 10.1094/Phyto-71-12

Lee, E., Helt, G., Reese, J., Munoz-Torres, M., Childers, C., Buels, R., et al. (2013). Web Apollo: a web-based genomic annotation editing platform. *Genome Biol.* 14:R93. doi: 10.1186/gb-2013-14-8-r93

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2013). The haustorial transcriptomes of Uromyces appendiculatus and Phakopsora pachyrhizi and their candidate effector families. *Mol. Plant Pathol.* n/a-n/a.

Link, T. I., and Voegele, R. T. (2008). Secreted proteins of *Uromyces fabae*, similarities and stage specificity. *Mol. Plant Pathol.* 9, 59–66. doi: 10.1111/j.1364-3703.2007.00448.x

Lohse, M., Bolger, A., Nagel, A., Fernie, A., Lunn, J., Stitt, M., et al. (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* 40, W622–W627. doi: 10.1093/nar/gks540

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1:18. doi: 10.1186/2047-217X-1-18

McDowell, J. M. (2011). Genomes of obligate plant pathogens reveal adaptations for obligate parasitism. *Proc. Natl. Acad. Sci. U.S.A.* 108, 8921–8922. doi: 10.1073/pnas.1105802108

Mendgen, K., and Nass, P. (1988). The activity of powdery-mildew haustoria after feeding the host cells with different sugars, as measured with a potentiometric cyanine dye. *Planta* 174, 283–288 doi: 10.1007/BF00394782

Nemri, A., Barrett, L. G., Laine, A.-L., Burdon, J. J., and Thrall, P. H. (2012). Population processes at multiple spatial scales maintain diversity and adaptation in the *Linum marginale - Melampsora lini* association. *PLoS ONE* 7:e41366. doi: 10.1371/journal.pone.0041366

Nguyen Ba, A., Pogoutse, A., Provart, N., and Moses, A. (2009). NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10:202. doi: 10.1186/1471-2105-10-202

Nielsen, H., and Krogh, A. (1998). "Prediction of signal peptides and signal anchors by a hidden Markov model.," in *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (Menlo Park, CA: AAAI Press), 122–130.

Pardey, P. G., Beddow, J. M., Kriticos, D. J., Hurley, T. M., Park, R. F., Duveiller, E., et al. (2013). Right-sizing stem-rust research. *Science* 340, 147–148. doi: 10.1126/science.122970

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. *Nucleic Acids Res.* 37, 298–297. doi: 10.1093/nar/gkn916

Petersen, T. N., Brunak, S., Von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Meth* 8, 785–786. doi: 10.1038/nmeth.1701

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins—a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Rafiqi, M., Gan, P. H. P., Ravensdale, M., Lawrence, G. J., Ellis, J. G., Jones, D. A., et al. (2010). Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. *Plant Cell* 22, 2017–2032. doi: 10.1105/tpc.109.072983

Ravensdale, M., Nemri, A., Thrall, P. H., Ellis, J. G., and Dodds, P. N. (2010). Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. *Mol. Plant Pathol.* 12, 93–102. doi: 10.1111/j.1364-3703.2010.00657.x

Ridout, C. J., Skamnioti, P., Porritt, O., Sacristan, S., Jones, J. D. G., and Brown, J. K. M. (2006). Multiple avirulence paralogues in cereal powdery mildew fungi may contribute to parasite fitness and defeat of plant resistance. *Plant Cell* 18, 2402–2414. doi: 10.1105/tpc.106.043307

Saunders, D. G. O., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:14. doi: 10.1371/journal.pone.0029847

Slater, G., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. doi: 10.1186/1471-2105-6-31

Smeds, L., and Kunstner, A. (2011). ConDeTri - a content dependent read trimmer for Illumina data. *PLoS ONE* 6:e26314. doi: 10.1371/journal.pone.0026314

Smit, A., Hubley, R., and Green, P. (1996). *RepeatMasker Open-3.0*. Available online at: http://www.repeatmasker.org

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330, 1543–1546. doi: 10.1126/science.1194573

Stanke, M., Schoffmann, O., Morgenstern, B., and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* 7:62. doi: 10.1186/1471-2105-7-62

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., and Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res.* 18, 1979–1990. doi: 10.1101/gr.081612.108

Torto, T. A., Li, S., Styer, A., Huitema, E., Testa, A., Gow, N. A. R., et al. (2003). EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen Phytophthora. *Genome Res.* 13, 1675–1685. doi: 10.1101/gr.910003

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotech.* 28, 511–515. doi: 10.1038/nbt.1621

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M., Ellis, J., and Dodds, P. (2013). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol. Plant Microbe Interact.* 27, 255–264. doi: 10.1094/MPMI-07-13-0187-FI

Van Den Burg, H. A., Harrison, S. J., Joosten, M. H. A. J., Vervoort, J., and De Wit, P. J. G. M. (2006). Cladosporium fulvum Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. *Mol. Plant Microbe Interact.* 19, 1420–1430. doi: 10.1094/MPMI-19-1420

Van Der Merwe, M. M., Kinnear, M. W., Barrett, L. G., Dodds, P. N., Ericson, L., Thrall, P. H., et al. (2009). Positive selection in AvrP4 avirulence gene homologues across the genus Melampsora. *Proc. R. Soc. B* 276, 2913–2922. doi: 10.1098/rspb.2009.0328

Wang, S., and Fu, J. (2011). Insights into auxin signaling in plant-pathogen interactions. *Front. Plant Sci.* 2:74. doi: 10.3389/fpls.2011.00074

Wang, Z., Hobson, N., Galindo, L., Zhu, S., Shi, D., McDill, J., et al. (2012). The genome of flax (*Linum usitatissimum*) assembled de novo from short shotgun sequence reads. *Plant J.* 72, 461–473. doi: 10.1111/j.1365-313X.2012.05093.x

Wit, P. G. M., Joosten, M. A. J., Thomma, B. P. J., and Stergiopoulos, I. (2009). "Gene for gene models and beyond: the *Cladosporium fulvum*-Tomato pathosystem," in *Plant Relationships*, ed H. Deising (Berlin; Heidelberg: Springer), 135–156.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. doi: 10.1038/nrg3174

Yin, C., Chen, X., Wang, X., Han, Q., Kang, Z., and Hulbert, S. (2009). Generation and analysis of expression sequence tags from haustoria of the wheat stripe rust fungus *Puccinia striiformis* f. *sp. tritici. BMC Genomics* 10:626. doi: 10.1186/1471-2164-10-626

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4:2673. doi: 10.1038/ncomms3673

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 13 January 2014; accepted: 28 February 2014; published online: 24 March 2014.

*Citation: Nemri A, Saunders DGO, Anderson C, Upadhyaya NM, Win J, Lawrence GJ, Jones DA, Kamoun S, Ellis JG and Dodds PN (2014) The genome sequence and effector complement of the flax rust pathogen Melampsora lini. Front. Plant Sci. **5**:98. doi: 10.3389/fpls.2014.00098*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Nemri, Saunders, Anderson, Upadhyaya, Win, Lawrence, Jones, Kamoun, Ellis and Dodds. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 31 October 2014
doi: 10.3389/fpls.2014.00594

Annotation of a hybrid partial genome of the coffee rust (*Hemileia vastatrix*) contributes to the gene repertoire catalog of the Pucciniales

Marco A. Cristancho[1]*, David Octavio Botero-Rozo[1,2], William Giraldo[1], Javier Tabima[1,2], Diego Mauricio Riaño-Pachón[2†], Carolina Escobar[1], Yomara Rozo[1], Luis F. Rivera[1], Andrés Durán[1], Silvia Restrepo[2], Tamar Eilam[3], Yehoshua Anikster[3] and Alvaro L. Gaitán[1]

[1] Plant Pathology, National Center for Coffee Research – CENICAFÉ, Chinchiná, Colombia
[2] Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia
[3] Institute for Cereal Crops Improvement, Tel Aviv University, Tel Aviv, Israel

Edited by:
Sébastien Duplessis, National Institute of Agronomic Research, France

Reviewed by:
Guus Bakkeren, Agriculture and Agri-Food Canada, Canada
Diana Fernandez, Institut de Recherche Pour le Développement, France

***Correspondence:**
Marco A. Cristancho, Department of Plant Pathology, National Center for Coffee Research – CENICAFÉ, Km 4 vía a Manizales, Chinchiná 2427, Colombia
e-mail: marco.cristancho@cafedecolombia.com

† Present address:
Diego Mauricio Riaño-Pachón, Labóratorio Nacional de Ciência e Tecnologia do Bioetanol, Centro Nacional de Pesquisa em Energia e Materiais, Campinas, São Paulo, Brazil

Coffee leaf rust caused by the fungus *Hemileia vastatrix* is the most damaging disease to coffee worldwide. The pathogen has recently appeared in multiple outbreaks in coffee producing countries resulting in significant yield losses and increases in costs related to its control. New races/isolates are constantly emerging as evidenced by the presence of the fungus in plants that were previously resistant. Genomic studies are opening new avenues for the study of the evolution of pathogens, the detailed description of plant-pathogen interactions and the development of molecular techniques for the identification of individual isolates. For this purpose we sequenced 8 different *H. vastatrix* isolates using NGS technologies and gathered partial genome assemblies due to the large repetitive content in the coffee rust hybrid genome; 74.4% of the assembled contigs harbor repetitive sequences. A hybrid assembly of 333 Mb was built based on the 8 isolates; this assembly was used for subsequent analyses. Analysis of the conserved gene space showed that the hybrid *H. vastatrix* genome, though highly fragmented, had a satisfactory level of completion with 91.94% of core protein-coding orthologous genes present. RNA-Seq from urediniospores was used to guide the de novo annotation of the *H. vastatrix* gene complement. In total, 14,445 genes organized in 3921 families were uncovered; a considerable proportion of the predicted proteins (73.8%) were homologous to other Pucciniales species genomes. Several gene families related to the fungal lifestyle were identified, particularly 483 predicted secreted proteins that represent candidate effector genes and will provide interesting hints to decipher virulence in the coffee rust fungus. The genome sequence of Hva will serve as a template to understand the molecular mechanisms used by this fungus to attack the coffee plant, to study the diversity of this species and for the development of molecular markers to distinguish races/isolates.

Keywords: genome, coffee rust, coffee, plant pathogens diversity, RNA-seq, genetic variants

INTRODUCTION

Coffee rust caused by the fungus *Hemileia vastatrix* (Hva) leads to widespread damage to crops worldwide. The disease develops polycyclic epidemics in a season, which means there is an overlapping succession of infection cycles. Under tropical conditions and in semi-perennial plants, such as coffee, Hva poses a permanent threat for producers. In the absence of control methods, the pathogen has been reported to cause losses of up to 30% in susceptible varieties of the species *Coffea arabica* during mild epidemics (Monaco, 1977; Rivillas et al., 2011). The fungus, which is a biotroph that targets *Coffea* as the single known host genus, spread from Africa and was responsible for the collapse of coffee production in India and Ceylon by the mid XIX century. It arrived in America in 1970, and since then, it has quickly disseminated to all the other coffee-producing areas of the continent. In Colombia, Hva was reported for the first time in 1983 in the central coffee-producing zone of the country (Leguizamón et al., 1984). The presence of Hva in almost every coffee plantation in the world has been one of the main drivers for plant breeders to release rust-resistant varieties. Recent outbreaks of the disease have affected major areas of coffee production in Colombia (Rozo et al., 2012) and Central America (Cressey, 2013), and the evidence linked these new epidemics to changes in weather patterns, including rainfall distribution and quantity (Cristancho et al., 2012).

Multiple Hva races have been reported throughout the world (Rodrigues et al., 1975; Carvalho et al., 1987), and studies from the CIFC (Coffee Rust Research Center) in Portugal have identified over 30 races of the pathogen using a series of more than 40 differential coffee genotypes (Rodrigues et al., 1993). Historically,

race II has been predominant in most countries, and it attacks all cultivated varieties of the species *C. arabica* that have not been bred for disease resistance (Rodrigues et al., 1975). In addition to race II, 6 other physiological races have been identified in Colombia using a set of differential plants developed at CIFC (Oeiras, Portugal), attacking some lines of the resistant cultivars (Castillo and Leguizamón, 1992; Gil and Ocampo, 1998; Alvarado and Moreno, 2005; Rozo et al., 2012). At least 10 more isolates not differentiated by CIFC differential plants, remain to be characterized in Colombia, and several other unknown isolates have also been detected elsewhere (Gouveia et al., 2005).

The emergence of new races and more aggressive isolates in plant pathogens threaten agriculture worldwide as recently observed with the wheat stem rust fungus and the new epidemics of coffee rust, which clearly indicate that further detailed studies and continuous monitoring are needed to improve integrated disease management strategies that mitigate their destructive effect (Aime et al., 2006). Being obligate pathogens, the study of rust fungi biology is particularly challenging and needs substantial investments given the fact that a large set of differential plants have to be employed for the classification of races. The development of novel tools for the identification of isolates is critical to study the biology of these major plant pathogens and genomics might offer such tools. Differential Hva genes expressed in urediniospore, appressoria, and haustoria have already been identified (Fernandez et al., 2012; Talhinhas et al., 2014).

Genomic studies of plant pathogens have provided insights into their evolution, the mechanisms that generate genetic variability and the repertoire of genes that are involved in pathogenesis. These studies have shown that rust fungi exhibit very large genome sizes [*Melampsora lini* = Mli (Nemri et al., 2014) is the largest fungal genome so far] compared to other fungi, containing very large numbers of genes, over 16,000 in most cases, compared to other fungi groups such as Ustilaginomycotina [*U. maydis* = 6786 protein coding genes (Schirawski et al., 2010)] or other non-rust Pucciniomycotina such as *Mixia osmundae* (Toome et al., 2014). Rusts also show a large content in transposable elements (i.e., nearly 50%, in the genomes analyzed so far, Duplessis et al., 2014). All these features indicate that rust genomes in general are complex to sequence and assemble due to the repetitive content and large genome size.

The discovery of predicted secreted virulence determinants in plant pathogens has also been possible through genomic analysis. Secreted proteins have been linked to the virulence of plant-pathogenic fungi (Spanu, 2012) and many have been predicted in *Melampsora* spp. (Joly et al., 2010; Hacquard et al., 2012), *P. striiformis* (Cantu et al., 2011, 2013), *P. graminis* (Duplessis et al., 2011), and *H. vastatrix* (Fernandez et al., 2012). Thus, the discovery of the secreted protein genes (Saunders et al., 2012) and the functional demonstration of their decisive role in the infection process help in unraveling previously unknown mechanisms of pathogenicity that operate in biotrophic fungi.

We have obtained genome and transcriptome sequences of the coffee rust fungus, and we expect these data to allow the identification of potential molecular markers for the study of rust isolates/races. The knowledge of the Hva genome and particularly of its secretome is a critical point for understanding the mechanisms used by the fungus during the colonization of coffee tissues and allows for comparisons of pathogenesis processes in other rust fungus-plant interactions. The chimeric genome assembly obtained was further used to define polymorphism between isolates and to analyse its basic contents such as the gene complement and TE families. The predicted proteome was additionally supported by a transcriptome analysis of Hva urediniospores. Within the predicted gene complement, a more precise analysis was performed on predicted secreted proteins, likely containing Hva candidate effectors.

RESULTS

NUCLEAR DNA CONTENT

The nuclear DNA content of 10 Hva samples was measured by Flow cytometry analysis (**Table 1**). Two groups could be distinguished among the 10 samples: one contains four samples that showed a lower content of 1.17–1.29 pg of DNA and a second with the six remaining samples showed a higher content of 1.55–1.76 pg of DNA per urediniospore (∼30% more). Based on these results and compared to the genome size of *P. triticina* used as a control (135 Mb, Puccinia Group Genomes Database, Broad Institute), we estimate the genome size of Hva to be 243–324 Mb.

GENOME ASSEMBLIES

We sequenced the genomes of the following isolates: HvCat, Hv387, Hv949, HvDQ952, HvH179, HvH569, HvH701, and

Table 1 | Nuclear DNA content of Hva urediniospore samples measured by FCM.

Coffea species and genotypes	Geographical location	DNA Content (pg)	
		Group 1	Group 2
C. arabica var. Caturra	La Alcancía, Antioquia	–	1.63
C. arabica var. Caturra	El Cedral, Pereira, Risaralda	–	1.55
C. arabica var. Caturra	Santa María, Antioquia	1.29	–
C. arabica var. Caturra (Acc. 1421)	Chinchiná, Caldas	1.17	–
C. arabica BA-13	Chinchiná, Caldas	–	1.76
C. arabica × *C. canephora*—Timor Hybrid H-419/2	Chinchiná, Caldas	–	1.62
C. arabica × *C. canephora*—Timor Hybrid H-584	Chinchiná, Caldas	1.21	–
C. arabica var. Tipica	Chinchiná, Caldas	1.18	–
C. arabica var. Mundo Novo	Chinchiná, Caldas	–	1.70
C. liberica	Chinchiná, Caldas	–	1.70
	Mean	1.21	1.66
	Standard deviation	0.05	0.08
	CV%	5.87	6.41

Hva urediniospore samples were stained with Propidium Iodide for Flow Cytometry fluorescence measures following the protocol described by Eilam et al. (1994).

HvMar; for isolate HvCat, Illumina and 454 sequencing were combined. We obtained a total of 412 million short-reads from Illumina and 5.8 million reads from 454. The fraction of reads that passed quality filtering was over 85% in all Hva Illumina sequenced samples but only 52% for the 454 Hva sequenced sample (Table S1, see methods section for filtering parameters). A hybrid 454-Illumina assembly was obtained, combining all genomic Hva sequences with the script clc_novo_assembly from the assembler suite CLC Assembly Cell v4.0.1 (CLC bio, Aarhus, Denmark); we have also performed separate assemblies of the genomes of each isolate. Unfortunately due to inherent characteristics and composition of the genomes (low GC content and richness in TE, detailed in the following sections), we were only able to obtain partial genome assemblies. In order to improve the overall genome assembly, our strategy was to generate an hybrid assembly that takes into account all reads obtained from the 8 isolates together to produce a unique sequence that is a chimera of the sequenced isolates. We were able to define a genome sequence of 333 Mb (129X sequencing depth) composed of 396,264 contigs and 302,466 scaffolds.

We assessed the completeness of the hybrid and individual genomes by running CEGMA with a set of 248 ultra-conserved Core Eukaryotic Genes (CEGs) (Parra et al., 2007) (**Table 2**). Statistics for the hybrid assembly are shown in **Table 3**; based on the Hva chimeric assembly data, and using Jellyfish (Marçais and Kingsford, 2011) to compute the number of distinct k-mers of different lengths and their relationship to coverage, we estimated the size of the Hva genome to be 333 Mb. Considering that the HvaHybrid genome sequence was the only one with a fairly good coverage of conserved core genes, we decided to use it as the reference genome for further analysis.

A total of 23.2% of the paired-end reads mapped in the same contig of the HvaHybrid genome assembled. Most unpaired reads (66.8%) matched two different contigs (useful for scaffolding, data not shown). Several contigs displayed coverage greater than 100X, but most of the contigs exhibited low coverage (**Figure 1**); however, over-coverage was pronounced in the short contigs (**Figure 2**). We also illustrated the range of coverage of the contigs and its association to the contigs size (**Figure 3**). The largest contig, Hvcontig_23458 (85 Kbp) showed good coverage. However, the next two contigs in size, Hvcontig_171 (45 kbp) and Hvcontig_161 (71 kbp), showed over-coverage and belong to the Hva mitochondria (see below). Bacterial contamination was very low representing less than 1% of the sequences (**Figure 4**).

Table 3 | Summary of the Hva genome hybrid assembly.

N° Contigs assembled		396,264
N° Scaffolds assembled		302,466
Total residues assembled		333,481,311
Length	Max	85,126
	Average	841.56
	N50	1,59
Reads	Total	336,649,188
	Unassembled	197,88,611
	Assembled	316,860,577
	Multihit	37,520,793
	Potential pairs	
	Paired	78,105,740
	Not Paired	255,469,308

The 454 and Illumina clean reads were assembled and the same reads were mapped against this set of assembled contigs using the software CLC Assembly Cell v4.0.1.

Table 2 | Statistics gathered from the genome assemblies of Hva individual isolates and the hybrid assembly.

Sample IDs	*Coffea* species and genotypes	Raw reads	Clean reads[a]	Contigs	Assembly size	Completeness[b]
HvHybrid[c]		412,417,464	359,076,496	396,264	333,258,024	91.94%
HvCat 454[d]		5,860,446				
HvCat Illumina	*C. arabica* var. Caturra	48,396,016	43,704,716	254,645	122,820,521	57.26%
Hv387	*C. canephora* CII56	58,593,986	50,782,526	211,495	150,707,107	44.35%
Hv494	H89: *C. arabica* var. Bourbon resistant × *C. arabica* CaRCV3	55,326,774	47,738,686	211,728	138,293,025	35.08%
HvDQ952	F2 – *C. arabica* var. Caturra × HdT 1343 (*C. arabica* × *C. canephora*)	43,875,056	38,844,164	197,927	121,119,448	31.85%
HvH_179	H3101: (*C. arabica* CaCV1 × Hdt (*C. arabica* × *C. canephora*) 1343 574CV2) × *C. arabica* CtyR	49,025,718	42,033,780	203,770	131,574,289	31.05%
HvH_569	H3041: (*C. arabica* × HarrarR3) × HdT(*C. arabica* × *C. canephora*) 1343 Africa 1386	51,960,392	45,000,606	202,168	133,358,100	37.50%
HvH_701	H2094: (*C. arabica* MundoNovo) × F502 (*C. arabica* accession from Tanganica)	60,634,018	53,080,264	215,628	158,292,515	56.86%
HvMar_1	*C. arabica* var. Caturra	44,605,504	39,408,690	203,360	125,814,765	22.18%

[a] *Clean reads were obtained after quality trimming and removal of duplicates.*

[b] *Completeness of the genome was calculated running the software CEGMA with a set of 248 ultra-conserved Core Eukaryotic Genes (CEGs) (Parra et al., 2007).*

[c] *The hybrid assembly was generated by the combination of all short reads from the eight isolates.*

[d] *Eight and a half plates were sequenced with 454 technology.*

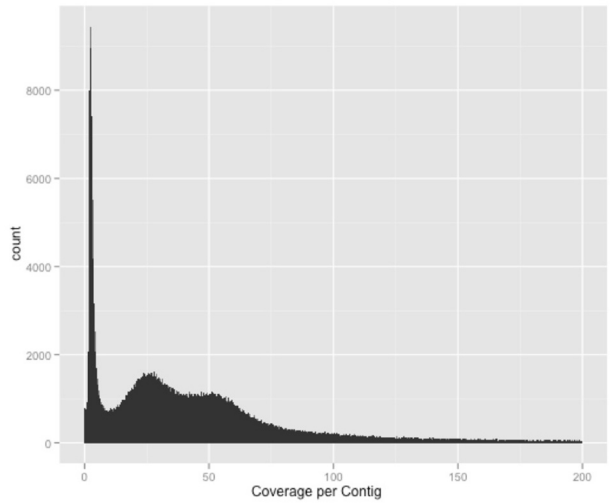


FIGURE 1 | Plot representing the reads coverage per contig in the HvHybrid assembly. Low read coverage was identified in most contigs.

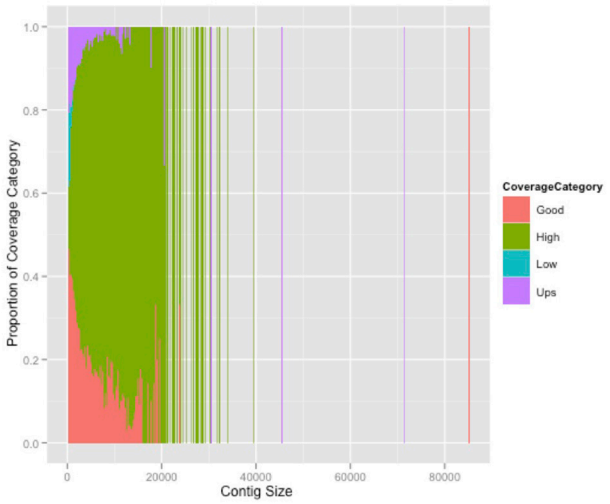


FIGURE 3 | Association between contig size and coverage. Contigs were clustered by color by running an R script: 0–5 low coverage, 5–45 good coverage, 45–100 high coverage and more than 100 over-coverage.

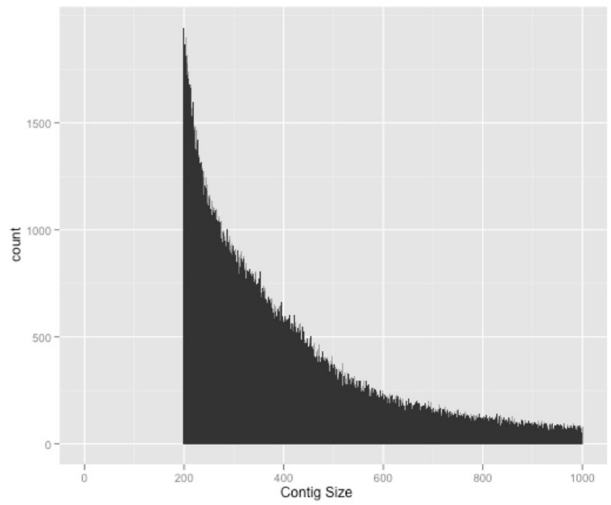


FIGURE 2 | Association between number of reads and size of contigs in the HvHybrid Assembly. Over-coverage was identified in short contigs.

MITOCHONDRIAL GENOME ANNOTATION

We used the *P. graminis* f.sp. *tritici* (PGT) (Puccinia Group Genomes Database, Broad Institute) mitochondrial genome to find homologs in the *HvHybrid* assembly by BLAST ($e = 1e\text{-}5$). BLAST sequence similarities were pictured using the visualizing tool Circoletto (Darzentas, 2010). As shown in **Figure 5**, Hvcontig_161 (71,379 bp, %GC = 33%) and Hvcontig_171 (45,581 bp, %GC = 35%) cover over 70% of the mitochondrial genome of PGT. We also found that 7 contigs of the HvCat assembly entirely covered the mitochondrial genomes of PGT (79,748 bp) and the soybean rust *Phakopsora pachyrhizi* (31,825 bp; Stone et al., 2010) (results not shown). From this analysis we estimate that the Hva mitochondrial genome it is at least of the size of the PGT mitochondrial genome.

DE NOVO IDENTIFICATION OF TRANSPOSONS

The genomes of rust fungi sequenced to date all contain large numbers of transposable elements, which is a major problem for proper assembly (Duplessis et al., 2011; Zheng et al., 2013). A careful annotation of TEs was performed in the chimeric assembly and it showed that the Hva genome also contained a large proportion of repeated sequences. Interspersed repeats were identified in 74.4% of the Hva assembled hybrid contigs using the algorithm RepeatMasker. A similar fraction of repeats was identified in the individual assemblies (71–74%). A large proportion of repeats were classified as LTR elements (38.7%), a smaller proportion was classified as DNA elements (7.2%), only 2.1% were classified as LINEs, and 26.3% of repeats were unclassified.

The sequences were annotated for the presence of LTRs, direct repeats and inverted repeats as well as sequence similarities to repeat sequences from Pucciniales. Four novel retrotransposon families were identified in the Hva genome (Table S2).

PREDICTING PROTEIN-CODING GENES

In order to capture the gene space of the coffee rust genome, we performed a transcriptome analysis of freshly harvested urediniospores based on Illumina RNA-Seq. A total of 44,297–64,752 transcripts could be identified in the three RNA-seq based libraries with the program Trinity, with the HvCatNor normalized library holding the smallest number of genes (Table S3). We aligned those transcripts onto the chimeric and individual samples assemblies and showed that the normalized library contains the largest fraction of Hva expressed mapped genes (Table S4). The normalization approach decreases the prevalence of high abundance transcripts and equalizes transcript concentrations in a cDNA sample, thereby increasing the discovery of low abundance transcripts. The level of contamination of the Hva

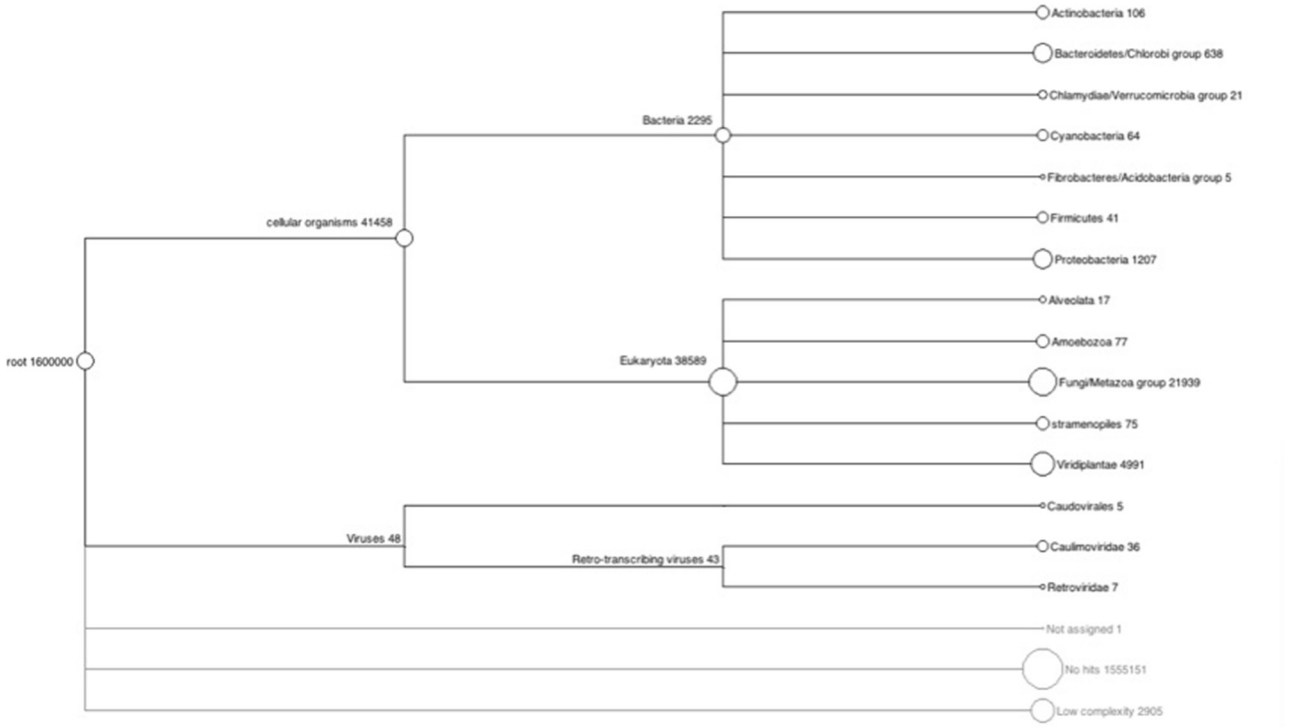


FIGURE 4 | Analysis of sequence contamination in the HvHybrid assembly. The HvHybrid assembled contigs were searched for contaminants with the program Megan4 (Huson et al., 2001). Each node is labeled by a taxon and the number of reads assigned to the taxon. The size of a node (circle size) is scaled logarithmically to represent the number of assigned reads. A low bacterial and plant sequence contamination was identified in the assembly.

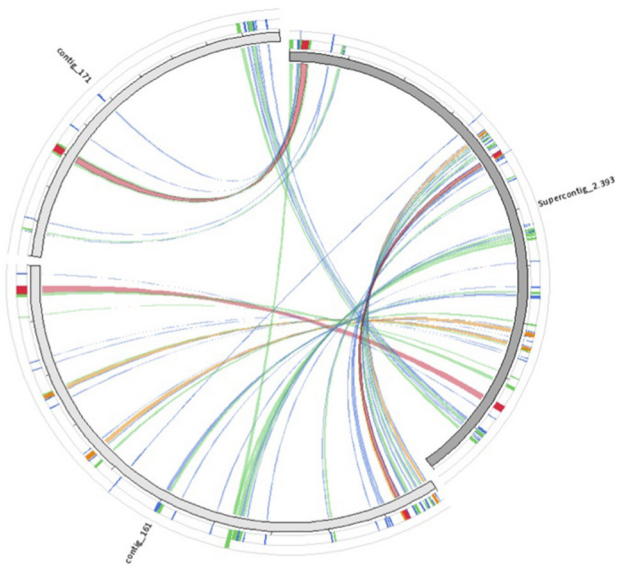


FIGURE 5 | Homology of Hva mitochondrial sequences to the mitochondrial genome of PGT. Clone_161 and clone_171 sequences from HvHybryd assembly were compared to the mitochondrial genome of *P. graminis* f.sp *tritici* (Supercontig_2.393; 79,2Kb) by running BLAST sequence similarities ($e = 1e\text{-}5$). Homologous regions are illustrated with Circoletto (Darzentas, 2010).

RNA-seq datasets with plant, bacterial and other contaminant sequences was moderate; the fraction of contamination for each sample was 13.6% for the normalized library HvCatNor, 12,3% for the HvH420_701 library, and 18,9% for the HvCat955 library (Table S5). We carried out homology annotation with BLAST against Hva germinating urediniospore transcripts dataset and other rusts predicted proteins datasets (**Table 4**). Our Hva urediniospores dataset is as expected very similar to the Hva germinating urediniospores transcripts identified by Talhinhas et al. (2014).

The predictions of the gene coding space was performed using TopHat (Trapnell et al., 2009) for mapping of RNA-seq data against the HvHybrid genome assembly and proteins were predicted with Augustus (Stanke and Waack, 2003; size filter = 70 amino-acids), using the RNA-Seq data as a guide. We identified a total of 21,345 contigs that matched the RNA-seqs and we predicted a total of 18,234 protein sequences with an average length of 1047 bp for the gene models. We identified 13,796 Hva protein homologs (73.86%) in the Pucciniales order (67,118 sequences) using blastp ($e = 1e\text{-}3$).

The total set of sequences was filtered to remove repeats identified in RepBase Release 17.01, and this resulted in a final set containing 14,445 predicted protein-coding gene sequences. Over 96% of this set of gene models was identified in the individual assemblies (Table S6). We explored this set of predicted proteins searching for KOGs in the NCBI Conserved Domain Database; 8458 Hva protein-coding genes having a KOG homolog were

Table 4 | Homology of Hva transcript datasets to Pucciniales predicted proteins.

		Hva urediniospore transcripts		
Hva samples[a]	**Assembled transcripts**	**HvCatNor**	**HvH420_701**	**HvCat955**
I				
Hva (gU)	4267	91.0% (3884)	93.1% (3973)	93.9% (4007)
H. vastatrix (Ap)	3627	59.2% (2147)	63.1% (2290)	63.8% (2315)
H. vastatrix (H)	4465	50.0% (2232)	49.1% (2202)	49.6% (2229)
Rust species	**Predicted proteins**			
II				
Pt	11,630	42.2% (18,703)	40.0% (22,240)	36.6% (23,705)
Pgt	15,979	43.8% (19,411)	41.5% (23,129)	37.8% (24,480)
Pst	22,815	43.5% (19,257)	40.4% (22,544)	36.9% (23,874)
Mlp	16,694	39.2% (17,385)	40.8% (22,765)	37.8% (21,302)

I. Homology sequence analysis between Hva transcriptomes datasets (this study) and Hva germinating urediniospores, appresoria, and haustoria transcript sequences (gU, Ap, H, Talhinhas et al., 2014) was performed with the program BLASTn and an E = 1e^{-20}. II. BLASTx sequence similarity analysis of H. vastatrix RNA-seq sequences and other rust predicted protein datasets (E = 1e^{-3}). Numbers represent fraction (%) of hits found.

[a] Hva samples described in Fernandez et al. (2012).

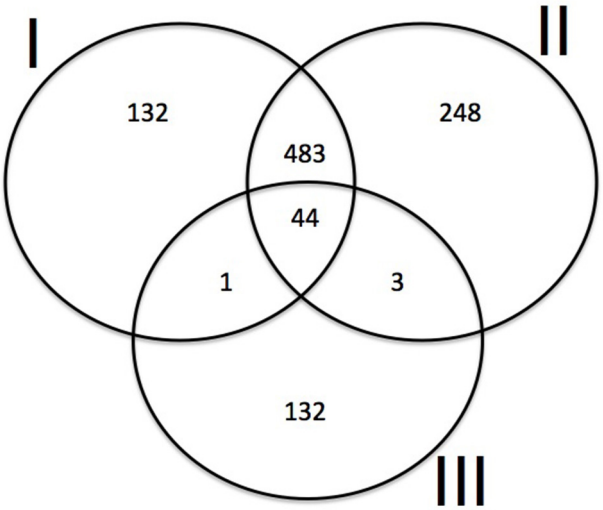


FIGURE 6 | Venn diagram showing the consensus set of Hva secreted predicted proteins. PProwler (0.9 probability cut-off) and SignalP were used to predict secreted proteins. A set of 14,445 putative proteins was used for classification into secreted and non-secreted proteins. The results were compared with Hva secreted proteins predicted by Fernandez et al. (2012). Hva secreted proteins predictions: **(I)** PProwler (this study). **(II)** SignalP (this study). **(III)** Secreted proteins predicted by Fernandez et al. (2012).

functionally annotated (Table S7). A total of 3921 gene families with 2–66 gene members were identified with OrthoMCL (Table S8); we also identified 2103 orphan genes.

SECRETOME ANNOTATION

We predicted 659 secreted proteins using PProwler and 775 secreted proteins with the SignalP algorithm. A total of 180 proteins in our Hva set presented homologs with the secreted proteins already predicted in Hva (Fernandez et al., 2012). A Venn diagram (**Figure 6**) showed shared and unique coincidences between the three sets of data, including 44 proteins extracted by comparison with the dataset predicted by Fernandez et al. (2012). Most of the secreted proteins predicted are organized in gene families and they were mapped with tblastx to at least one contig from each of the individual assemblies; a final set of 28 predicted proteins was obtained after filtering those belonging to the same gene family and they were functionally annotated with blastp against swissprot, RefSeq, Uniref100 and the non-redundant protein sequences databases (Table S9). Only five sequences did not have a homolog sequence with other Pucciniales fungi. Six sequences had a homolog already identified as a secreted protein in *M. larici-populina*. We did not identify in the genome of Hva a homolog of ps87 of *P. striiformis* f.sp. *tritici*, a conserved secreted protein in several fungal plant pathogens (Gu et al., 2011). However, an identical copy of the Hva RTP1 gene (GenBank: FR851895), a transferred protein belonging to the family of effectors in rusts (Pretsch et al., 2013) was identified, suggesting that some effectors are very well conserved between different rust species (Spanu, 2012).

We identified homologs of the predicted secreted proteins in all but one of the Hva individual assemblies; a homolog of protein KF018005 was not identified in the assembly of HvCat. The secreted proteins mapping to the individual assemblies showed that 6 proteins were identical in every Hva sample. The remaining 22 predicted proteins displayed polymorphism in at least two isolates (Table S10).

For Hva proteins KF018008, KF018015, KF018016, KF018020, KF018028 we did not find evidence of predicted homologs in other rusts, representing unique genes of the coffee-rust interaction not represented in other pathosystems. Interestingly, protein KF018020 had no homolog in any database and we could not detect polymorphisms of this protein-coding gene in the Hva individual assemblies, but the Hva genome holds 32 copies of this gene. Protein KF018028, represented by a single copy in the genome, is surprisingly the most diverse of the Hva secreted proteins. It is worth noting that it does not have homologs in any other organism.

PROTEIN KINASES (PKs)

Given the fact that PKs are involved in essential pathways related to development and adaptation to different environments (Miranda and Barton, 2007), we determined differences between PKs families present in pathogen and non-pathogen basidiomycetes. A total of 210 PKs were identified within the set of Hva predicted proteins using HMM3 (Eddy, 2011).

This set of sequences was compared against the predicted PKs of other Pucciniales showing that most protein kinases, including gene families coding for signal transduction pathways, are shared between rust genomes (Table S11). We wanted to know the protein kinases exclusively found in pathogenic basidiomycetes and for that we run a BLASTp homology search of PKs of five pathogenic species, *H. vastatrix, P. graminis* f.sp *tritici*, *P. triticina*,

M. larici-populina, and *U. maydis* against predicted PKs of the non-pathogenic basidiomycetes *Coprinopsis cinerea* (Stajich et al., 2010) and *Laccaria bicolor* (Martin et al., 2008). There are 18 PKs unique to the genomes of plant pathogenic fungi (Table S12) and we identified 236 PKs sequences present in *C. cinerea* and *L. bicolor* but not in pathogenic fungi. In the later group we highlight functions TKL/TKL, PKL/ccin9, PKL/CAK/Fmp29, FunK1, AgaK1, atypical/PIKK/TRRAP, and atypical/HisK PKs families because they were not identified in any of the pathogen species although they are expanded in unique families in *C. cinerea* (Stajich et al., 2010) and *L. bicolor* (Martin et al., 2008).

DISCUSSION

We have generated a de novo hybrid genome assembly of the coffee rust from the sequence of 8 rust samples. We also assembled transcript sequences obtained from normalized and non-normalized RNA-seq libraries representing the urediniospore stage of the fungus. The hybrid genome assembly was the most comprehensive in terms of capturing the largest proportion of the gene space in Hva, therefore offering a picture of a chimeric genome of this species. There appears to be a large extent of repetitive sequences in this chimeric genome, which was evident in our hybrid assembly, as shown by over-covered contigs. For example, two of the three largest contigs showed over-coverage and this event was also found in most contigs shorter than 400 bp. Contamination analysis showed the presence of Bacteria and Viridiplantae sequences in the sequencing reads but the fraction of these sequences was very low. The limitation of having to collect spore samples from plant tissue renders it impossible to have samples free of other organisms; consequently, finding bacterial, plant and other DNA sequences was not unexpected.

We estimated the Hva genome size to be 243–324 Mb by FCM and the assembled scaffolds size was close to the larger figure (333 Mb). Differences between the relative amounts of DNA measured by FCM might be due to the presence of a mixture of different cells containing different number of nuclei in the Hva samples tested; Hva urediniospores in coffee leaves carry out meiosis giving rise to spores at different stages of development containing unequal numbers of nuclei in a process referred as cryptosexuality (Carvalho et al., 2011). A similar mechanism of parasexual recombination has been described in *P. triticina* (Wang and McCallum, 2009). The haplophase has not been recognized in Hva and only urediniospores and teliospores representing the dikaryophase have been identified (De Castro et al., 2009).

Our Hva genome size estimates fall remarkably short of recent measurements of 733 Mbp (Carvalho et al., 2014) and 796.8 Mbp (Tavares et al., 2014) obtained by FCM of Hva nuclei isolated from urediniospores. Highly repetitive genomes such as the Hva genome are complex to sequence and analyze and genomes with a high content of repeats are difficult to sequence completely (Sun et al., 2003). Overall, the Hva genome size is larger compared with other fungal genomes, including the Basidiomycetes *M. larici-populina* (101.1 Mb), *P. graminis* f.sp. *tritici* (88.6 Mb) (Duplessis et al., 2011), and *Laccaria bicolor* (68.9 Mb) (Martin et al., 2008), the arbuscular mycorrhizal fungus *Rhizophagus irregularis* (153 Mb) (Tisserant et al., 2013), and the Ascomicota *Tuber melanosporum* (125 Mb) (Martin et al., 2010), and *Blumeria graminis* f.sp. *tritici* (174 Mb) (Parlange et al., 2011).

Our analysis indicated that the Hva mitochondrial genome is at least the size of the *P. graminis* mitochondria. The different genome size estimates obtained so far make imperative the assembly of an Hva genome from a single Hva isolate to clearly elucidate the real nuclear genome size, mitochondrial genome size and fraction of repetitive sequences for this fungus.

We identified a large fraction of repetitive sequences in the hybrid genome; 74.4% of the assembled contigs contain repetitive sequences, with most of them representing transposable elements. Because of the hybrid nature of the assembly, this might be and over-estimate of the real fraction of repetitive sequences present in the genome. However, given the large estimates for the Hva genome size, we expect the genome sequence to contain a large proportion of repeats. A high proportion of transposable elements have also been identified in the genomes of other rusts (Duplessis et al., 2011; Zheng et al., 2013) and the plant pathogen *Blumeria graminis* (Spanu et al., 2010). Genome expansion caused by the replication of TEs has been shown to occur in filamentous fungal and oomycete pathogens of plants, and some expansion of virulence-related genes are associated with their large genome size (Kemen and Jones, 2012). The high diversity of many *Avr*-genes in the rice blast fungus *Magnaporthe grisea* is related to their association with repeated sequences (Huang et al., 2014). On the other hand, the non-pathogen basidiomycetes *L. bicolor* (Martin et al., 2008) and *C. cinerea* (Stajich et al., 2010) harbor a much-reduced proportion of repeated sequences. Whether the genome of Hva has suffered an expansion of virulence-related genes mediated by transposition events should be investigated in further detail. Given the fact that a hybrid genome might contain an over-representation of the fraction of repetitive elements present in single genomes, there is still need to be cautious about the final proportion of repeats in the Hva genome.

The assembly exhibited a high level of fragmentation as shown by the large number of scaffolds obtained in the final assembly and the low N50 value. This fragmentation can be explained by the highly repetitive nature of the Hva genome. It should be possible in the future to improve this assembly by sequencing large insert libraries that will aid in resolving the repetitive nature and to enlarge scaffolds of the Hva assembly (Raffaele and Kamoun, 2012). An additional approach that might be implemented to improve our current hybrid assembly would be to use a "fosmid-to-fosmid" strategy as that followed by Zheng et al. (2013), who significantly improved an earlier assembly of the *P. striiformis* f.sp. *triticina* genome (Cantu et al., 2011). The GC content of the Hva genome (33%) was lower than *M. larici-populina* (41%), and *P. graminis* f.sp *tritici* (43.3%) (Duplessis et al., 2014). This difference could be explained by GC repetitive sequences collapsing into contigs, therefore yielding a GC content reduction because GC sequences are underrepresented. Though the hybrid Hva genome assembly was highly fragmented, the CEGMA analysis indicated that a significant amount of the genome's gene space was revealed and we consider the current hybrid assembly to be representative of the gene space of a chimeric Hva genome.

Comparative genomics showed considerable similarities between Hva and other rust fungal genomes; over 73% of Hva

predicted proteins had homologs among Pucciniales protein datasets. Although rust genomes vary in size, they are very similar in gene content suggesting the presence of a large core set of rust fungus specific genes needed for their pathogenicity. It will also be significant to study the function and specificity of Hva predicted proteins not found in other rusts and study their virulence species-specific adaptations. All in all this set of Hva predicted proteins represent a valuable resource that contributes to the Pucciniales gene repertoire.

In order to capture the gene space of the coffee rust genome, we performed a transcriptome analysis of freshly harvested urediniospores based on Illumina RNA-Seq. The number of secreted proteins predicted in this study is smaller than the number found in *M. larici-populina* (1184 SSPs) and *P. graminis* f.sp. *tritici* (1106 SSPs) genomes (Duplessis et al., 2011), perhaps reflecting the Hva partial genome assembled. Also, it has to be considered that *M. larici-populina* and *P. graminis* f.sp. *tritici* SSPs were predicted with the SignalP, TargetP, and TMHMM algorithms while we did not include TMHMM in our predictions. The non-inclusion of TMHMM transmembrane protein predictions in some way renders our current set of secreted proteins incomplete. On the whole, this set of Hva predicted secreted-proteins is a basic tool for the identification of pathogenicity-related genes as shown for other rusts (Joly et al., 2010; Cantu et al., 2011). It is possible that avirulence elicitors be present among the set of predicted secreted proteins, as it has been found in flax rust (Catanzariti et al., 2006). For Hva proteins KF018008, KF018015, KF018016, KF018020, KF018028 we did not find evidence of predicted homologs in other rusts, representing unique genes of the coffee-rust interaction not represented in other pathosystems; secreted proteins have been found to be lineage-specific in other rusts as well (Duplessis et al., 2011). Overall analysis of the Hva predicted secretome shows that secreted proteins are well conserved among plant rusts and that they include functions most likely involved in the pathogenesis of the fungus. Therefore, this group of annotated secreted proteins suits well as prime candidates for functional testing.

The Hva genome contains most of the gene families coding for signal transduction pathways identified in the genomes of other rust fungi. It is assumed that these gene families are involved in signal perception mechanisms of rust urediniospores (Duplessis et al., 2011) and gives them a highly specialized mechanism for the detection of stomata (Kemen and Jones, 2012). We have grouped the candidate PKs with signal perception roles related to pathogenesis in 18 gene families, those identified in pathogen basidiomycetes but absent in non-pathogenic species.

Illumina and 454 sequencing was used to generate a draft genome in different Hva isolates. Due to the complexity of the genome sequence—similar to other rust fungi- a minimal draft chimeric genome was defined by considering the genome of the different isolates altogether. The genome sequence is a novel resource in Pucciniales, a group that includes many species that are economically major diseases of several crops. It provides data to study the evolution of this important group of plant pathogens. The draft genome sequence of Hva will serve as a template for future assemblies of isolates of this fungus and to understand the molecular mechanisms used by this pathogen to attack the coffee plant and to study its diversity. It will also be the basis for the development of molecular markers to distinguish races/isolates given the enormous difficulties of trying to identify coffee rust races by the use of differential plants. The genomic data of the coffee leaf rust presented here are a reference to track changes in field populations, to characterize the decline in sensitivity against widely used fungicides such as triazoles and strobilurins that are used in coffee rust disease management, and to preserve genotype identity in fungal collections. The increased use of coffee rust-resistant varieties will increase the selective pressure to favor complex fungal genotypes, and resources such as the secretome set is the cornerstone for the development of innovative resistance mechanisms to control this pathogen.

MATERIALS AND METHODS

NUCLEAR DNA CONTENT ESTIMATED BY FCM

Flow cytometry (FCM) was used to estimate nuclear DNA content in urediniospores of *H. vastatrix*, following the protocol described by Eilam et al. (1994), modified for the uredinial stage. Urediniospores were suspended in 0.1% Tween 20 in water for 20 min. The suspension was incubated for 2 min. in a 1000-watt microwave on 50% power level, adding Propidium Iodide and RNase to final concentrations of 4 μg/ml and 50 μg/ml respectively, and incubated for 1 h at 37°C. The urediniospores samples were run on a Bacton Dickinson FACS IV flow cytometer. Urediniospores of *P. triticina* were used as a control. Data from the flow cytometer were analyzed using the Flowing Software v2.5 at the Centre for Biotechnology University of Turku, Finland. Urediniospores of Hva and *P. triticina* were stained and analyzed simultaneously, with the standard control positioned on channel 200. The *C*-Value (pg) was converted to base pairs (bp), considering that 1pg = 978 Mb (Dolezel et al., 2003). *H. vastatrix* samples used for FCM are described in **Table 1**.

GENOME AND TRANSCRIPTOME SEQUENCING AND ASSEMBLY

Hva urediniospore samples were scraped from infected coffee leaves taking care to sample very young pustules with no evidence of the presence of the hyperparasitic fungus *Lecanicillium lecanii*. The coffee genotypes sampled for Hva and the sequencing technologies used are described in **Table 2**. DNA was extracted using the DNeasy Plant Mini-Kit (Qiagen, Hilden, Germany); *H. vastatrix* DNA samples were used to construct 100 bp paired-end libraries and sequenced by Illumina™ HiSeq 2000 at BGI in China. Single-end libraries were sequenced by ROCHE™ 454 GS FLX Titanium method at Macrogen in Korea.

Reads were subjected to quality control checks using FastQC (Babraham Bioinformatics, Babraham Institute), trimmed using the CLC quality_trim script (CLC bio, Aarhus, Denmark), masked or filtered by low complexity end regions, and exclusion of reads shorter than 70 nucleotides. Mdust and SeqClean were used for the cleaning process (The Gene Index Project, Harvard University—http://sourceforge.net/projects/seqclean/files/). Trimmed and filtered reads were assembled with the CLC Assembly Cell v4.0.1 (CLC bio, Aarhus, Denmark) with the following parameters: deletions penalty = 3, no global alignment, remove duplicates, min contig length = 200 bp, paired-end distance = 200–400 bp. The quality of the

assembly was assessed with CLC tools (clc_assembly_viewer, assembly_info) and in-house R scripts available at (http://bioinformatics.cenicafe.org/index.php/wiki/Third_Hybrid_Assembly_of_454_and_Illumina_data_with_CLC).

The hybrid assembly was analyzed using MEGAN 4 (Huson et al., 2001) to assess the level of possible contamination and to perform a first approximation of the biological communities associated with Hva on the coffee leaf. Blastx was performed using the contigs from the hybrid assembly (Illumina + 454 short reads) (396,264 contigs) against the NCBI non-redundant protein database. An *E*-value of $10e^{-3}$ was used as a cut-off following the recommendation from the MEGAN developers. MEGAN was used to map and visualize the Low Common Ancestor (LCA) in the NCBI tree taxonomy for each contig. With the aim of filtering out putative contaminated sequences, contigs that presented similarities to reported fungal sequences were extracted to form a reliable set of Hva genome contigs. The reliable set of *H. vastatrix* genome contigs was compared against the *P. graminis* f.sp. *tritici* and *P. pachyrhizi* mitochondrial genomes using Blastn (with an *E*-value threshold of $1E^{-5}$).

For RNA-seq sample preparation, Hva urediniospores were scraped from infected coffee leaves of the coffee genotypes described in Table S3, taking care to sample very young pustules with no evidence of the presence of the hyperparasitic fungus *Lecanicillium lecanii*. RNA was extracted from urediniospores using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA). Normalized library construction was performed at Evrogen, Moscow, Russia using Kamchatka crab duplex-specific nuclease (Zhulidov et al., 2004). First-strand cDNA was prepared from poly(A)+ *H. vastatrix* urediniospores RNA using a SMART™ PCR cDNA Synthesis Kit (Clontech), according to the manufacturer's protocol. SMART™ Oligo II and CDS primers (Clontech) were used for first-strand cDNA synthesis. A 1.5 ml aliquot of a 100 ng/ml of the first-strand cDNA solution was incubated for normalization with 0.25 Kunitz units of duplex-specific nuclease from kamchatka crab and amplified by PCR. Sequencing of amplified cDNA products was performed on an Illumina™ HiSeq 2000 system (BGI, Shenzhen, 518083, China).

RNA-seq data were filtered before assembly. The quality of the transcripts was measured using the FASTX-Toolkit, reads were trimmed by quality and duplicates were removed. Clean reads longer than 200bp were assembled using the Trinity package (Grabherr et al., 2011). First, the reads were run through Trinity's Inchworm module, which assembles the read data set into different pools of reads, and the Chrysalis module was used to construct de Brujin graphs for all the read pools obtained using Inchworm. We used the module Butterfly that reconciles de Brujin graphs using the read pools from the former modules and output assembled contigs. We mapped Hva transcript datasets to the HvHybrid 454-Illumina assembly and we also compared transcripts against the NR database to identify plant, bacterial and other contaminant sequences.

TRANSPOSABLE ELEMENTS PREDICTION

We surveyed the frequency and classes of TE-like elements present in the HvHybrid assembly using the algorithm RepeatMasker (Smit et al., 1996-2004) and the RepBase12.12 and fngrep.ref databases. The fngrep.ref database included 1726 transposable elements identified in fungi. Novel retrotransposon families were manually annotated from the gene families identified with OrthoMCL (see below).

GENE PREDICTION

For the prediction of gene models, we followed the "align then assemble" approach (Martin and Wang, 2011). We mapped RNA-seq short reads to the genome using TopHat (Trapnell et al., 2009), and we identified putative transcriptional units using Augustus (Stanke and Waack, 2003). Protein sequences were computationally deduced from the transcriptional units. Gene families from predicted proteins larger than 70 amino acids were identified with OrthoMCL using a default MCL inflation value of 1.5 and a blastp *e*-value of $10e^{-5}$ (Li et al., 2003). We explored the set of predicted proteins, searching for KOGs using the CD-Search Tool and the Conserved Domain Database (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

COMPARATIVE GENOMICS

The Hva genome contigs were aligned against the genomes of *P. graminis*, *M. larici-populina* and *U. maydis* using Mauve (Darling et al., 2004). The Low Collinear Block (LCB) values were set through visual inspection by searching the best block size for each pair of alignments (largest coverage of both genomes). Finally, values used for LCB were as follows: *P. graminis* 12,154, *M. larici-populina* 10,409, and *U. maydis* 1203.

For genome annotation, we used custom Perl scripts and basic bioinformatics software such as BLAST (Altschul et al., 1990). The databases we used for comparisons corresponded to 67,118 Pucciniales sequences comprising 16,694 protein-coding genes from *M. larici-populina* (Duplessis et al., 2011), 22,815 *P. striiformis* f.sp. *tritici* sequences (Cantu et al., 2011), 15,979 *P. graminis* f.sp. *tritici* sequences (Duplessis et al., 2011), and 11,630 *P. triticina* sequences (Xu et al., 2011).

For homology searches of protein kinases (PKs) we run BLASTp (version 2.2.28) with an $e = 1e^{-10}$. We searched *H. vastatrix* predicted proteins against 131 predicted PKs from *Sacharomyces cerevisiae* and then we compared the coffee rust PKs against *M. laricis-populina*, *P. striiformis* f.sp *tritici*, *P. graminis* f.sp *tritici*, *P. triticina,* and *U. maydis* predicted PKs.

SECRETED PROTEINS

The *H. vastatrix* predicted proteins were classified into secreted and non-secreted proteins. For this task, the programs SignalP 4.0 (Petersen et al., 2011) and PProwler (Hawkins and Boden, 2006) were used to predict putatively secreted proteins. A 0.9 probability cut-off was used for PProwler predictions. A set of secreted proteins predicted previously for *H. vastatrix* by Fernandez et al. (2012) was used for comparison with our predictions. Briefly, a Blastp was performed between our set of *H. vastatrix* proteins and the predictions by Fernandez et al. (2012). Finally, a set of proteins that showed similarity (Blastp $e = 1e^{-5}$) with the secreted proteins predicted by Fernandez et al. (2012) was obtained. Reciprocal comparisons of the three sets of secreted proteins were performed (SignalP, PProwler and

Fernandez-Blastp) to establish the proteins shared by the three predictions.

AVAILABILITY

Raw data and metadata for the Genome project is available at NCBI, BioProject ID: PRJNA188788 and the Transcriptome project ID: PRJEB2960. Predicted and secreted proteins are available at http://bioinformatics.cenicafe.org/index.php/wiki/CoffeeRustPredictedProteins.

The hybrid reference assembled genome contigs are available for download at: http://bioinformatics.cenicafe.org/index.php/wiki/CoffeeRustHybridDraftAssembly_Contigs.

ACKNOWLEDGMENTS

The authors wish to thank the National Federation of Coffee Growers of Colombia, the Ministry of Agriculture and Rural Development and the Colombian Administrative Department of Science, Technology and Innovation—Colciencias for financial support. Authors would also like to thank Perttu Terho from the Centre for Biotechnology, University of Turku, Finland for data analysis from the flow cytometer experiment. Diego Mauricio Riaño-Pachón acknowledges funding provided by the Faculty of Sciences at Universidad de los Andes through the Assistant Professor Support Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00594/abstract

REFERENCES

Aime, C., Matheny, P. B., Henk, D. A., Frieders, E. M., Nilsson, R. H., Piepenbring, M., et al. (2006). An overview of the higher-level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences. *Mycologia* 98, 896–905. doi: 10.3852/mycologia.98.6.896

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Alvarado, G., and Moreno, G. (2005). Cambio de la virulencia de *Hemileia vastatrix* en progenies de Caturra x Híbrido de Timor. *Cenicafe* 56, 110–126.

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f.sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f.sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Carvalho, A., Eskes, A. B., Castillo, J., Sreenivasan, M., Echeverri, J., Fernandez, C., et al. (1987). "Breeding programs," in *Coffee Rust: Epidemiology, Resistance, and Management*, eds A. C. Kushalappa and A. B. Eskes (Boca Raton, FL: CRC Press), 293–336

Carvalho, C. R., Fernandes, R. C., Carvalho, G. M. A., Barreto, R. W., and Evans, H. C. (2011). Cryptosexuality and the genetic diversity paradox in coffee rust, *Hemileia vastatrix*. *PLoS ONE* 6:e26387. doi: 10.1371/journal.pone.0026387

Carvalho, G. M. A., Carvalho, C. R., Barreto, R. W., and Evans, H. C. (2014). Coffee rust genome measured using flow cytometry: does size matter? *Plant Pathol.* 63, 1022–1026. doi: 10.1111/ppa.12175

Castillo, J., and Leguizamón, J. (1992). Virulencia de *Hemileia vastatrix* determinada por medio de plantas diferenciales de café en Colombia. *Cenicafe* 43, 114–124

Catanzariti, A. M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Cressey, D. (2013). Coffee rust regains foothold. *Nature* 493, 587. doi: 10.1038/493587a

Cristancho, M. A., Rozo, Y., Escobar, Y., Rivillas, C. A., and Gaitán, A. L. (2012). Outbreak of coffee leaf rust (*Hemileia vastatrix*) in Colombia. *New Dis. Rep.* 25, 2044–0588. doi: 10.5197/j.2044-0588.2012.025.019

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1397. doi: 10.1101/gr.2289704

Darzentas, N. (2010). Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* 26, 2620–2621. doi: 10.1093/bioinformatics/btq484

De Castro, R., Evans, H. C., and Barreto, R. W. (2009). Confirmation of the occurrence of teliospores of *Hemileia vastatrix* in Brazil with observations on their mode of germination. *Trop. Plant Pathol.* 34, 108–113. doi: 10.1590/S1982-56762009000200005

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry A* 51, 127–128. doi: 10.1002/cyto.a.10013

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). Advancing knowledge on biology of rust fungi through genomics. *Adv. Bot. Res.* 70, 173–209. doi: 10.1016/B978-0-12-397940-7.00006-9

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comp. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Eilam, T., Bushnell, W. R., and Anikster, Y. (1994). Relative nuclear DNA content of rust fungi estimated by flow cytometry of Propydium iodide-stained pycniospores. *Phytopathology* 84, 728–735. doi: 10.1094/Phyto-82-1212

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A. S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant–rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Gil, L. F., and Ocampo, J. D. (1998). Identificación de la raza XXII (V5-6) de *Hemileia vastatrix* Berk. y Br. en Colombia. *Cenicafé* 49, 340–344

Gouveia, M., Ribeiro, A., Várzea, V., and Rodrigues, C. J. Jr. (2005). Genetic diversity in *Hemileia vastatrix* based on RAPD markers. *Mycologia* 97, 396–404. doi: 10.3852/mycologia.97.2.396

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Gu, B., Shiv, D., Kale, Q. W., Dinghe, W., Qiaona, P., Hua, C., et al. (2011). Rust secreted protein Ps87 is conserved in diverse fungal pathogens and contains a RXLR-like motif sufficient for translocation into plant cells. *PLoS ONE* 6:e27217. doi: 10.1371/journal.pone.0027217

Hacquard, S., Joly, D. L., Lin, Y. G., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (Poplar Leaf Rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hawkins, J., and Boden, M. (2006). Detecting and sorting targeting peptides with recurrent networks and support vector machines. *J. Bioinform. Comput. Biol.* 4, 1–18. doi: 10.1142/S0219720006001771

Huang, J., Si, W., Deng, Q., Li, P., and Yang, S. (2014). Rapid evolution of avirulence genes in rice blast fungus *Magnaporthe oryzae*. *BMC Genet.* 15:45. doi: 10.1186/1471-2156-15-45

Huson, D. H., Mitra, S., Weber, N., Ruscheweyh, H., and Schuster, S. C. (2001). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Joly, D. L., Feau, N., Tanguay, P., and Hamelin, R. C. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11:422. doi: 10.1186/1471-2164-11-422

Kemen, E., and Jones, J. D. (2012). Obligate biotroph parasitism: can we link genomes to lifestyles? *Trends Plant Sci.* 17, 448–457. doi: 10.1016/j.tplants.2012.04.005

Leguizamón, J. E., Baeza, C. A., Fernández, O., Moreno, G., Castillo, Z. J., and Orozco, F. J. (1984). Identification of race II of *Hemileia vastatrix* Berk y Br. in Colombia. *Cenicafé* 35, 26–28

Li, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, F., Aerts, A., Ahrén, D., Brun, A., Danchin, E. G., Duchaussoy, F., et al. (2008). The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452, 88–92. doi: 10.1038/nature06556

Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P. M., Jaillon, O., et al. (2010). Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* 464, 1033–1038. doi: 10.1038/nature08867

Martin, J. A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet*. 12, 671–682. doi: 10.1038/nrg3068

Miranda, D., and Barton, G. J. (2007). Classification and functional annotation of eukaryotic protein kinases. *Proteins* 68, 893–914. doi: 10.1002/prot.21444

Monaco, L. C. (1977). Consequences of the introduction of coffee rust into Brazil. *Ann. N.Y. Acad. Sci*. 287, 57–71. doi: 10.1111/j.1749-6632.1977.tb34231.x

Nemri, A., Saunders, D. G., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Parlange, F., Oberhaensli, S., Breen, J., Platzer, M., Taudien, S., and Šimková, H., et al. (2011). A major invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome. *Funct. Integr. Genomics* 11, 671–677. doi: 10.1007/s10142-011-0240-5

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Petersen, T., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins-a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Puccinia Group Genomes Database. Available online at: http://www.broadinstitute.org/annotation/genome/puccinia_group/GenomeDescriptions.html#P_triticina_1_1_V1.

Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10, 417–430. doi: 10.1038/nrmicro2790

Rivillas, C., Serna, C., Cristancho, M., and Gaitán, A. (2011). *Roya del Cafeto en Colombia: Impacto, Manejo y Costos del Control*. Chinchiná: Boletín Técnico No. 36, Cenicafe.

Rodrigues, C. J. Jr., Bettencourt, A. J., and Rijo, L. (1975). Races of the pathogen and resistance to coffee rust. *Ann. Rev. Phytopathol.* 13, 49–70. doi: 10.1146/annurev.py.13.090175.000405

Rodrigues, C. J. Jr., Várzea, V., Godinho, I. L., Palma, S., and Rato, R. C. (1993). "New physiologic races of *Hemileia vastatrix*," in *Proceedings of the 15th International Conference on Coffee Science* (Montpellier), 318–321.

Rozo, Y., Escobar, C., Gaitán, A. L., and Cristancho, M. A. (2012). Aggressiveness and genetic diversity of *Hemileia vastatrix* during an epidemic in Colombia. *J. Phytopathol*. 160, 732–740. doi: 10.1111/jph.12024

Saunders, D. G. O., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Rafaelle, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:e29847. doi: 10.1371/journal.pone.0029847

Schirawski, J., Mannhaupt, G., Münch, K., Brefort, T., Schipper, K., Doehlemann, G., et al. (2010). Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* 330, 1546–1548. doi: 10.1126/science.1195330

Smit, A., Hubley, R., and Green, P. (1996-2004). *RepeatMasker Open-3.0*. Available online at: http://www.repeatmasker.org.

Spanu, P. D. (2012). The genomics of obligate (and nonobligate) biotrophs. *Annu. Rev. Phytopathol.* 50, 91–109. doi: 10.1146/annurev-phyto-081211-173024

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330, 1543–1546. doi: 10.1126/science.1194573

Stajich, J. E., Wilke, S., Ahrén, D., Au, C. A., Birren, B. W., Borodovsky, M., et al. (2010). Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc. Natl. Acad. Sci. U.S.A*. 107, 11889–11894. doi: 10.1073/pnas.1003391107

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl. 2), ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Stone, C. L., Buitrago, M. L., Boore, J. L., and Frederick, R. D. (2010). Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *Phakopsora pachyrhizi* and *P. meibomiae*. *Mycologia* 102, 887–897. doi: 10.3852/09-198

Sun, X., Le, H. D., Wahlstrom, J. M., and Karpen, G. H. (2003). Sequence analysis of a functional Drosophila centromere. *Genome Res.* 13, 182–194. doi: 10.1101/gr.681703

Talhinhas, P., Azinheira, H., Vieira, B., Loureiro, A., Tavares, S., Batista, D., et al. (2014). Overview of the functional virulent genome of the coffee leaf rust pathogen *Hemileia vastatrix* with an emphasis on early stages of infection. *Front. Plant Sci.* 5:88. doi: 10.3389/fpls.2014.00088

Tavares, S., Ramos, A. P., Pires, A. S., Azinheira, H. G., Caldeirinha, P., Link, T., et al. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. *Front. Plant Sci.* 5:422. doi: 10.3389/fpls.2014.00422

Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., et al. (2013). Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20117–20122. doi: 10.1073/pnas.1313452110

Toome, M., Ohm, R. A., Riley, R. W., James, T. Y., Lazarus, K. L., Henrissat, B., et al. (2014). Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. *New Phytol.* 202, 554–564. doi: 10.1111/nph.12653

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Wang, X., and McCallum, B. (2009). Fusion body formation, germ tube anastomosis, and nuclear migration during the germination of urediniospores of the wheat leaf rust fungus, *Puccinia triticina*. *Phytopathology* 99, 1355–1364. doi: 10.1094/PHYTO-99-12-1355

Xu, J., Linning, R., Fellers, J., Dickinson, M., Zhu, W., Antonov, I., et al. (2011). Gene discovery in EST sequences from the wheat leaf rust fungus *Puccinia triticina* sexual spores, asexual spores and haustoria, compared to other rust and corn smut fungi. *BMC Genomics* 12:161. doi: 10.1186/1471-2164-12-161

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4:2673. doi: 10.1038/ncomms3673

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32:e3. doi: 10.1093/nar/gnh031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 11 October 2014; published online: 31 October 2014.
Citation: Cristancho MA, Botero-Rozo DO, Giraldo W, Tabima J, Riaño-Pachón DM, Escobar C, Rozo Y, Rivera LF, Durán A, Restrepo S, Eilam T, Anikster Y and Gaitán AL (2014) Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales. Front. Plant Sci. ***5****:594. doi: 10.3389/fpls.2014.00594*
This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.
Copyright © 2014 Cristancho, Botero-Rozo, Giraldo, Tabima, Riaño-Pachón, Escobar, Rozo, Rivera, Durán, Restrepo, Eilam, Anikster and Gaitán. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 26 August 2014
doi: 10.3389/fpls.2014.00422

Genome size analyses of Pucciniales reveal the largest fungal genomes

Sílvia Tavares[1,2], Ana Paula Ramos[3], Ana Sofia Pires[1,2], Helena G. Azinheira[1,3], Patrícia Caldeirinha[4], Tobias Link[5], Rita Abranches[2], Maria do Céu Silva[1,3], Ralf T. Voegele[5], João Loureiro[4] and Pedro Talhinhas[1,2,3]*

[1] Centro de Investigação das Ferrugens do Cafeeiro, BioTrop, Instituto de Investigação Científica Tropical, Oeiras, Portugal
[2] Plant Cell Biology Laboratory, Instituto de Tecnologia Química e Biológica António Xavier, Universidade Nova de Lisboa, Oeiras, Portugal
[3] CEER-Biosystems Engeneering, Instituto Superior de Agronomia, Universidade de Lisboa, Lisbon, Portugal
[4] Department of Life Sciences, Centre for Functional Ecology, University of Coimbra, Coimbra, Portugal
[5] Institut für Phytomedizin, Universität Hohenheim, Stuttgart, Germany

Edited by:
Sébastien Duplessis, INRA, France

Reviewed by:
Leen Leus, ILVO, Belgium
Merje Toome, Ministry of Primary Industries, New Zealand

***Correspondence:**
Pedro Talhinhas, Centro de Investigação das Ferrugens do Cafeeiro, BioTrop, Instituto de Investigação Científica Tropical, Quinta do Marquês, 2784-505 Oeiras, Portugal
e-mail: ptalhinhas@iict.pt

Rust fungi (Basidiomycota, Pucciniales) are biotrophic plant pathogens which exhibit diverse complexities in their life cycles and host ranges. The completion of genome sequencing of a few rust fungi has revealed the occurrence of large genomes. Sequencing efforts for other rust fungi have been hampered by uncertainty concerning their genome sizes. Flow cytometry was recently applied to estimate the genome size of a few rust fungi, and confirmed the occurrence of large genomes in this order (averaging 225.3 Mbp, while the average for Basidiomycota was 49.9 Mbp and was 37.7 Mbp for all fungi). In this work, we have used an innovative and simple approach to simultaneously isolate nuclei from the rust and its host plant in order to estimate the genome size of 30 rust species by flow cytometry. Genome sizes varied over 10-fold, from 70 to 893 Mbp, with an average genome size value of 380.2 Mbp. Compared to the genome sizes of over 1800 fungi, *Gymnosporangium confusum* possesses the largest fungal genome ever reported (893.2 Mbp). Moreover, even the smallest rust genome determined in this study is larger than the vast majority of fungal genomes (94%). The average genome size of the Pucciniales is now of 305.5 Mbp, while the average Basidiomycota genome size has shifted to 70.4 Mbp and the average for all fungi reached 44.2 Mbp. Despite the fact that no correlation could be drawn between the genome sizes, the phylogenomics or the life cycle of rust fungi, it is interesting to note that rusts with Fabaceae hosts present genomes clearly larger than those with Poaceae hosts. Although this study comprises only a small fraction of the more than 7000 rust species described, it seems already evident that the Pucciniales represent a group where genome size expansion could be a common characteristic. This is in sharp contrast to sister taxa, placing this order in a relevant position in fungal genomics research.

Keywords: flow cytometry, *Gymnosporangium confusum*, mycological cytogenomics, nuclear DNA content, rust fungi

INTRODUCTION

The Pucciniales (Fungi, Basidiomycota, Pucciniomycotina) represent the largest group of fungal plant pathogens. They are characterized by orange, brown or red colored spore masses (**Figure 1**) appearing on the host tissue surface. Rust fungi are obligate biotrophs, depending entirely on living host cells to complete their biological cycle (Cummins and Hiratsuka, 2003). Their life cycles are diverse, both in terms of the number of spore types produced (micro-, hemi-, demi-, or macrocyclic) and their requirement (or not) of alternate hosts for life cycle completion (autoecious or heteroecious) (for a recent review see Fernandez et al., 2013). Karyogamy occurs in teliospores that germinate to produce basidia, the structure where meiosis takes place. Teliospores are thus responsible for sexual reproduction (Aime, 2006). Rust fungi are generally highly specialized pathogens frequently having narrow host ranges, and consequently they share a common evolutionary history with their host plants (Duplessis et al., 2011b). Rust fungi are able to infect plants from most families, including conifers, ferns and mosses, and are responsible for major diseases on agricultural and forest crops worldwide. Rust epidemics have impacted the development of human society, such as the early accounts of cereal rusts coming from the Bible and from Greek and Roman literatures (Park and Wellings, 2012), or the reports of coffee leaf rust epidemics in Sri Lanka in the 19th century (Silva et al., 2006).

Genome sequencing of some rust species provided evidence for their large genome sizes (Cantu et al., 2011; Duplessis et al., 2011a) especially when compared to non-biotrophic fungi (Spanu, 2012). Genome sequencing in additional rust species confirms this (Nemri et al., 2014; Tan et al., 2014). Nevertheless,

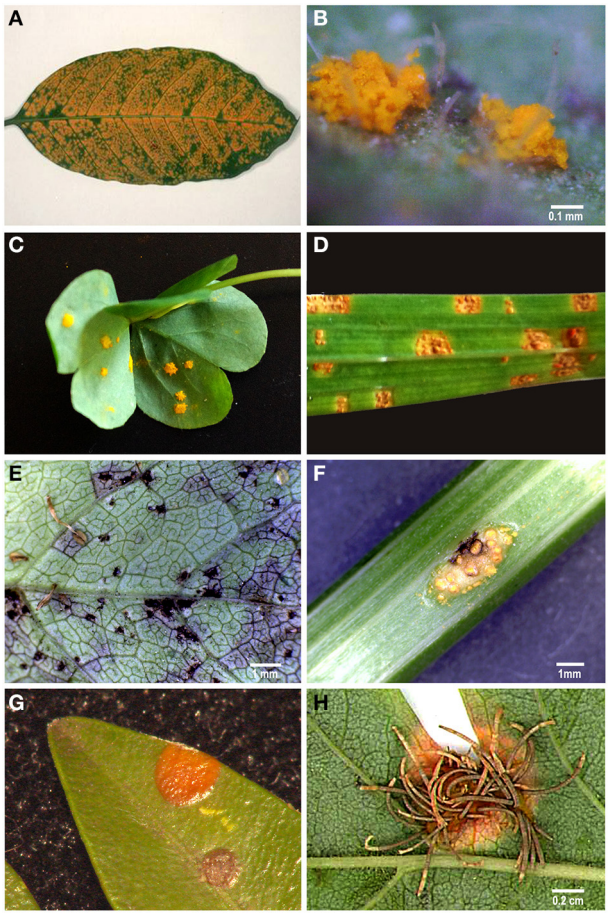


FIGURE 1 | Examples of rust sporulation. (A) *Hemileia vastatrix* uredinia on a *Coffea arabica* leaf; **(B)** *Phragmidium mexicanum* uredinia on a *Duchesnea indica* leaf; **(C)** *Puccinia oxalidis* uredinia on a *Oxalis articulata* leaf; **(D)** *Uromyces transversalis* uredinia on a *Gladiolus* sp. leaf; **(E)** *Puccinia smyrnii* telia on a *Smyrnium olusatrum* leaf; **(F)** *Puccinia smyrnii* aecia on a *Smyrnium olusatrum* stem; **(G)** *Puccinia buxi* telia on a *Buxus sempervirens* leaf; **(H)** *Gymnosporangium confusum* aecia on a *Crataegus monogyna* leaf.

sequencing efforts of other rusts species have been hampered by uncertainty concerning the genome size of the species of sequencing interest. Genome size records for 11 rust species (mostly from *Puccinia*, *Melampsora* and *Uromyces* genera) can be found at the Fungal Genome Size database (Kullman et al., 2005) and in the literature (Supplementary Data). With an average of 225.3 Mbp, available genome size values of rust species range from 77 Mbp (*Cronartium quercuum* f. sp. *fusiforme* Burds., and G.A. Snow; Anderson et al., 2010) to 733 Mbp (*Hemileia vastatrix* (733 Mbp; Carvalho et al., 2014).

Although considerably smaller than most other eukaryotes, fungi exhibit a remarkable variation in their genome sizes. The average fungal genome size is 37.7 Mbp overall, and 49.9 Mbp for the Basidiomycota (Kullman et al., 2005). The two largest fungal genomes reported so far are those of *Neottiella vivida* (Nyl.) Dennis (Ascomycota, Pezizales; Kullman, 2002) and *Scutellospora castanea* Walker (Glomeromycota, Diversisporales; Zeze et al., 1996; Hijri and Sanders, 2005), with 750 and 795 Mbp/1C, respectively. Variations in chromosome number and size are far from being an exception and ploidy levels ranging from 1x to 50x have already been found (Gregory et al., 2007). Nevertheless, Basidiomycota cells are more frequently dikaryotic with haploid nuclei for most of their life cycles. Such variations are often considered to be adaptive (Kelkar and Ochman, 2012), since variations in genome size of plant pathogens can have a direct impact in their pathogenicity (D'Hondt et al., 2011). This occurs namely through the diversity-creating effect of the activity of transposable elements and/or of polyploidization, or through the presence (or absence) of supernumerary/dispensable chromosomes (Aguileta et al., 2009; Albertin and Marullo, 2012).

Most probably due to technical constraints related with their smaller genome sizes in comparison with other organisms, only in the last two decades flow cytometry was considered the method of choice for genome size determination studies in fungi, with important impacts on plant pathology (D'Hondt et al., 2011). Using this technique, the size of the genome is estimated by comparing the fluorescence emitted by an intercalating DNA fluorochrome of a sample together with a reference standard with known genome size. Given that a flow cytometer is available, the method provides reliable estimates of genome size in a very short period of time (10 min.) and can be considered a fast and relatively cheap alternative to other molecular tools (D'Hondt et al., 2011). Still, rust species may pose technical constraints in the determination of genome sizes, as it can be especially difficult to extract nuclei in good quantity and quality from spores of obligate parasites.

Therefore, the objective of this study was to elucidate the apparent genome size expansion in the Pucciniales suggested by the genome size information available for a few species, by addressing a larger number of rust fungi with distinct life cycles, hosts and types of spores produced. For such, this work describes an innovative approach for obtaining nuclear suspensions from fungi, in particular from obligate parasites, such as the rusts. The chopping procedure of Galbraith et al. (1983) developed for plant tissues was applied for the first time for plant pathogenic fungi, enabling the analyses to be carried out directly on infected samples and not on spores.

MATERIALS AND METHODS

BIOLOGICAL MATERIAL

A total of 23 rust samples were obtained from field surveys (during 2013 and 2014 in the Lisbon area, Portugal) as infected plant material, being subsequently identified by microscopic observation. Infected plant material was preserved as dry herbarium specimens at the "João de Carvalho e Vasconcellos" Herbarium (LISI; Lisbon, Portugal). Another nine samples were retrieved as urediniospores from active collections. Thus, 32 rust samples were subjected to analysis, as detailed in **Table 1**. Infected plant material was employed directly for fungal (and plant) nuclear isolation and subsequently for flow cytometric analysis. Urediniospores (ca. 50 mg) were spread on sterile water in Petri dishes and incubated over-night at 25°C to obtain germ tubes, or used directly for flow cytometry.

Table 1 | List of 32 rust samples analyzed for genome size determination, with reference to (and sorted by) family and species, source of material [host plant (botanical name, family, location and infection stage), or spores in collection], plant reference standard used, average (GS, in pg, and Mbp), standard deviation (SD, in Mbp) and coefficient of variation (CV, in %) of the monoploid genome size, number of samples (n) and typical life cycle.

Family	Rust species	Host species[a]	Material in collection[b]	Type of spores[c]	Genome size (1C)					Reference standard[d]	Cycle[e]
					Mean pg	Mean Mbp	*SD* Mbp	CV %	*n*		
Coleosporiaceae	*Coleosporium inulae* Rabenh.	*Dittrichia viscosa* (As)	LISI-Fungi-00001	II	0.3991	390.3	37.0	9.48	2	*Sl*	Ma, He
Incertae sedis	*Hemileia vastatrix* Berk. and Broome	*Coffea arabica* (Ru)	CIFC/IICT isolate 178a	II	0.8147	796.8	6.9	0.87	5	H	Hc
Melampsoraceae	*Melampsora euphorbiae* (Ficinus and C.Schub) Castagne	*Euphorbia pterococca* (Eu)	LISI-Fungi-00002	II	0.2391	233.8	15.6	6.65	4	*Rs*	Ma, Ae
	Melampsora hypericorum (DC.) J. Schröt.	*Hypericum calycinum* (Hy)	LISI-Fungi-00003	II	0.2118	207.1	12.7	6.13	3	*Rs*	Hc
	Melampsora hypericorum (DC.) J. Schröt.	*Hypericum androsaemum* (Hy)	LISI-Fungi-00004	II	0.2520	246.5	6.9	2.81	5	*At*	Hc
	Melampsora larici-populina Kleb.	*Populus trichocarpa* x *deltoides* (Sa)	INRA isolate 98AG31	II	0.1204	117.8	7.4	6.31	7	*Rs*	Ma, He
	Melampsora ricini Pass. ex E.A. Noronha	*Ricinus communis* (Eu)	LISI-Fungi-00005	II	0.3403	332.8	20.3	6.11	3	*Rs*	He
Phakopsoraceae	*Phakopsora pachyrhizi* Syd. and P. Syd.	*Glycine max* (Fa)	UoH isolate Thai1	II	0.7364	720.2	47.4	6.59	3	H	Hc
Phragmidiaceae	*Phragmidium mexicanum* (Mains) H.Y. Yun	*Duchesnea indica* (Ro)	LISI-Fungi-00006	II	0.6181	604.5	0.8	0.14	2	*Rs*	Ma, Ae
	Phragmidium mucronatum (Pers.) Schltdl.	*Rosa* sp. (Ro)	LISI-Fungi-00007	II	0.1488	145.5	–	–	1	*Sl*	Ma, Ae
Pucciniaceae	*Gymnosporangium confusum* Dietel	*Crataegus monogyna* (Ro)	LISI-Fungi-00008	I	0.9133	893.2	3.8	0.43	4	H, *Rs*	Dc
	Puccinia allii (DC.) F. Rudolphi	*Allium ampeloprasum* (Am)	LISI-Fungi-00009	II	0.3596	351.7	8.2	2.33	3	*Rs*	Ma, Ae
	Puccinia buxi Sowerby	*Buxus sempervirens* (Bu)	LISI-Fungi-00010	III	0.6719	657.1	3.9	0.59	3	H, *Rs*	Mi
	Puccinia chrysanthemi Roze	*Dendranthema* sp. (As)	LISI-Fungi-00011	II	0.8246	806.5	15.7	1.95	4	*Rs*	Ma
	Puccinia coronata Corda	*Avena sterilis* (Po)	LISI-Fungi-00012	II	0.2491	243.6	4.6	1.91	4	*Rs*	Ma, He
	Puccinia cymbopogonis Massee	*Cymbopogon citratus* (Po)	LISI-Fungi-00013	II	0.2227	217.8	13.2	6.08	3	*Rs*	Hc
	Puccinia graminis f. sp. *tritici* Erikss. and Henning	*Triticum aestivum* (Po)	UoH; isolate ANZ	II	0.0791	77.4	–	–	1	*At*	Ma, He
	Puccinia hordei G.H. Otth	*Hordeum vulgare* (Po)	LISI-Fungi-00014	II	0.2425	237.2	1.9	0.78	3	*Rs*	Ma, He
	Puccinia malvacearum Bertero ex Mont.	*Lavatera cretica* (Ma)	LISI-Fungi-00015	III	0.1818	177.8	13.6	7.63	3	*Rs*	Ma, Ae
	Puccinia oxalidis Dietel and Ellis	*Oxalis articulata* (Ox)	LISI-Fungi-00016	II	0.3629	354.9	25.0	7.04	3	*Rs, Sl*	Ma, He
	Puccinia pelargonii-zonalis Doidge	*Pelargonium zonale* (Ge)	LISI-Fungi-00017	II	0.1877	183.6	6.8	3.73	3	*Rs*	Hc
	Puccinia smyrnii Corda	*Smyrnium olusatrum* (Ap)	LISI-Fungi-00018	I,III	0.2647	258.9	14.2	5.47	4	*Rs*	Dc
	Puccinia triticina Erikss.	*Triticum aestivum* (Po)	UoH	II	0.0786	76.9	1.3	1.64	2	*At*	Ma, He
	Uromyces appendiculatus F. Strauss	*Phaseolus vulgaris* (Fa)	UoH isolate SWBR1	II	0.6947	679.4	26.4	3.89	6	H, *Sl*	Ma, Ae
	Uromyces fabae de Bary ex Cooke	*Vicia faba* (Fa)	UoH race I2	II	0.3879	379.4	2.1	0.56	4	*Rs*	Ma, Ae
	Uromyces fabae f. sp. *pisi-sativae* Hirats. f.	*Vicia sativa* (Fa)	LISI-Fungi-00019	II	0.4427	433.0	31.8	7.34	4	*Rs*	Hc
	Uromyces rumicis (Schumach.) G. Winter	*Rumex crispus* (Pl)	LISI-Fungi-00020	II	0.2830	276.8	7.8	2.81	3	*Rs*	Ma, He
	Uromyces striatus J. Schröt.	*Medicago sativa* (Fa)	UoH	II	0.4245	415.2	5.5	1.31	6	*Sl*	Ma, He
	Uromyces transversalis (Thüm.) G. Winter	*Gladiolus* sp. (Ir)	LISI-Fungi-00021	II	0.3852	376.7	12.8	3.40	3	*Sl*	Hc
	Uromyces vignae Barclay	*Vigna unguiculata* (Fa)	UoK isolate CPR-1	II	0.7282	712.2	6.7	0.95	2	*Rs*	Ma, He

(Continued)

Table 1 | Continued

Family	Rust species	Host species[a]	Material in collection[b]	Type of spores[c]	Genome size (1C)				Reference standard[d]	Cycle[e]
					Mean pg	Mean Mbp	*SD* Mbp	CV %	*n*		
Pucciniastraceae	*Pucciniastrum epilobii* G.H. Otth	*Fuchsia* sp. (On)	LISI-Fungi-00022	II	0.2892	282.8	5.9	2.08	3	*Rs*	Hc
Uropyxidaceae	*Tranzschelia discolor* (Fuckel) Tranzschel and M.A. Litv.	*Prunus dulcis* (Ro)	LISI-Fungi-00023	II	0.3094	302.6	–	–	1	H	Hc

[a] *Acronyms in brackets refer to host family: Am, Amaryllidaceae; Ap, Apiaceae; As, Asteraceae; Bu, Buxaceae; Eu, Euphorbiaceae; Fa, Fabaceae; Ge, Geraniaceae; Hy, Hypericaceae; Ir, Iridaceae; Ma, Malvaceae; On, Onagraceae; Ox, Oxalidaceae; Pl, Polygonaceae; Po, Poaceae; Ro, Rosaceae; Ru, Rubiaceae; Sa, Salicaceae.*

[b] *Institutional acronyms: CIFC/IICT, Centro de Investigação das Ferrugens do Cafeeiro/Instituto de Investigação Científica Tropical (Oeiras, Portugal); INRA, Institut National de la Recherche Agronomique, center Nancy (Nancy, France); LISI, "João de Carvalho e Vasconcellos" Herbarium (Lisbon, Portugal); UoH, Universität Hohenheim (Stuttgart, Germany); UoK, Universität Konstanz (Konstanz, Germany).*

[c] *Type of spores present in sampled material: 0, Pycniospores; I, Aeciospores; II, Urediniospores; III, Teliospores; IV, Basidiospores (Laundon, 1967).*

[d] *Plant reference standards used: At, Arabidopsis thaliana; Rs, Raphanus sativus; Sl, Solanum lycopersicum; H, host plant (genome sizes retrieved from the Plant DNA C-values Database release 6.0, Royal Botanical Gardens, Kew: Coffea arabica, 2C = 2.40 pg or 2347 Mbp; Glycine max, 2C = 2.25 pg or 2201 Mbp; Crataegus monogyna, 2C = 1.52 pg or 1487 Mbp; Buxus sempervirens, 2C = 1.62 pg or 1584 Mbp; Phaseolus vulgaris, 2C = 1.20 pg or 1174 Mbp; Prunus dulcis, 2C = 0.66 pg or 645 Mbp).*

[e] *Typical life cycle: Dc, demicyclic; Hc, hemicyclic; Ma, macrocyclic; Mi, microcyclic; Ae, autoecious; He, heteroecious.*

Plants used as reference for flow cytometry were grown from seeds (*Arabidopsis thaliana* 'Col-0,' *Raphanus sativus* 'Saxa' and *Solanum lycopersicum* 'Stupické') and were maintained at the CFE/FCT/UC.

FLUORESCENCE MICROSCOPY

Fungal material was stained using an aqueous solution of 1 μg/ml 4′,6-diamidino-2-phenylindole (DAPI; Sigma-Aldrich, St. Louis, USA) and slides were mounted in vectashield® (Vector Laboratories, Burlingame, USA), an antifading agent. Samples were observed in an epifluorescence microscope (Leica DMRB, DFC 340FX) equipped with a BP470/40 cube and an excitation wavelength of 340–380 nm. Pictures were captured with MetaMorph® software.

Infected leaf pieces, about 2–4 cm^2, were fixed overnight in a 2% solution of glutaraldehyde in 0.1 M sodium phosphate buffer, pH 7.2. Leaf pieces were then sectioned with a freezing microtome (Leica CM1850) and the sections (20–25 μm) were stained with DAPI (as before), for 2 h. The sections were then washed with distilled water, stained with an aqueous solution of 0.3% w/v diethanol for 2–3 s, washed again with distilled water and mounted in 50% v/v glycerol (adapted from Stark-Urnau and Mendgen, 1993; Skalamera and Heath, 1998).

Leaf material was examined with bright field microscopes (Leitz Dialux 20 and Leica DM-2500) equipped with mercury bulbs HB 100W, ultra-violet light (excitation filter BP 340–380; barrier filter LP 430).

FLOW CYTOMETRY

The nuclear DNA content of rust fungi was estimated by flow cytometry using infected host tissue samples (occasionally spores or germ tubes only), by comparison with the host plant genome size (as given in **Table 1**) and/or, when the latter was either unknown, uncertain or out of range, with healthy leaves of plant DNA reference standards: *Arabidopsis thaliana* 'Col-0' (2C = 0.32 pg or 313 Mbp; this study after calibration with *Raphanus sativus* 'Saxa'); *Raphanus sativus* 'Saxa' (2C = 1.11 pg or 1086 Mbp; Doležel et al., 1992); or *Solanum lycopersicum* 'Stupické' (2C = 1.96 pg or 1917 Mbp; Doležel et al., 1992) (**Table 1**).

Nuclei were released from infected host tissues, fungal germ tubes and/or leaves of the reference standards following the procedure of Galbraith et al. (1983). In brief, ∼50 mg of both fungus and plant (internal standard) were chopped with a razor blade in a Petri dish with 1 mL of Woody Plant Buffer (WPB; 0.2 M Tris–HCl, 4 mM $MgCl_2$, 1% Triton X-100, 2 mM Na_2EDTA, 86 mM NaCl, 20 mM sodium metabisulfite, 1% PVP-10, pH 7.5; Loureiro et al., 2007). Nuclei from spores were released by grinding ∼10 mg of spores in a mortar in the presence of 1 mL of WPB. For the latter, nuclei from the plant DNA reference standard were added afterwards (pseudo-internal standardization).

The nuclear suspension was then filtered through a 30 μm nylon filter to remove plant and fungal debris, and 50 μg/mL of propidium iodide (PI; Fluka, Buchs, Switzerland) and 50 μg/mL of RNase (Fluka), both suspended in water, were added to stain the DNA only. After incubation for 5 min. at room temperature, the fluorescence intensity of at least 3000 nuclei per sample was

analyzed using a Partec CyFlow Space flow cytometer (Partec GmbH, Görlitz, Germany), equipped with a 30 mW green solid-state laser emitting at 532 nm for optimal PI excitation. The assignment of each peak to rust fungi, host plant and plant reference standard was confirmed by separately analysing healthy plant samples and fungal spores or germ tubes. For each rust species, the G_1 peak of the plant species used as internal reference standard was set to a specific channel (usually between channel positions 500 and 750 on a 0-1028 scale), with the amplification system kept at a constant voltage and gain throughout the analyses. Each day, prior to analysis, the overall instrument quality was assessed using calibration beads green concentrate (Partec GmbH). For each sample, when possible at least three independent replicate measurements were performed.

FLOW CYTOMETRY DATA ANALYSIS

Data were acquired using Partec FloMax software v2.4d (Partec GmbH) in the form of four graphics: fluorescence pulse integral in linear scale (FL); forward light scatter (FSC) vs. side light scatter (SSC), both in logarithmic (log) scale; time vs. FL in linear scale; and SSC in log scale vs. FL in linear scale. To analyse only intact nuclei, the FL histogram was gated with a polygonal region defined in the FL vs. SSC dot-plot (Loureiro et al., 2006). Afterwards, using FloMax gating tools, linear regions were created in the FL histogram to gate the nuclei and obtain descriptive statistics of each peak, including number of nuclei, mean channel position and coefficient of variation (CV).

The genome size in mass units (1C in pg for fungi) was assessed using the formula:

$$\frac{\text{Mean G1 fluorescence of sample nuclei}}{\text{Mean G1 fluorescence of reference standard}} \times \text{2C genome size of the reference standard}$$

Conversion of mass values into numbers of base pairs was done according to the factor 1 pg = 978 Mbp (Doležel and Bartoš, 2005).

The reliability of the genome size measurements was verified by evaluating the quality of the flow cytometry histograms based on the CV of the G_1 peaks and on the background debris, and by the CV of the genome size estimation of each isolate based on the independent measurements. According with the criteria established by Bourne et al. (2014), only CV values of DNA peaks below 10% were considered in the analyses.

STATISTICAL ANALYSIS

Statistical analyses were performed using R (R Core Team, 2014). Comparison of genome size values for the most outstanding phylogenetic groups was performed using the Wilcoxon test ($\alpha = 0.05$). Comparison of individual data was performed using the χ^2-test. A total of 1820 fungal genome sizes were compiled from information publically available at the Fungal Genome Size Database (http://www.zbi.ee/fungal-genomesize; Kullman et al., 2005), the JGI Genome Portal (http://genome.jgi-psf.org), the Broad Institute (http://www.broadinstitute.org/), and from the literature (Supplementary Data, database sheet).

RESULTS

Field surveys conducted over 1 year enabled the identification of several rust-infected plants. Both plants and fungi were identified by experienced botanists and mycologists. The collected samples (**Table 1**) comprised several botanical and mycological families. Most rusts were retrieved at the urediniosporic infection cycle, but telia and aecia were also readily found for certain rusts (**Figures 1**, **2**), further confirming the diagnosis and allowing the identification of the pathogen. Staining of spores with DAPI enabled the visualization of two nuclei per rust cell in

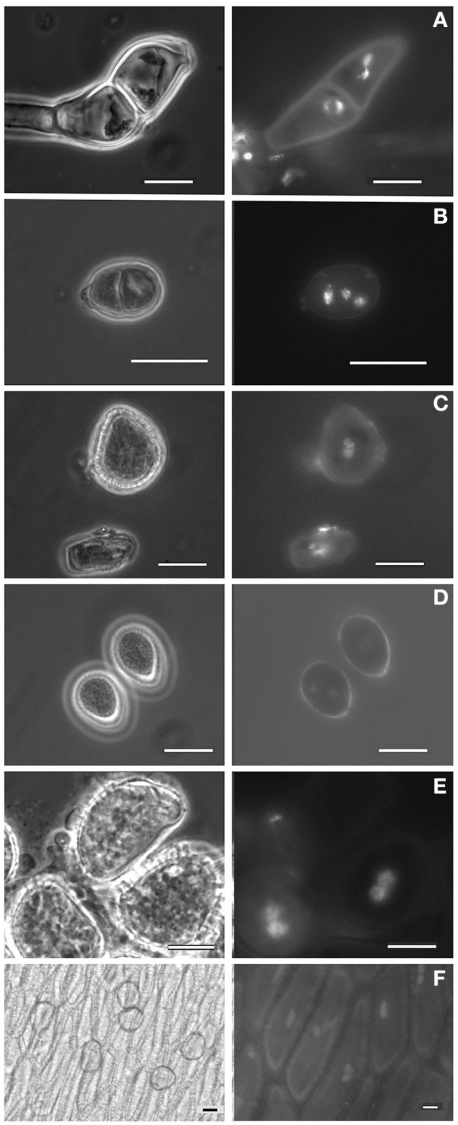


FIGURE 2 | Spores (and other cells) from Pucciniales under phase-contrast microscopy and showing the DAPI-stained nuclei under fluorescence light. (A) Teliospore from *Puccinia malvacearum*; **(B)** Teliospore from *Puccinia smyrnii*; **(C)** Aeciospore from *Puccinia smyrnii*; **(D)** Urediniospores from *Phragmidium mexicanum*; **(E)** Urediniospore from *Hemileia vastatrix*; **(F)** Pseudoperidia cells and aeciospores from *Gymnosporangium confusum*. Bars, 10 μm.

aeciospores and urediniospores. In teliospores, some cells contained two nuclei, while others, following karyogamy, contained a single nucleus (**Figure 2**). The application of the nuclear isolation protocol to rust-infected host tissues (**Figure 1**) enabled the release of intact nuclei from both the host plant and the fungal cells. Nuclei were efficiently stained with PI, according to the clearly defined G_1 peaks of both organisms (**Figure 3**). Following this innovative approach, after identification of the fungal peaks, 32 rust samples representing 30 species were analyzed by flow cytometry (**Table 1**).

The genome size determinations based on the fungal G_1 fluorescence peaks had CV values below 10% (usually between 4 and 7%), which is within the range of accepted values for fungal species (Bourne et al., 2014), and, for each sample, CV measures of genome size estimations never exceeded 10% (**Table 1**). Polygonal regions in dot-plots of SSC vs. FL enabled to gate and present in a histogram nuclei that were uniform in size and shape, eliminating partial nuclei and other types of debris (**Figure 3A1**). This strategy improved the CV values of DNA peaks and high-quality histograms were obtained for the analyses of genome sizes.

When the genome size of the host plant was known and appropriate (i.e., when it appeared in the same scale set as the fungal species), the host plant itself was used as primary reference standard. Otherwise, according with the genome size of the fungal species, *Arabidopsis thaliana*, *Raphanus sativus*, or *Solanum lycopersicum* were used as reference genomes. The analysis was not affected by the endopolyploid nature of *R. sativus* and especially of *A. thaliana* (Kudo and Kimura, 2001) since the only visible peak of plant DNA reference standard in the scale set was that of 2C nuclei and thus the three plant species were considered adequate for the analysis. Even when the rust and the host genome sizes were within the same size range, the host genome was always larger than that of the rust (as exemplified in **Figures 3A,D,E**).

In this study we have analyzed 12 *Puccinia* spp., six *Uromyces* spp., four *Melampsora* spp., and two *Phragmidium* spp. The remaining six genera analyzed were represented by a single species (**Table 1**). The average genome size of species of *Melampsora*, *Puccinia* and *Uromyces* was 227.6, 303.6, and 467.5 Mbp respectively. While the five *Melampsora* genomes (four species) were all below the overall average and varied by less than 3x, from 117.8 Mbp (for *M. larici-populina*) to 332.8 Mbp (for *M. ricini*),

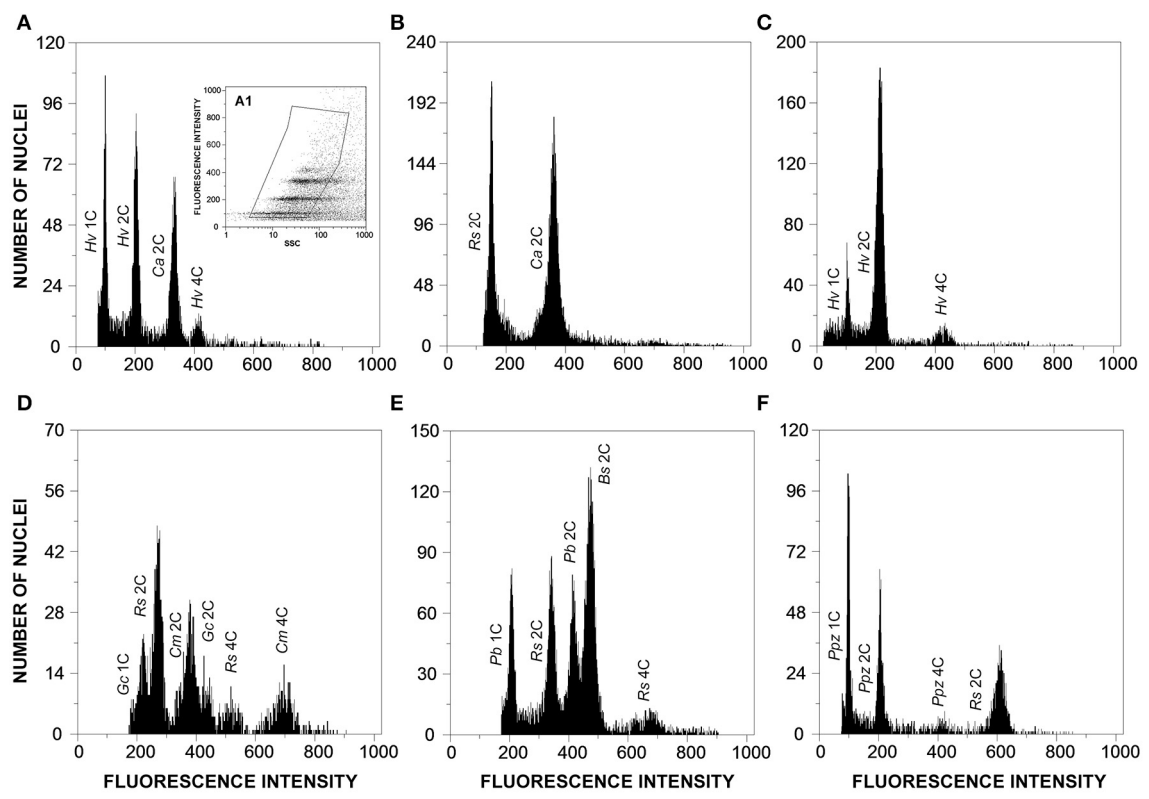


FIGURE 3 | Flow cytometric histograms of relative fluorescence intensities of propidium iodide-stained nuclei simultaneously isolated from: (A) *Hemileia vastatrix* (*Hv*) and its host plant, *Coffea arabica* (*Ca*; 2C = 2.49 pg DNA); **(B)** *Coffea arabica* (Ca) and the plant DNA reference standard, *Raphanus sativus* (*Rs*, 2C = 1.11 pg DNA); **(C)** *Hemileia vastatrix* (*Hv*) hyphae obtained upon germination of urediniospores in water; **(D)** *Gymnosporangium confusum* (*Gc*), its host plant, *Crataegus monogyna* (*Cm*; 2C = 1.50 pg DNA), and the plant DNA reference standard, *Raphanus sativus* (*Rs*); **(E)** *Puccinia buxi* (*Pb*), the plant DNA reference standard (*Rs*), and its host plant, *Buxus sempervirens* (*Bs*; 2C = 1.60 pg DNA); and **(F)** *Puccinia pelargonii-zonalis* (*Ppz*) and the plant DNA reference standard, *Raphanus sativus* (*Rs*). The inset (A1) in histogram A represents the gating made in the dot-plot of SSC vs. FL to exclude as much as possible partial nuclei and other types of debris.

the 11 *Puccinia* species varied by more than 10x, and included the smallest genomes analyzed in this study (76.9 and 77.4 Mbp for *P. triticina* and *P. graminis* f. sp. *tritici*) and the second largest one, *P. chrysanthemi* with 806.5 Mbp. The seven *Uromyces* genomes (six species) varied also by less than 3x, but in most cases their genome sizes were higher than the overall average, from 276.8 Mbp in *U. rumicis* to 712.2 Mbp in *U. vignae*. No statistically significant differences were identified when comparing rust genera or families.

When rusts are clustered according to their hosts family, it is clear that rusts with Poaceae hosts (*Puccinia coronata*, *P. cymbopogonis*, *P. graminis* f. sp. *tritici*, *P. hordei*, and *P. triticina*) have considerably smaller genomes (170.6 Mbp on average) than rusts with Fabaceae hosts (556.6 Mbp on average; *Phakopsora pachyrhizi*, *Uromyces appendiculatus*, *U. fabae*, *U. fabae* f. sp. *pisi-sativae*, *U. striatus*, and *U. vignae*). This difference is statistically supported ($P < 0.05$). The four rust species with Rosaceae hosts (*Gymnosporangium confusum*, *Phragmidium mexicanum*, *Phr. mucronatum*, and *Tranzschelia discolor*) presented vastly different genome sizes, with estimates below the average (145.5 Mbp in *Phr. mucronatum*) to the largest estimate discovered so far (893.2 Mbp in *G. confusum*), with an average of 486.4 Mbp.

Adding these values to the available genome sizes for other Pucciniales (Supplementary Data, database sheet) results in a global average for all rust fungi of 305.5 Mbp, a value that is significantly higher ($P < 0.001$) than that of any other order in Fungi for which three or more genome sizes are available (**Figure 4**), with the single exception of the Diversisporales (Glomeromycota), which contains three genomes in the 723–795 Mbp range. The smallest rust genome size in this list (*Cronartium quercuum* f. sp. *fusiforme*, with 76.6 Mbp; Anderson et al., 2010) is larger than 94% of all the fungi analyzed so far. The global average for all fungi (including the results obtained in this work) is of 44.2 Mbp (genome sizes from 1852

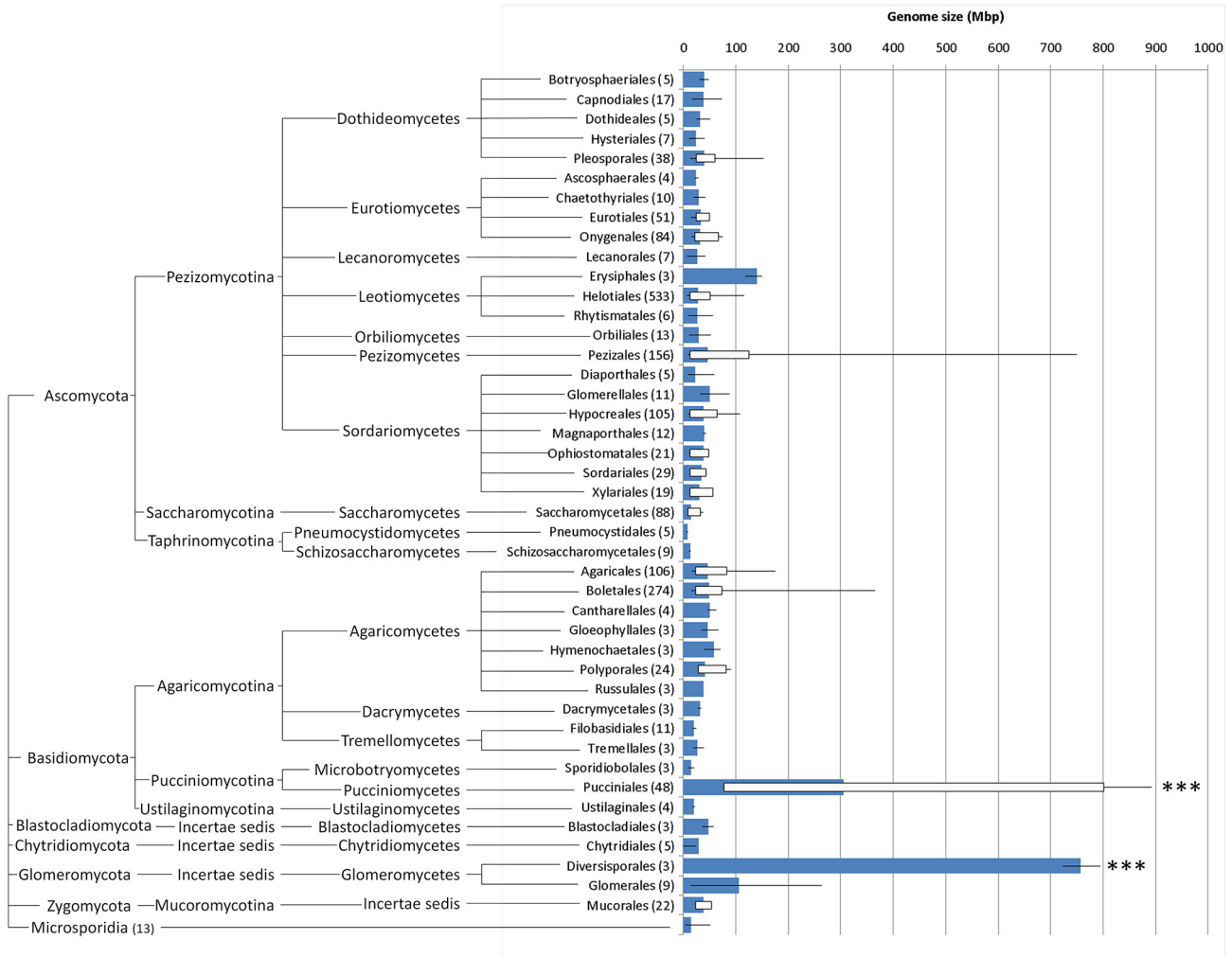


FIGURE 4 | Whisker box plots for genome sizes (Mbp) for every fungal order for which at least three values were available (number of organisms in brackets; further details in Supplementary Data, database sheet), including the results obtained in this study; blue bars represent average, lines denote minimum and maximum values and white boxes represent 5 and 95 percentiles; fungal orders are arranged phylogenetically (http://tolweb.org/Fungi/). *** $P < 0.001$.

organisms), a value that is to a large extent influenced by the large amount of estimations available for the Ascomycota (1278 organisms, i.e., 69% of all organisms; global average of 31.8 Mbp), and Basidiomycota fungi (516 organisms; 28%; global average of 70.4 Mbp (Supplementary Data, analysis sheet).

DISCUSSION

The Pucciniales represent an important group of plant pathogens, with several unifying biological characteristics. Recently, the occurrence of a genome size expansion in the Pucciniales was suggested based on the first completed rust genome sequences (Spanu, 2012). With the purpose of providing a broader set of data to support or refute this hypothesis, the genome size of 32 rust samples (comprising 30 rust species) was estimated by flow cytometry. The chopping procedure of fresh tissue for isolating intact nuclei in the presence of a nuclear isolation buffer (developed for plant tissues) was successfully employed for flow cytometric analysis of the genome sizes of rust-infected plant material. This approach proved to be very efficient and may be applied on a variety of infected plant tissues in the future, as it circumvents the need to isolate basidiospores or pycniospores as previously reported for flow cytometric estimation of genome size in rust fungi (Williams and Mendgen, 1975; Eilam et al., 1992, 1994).

A collection of rust fungi found in nature, together with some of the economically most important rust species, revealed that the variability of the genomes sizes was very high, ranging from 76.9 to 893.2 Mbp. These estimates are even higher than those already made available through the fungal genome size database. Two rust fungi, *Puccinia chrysanthemi* and *Gymnosporangium confusum*, with genome sizes of 806.5 and 893.2 Mbp/1C respectively, constitute the two largest fungal genomes reported to date. Both genomes are larger than the so far largest rust genome, *Hemileia vastatrix* (Carvalho et al., 2014 and in the present study). These genome sizes also surpass the two largest fungal genomes reported so far, *Neottiella vivida* and *Scutellospora castanea.* Remarkably both of these fungi also interact closely with plants.

Comparing the results obtained in this study with the 1820 fungal genome sizes available in databases and in the literature, it is evident that even the smallest rust genome is larger than the genome size found in 94% of all fungi. The inclusion of these 32 genome sizes shifts the global genome size average of all fungi from 37.7 to 44.2 Mbp, and that of Basidiomycota from 49.9 to 70.4 Mbp. The average genome size for the Pucciniales reaches 305.5 Mbp, a value that is significantly higher than any other Ascomycota or Basidiomycota order. The few genome sizes available for species of other orders in the Pucciniomycotina besides the Pucciniales are much smaller, with estimates of 13 Mbp for *Mixia osmundae* (Nishida) C.L. Kramer, of 26 Mbp for *Microbotryum violaceum* (Pers.) G. Deml and Oberw. and of 21 Mbp for *Rhodotorula graminis* Di Menna and *Sporobolomyces roseus* Kluyver and C.B. Niel.

The collection of rusts under study represents different hosts and life cycles and comprises 10 rust genera, enabling the analysis of correlations between these characteristics and the genome size estimates obtained. The two *Phragmidium* species analyzed in this study differ clearly in their genomes sizes, despite both being macrocyclic and autoecious. *Phragmidium mexicanum* infects *Potentilla/Duchesnea* hosts (Yun et al., 2011), while *Phr. mucronatum* colonizes species of *Rosa* (Helfer, 2005). Two *Melampsora hypericorum* samples obtained from *Hypericum calycinum* or *H. androseamum* also exhibited distinct genome sizes. The latter supports other reports that have shown the occurrence of host-depended intra-specific variation in nuclear content of *Puccinia hordei* and of *P. recondita* isolates (Eilam et al., 1994).

The genome size of *Puccinia graminis* f. sp. *tritici* is estimated to be 77.4 Mbp. Eilam et al. (1994) estimated a value of 67 Mbp, while the genome sequence yielded a value of 88.6 Mbp (Duplessis et al., 2011a). These differences could also be attributed to intra-specific variability, although the distinct methodologies adopted may also account for some variation. A difference of 16.7 Mbp was observed between the flow cytometric estimate and the value obtained from genome sequencing (Duplessis et al., 2011a) for *Melampsora larici-populina* isolate 98AG31. *Uromyces appendiculatus* and *U. vignae* have been reported to have some of the largest rust genomes (Eilam et al., 1994; Kullman et al., 2005), with 400 to 418 Mbp. In this work, however the genome size of laboratory strains of these two species was estimated as 679.4 and 712.2 Mbp, respectively. Such discrepancies could be due to the employment of a different technique.

All *Uromyces* species analyzed with Fabaceae hosts presented genome sizes above 300 Mbp. Moreover, *Phakopsora pachyrhizi*, another rust that infects a member of the Fabaceae, also possesses a large genome size. This markedly contrasts with the smaller genome sizes of rusts with Poaeceae hosts (all below 250 Mbp and all in the genus *Puccinia*). Interestingly, *Uromyces* species with Fabaceae hosts constitute a monophyletic group that probably evolved together, and are all autoecious (van der Merwe et al., 2008). Considering the relationship between genome sizes and life cycles, the species with the largest genome sizes are either autoecious or hemicyclic with no known alternate host, with the exception of the heteroecious *Gymnosporangium confusum*.

As a microcyclic rust, *Puccinia buxi* depends strictly in sexual reproduction for multiplication. This fungus is only found in a limited number of locations, most likely due to its specific requirements of shaded and humid microclimatic conditions (Preece, 2000; Durrieu, 2001), thus suggesting low population size. In this study, this species was also found to possess a large genome, which could be linked to its populational and reproductive characteristics. In fact, a major force conditioning genome size seems to be genetic drift, which was negatively correlated with effective population size (Kelkar and Ochman, 2012).

The Pucciniales share some common features, such as biotrophy and obligate parasitism. Biotrophy has been highlighted as a lifestyle that leads to increasing genome size as compared to non-biotrophs (Spanu, 2012). The very large genome sizes of the 30 rust fungal species revealed by our study strongly reinforce the view that expanded genome sizes occur among biotrophs, and that large genomes are a common characteristic of the Pucciniales. From the genome sequencing of rust fungi (e.g.,

Duplessis et al., 2011a; Nemri et al., 2014) and other biotrophs it is now clear that larger genomes do not imply higher numbers of structural genes, resulting invariably in an increased proliferation of transposable elements (TE) and repetitive DNA. Such a genomic environment can create genetic polymorphisms, especially in the case of sexual abstinence (Spanu, 2012). As in plants, it would be interesting to evaluate in the future if there are costs for the fungi associated with the accumulation and replication of this excess DNA (large genome constraint; for a review see Knight et al., 2005).

Although the effect of sex on genome size evolution is still unclear (Raffaele and Kamoun, 2012), three of the rust fungi with large genome sizes, *Hemileia vastatrix*, *Phakospora pachyrhizi* and to some extent *Puccinia chrysanthemi,* all rely on asexual reproduction. The first two species are hemicyclic or at least the aecial host is unknown and the third also reproduces mainly asexually, although it was reported to be autoecious in Japan (Alaei et al., 2009). Even for those species which are capable of sexual reproduction, it is expectable that urediniosporic infection cycles may well represent a very important fraction of reproduction, for which TE activity would be potentially an important source for the generation of diversity. In this sense, rust species that do not produce urediniospores (demicyclic rusts), such as the autoecious *Puccinia buxi* and *P. smyrnii*, and therefore strictly depend on sexual reproduction for life cycle completion, are of great interest for studying the relation between genome size and reproduction/diversity creation strategies.

A very large fraction (up to 50%) of the rust genome sequences published so far is composed of repetitive elements. Those genomes, however, are all below 200 Mbp. In this work we have revealed genome sizes several times larger. One can speculate that such genome size expansions could be due to an even higher proportion of non-coding regions, but also to genome duplication/polyploidy. Although still largely overlooked in fungi, polyploidy is a major evolutionary process in eukaryotes (Albertin and Marullo, 2012), playing a role on the wide capacity of fungi to evolve adaptability to virtually all ecosystems and modes of heterotrophic nutrition (Aguileta et al., 2009). Polyploidy events may have occurred in other Basidiomycota (such as the Agaricales), and tolerance for genome merging has been suggested in the Microbotryales (Pucciniomycotina) (Albertin and Marullo, 2012). Although genome sizes in the Pucciniales are clearly expanded as compared to neighboring clades, variation in genome sizes across the Pucciniales suggests little correlation to phylogeny. *Hemileia vastatrix*, which was considered to represent an ancestral clade in rusts phylogeny (Aime, 2006), presents one of the largest genome sizes determined in this study. Also, species within the same genus (e.g., in *Puccinia* or *Uromyces*) presented very divergent genome sizes. These finds suggest that variation in genome sizes is rapidly occurring along the evolution of Pucciniales.

Gymnosporagium spp. are unique rust fungi since they comprise the only genus forming teliospores on members of the Cupressaceae. Molecular data (18S and 28S rDNA sequences) question their placement within the Pucciniaceae (Aime, 2006). Now due to its highest genome size, this group is likely to gain more attention from the scientific community.

A unifying characteristic amongst the species with a larger genome size within the Pucciniales was not found. It seems more likely that different events have driven the evolution of genome size of particular species or groups of species. Genome variability is considered to be adaptive and host driven resulting in a high capability to overcome the host defenses (Stukenbrock and Croll, 2014). Relationships between genome size and biological parameters are of special interest because they can be linked to the ability of an organism to overcome selection pressure (D'Hondt et al., 2011). Although the rust genome sizes determined in this study surpass most other fungi and are within the range of genome sizes of many plants, it is interesting to note that all rust genome sizes in this study are smaller than those of the hosts from where they were obtained.

In conclusion, in this work the analysis of the genome size of 30 rust species (representing eight families) revealed the occurrence of very large genome sizes, including the two largest fungal genomes ever reported, *Gymnosporangium confusum* (893.2 Mbp) and *Puccinia chrysanthemi* (806.5 Mbp). Although comprising only a very small fraction of the more than 7000 rust species described, with many genera and some families not represented, this work suggests that the Pucciniales represent a group where genome size expansion could be a common characteristic, in sharp contrast with sister taxa, making this group of organisms a subject of utmost interest for genomic research and for further studies.

AUTHOR CONTRIBUTIONS

This study was conceived and directed by Sílvia Tavares, Ana Paula Ramos, Ana Sofia Pires, Helena G. Azinheira, Tobias Link, Rita Abranches, Ralf T. Voegele, João Loureiro, and Pedro Talhinhas. Collection and identification of field material was performed by Ana Paula Ramos, Helena G. Azinheira, and Pedro Talhinhas. Sample preparation, nuclei isolation and flow cytometry analyses were performed by Sílvia Tavares, Patrícia Caldeirinha, João Loureiro, and Pedro Talhinhas. Microscopy observations and image acquisition were conducted by Sílvia Tavares, Ana Sofia Pires, and Maria do Céu Silva. Data analysis and biological interpretation of results were conducted by Sílvia Tavares, Ana Paula Ramos, João Loureiro, and Pedro Talhinhas. Sílvia Tavares, Ana Paula Ramos, João Loureiro, and Pedro Talhinhas wrote the paper. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

Prof. Arlindo Lima, Eng. Filomena Caetano, Ms. Ana Paula Paes, and Eng. Teresa Vasconcelos (ISA/UL, Portugal) are acknowledged for support on the identification of rust and host species. The *Melamspora larici-populina* isolate 98AG31 was kindly provided by Dr. Pascal Frey (INRA Nancy, France). The *Uromyces vignae* isolate CPR-1 was kindly provided by Prof. Kurt Mendgen (Univ. Konstanz, Germany). This work was supported by Fundação para a Ciência e a Tecnologia (FCT, Portugal) through PEst-OE/EQB/LA0004/2011 (at ITQB/UNL) and PTDC/AGR-GPL/114949/2009 (at CIFC/IICT and ITQB/UNL). Sílvia Tavares, Ana Sofia Pires, and Pedro Talhinhas received postdoctoral grants from FCT (SFRH/BPD/65965/2009, SFRH/BPD/65686/2009 and SFRH/BPD/88994/2012, respectively).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00422/abstract

REFERENCES

Aguileta, G., Hood, M. E., Refrégier, G., and Giraud, T. (2009). Genome evolution in plant pathogenic and symbiotic fungi. *Adv. Bot. Res.* 49, 151–193. doi: 10.1016/S0065-2296(08)00603-4

Aime, M. C. (2006). Toward resolving family-level relationships in rust fungi (Uredinales). *Mycoscience* 47, 112–122. doi: 10.1007/s10267-006-0281-0

Alaei, H., De Backer, M., Nuytinck, J., Maes, M., Hofte, M., and Heungens, K. (2009). Phylogenetic relationships of *Puccinia horiana* and other rust pathogens of *Chrysanthemum* x *morifolium* based on rDNA ITS sequence analysis. *Mycol. Res.* 113, 668–683. doi: 10.1016/j.mycres.2009.02.003

Albertin, W., and Marullo, P. (2012). Polyploidy in fungi: evolution after whole-genome duplication. *Proc. R. Soc. B* 279, 2497–2509. doi: 10.1098/rspb.2012.0434

Anderson, C. L., Kubisiak, T. L., Nelson, C. D., Smith, J. A., and Davis, J. M. (2010). Genome size variation in the pine fusiform rust pathogen *Cronartium quercuum* f. sp. fusiforme as determined by flow cytometry. *Mycologia* 102, 1295–1302. doi: 10.3852/10-040

Bourne, E. C., Mina, D., Gonçalves, S. C., Loureiro, J., Freitas, H., and Muller, L. A. H. (2014). Large and variable genome size unrelated to serpentine adaptation but supportive of cryptic sexuality in *Cenococcum geophilum*. *Mycorrhiza* 24, 13–20. doi: 10.1007/s00572-013-0501-3

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Carvalho, G. M. A., Carvalho, C. R., Barreto, R. W., and Evans, H. C. (2014). Coffee rust genome measured using flow cytometry: does size matter? *Plant Pathol.* doi: 10.1111/ppa.12175. [Epub ahead of print].

Cummins, G. B., and Hiratsuka, Y. (2003). *Illustrated Genera of Rust Fungi*. 3rd Edn. St. Paul, MN: American Phytopathological Society, 240.

D'Hondt, L., Hofte, M., Van Bockstaele, E., and Leus, L. (2011). Applications of flow cytometry in plant pathology for genome size determination, detection and physiological status. *Mol. Plant Pathol.* 12, 815–828. doi: 10.1111/j.1364-3703.2011.00711.x

Doležel, J., and Bartoš, J. A. N. (2005). Plant DNA flow cytometry and estimation of nuclear genome size. *Ann. Bot.* 95, 99–110. doi: 10.1093/aob/mci005

Doležel, J., Sgorbati, S., and Lucretti, S. (1992). Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. *Physiol. Plantarum* 85, 625–631. doi: 10.1111/j.1399-3054.1992.tb04764.x

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Joly, D. L., and Dodds, P. N. (2011b). "Rust effectors," in *Effectors in Plant-Microbe Interactions*, *1st Edn.*, eds F. Martin and S. Kamoun (Oxford: Wiley-Blackwell), 155–193.

Durrieu, G. (2001). More about box rust. *Mycologist* 15, 144. doi: 10.1016/S0269-915X(01)80046-X

Eilam, T., Bushnell, W. R., and Anikster, Y. (1994). Relative nuclear DNA content of rust fungi estimated by flow cytometry of propidium iodide-stained pycniospores. *Phytopathology* 84, 728–735.

Eilam, T., Bushnell, W. R., Anikster, Y., and McLaughlin, D. J. (1992). Nuclear DNA content of basidiospores of selected rust fungi as estimated from fluorescence of propidium iodide-stained nuclei. *Phytopathology* 82, 705–712.

Fernandez, D., Talhinhas, P., and Duplessis, S. (2013). "Rust fungi: new advances on genomics and host-parasite interactions," in *The Mycota, Vol. XI. 2nd Edn., Application in Agriculture*, ed F. Kempken (Berlin: Springer Verlag), 315–341.

Galbraith, D. W., Harkins, K. R., Maddox, J. M., Ayres, N. M., Sharma, D. P., and Firoozabady, E. (1983). Rapid flow cytometric analysis of the cell-cycle in intact plant-tissues. *Science* 220, 1049–1051. doi: 10.1126/science.220.4601.1049

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., et al. (2007). Eukaryotic genome size databases. *Nucleic Acids Res.* 35, D332–D338. doi: 10.1093/nar/gkl828

Helfer, S. (2005). Overview of the rust fungi (Uredinales) occurring on Rosaceae in Europe. *Nova Hedwigia* 81, 325–370. doi: 10.1127/0029-5035/2005/0081-0325

Hijri, M., and Sanders, J. R. (2005). Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. *Nature* 433, 160–163. doi: 10.1038/nature03069

Kelkar, Y. D., and Ochman, H. (2012). Causes and consequences of genome expansion in fungi. *Genome Biol. Evol.* 4, 13–23. doi: 10.1093/gbe/evr124

Knight, C. A., Molinari, N. A., and Petrov, D. A. (2005). The large genome constraint hypothesis: evolution, ecology and phenotype. *Ann. Bot.* 95, 177–190. doi: 10.1093/aob/mci011

Kudo, N., and Kimura, Y. (2001). Flow cytometric evidence for endopolyploidy in seedlings of some *Brassica* species. *Theor. Appl. Genet.* 102, 104–110. doi: 10.1007/s001220051624

Kullman, B. (2002). Nuclear DNA content, life cycle and ploidy in two *Neottiella* species (Pezizales, Ascomycetes). *Persoonia* 18, 103–115.

Kullman, B., Tamm, H., and Kullman, K. (2005). *Fungal Genome Size Database*. Available online at: http://www.zbi.ee/fungal-genomesize/

Laundon, G. F. (1967). Terminology in the rust fungi. *Trans. Br. Mycol. Soc.* 50, 189–194.

Loureiro, J., Rodriguez, E., Doležel, J., and Santos, C. (2006). Flow cytometric and microscopic analysis of the effect of tannic acid on plant nuclei and estimation of DNA content. *Ann. Bot.* 98, 515–527. doi: 10.1093/aob/mcl140

Loureiro, J., Rodriguez, E., Doležel, J., and Santos, C. (2007). Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. *Ann. Bot.* 100, 875–888. doi: 10.1093/annbot/mcm152

Nemri, A., Saunders, D. G., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Park, R. F., and Wellings, C. R. (2012). Somatic hybridization in the Uredinales. *Annu. Rev. Phytopathol.* 50, 219–239. doi: 10.1146/annurev-phyto-072910-095405

Preece, T. F. (2000). The strange story of box rust. *Mycologist* 14, 104–106. doi: 10.1016/S0269-915X(00)80018-X

Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10, 417–430. doi: 10.1038/nrmicro2790

R Core Team. (2014). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: http://www.R-project.org/

Silva, M. C., Várzea, V. M., Guerra-Guimarães, L., Azinheira, H., Fernandez, D., Petitot, A. S., et al. (2006). Coffee resistance to the main diseases: leaf rust and coffee berry disease. *Braz. J. Plant Physiol.* 18, 119–147. doi: 10.1590/S1677-04202006000100010

Skalamera, D., and Heath, M. C. (1998). Change in the cytoskeleton accompanying infection induced nuclear movements and the hypersensitive response in plant cells invaded by rust fungi. *Plant J.* 16, 191–200. doi: 10.1046/j.1365-313x.1998.00285.x

Spanu, P. D. (2012). The genomics of obligate (and nonobligate) biotrophs. *Annu. Rev. Phytopathol.* 50, 91–109. doi: 10.1146/annurev-phyto-081211-173024

Stark-Urnau, M., and Mendgen, K. (1993). Differentiation of aecidiospore and uredospore-derived infection structures on cowpea leaves and on artificial surfaces by *Uromyces vignae*. *Can. J. Bot.* 71, 1236–1242.

Stukenbrock, E. H., and Croll, D. (2014). The evolving fungal genome. *Fungal Biol. Rev.* 28, 1–12. doi: 10.1016/j.fbr.2014.02.001

Tan, M.-K., Collins, D. I., Chen, Z., Englezou, A., and Wilkins, M. R. (2014). A brief overview of the size and composition of the myrtle rust genome and its taxonomic status. *Mycology* 5, 52–63. doi: 10.1080/21501203.2014.919967

van der Merwe, M. M., Walker, J., Ericson, L., and Burdon, J. J. (2008). Coevolution with higher taxonomic host groups within the *Puccinia/Uromyces* rust lineage obscured by host jumps. *Mycol. Res.* 112, 1387–1408. doi: 10.1016/j.mycres.2008.06.027

Williams, P. G., and Mendgen, K. W. (1975). Cytofluorometry of DNA in uredospores of *Puccinia graminis* f. sp. *tritici*. *Trans. Br. Mycol. Soc.* 64, 23–28.

Yun, H. Y., Minnis, A. M., Castlebury, L. A., and Aime, M. C. (2011). The rust genus *Frommeëlla* revisited: a later synonym of *Phragmidium* after all. *Mycologia* 103, 1451–1463. doi: 10.3852/11-120

Zeze, A., Hosny, M., Gianinazzi-Pearson, V., and Dulieu, H. (1996). Characterization of a highly repeated DNA sequence (SC1) from the arbuscular mycorrhizal fungus *Scutellospora castanea* and its detection in planta. *Appl. Environ. Microbiol.* 62, 2443–2448.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 07 August 2014; published online: 26 August 2014.

Citation: Tavares S, Ramos AP, Pires AS, Azinheira HG, Caldeirinha P, Link T, Abranches R, Silva MC, Voegele RT, Loureiro J and Talhinhas P (2014) Genome size analyses of Pucciniales reveal the largest fungal genomes. Front. Plant Sci. ***5****:422. doi: 10.3389/fpls.2014.00422*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Tavares, Ramos, Pires, Azinheira, Caldeirinha, Link, Abranches, Silva, Voegele, Loureiro and Talhinhas. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

PERSPECTIVE ARTICLE
published: 27 August 2014
doi: 10.3389/fpls.2014.00377

On the current status of *Phakopsora pachyrhizi* genome sequencing

Marco Loehrer[1], Alexander Vogel[2], Bruno Huettel[3], Richard Reinhardt[3], Vladimir Benes[4], Sébastien Duplessis[5,6], Björn Usadel[2,7] and Ulrich Schaffrath[1]*

[1] Department of Plant Physiology, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany
[2] Institute for Botany and Molecular Genetics, Institute for Biology I, Rheinisch-Westfälische Technische Hochschule Aachen University, Aachen, Germany
[3] Max Planck Institute for Plant Breeding Research, Köln, Germany
[4] Genomics Core Facility, European Molecular Biology Laboratory, Heidelberg, Germany
[5] Institut National de la Recherche Agronomique, Interactions Arbres/Microorganismes, UMR 1136, Champenoux, France
[6] Université de Lorraine, Interactions Arbres/Microorganismes, UMR 1136, Vandoeuvre-lès-Nancy, France
[7] Institute of Bio- and Geosciences-2 Plant Sciences, Institute for Bio- and Geosciences, Forschungszentrum Jülich, Jülich, Germany

Edited by:
David L. Joly, Université de Moncton, Canada

Reviewed by:
John Fellers, United States Department of Agriculture – Agricultural Research Service, USA
Ralf Thomas Voegele, Universität Hohenheim, Germany

***Correspondence:**
Ulrich Schaffrath, Department of Plant Physiology, RWTH Aachen University, Worringerweg 1, 52056 Aachen, Germany
e-mail: schaffrath@bio3.rwth-aachen.de

Recent advances in the field of sequencing technologies and bioinformatics allow a more rapid access to genomes of non-model organisms at sinking costs. Accordingly, draft genomes of several economically important cereal rust fungi have been released in the last 3 years. Aside from the very recent flax rust and poplar rust draft assemblies there are no genomic data available for other dicot-infecting rust fungi. In this article we outline rust fungus sequencing efforts and comment on the current status of *Phakopsora pachyrhizi* (Asian soybean rust) genome sequencing.

Keywords: fungal genomics, rust fungi, Asian soybean rust, next-generation sequencing, herterozygosity, genome size, k-mer analysis

Sequencing of fungal genomes represented a significant milestone in the emerging era of "genomics." In fact, the first eukaryotic genome ever sequenced was that of baker's yeast, *Saccharomyces cerevisiae*, which consequently strengthened its position as a fungal model organism after the release of the 12 Mb genome with approximately 6000 genes in 1996 (Goffeau et al., 1996). Some time thereafter the genomes of the fission yeast *S. pombe* (14 Mb) and the filamentous ascomycete *Neurospora crassa* (40 Mb) were released in Wood et al. (2002) and Galagan et al. (2003), respectively. Accelerated progress in sequencing technology from early clone-by-clone approaches through Sanger-based whole-genome shotgun sequencing (WGS) to today's next-generation sequencing (NGS) shortened the periods between releases of novel genomes considerably (Grigoriev, 2014). This paved the way for comparative genomics which opened new possibilities for people working in the field of agriculture and biotechnology or combating human, animal or plant diseases (Vebø et al., 2009; Manning et al., 2013; Bolger et al., 2014).

In the latter field, the sequencing of the genome of the ascomycete *Magnaporthe oryzae* was achieved by Dean et al. (2005). Along with the genome of rice (Goff et al., 2002), the *M. oryzae* host plant, an understanding of the plant–pathogen interaction became possible at the genome level. Since then, several plant-pathogenic fungi were sequenced; however, a group of pathogens that exclusively feed from living plant tissue, so-called obligate biotrophs, remained recalcitrant. This was disappointing particularly because some of the most economically serious threats to human nutrition, such as powdery mildew fungi and rust fungi, are among this group.

Rust fungi have long been in the focus of plant pathologists. Already in the 19th century, Anton de Bary, who is considered as a founder of plant pathology, picked up *Puccinia graminis* with its various *formae speciales* that are specialized for parasitism on particular cereal hosts, as subject for his groundbreaking studies. Later Harold Henry Flor developed the famous "gene-for-gene" concept based on his work on the interaction of flax rust (*Melampsora lini*) with its host plant flax (*Linum usitatissimum*; Flor, 1955). Despite considerable interest, sequencing of rust genomes was not achieved until most recently. Thus, the 101 Mb genome of *Melampsora larici-populina* and the 89 Mb draft genome of *Puccinia graminis* f. sp. *tritici* were sequenced in a common effort by the Joint Genome Institute and the Broad Institute, respectively, and published in Duplessis et al. (2011). Following, more or less advanced draft genomes of other rust fungi were sequenced and published by the community, such as several *Puccinia striiformis* f. sp. *tritici* races (56–110 Mb) and the flax rust genome *M. lini* (Cantu et al., 2011, 2013; Zheng et al., 2013; Nemri et al., 2014; see **Table 1**). Although Pucciniales is an order with a lesser coverage compared to other fungi[1], more genomic resources are becoming accessible. A major drawback encountered during sequencing efforts of rust genomes was their unexpected large sizes, a fact that also hampered attempts of sequencing the genome of the Asian soybean rust fungus *Phakopsora pachyrhizi*, an economically important threat to soybean cultivation. The following commentary is written to give an overview on the current status of *P. pachyrhizi* genome

[1] http://genome.jgi.doe.gov/programs/fungi/1000fungalgenomes.jsf

Table 1 | Published rust fungi draft genomes and genome size estimations in alphabetical order.

Organism	Genome size (Mb)	Estimation/sequencing method	Reference
Cronartium quercuum f. sp. *fusiforme*	90	Flow cytometry	Anderson et al. (2010)
Hemileia vastatrix	733.5	Flow cytometry	Carvalho et al. (2013)
Melampsora larici-populina	101	Sanger sequencing	Duplessis et al. (2011)
Melampsora lini	189	Next-generation sequencing	Nemri et al. (2014)
Puccinia coronata	77	Flow cytometry	Eilam et al. (1994)
Puccinia graminis f. sp. *tritici*	89	Sanger sequencing	Duplessis et al. (2011)
Puccinia hordei	121	Flow cytometry	Eilam et al. (1994)
Puccinia recondita	127	Flow cytometry	Eilam et al. (1994)
Puccinia sorghi	102	Flow cytometry	Eilam et al. (1994)
Puccinia striiformis f. sp. *tritici*, race PST-130	65	Next-generation sequencing	Cantu et al. (2011)
Puccinia striiformis f. sp. *tritici*, races PST-21, PST-43, PST-87/7, PST-08/21	73, 71, 53, 56	Next-generation sequencing	Cantu et al. (2013)
Puccinia striiformis f. sp. *tritici*, isolate CY32	110	Next-generation sequencing, "fosmid-to-fosmid" sequencing	Zheng et al. (2013)
Puccinia triticiana 1-1 BBDB race 1	135	Assembly	Fellers et al. (2013)
Uromyces appendiculatus	418	Flow cytometry	Eilam et al. (1994)
Uromyces vignae	407	Flow cytometry	Eilam et al. (1994)

sequencing and is intended to initiate combined activities toward this goal.

What makes *P. pachyrhizi* so interesting? For sure it is a devastating fungal disease of the important crop plant soybean. The origin of the pathogen can be traced back to Asia and most likely it spread alongside with the propagation of soybean cultivation. *P. pachyrhizi* is able to infect more than 31 species from 17 genera of legumes, which is a rather unusual feature for rust fungi that usually are highly specialized for particular hosts (Goellner et al., 2010). *P. pachyrhizi* differs in a further important aspect from the majority of rusts: it directly penetrates leaf cells rather than entering the leaf via stomata at the uredinial stage. On the contrary, most rust fungi use stomata to get inside the host tissues at this stage and a direct penetration is only observed for some rust fungi when basidiospores infect the aecial host at later stages of the rust life cycle (Heath, 1997). Recent studies imply that generation of high turgor pressure of around 5 MPa in the non-melanized appressoria supports penetration (Loehrer et al., 2014). Penetrated epidermal cells undergo a cell death response, again an unexpected property for a biotrophic pathogen. Experiments with non-host plants such as barley and Arabidopsis showed that during penetration and concomitant epidermal cell death, marker genes associated with responses to necrotrophic pathogens are switched on and that cell death suppression had a negative influence on infection success of *P. pachyrhizi* (Loehrer et al., 2008; Hoefle et al., 2009). Regarding its lifestyle, *P. pachyrhizi* which forms so far only a single spore type in the wild, i.e. urediospores, is a minimalist compared to, e.g., *Puccinia graminis* f. sp. *tritici* which has five distinct spore types and performs a host jump (Leonard and Szabo, 2005). Despite the unknown or missing sexual life cycle the genetic diversity of *P. pachyrhizi* seems not to be impaired. This may be explained by parasexual nuclear recombination occurring between different isolates after germ tube fusion or hyphal anastomosis, a feature also reported for cereal rusts (Wang and McCallum, 2009; Vittal et al., 2012).

Public information about the *P. pachyrhizi* genome sequencing project is rare. In the DoE JGI Community Sequencing Program of 2004, a project was launched to sequence the genome *of P. pachyrhizi* (isolate Taiwan 72-1) based on a fosmid shotgun sequencing approach. The genome size prediction with 50 Mb at that time was much underestimated. The sequencing project has now a "permanent draft" status at the JGI[2]. Besides the recently released mitochondrial genome sequence (Stone et al., 2010), information on assembly attempts of the nuclear genome have not been published. The major drawback for progress in

[2]http://genome.jgi.doe.gov/genome-projects/pages/project-status.jsf?projectId=16847

P. pachyrhizi genome sequencing seems to be its huge size. An update on this topic was given at the National Soybean Rust Symposium 2005 in Nashville (TN, USA). Genome size estimations ranged from 300 to 950 Mb depending on the analysis method used (Posada-Buitrago et al., 2005). A similar statement was provided by Igor Grigoriev (Head of the JGI Fungal Program) suggesting a genome size above 850 Mb (Duplessis et al., 2012). Besides, other general features of rust fungi genomes unraveled since then, such as expanded multigene families and very large amount of transposable elements (>45%), pose serious problems for proper genome assembly.

We started our own efforts toward uncovering the genome size of *P. pachyrhizi* by using our lab isolate (Brazil 05-1) and we followed a strategy based on k-mer analysis. By breaking down the reads obtained by Illumina sequencing into short nucleotide sequences of defined length k (k-mers), several characteristics of genomes, like size, heterozygosity and repeat content, can be analyzed, that would normally require a complete *de novo* assembly. As basis for our analysis, DNA was generated from urediospores of the *P. pachyrhizi* isolate Brazil 05-1. A total of 47 Gb Illumina whole-genome sequencing data (100 bp paired-end reads) were then subjected to analysis using the program JELLYFISH (Marçais and Kingsford, 2011). In the 17-mer distribution depicted in **Figure 1**, two peaks could be differentiated at a depth of 37 and 75. This can be explained by the dikaryotic nature of the urediospores of rust fungi, which means that these organisms maintain two haploid nuclei separately during prolonged stages of their lifecycle. The two peaks in the k-mer histogram point to a high degree of heterozygosity between the two nuclei or to largely heterozygotic regions within the haploid nuclei. Similar results were also observed by Zheng et al. (2013) in the case of the wheat stripe rust fungus.

By adding up the products of k-mer depth coverage and frequency for each pair of values in **Figure 1**, divided by the depth coverage of the first peak (=37), the size of the genome in bp was computed, similarly as in (Li et al., 2010). Values, smaller than the first minimum in **Figure 1**, were considered noise caused by sequencing errors and were excluded from the calculation. Based on this analysis, the overall size of the dikaryotic genome of *P. pachyrhizi* is at most around 1 Gb. However, due to the unknown degree of heterozygosity between or within the genomes of both nuclei, this might be an overestimation (see above). The minimal size of the haploid genome can be estimated to be around 500 Mb, based on the second peak in the k-mer analysis (**Figure 1**). This would place the genome of the Asian soybean fungus in the same range as published rust genomes, e.g., *Hemileia vastatrix* (733.5 Mb) and *Uromyces* spp. (420 Mb; **Table 1**). It should be noted, however, that the analysis method might considerably influence the outcome of such genome size estimations. The genome size of *H. vastatrix*, e.g., was estimated by DNA-staining in combination with flow cytometry which itself is prone to errors but has the advantage of not being sequencing-dependent (Bainard et al., 2010). Phenomena related to the partial heterozygosity of the *P. pachyrhizi* genome are only detectable by assembly or k-mer analysis as described above. Since we did not use a large insert size sequencing approach for genome size estimation, we obtained a N50 value of 569 bp after assembly and scaffolding with SOAPdenovo. This allowed no prediction on gene number or length. In future studies a combined BAC- and third generation sequencing approach hopefully will increase the assembly quality to a point at which comprehensive gene predictions become possible.

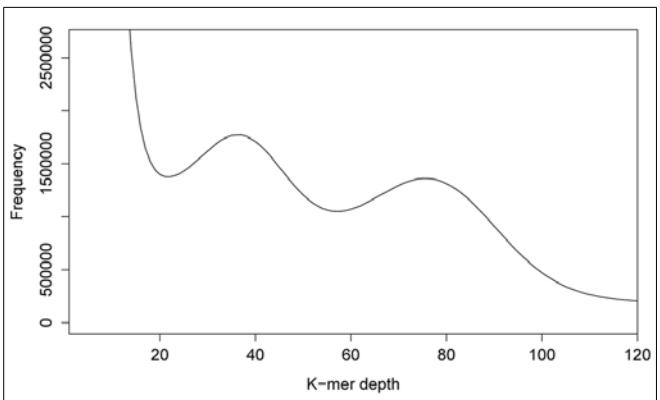


FIGURE 1 | K-mer analysis for *P. pachyrhizi* whole-genome sequencing data. The 17-mer distribution for 47 Gb of 100 bp paired-end Illumina whole-genome sequencing data indicates two peaks at a depth of 37 and 75. These findings point to a possibly highly repetitive genome with a high degree of heterozygosity between the genomes of the two haploid nuclei.

Working with organisms, whose genome has been sequenced provides many advantages over working with non-sequenced species. Besides the comprehensive prediction of all genes, intra-genomic structural analyses or comparative genome analyses between different species become possible. An alternative to genomic-based approaches in large-scale analyses of plant-pathogen-interactions, however, is the use of transcriptomics, proteomics, or metabolomics (Tan et al., 2009). Up to now, only limited information is available on *P. pachyrhizi* transcriptomics, though very recent publications have broadened the view on particular aspects of the infection process of *P. pachyrhizi* (Tremblay et al., 2010, 2012, 2013; Link et al., 2013). For instance, Illumina-based transcriptome profiling at several stages of soybean leaf infection has led to the identification of nearly 19,000 transcripts not previously identified in other rust fungi (Tremblay et al., 2013). This would imply a much larger gene complement in the soybean rust than in other rust fungi. So far, the numbers of genes reported in rust fungi are between 15,000 and 20,000 genes (Duplessis et al., 2014). Although biases in the RNA-Seq approach can not be excluded, it is possible that the *P. pachyrhizi* genome has experienced a high level of gene duplication during its evolution along with important transposable element activity that could explain the huge genome size predicted for this species. There is an urgent need for genome sequences as prerequisite for accurate large scale expression analysis and more RNA-seq efforts are needed. Without a genome, transcript reads have to be assembled first and not only RNA quality and sequencing technique used will influence the resulting assembly quality but also the algorithms used for assembly. And even if these problems could be sufficiently solved, the resulting contigs are much smaller than transcribed ORFs, limiting for example predictions of putatively secreted proteins. Also, redundancy within gene families

could be better resolved when compared to a reference genome sequence.

Hopefully in the near future, the development of novel sequencing and assembly strategies, together with dropping costs for NGS, will make the sequencing of large and complex genomes more affordable and will help to unravel the secrets of the genome of *P. pachyrhizi*.

ACKNOWLEDGMENTS

We thank Ralph Panstruga (Biology I, RWTH Aachen) for helpful discussions. We thank Anthony Bolger (Biology I, RWTH Aachen University) for helpful advice on k-mer analysis. Sébastien Duplessis acknowledges the ANR "Investissements d'Avenir" program (ANR-11-LABX-0002-01, Lab of Excellence ARBRE).

REFERENCES

Anderson, C. L., Kubisiak, T. L., Nelson, C. D., Smith, J. A., and Davis, J. M. (2010). Genome size variation in the pine fusiform rust pathogen Cronartium quercuum f.sp. fusiforme as determined by flow cytometry. *Mycologia* 102, 1295–1302. doi: 10.3852/10-040

Bainard, J. D., Fazekas, A. J., and Newmaster, S. G. (2010). Methodology significantly affects genome size estimates: quantitative evidence using bryophytes. *Cytometry* 77, 725–732. doi: 10.1002/cyto.a.20902

Bolger, M. E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., and Mayer, K. F. (2014). Plant genome sequencing — applications for crop improvement. *Curr. Opin. Biotechnol.* 26, 31–37. doi: 10.1016/j.copbio.2013.08.019

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Carvalho, G. M. A., Carvalho, C. R., Barreto, R. W., and Evans, H. C. (2013). Coffee rust genome measured using flow cytometry: does size matter? *Plant Pathol.* doi: 10.1111/ppa.12175

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., et al. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* 434, 980–986. doi: 10.1038/nature03449

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). "Advancing knowledge on biology of rust fungi through genomics," in *Advances in Botanical Research*, 1st Edn, Vol. 70, ed. F. Martin (London: Elsevier), 173–209.

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Joly, D. J., and Dodds, P. N. (2012). "Rust Effectors," in *Effectors in Plant-Microbe Interactions*, 1st Edn, eds F. Martin and S. Kamoun (Chichester: John Wiley and Sons, Ltd), 155–193.

Eilam, T., Bushnell, W. R., and Anikster, Y. (1994). Relative nuclear DNA content of rust fungi estimated by flow cytometry of propidium iodide-stained pycniospores. *Phytopathology* 84, 728–735. doi: 10.1094/Phyto-84-728

Fellers, J. P., Soltani, B. M., Bruce, M., Linning, R., Cuomo, C. A, Szabo, L. J., et al. (2013). Conserved loci of leaf and stem rust fungi of wheat share synteny interrupted by lineage-specific influx of repeat elements. *BMC Genomics* 14:60. doi: 10.1186/1471-2164-14-60

Flor, H. H. (1955). Host-parasite interaction in flax rust - its genetics and other implications. *Phytopathology* 45, 680–685.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., et al. (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422, 859–868. doi: 10.1038/nature01554

Goellner, K., Loehrer, M., Langenbach, C., Conrath, U., Koch, E., and Schaffrath, U. (2010). *Phakopsora pachyrhizi*, the causal agent of Asian soybean rust. *Mol. Plant Pathol.* 11, 169–177. doi: 10.1111/j.1364-3703.2009.00589.x

Goff, S. A., Ricke, D., Lan, T.-H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). *Science* 296, 92–100. doi: 10.1126/science.1068275

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 546–567. doi: 10.1126/science.274.5287.546

Grigoriev, I. V. (2014). "A changing landscape of fungal genomics," in *Ecological Genomics of the Fungi*, 1st Edn, ed. F. Martin (Chichester: John Wiley and Sons, Ltd), 3–20.

Heath, M. C. (1997). Signalling between pathogenic rust fungi and resistant or susceptible host plants. *Ann. Bot.* 80, 713–720. doi: 10.1006/anbo.1997.0507

Hoefle, C., Loehrer, M., Schaffrath, U., Frank, M., Schultheiss, H., and Hückelhoven, R. (2009). Transgenic suppression of cell death limits penetration success of the soybean rust fungus *Phakopsora pachyrhizi* into epidermal cells of barley. *Phytopathology* 99, 220–226. doi: 10.1094/PHYTO-99-3-0220

Leonard, K. J., and Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol. Plant Pathol.* 6, 99–111. doi: 10.1111/j.1364-3703.2005.00273.x

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463, 311–317. doi: 10.1038/nature08696

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2013). The haustorial transcriptomes of *Uromyces appendiculatus* and *Phakopsora pachyrhizi* and their candidate effector families. *Mol. Plant Pathol.* 15, 379–393. doi: 10.1111/mpp.12099

Loehrer, M., Botterweck, J., Jahnke, J., Mahlmann, D. M., Gaetgens, J., Oldiges, M., et al. (2014). In vivo assessment of the invasive force exerted by the Asian soybean rust fungus by mach-zehnder double-beam interferometry. *New Phytol.* 203, 620–631. doi: 10.1111/nph.12784

Loehrer, M., Langenbach, C., Goellner, K., Conrath, U., and Schaffrath, U. (2008). Characterization of nonhost resistance of *Arabidopsis* to the Asian soybean rust. *Mol. Plant Microbe Interact.* 21, 1421–1430. doi: 10.1094/MPMI-21-11-1421

Manning, V. A., Pandelova, I., Dhillon, B., Wilhelm, L. J., Goodwin, S. B., Berlin, A. M., et al. (2013). Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3 (Bethesda)* 3, 41–63. doi: 10.1534/g3.112.004044

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Nemri, A., Saunders, D. G. O., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Posada-Buitrago, M. L., Boore, J. L., and Frederick, R. D. (2005). "Soybean Rust Genome Sequencing Project," in *Proceedings of the National Soybean Rust Symposium*, Nashville, TN. Available at: http://www.plantmanagementnetwork.org/infocenter/topic/soybeanrust/symposium/posters/3.pdf [accessed November 14–16, 2005].

Stone, C. L., Posada-Buitrago, M. L., Boore, J. L., and Frederick, R. D. (2010). Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *Phakopsora pachyrhizi* and *P. meibomiae*. *Mycologia* 102, 887–897. doi: 10.3852/09-198

Tan, K.-C., Ipcho, S. V. S., Trengove, R. D., Oliver, R. P., and Solomon, P. S. (2009). Assessing the impact of transcriptomics, proteomics and metabolomics on fungal phytopathology. *Mol. Plant Pathol.* 10, 703–715. doi: 10.1111/j.1364-3703.2009.00565.x

Tremblay, A., Hosseini, P., Alkharouf, N. W., Li, S., and Matthews, B. F. (2010). Transcriptome analysis of a compatible response by Glycine max to *Phakopsora pachyrhizi* infection. *Plant Sci.* 179, 183–193. doi: 10.1016/j.plantsci.2010.04.011

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2012). Identification of genes expressed by *Phakopsora pachyrhizi*, the pathogen causing soybean rust, at a late stage of infection of susceptible soybean leaves. *Plant Pathol.* 61, 773–786. doi: 10.1111/j.1365-3059.2011.02550.x

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2013). Analysis of *Phakopsora pachyrhizi* transcript abundance in critical pathways at four time-points during infection of a susceptible soybean cultivar using deep sequencing. *BMC Genomics* 14:614. doi: 10.1186/1471-2164-14-614

Vebø, H. C., Snipen, L., Nes, I. F., and Brede, D. A. (2009). The transcriptome of the nosocomial pathogen *Enterococcus faecalis* V583 reveals adaptive responses to growth in blood. *PLoS ONE* 4:e7660. doi: 10.1371/journal.pone.0007660

Vittal, R., Yang, H.-C., and Hartman, G. L. (2012). Anastomosis of germ tubes and migration of nuclei in germ tube networks of the soybean rust pathogen, *Phakopsora pachyrhizi. Eur. J. Plant Pathol.* 132, 163–167. doi: 10.1007/s10658-011-9872-5

Wang, X., and McCallum, B. (2009). Fusion body formation, germ tube anastomosis, and nuclear migration during the germination of urediniospores of the wheat leaf rust fungus, *Puccinia triticina. Phytopathology* 99, 1355–1364. doi: 10.1094/PHYTO-99-12-1355

Wood, V., Gwilliam, R., Rajandream, M.-A., Lyne, M., Lyne, R., Stewart, A., et al. (2002). The genome sequence of *Schizosaccharomyces pombe. Nature* 415, 871–880. doi: 10.1038/nature724

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4, 2673. doi: 10.1038/ncomms3673

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2014; accepted: 14 July 2014; published online: 27 August 2014.
Citation: Loehrer M, Vogel A, Huettel B, Reinhardt R, Benes V, Duplessis S, Usadel B and Schaffrath U (2014) On the current status of Phakopsora pachyrhizi genome sequencing. Front. Plant Sci. ***5****:377. doi: 10.3389/fpls.2014.00377*
This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.
Copyright © 2014 Loehrer, Vogel, Huettel, Reinhardt, Benes, Duplessis, Usadel and Schaffrath. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

PERSPECTIVE ARTICLE
published: 29 October 2014
doi: 10.3389/fpls.2014.00587

Early insights into the genome sequence of *Uromyces fabae*

Tobias Link*, Christian Seibel and Ralf T. Voegele*

Fachgebiet Phytopathologie, Institut für Phytomedizin, Fakultät Agrarwissenschaften, Universität Hohenheim, Stuttgart, Germany

Edited by:
Sébastien Duplessis, Institut National de la Recherche Agronomique, France

Reviewed by:
Claude Murat, Institut National de la Recherche Agronomique, France
Yao-Cheng Lin, Vlaams Instituut voor Biotechnologie/Ghent University, Belgium

****Correspondence:***
Tobias Link and Ralf T. Voegele, Fachgebiet Phytopathologie, Institut für Phytomedizin, Fakultät Agrarwissenschaften, Universität Hohenheim, Otto-Sander-Str. 5, 70599 Stuttgart, Germany
e-mail: tobias.link@uni-hohenheim.de; ralf.voegele@uni-hohenheim.de

Uromyces fabae is a major pathogen of broad bean, *Vicia faba*. *U. fabae* has served as a model among rust fungi to elucidate the development of infection structures, expression and secretion of cell wall degrading enzymes and gene expression. Using *U. fabae*, enormous progress was made regarding nutrient uptake and metabolism and in the search for secreted proteins and effectors. Here, we present results from a genome survey of *U. fabae*. Paired end Illumina sequencing provided 53 Gb of data. An assembly gave 59,735 scaffolds with a total length of 216 Mb. K-mer analysis estimated the genome size to be 329 Mb. Of a representative set of 23,153 predicted proteins we could annotate 10,209, and predict 599 secreted proteins. Clustering of the protein set indicates families of highly likely effectors. We also found new homologs of RTP1p, a prototype rust effector. The *U. fabae* genome will be an important resource for comparative analyses with *U. appendiculatus* and *P. pachyrhizi* and provide information regarding the phylogenetic relationship of the genus *Uromyces* with respect to other rust fungi already sequenced, namely *Puccinia graminis* f. sp. *tritici*, *P. striiformis* f. sp. *tritici*, *Melampsora lini*, and *Melampsora larici-populina*.

Keywords: *Uromyces fabae*, rust fungus, genome survey, genome size, candidate effectors, RTP1 homologs

INTRODUCTION

The order Pucciniales is dominated by two genera: *Puccinia* with about 4,000 species and *Uromyces* with around 600 species (Maier et al., 2003). Though the phylogeny of the two genera has not been entirely disentangled, there is a clear tendency that *Puccinia* species mainly infect grass hosts, whereas *Uromyces* species seem to be concentrated on legumes. There is now sequence information available for two species within the genus *Puccinia* [*Puccinia graminis* f. sp. *tritici* (*Pgt*), and *Puccinia striiformis* f. sp. *tritici* (*Pst*)] (Duplessis et al., 2011; Cantu et al., 2013; Zheng et al., 2013). Recently, re-sequencing of several *Pst* strains was undertaken, enabling the search for accelerated evolution among secreted proteins, an interesting means for identifying highly likely effector candidates (Cantu et al., 2013). At the same time no genome information is available for the second largest genus *Uromyces*. Therefore, a reference genome for *Uromyces* enabling comparisons between these two closely related genera would be highly desirable.

Rust fungi cannot truly be called model species since their obligate biotrophic lifestyle makes basic research on these species very difficult. General information regarding fungi or basidiomycetes can much easier be obtained with other species. Among the 7,000 rust fungi, research is concentrated on a very limited selection of species. One reason that brought these species into focus is the high economic losses associated with them. This is true for the cereal rusts *Pgt* and *Pst*, or the Asian Soybean Rust, *Phakopsora pachyrhizi*. The other reason why some rust species are prominent in research is their historical significance. The most important example here is *Melampsora lini* on *Linum usitatissimum*, the system upon which Flor developed the gene-for-gene hypothesis (Flor, 1956). These early successes have been picked up in modern molecular research, for example making the connection between Avr genes and proteins secreted from haustoria (Catanzariti et al., 2006), or providing information on the interaction between Avr and R-proteins (Dodds et al., 2006), up to contributions as to how effectors may reach their targets (Rafiqi et al., 2010).

Among *Uromyces* species a similar role can be assigned to *U. fabae*. Over the years especially morphologic studies using light and electron microscopy contributed to the elucidation of infection structures of rust fungi (Kapooria and Mendgen, 1985). Later, biochemical studies contributed to the understanding of the physiology of early infection (Deising et al., 1991). The development of a method to isolate haustoria from infected leaves by Hahn and Mendgen (1992) made it possible to study these hallmark structures of obligate biotrophic pathogens. Building on the study of PIGs (in planta induced genes, genes found highly expressed in haustoria) (Hahn and Mendgen, 1997), more molecular and biochemical studies followed. These studies provided proof for the importance of haustoria in nutrient uptake (Voegele et al., 2001), as well as the generation of energy (Sohn et al., 2000). Among the PIGs also the first non-avirulence protein shown to be transferred from the fungus into the host cytoplasm (*Uf*-RTP1p) was discovered (Kemen et al., 2005). Current research on *U. fabae* is focused on searching novel candidate effectors among secreted proteins (Link and Voegele, 2008), augmenting the knowledge on carbohydrate uptake and metabolism, and the quest for a generally applicable method to stably transform rust fungi (Djulic et al., 2011). Keeping up this tradition of research

on this species, *U. fabae* was the logical choice for generating a reference genome for the genus *Uromyces*. However, *U. fabae* could serve as a model not only for *Uromyces* species but also for other legume rusts, the most important of which at the moment is *P. pachyrhizi*.

Here, we report first results from a genome survey on *U. fabae*. Based on the results of this survey we expect to get a better understanding of the physiology of *U. fabae*. We found more candidate effectors, and we did and will do more comparisons of gene content against *Pgt* and *Pst* and also *Melampsora larici-populina* (*Mlp*). We plan to expand this survey into a full genome sequencing project.

SEQUENCING

DNA was prepared from urediospores of *U. fabae* isolate I2. This isolate has been in use in the Mendgen lab (Universität Konstanz, Germany) and the Voegele lab for many years. Virtually all experiments published on *U. fabae* were made using this strain. DNA was isolated from germinated urediospores using a protocol modified from Kolmer et al. (1995). Urediospores were washed for 30 min and germinated for 3.5 h. Germinated spores were homogenized by grinding in liquid N_2 and acid washed sand, and incubated in CTAB solution. Phenol-chloroform extraction, chloroform extraction, precipitation with 2-propanol, and RNaseA digest were performed to purify the DNA. Quality assessments showed only minor degradation and a slight bacterial contamination.

Paired end sequencing using Illumina HiSeq2000 with a 500 nt library was performed by BGI TECH SOLUTIONS (HONGKONG) CO. LIMITED (16 Dai Fu Street, Tai Po Industrial Estate, Tai Po, N.T., Hong Kong) who also supplied the assembly that is presented. 593,062,170 reads with a length of 90 bp were produced giving 53,375 Mb in raw data. 8.8% of the data were removed during filtering, leaving 48,661 Mb of clean data. The sequence reads were deposited in NCBI SRA in experiment SRX547322 corresponding to BioProject PRJNA248166.

Using SOAPdenovo reads were assembled into 59,735 scaffolds with a total length of 215,710,123 bp. N50 for the scaffolds is 5873 bp, the longest scaffold spans 72,118 bp, the shortest one 1,000 bp. These scaffolds were built from 95,847 contigs with a total length of 209,504,160 bp. N50 for the contigs is 4,171 bp, the longest contig is 45,252 bp, the shortest 200 bp. This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession JNCO00000000. The version described in this paper is version JNCO01000000.

The original assembly with 1,191,649 scaffolds that was used for prediction of proteins, is very fragmented, so it is not possible to draw definite conclusions. We hope to improve our data by further sequencing (i.e., pacbio sequencing) and performing new assemblies for example with realizing mate pair sequence. One objective of this publication is to spark interest in this genome so that other groups or organizations could add their expertise into the project. However, as mentioned above, we were most interested in the gene complement, especially in distinct metabolic pathways and secreted proteins as effector candidates, so nevertheless, we set out to do further analyses that are presented below. Here we want to supply the reader with this caveat: All analyses are based on a provisional assembly.

GENOME SIZE

Our data so far gave us three estimates on the size of the *U. fabae* genome. On the one hand we have two different k-mer analyses, a 15mer analysis that estimates the genome size to 329 Mb and a 17mer analysis that calculates to 330 Mb. On the other hand, there is the original assembly size of 422 Mb and the filtered assembly with 216 Mb. Apart from this, the genome size of *U. fabae* isolate I2 was recently measured using flow cytometry on isolated nuclei from germinated urediospores, giving an estimate of 379 Mb (Tavares et al., 2014). Given the preliminary nature of the assemblies we consider the k-mer analysis and the flow cytometric data as most reliable and thus estimate the actual genome size in the range between 330 and 379 Mb.

Analysis on other rust fungi has shown that compared to other fungi the genomes of rusts are fairly inflated. *U. fabae* seems to be no exception. The pioneering genome sequences of *Mlp* and *Pgt* also revealed a reason or the mechanism for these big genome sizes—a high amount of transposable elements (TE) (Duplessis et al., 2011). We assume that *U. fabae* likewise has a large amount of TE, probably more than *Mlp* and *Pgt*. The high amount of predicted proteins that were annotated as transposon related (see below) also seems to point in this direction. So far, an analysis of repeats and TE could not be performed, but given the large genome size, this will be an important part in the analysis of later assemblies.

A PRELIMINARY VIEW ON THE GENE COMPLEMENT

All data regarding predicted proteins (annotation, prediction as secreted, clustering) as described below are integrated in Supplementary Table 1.

PROTEIN PREDICTION AND ANNOTATION

To have an idea, how useful the original assembly could be despite its fragmented state, we analyzed it with CEGMA (core eukaryotic genes mapping approach, Parra et al. (2009). This analysis showed that of the 248 highly conserved CEGs 95% were at least partially, and 89% completely present. Thus, this indicates that a large portion of the gene complement should be represented in the current assembly. Without preceding prediction and masking of repeats we used the Augustus Web server (Hoff and Stanke, 2013) with the gene structure file from the CEGMA output as a training set and 590 available ESTs (Jakupovic et al., 2006; Link and Voegele, 2008) as "hints" to predict 70,913 proteins. Compared to the 17,773 protein coding genes predicted for *Pgt*, and the 16,399 for *Mlp* (Duplessis et al., 2011), this is a gross over-prediction—most likely due in large part to TEs. For a more accurate gene prediction a better assembly, prediction of repeats, and more information on gene structure and especially more cDNA sequence information will be necessary.

To get a workable dataset steps were taken to reduce the set of predicted proteins closer to realistic numbers. First, all predicted sequences were truncated to the first methionine and all sequences shorter than 80 aa were removed. To remove redundancy among the remaining 56,594 predicted proteins they were

clustered using the cd-hit-suite web server. Clustering with 0.7% ID as cutoff yielded 23,153 clusters, which seemed adequate. The representative proteins from this clustering were used for subsequent analyses.

Proteins were annotated using the Blast2GO suite. BLAST results could be obtained for 20,153 proteins, Gene Ontology (GO) terms were mapped for 14,085 proteins, and after integrating the InterProScan results and running Annex, 10,209 proteins could be annotated according to the Blast2GO rule.

The most important species among the BLAST hits in the NCBI nr database is *Puccinia graminis*, both for all hits and for best hits. The rest of the list is dominated by other fungal species, though surprisingly also plant and animal species are represented. Despite the result of the PCR on rDNA that predicted a bacterial contamination (and despite the omission of steps to remove this contamination), no bacterial species was prominent among the best hits. Almost half of the annotated proteins are transposon related indicating again that a new assembly will be necessary that should be masked against repeats with RepeatMasker.

FAMILIES OF SECRETED PROTEINS/CANDIDATE EFFECTORS

Using SignalP4 760 signal peptides were predicted. Of the proteins carrying a signal peptide 135 had additional transmembrane domains (predicted by TMHMM), two were predicted as mitochondrial by TargetP, and 33 carry a predicted glycophosphatidylinositol (GPI) anchor (predGPI); six proteins have both transmembrane domains and a GPI anchor. 599 predicted secreted proteins remained.

To identify candidate effectors, which we assume to be specific to rust fungi or even to a single species, all proteins were blasted (blast+) against the protein complement of 10 basidiomycete species, among them four rust fungi, a hemibiotrophic smut fungus, a biotrophic mutualistic symbiont, a close relative of rust and smut fungi and three saprotrophic fungi (see Supplementary Tables 1, 3). Using spectral clustering (SCPS), 1,315 clusters containing at least two proteins were formed. 18,908 proteins fell into these clusters. This way, several families of secreted proteins, and also specific to rust fungi or lineage specific could be identified. For a better overview a smaller clustering just for proteins with predicted signal peptide was performed, including secretome results for *U. fabae* and predicted secreted proteins of *U. appendiculatus* and *P. pachyrhizi* (Link and Voegele, 2008; Link et al., 2014). Of the clusters that were formed 191 contained predicted secreted proteins from *U. fabae*. 44 of these families contained proteins that were found secreted with the signal sequence trap (Link and Voegele, 2008), 121 also contained *U. appendiculatus* proteins, 66 *P. pachyrhizi* proteins. **Table 1** shows an overview of the 10 biggest families. Remarkably, only one of these families could be assigned a function, cluster 9, which is a family of expansins.

Table 1 | The 10 largest clusters of predicted secreted proteins.

No.	*U. fabae*	*U. appendiculatus*	*P. pachyrhizi*	YSST
1	34	12	0	1
2	3	15	14	1
3	4	4	19	0
4	5	13	5	3
5	15	4	3	2
6	24	0	0	0
7	1	22	0	1
8	6	14	0	2
9	9	7	3	1
10	7	7	2	3

Numbers indicate how many proteins of each species or how many secreted proteins determined by the yeast signal sequence trap (YSST) are present in the cluster.

One effector family that has held our interest for some time now is the RTP (rust transferred protein) family. Our latest findings on RTP1p indicate that the protein has proteinase inhibitor function (Pretsch et al., 2013). Other findings show, that the protein forms fibrils in the extrahaustorial matrix as well as in the host cytoplasm (Kemen et al., 2013). This may be associated with slowing down cyclosis in the host cell and/or keeping the plant cell nucleus in the immediate vicinity of the haustorium, thus ensuring close contact and a better influence of the pathogen on its host. While these two functions are not mutually exclusive, as the fibrils around the haustorium could have a protective function, it is highly unusual that a protein should have both a structural and an enzymatic function. So far, structural and functional analyses have been limited to *Uf*-RTP1p.

That *Uf*-RTP1p is a member of an extended gene family was shown in recent sequencing projects. Search for *RTP1* homologs using degenerate primers yielded additional information. The most recent summary of these homologs can be found in Pretsch et al. (2013). Using tblastn to search *RTP1* homologs against our assembly we could find—in addition to *Uf-RTP1*—two more homologs. One, located on scaffold6572, shows highest similarity to *Ua-RTP2* (*Ua*: *Uromyces appendiculatus*). According to the nomenclature proposed by Puthoff et al. (2008), it will be designated *Uf-RTP2*. The second homolog showed low similarity to several *RTP1* homologs and could not be clearly assigned a best match. Using as query *RTP* homologs from other rust fungi, we also found seven homologs to *Ua-RTP9*, the highest scoring of these is located on scaffold2295 and was designated *Uf-RTP9*. The similarity between the *Uf-RTP9* homologs and *Uf-RTP1* was so low however, that they did not cluster. Therefore, we find it reasonable to designate a new family, the *RTP9* family.

It will now be interesting to do corresponding experiments with the newly identified RTP homologs to check whether these have a similar duality of functions as *Uf*-RTP1p. This phenomenon seems all the more fascinating now that *RTP* was shown to be a gene family in *U. fabae* as well as in other rust species. It seems reasonable to presume that the different proteins could be expressed at different stages and secreted from different structures, as was shown for the *Mlp-RTP*s by Hacquard et al. (2012).

In additional analyses we want to systemize the families of predicted secreted proteins into those that have functional annotation, and those that are also lineage specific, which makes them likely effector candidates. We will search the secreted proteins and the protein families for common motifs, both novel motifs and motifs already linked to effector function and/or transfer

into the host cytoplasm. We will also build phylogenies for both all secreted proteins and selected families of high interest. We will also build groups of orthologs and identify those genes that have formed a high number of paralogs. Alignments of protein families will also help us to sort out proteins that were predicted as secreted because of truncations. These predictions will lead to wet-lab experiments, i.e., phenotypical screens like cell death suppression assays, search for interaction partners, silencing experiments and localizations. As indicated earlier, we will try to improve the assemblies and gene prediction and, given the opportunity, expand this survey into a full genome sequencing project.

ACKNOWLEDGMENTS

We thank Sibylle Berger for technical assistance. We are also grateful to the reviewers and the editor Sébastien Duplessis for useful suggestions that will help us to improve our work in later analyzes.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00587/abstract

REFERENCES

Cantu, D., Segovia, V., Maclean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Catanzariti, A. M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Deising, H., Jungblut, P. R., and Mendgen, K. (1991). Differentiation-related proteins of the broad bean rust fungus *Uromyces viciae-fabae*, as revealed by high resolution two-dimensional polyacrylamide gel electrophoresis. *Arch. Microbiol.* 155, 191–198. doi: 10.1007/BF00248616

Djulic, A., Schmid, A., Lenz, H., Sharma, P., Koch, C., Wirsel, S. G., et al. (2011). Transient transformation of the obligate biotrophic rust fungus *Uromyces fabae* using biolistics. *Fungal Biol.* 115, 633–642. doi: 10.1016/j.funbio.2011.03.007

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Teh, T., Wang, C. I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8888–8893. doi: 10.1073/pnas.0602577103

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Flor, H. H. (1956). The complementary genic systems in flax and flax rust. *Adv. Genet.* 8, 29–54. doi: 10.1016/S0065-2660(08)60498-8

Hacquard, S., Joly, D. L., Lin, Y. C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hahn, M., and Mendgen, K. (1992). Isolation by ConA binding of haustoria from different rust fungi and comparison of their surface qualities. *Protoplasma* 170, 95–103. doi: 10.1007/BF01378785

Hahn, M., and Mendgen, K. (1997). Characterization of *in planta*-induced rust genes isolated from a haustorium-specific cDNA library. *Mol. Plant Microbe Interact.* 10, 427–437. doi: 10.1094/MPMI.1997.10.4.427

Hoff, K. J., and Stanke, M. (2013). WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucl. Acids Res.* 41, W123–W128. doi: 10.1093/nar/gkt418

Jakupovic, M., Heintz, M., Reichmann, P., Mendgen, K., and Hahn, M. (2006). Microarray analysis of expressed sequence tags from haustoria of the rust fungus *Uromyces fabae*. *Fungal Genet. Biol.* 43, 8–19. doi: 10.1016/j.fgb.2005.09.001

Kapooria, R. G., and Mendgen, K. (1985). Infection structures and their surface changes during differentiation in *Uromyces fabae*. *J. Phytopathol.* 113, 317–323. doi: 10.1111/j.1439-0434.1985.tb04832.x

Kemen, E., Kemen, A., Ehlers, A., Voegele, R., and Mendgen, K. (2013). A novel structural effector from rust fungi is capable of fibril formation. *Plant J.* 75, 767–780. doi: 10.1111/tpj.12237

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Kolmer, J. A., Liu, J. Q., and Sies, M. (1995). Virulence and molecular polymorphism in *Puccinia recondita* f. sp. *tritici* in Canada. *Phytopathology* 85, 276–285. doi: 10.1094/Phyto-85-276

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2014). The haustorial transcriptomes of *Uromyces appendiculatus* and *Phakopsora pachyrhizi* and their candidate effector families. *Mol. Plant Pathol.* 15, 379–393. doi: 10.1111/mpp.12099

Link, T. I., and Voegele, R. T. (2008). Secreted proteins of *Uromyces fabae*: similarities and stage specificity. *Mol. Plant Pathol.* 9, 59–66. doi: 10.1111/j.1364-3703.2007.00448.x

Maier, W., Begerow, D., Weiß, M., and Oberwinkler, F. (2003). Phylogeny of the rust fungi: an approach using nuclear large subunit ribosomal DNA sequences. *Can. J. Bot.* 81, 12–23. doi: 10.1139/b02-113

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. *Nucl. Acids Res.* 37, 289–297. doi: 10.1093/nar/gkn916

Pretsch, K., Kemen, A. C., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. T. (2013). The rust transferred proteins—a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Puthoff, D. P., Neelam, A., Ehrenfried, M. L., Scheffler, B. E., Ballard, L., Song, Q., et al. (2008). Analysis of expressed sequence tags from *Uromyces appendiculatus* hyphae and haustoria and their comparison to sequences from other rust fungi. *Phytopathology* 98, 1126–1135. doi: 10.1094/PHYTO-98-10-1126

Rafiqi, M., Gan, P. H., Ravensdale, M., Lawrence, G. J., Ellis, J. G., Jones, D. A., et al. (2010). Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. *Plant Cell* 22, 2017–2032. doi: 10.1105/tpc.109.072983

Sohn, J., Voegele, R. T., Mendgen, K., and Hahn, M. (2000). High level activation of vitamin B1 biosynthesis genes in haustoria of the rust fungus *Uromyces fabae*. *Mol. Plant Microbe Interact.* 13, 629–636. doi: 10.1094/MPMI.2000.13.6.629

Tavares, S., Ramos, A. P., Pires, A. S., Azinheira, H. G., Caldeirinha, P., Link, T., et al. (2014). Genome size analysis of Pucciniales reveal the largest fungal genomes. *Front. Plant Sci.* 5:422. doi: 10.3389/fpls.2014.00422

Voegele, R. T., Struck, C., Hahn, M., and Mendgen, K. (2001). The role of haustoria in sugar supply during infection of broad bean by the rust fungus *Uromyces fabae*. *Proc. Natl. Acad. Sci. U.S.A.* 98, 8133–8138. doi: 10.1073/pnas.131186798

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4, 2673. doi: 10.1038/ncomms3673

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2014; accepted: 09 October 2014; published online: 29 October 2014.
Citation: Link T, Seibel C and Voegele RT (2014) Early insights into the genome sequence of Uromyces fabae. Front. Plant Sci. ***5****:587. doi: 10.3389/fpls.2014.00587*
This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.
Copyright © 2014 Link, Seibel and Voegele. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 10 September 2014
doi: 10.3389/fpls.2014.00454

Genome-wide patterns of segregation and linkage disequilibrium: the construction of a linkage genetic map of the poplar rust fungus *Melampsora larici-populina*

Michaël Pernaci[1,2], Stéphane De Mita[1,2], Axelle Andrieux[1,2], Jérémy Pétrowski[1,2], Fabien Halkett[1,2], Sébastien Duplessis[1,2] and Pascal Frey[1,2]*

[1] *Interactions Arbres - Micro organismes, Institut national de la recherche agronomique, UMR1136, Champenoux, France*
[2] *Interactions Arbres - Micro organismes, Université de Lorraine, UMR1136, Vandoeuvre-lès-Nancy, France*

Edited by:
Ton Bisseling, Wageningen University, Netherlands

Reviewed by:
Peter Dodds, Commonwealth Scientific and Industrial Research Organisation, Australia
Eric Kemen, Max Planck Institute for Plant Breeding Research, Germany

***Correspondence:**
Pascal Frey, Interactions Arbres - Micro organismes, INRA, UMR1136, Rue d'Amance, F-54280 Champenoux, France
e-mail: pascal.frey@nancy.inra.fr

The poplar rust fungus *Melampsora larici-populina* causes significant yield reduction and severe economic losses in commercial poplar plantations. After several decades of breeding for qualitative resistance and subsequent breakdown of the released resistance genes, breeders now focus on quantitative resistance, perceived to be more durable. But quantitative resistance also can be challenged by an increase of aggressiveness in the pathogen. Thus, it is of primary importance to better understand the genetic architecture of aggressiveness traits. To this aim, our goal is to build a genetic linkage map for *M. larici-populina* in order to map quantitative trait loci related to aggressiveness. First, a large progeny of *M. larici-populina* was generated through selfing of the reference strain 98AG31 (which genome sequence is available) on larch plants, the alternate host of the poplar rust fungus. The progeny's meiotic origin was validated through a segregation analysis of 115 offspring with 14 polymorphic microsatellite markers, of which 12 segregated in the expected 1:2:1 Mendelian ratio. A microsatellite-based linkage disequilibrium analysis allowed us to identify one potential linkage group comprising two scaffolds. The whole genome of a subset of 47 offspring was resequenced using the Illumina HiSeq 2000 technology at a mean sequencing depth of 6X. The reads were mapped onto the reference genome of the parental strain and 144,566 SNPs were identified across the genome. Analysis of distribution and polymorphism of the SNPs along the genome led to the identification of 2580 recombination blocks. A second linkage disequilibrium analysis, using the recombination blocks as markers, allowed us to group 81 scaffolds into 23 potential linkage groups. These preliminary results showed that a high-density linkage map could be constructed by using high-quality SNPs based on low-coverage resequencing of a larger number of *M. larici-populina* offspring.

Keywords: fungal pathogen, linkage mapping, genome mapping, genome sequencing, Mendelian segregation, single-nucleotide polymorphism, selfing, progeny

INTRODUCTION

Poplar is an important product for the wood industry worldwide (Heilman, 1999) and its contribution to energy systems has increased recently (Covarelli et al., 2013). Poplar rust, caused by the pathogenic fungus *Melampsora larici-populina* (Basidiomycota, Pucciniales), is the main phytosanitary constraint for commercial poplar cultivation in Europe and other parts of the world (Gérard et al., 2006; Barrès et al., 2008). In the last 50 years many rust-resistant cultivars were bred and released, but all the qualitative resistance genes (i.e., major resistance genes) released were overcome by pathogen evolution within a short period (Xhaard et al., 2011). Qualitative resistance is particularly subject to breakdown by pathogen evolution for perennial hosts, such as poplar trees, because of the wide inequality between the pathogen's rapid generation time and the time needed to deploy new host varieties (Xu, 2012). Knowledge of the genetic determinism of the virulence factors leading to resistance breakdown would be beneficial both from an academic perspective (e.g., to decipher interactions between avirulence loci and resistance loci, Dangl et al., 2013) and from an applied perspective (e.g., for determining strategies of spatiotemporal management of qualitative resistance, McDonald and Linde, 2002).

The failure of qualitative resistance genes to control poplar rust has prompted poplar breeders to search for quantitative resistance, which is supposed to be more durable (Jorge et al., 2005; Brun et al., 2010; Fabre et al., 2012). Durable resistance is defined as a resistance remaining effective in a cultivar for a long period of time during its widespread cultivation (Johnson, 1979). Nevertheless, quantitative resistance can also be challenged by the evolution of aggressiveness, which is the quantitative component of pathogenicity, determined by several disease-associated traits (Andrivon et al., 2007; Pariaud et al., 2009; Dowkiw et al., 2010).

Thus, it is of primary importance to assess the potential evolution of aggressiveness traits in the pathogen and the potential trade-offs between these traits (Lannou, 2012). Knowledge of the genetic determinism of such aggressiveness life history traits (latency period, infection efficiency, sporulation capacity, lesion size, etc.) would be useful in order to guide breeding strategies toward durable resistance. To this aim, a genetic linkage map of the poplar rust fungus would allow examination of the genetic architecture of those traits and mapping the QTLs related to aggressiveness.

The genome of *M. larici-populina* was shotgun sequenced at a coverage of ∼6.9X, assembled and annotated recently by the US Department of Energy Joint Genome Institute and an international consortium (Duplessis et al., 2011). The 101.1 Mbp genome is assembled into 462 scaffolds and contains a total of 16,399 predicted gene models. About half of the genome is contained in 27 scaffolds all at least 1.1 Mbp in length. Therefore, the construction of a genetic map of *M. larici-populina* and its integration with the physical map would enable us to identify a chromosomal order of scaffolds and provide a valuable resource for fine mapping and positional cloning of QTLs associated with aggressiveness traits (Hahn et al., 2014).

Compared to plant and animal models, there has been much less interest in genetic mapping for fungi (for a review, see Foulongne-Oriol, 2012) and Oomycetes (Sicard et al., 2003). Among Ascomycota, genetic linkage maps have been developed for some major pathogens of economically important crops, such as *Blumeria graminis* (Pedersen et al., 2002), *Magnaporthe oryzae* (Kaye et al., 2003), *Leptosphaeria maculans* (Cozijnsen et al., 2000), *Mycosphaerella graminicola* (Kema et al., 2002), *Mycosphaerella fijiensis* (Manzo-Sánchez et al., 2008), and *Venturia inaequalis* (Broggini et al., 2011). Among Basidiomycota, most of the mapping efforts have been devoted to economically important edible mushrooms, such as *Agaricus bisporus* (Foulongne-Oriol et al., 2010), *Lentinula edodes* (Terashima et al., 2002), and *Pleurotus ostreatus* (Larraya et al., 2000), and some model ectomycorrhizal fungi (Doudrick et al., 1995; Labbé et al., 2008). Obtaining controlled crosses (either selfed or outcrossed) is even more challenging for rust fungi (Pucciniales), since (i) they are obligate biotrophs, which precludes crosses to be made *in vitro*, and (ii) most of the rust fungi are heteroecious, thus the completion of sexual crosses requires two non-related host plants (Leonard and Szabo, 2005). As a result there are very few reports of genetic linkage maps in rust fungi. To our knowledge, partial genetic maps have been built only for the fusiform rust fungus, *Cronartium quercuum* f. sp. *fusiforme* (Doudrick et al., 1993; Kubisiak et al., 2011), and the wheat stem rust fungus, *Puccinia graminis* f. sp. *tritici* (Zambino et al., 2000). A genetic linkage map is also being constructed for the wheat leaf rust fungus, *Puccinia triticina* (Duplessis et al., 2014).

Genetic mapping is primarily based on the genotyping of a large number of individuals from a controlled cross, using PCR-based (SSR, AFLP, SNP, etc.) molecular markers. For a genetic map of sufficient density, the development of a large number of markers and the genotyping of a large number of progeny represent significant costs, both in time and in money. With the advent of next generation sequencing (NGS) techniques, combined with genome-wide marker discovery techniques such as reduced-representation sequencing (RRS), restriction-site-associated DNA sequencing (RAD-seq) or multiplexed shotgun genotyping (MSG), it is now possible to overcome the technical and financial constraints of genetic mapping (Davey et al., 2011). New technologies for high throughput sequencing, such as Illumina sequencing, open the way to new genotyping and genetic mapping strategies based on re-sequencing at low coverage of a large number of progeny (Huang et al., 2009; Xie et al., 2010). In addition, tagging techniques for multiplex sequencing of numerous individuals on a single Illumina sequencing lane further reduces the cost of re-sequencing (Cronn et al., 2008). This new approach has been successfully applied to build a ultra-high density genetic map of rice, through sequencing of 150 Recombinant Inbred Lines (RILs) to a depth of 0.02X (Huang et al., 2009). This new methodology was found about 20 times faster in data collection, and the linkage map obtained was 35 times more precise in recombination breakpoint determination, compared to the use of PCR-based markers. The accuracy of QTL detection was also improved (Huang et al., 2009; Xie et al., 2010; Yu et al., 2011). This new genetic mapping strategy has been recently applied to plants (Huang et al., 2012; Zhou et al., 2014) and animals (Andolfatto et al., 2011; You et al., 2013). It was also used for the first time to build a high-density sequence-based genetic linkage map for a fungus, the Shiitake mushroom, *L. edodes* (Au et al., 2013). This example provides a proof-of-principle that low-coverage resequencing could allow rapid genotyping of basidiospore-derived progenies, thus facilitating the construction of high-density genetic linkage maps of Basidiomycota for QTL mapping (e.g., of aggressiveness traits in the case of phytopathogenic fungi) and improvement of whole-genome assembly.

The purpose of this work was (i) to generate S1 progeny of *M. larici-populina* through selfing of the reference strain 98AG31 on larch plants; (ii) to characterize the S1 progeny through a segregation analysis of polymorphic microsatellite loci; (iii) to test for linkage disequilibrium between these loci in the progeny; and (iv) to perform a pilot study of genetic mapping through whole-genome resequencing of a subset of 47 S1 individuals.

MATERIALS AND METHODS

PRODUCTION OF THE *M. LARICI-POPULINA* S1 PROGENY

Since *M. larici-populina* is a heteroecious and macrocyclic rust fungus, the completion of the life cycle requires two unrelated host plants, poplar and larch [for detailed life cycles of *Melampsora* spp. and *M. larici-populina* see Vialle et al. (2011) and Hacquard et al. (2011), respectively].

Production of M. larici-populina telia

Poplar plants (*Populus deltoides* × *P. nigra* 'Robusta') were grown from dormant cuttings in 5-l pots containing a sand-peat (50:50, v/v) mixture, with an initial fertilization of 3.5 $g.l^{-1}$ $CaCO_3$ and 6 $g.l^{-1}$ of slow release 13:13:13 N:P:K fertilizer (Nutricote T 100). The plants were grown for 4 months (June–September) in a non-heated greenhouse with natural photoperiod and were watered daily with deionized water. After 4 months, young trees were about 1.2 m high and exhibited 25–30 fully expanded leaves.

In late September, 20 poplar plants were spray-inoculated by *M. larici-populina* 98AG31. A urediniospore suspension (40,000 urediniospores.ml^{-1}) was sprayed (ca. 1.5 ml per leaf) on the abaxial (lower) surface of each fully expanded leaf with a fine atomization paint sprayer (Pico-Bel, Wagner, Germany). The inoculated plants were maintained under plastic bags overnight in order to ensure 100% relative humidity during the first steps of the infection process (Pinon et al., 2006). After 1 day, the plastic bags were removed and replaced by bags made of cotton fine mesh cloth (porosity $< 20\,\mu m$), in order to avoid dissemination of rust urediniospores in the greenhouse. The plants were maintained in the non-heated greenhouse throughout autumn (September–December) in order to induce the formation of telia during autumnal senescence of the plants. They were visually inspected every week to check the formation of uredinia and subsequently telia on the leaves (Hacquard et al., 2013). After 3 months, the fallen poplar leaves were collected. Leaves bearing telia were placed in 50×20 cm bags made of 1 cm plastic mesh. The bags were allowed to sit on the soil outdoors during a continental European winter (December–February) in order to break teliospore dormancy through natural alternation of freezing/thawing and wetting/drying (Leonard and Szabo, 2005). Starting in February, one leaf was sampled every week to assess whether teliospores would germinate at $19 \pm 1°C$ (see below). Once teliospore dormancy was broken (2–3 months), poplar leaves were collected and stored in a refrigerator ($7 \pm 1°C$) until used for larch inoculation.

Inoculation of larch plants

Larch (*Larix decidua*) seedlings were grown in spring in small ($8 \times 12 \times 5$ cm) plastic containers in a greenhouse. When they were 5–8 cm tall, larch plants were inoculated with *M. larici-populina* basidiospores. For this, poplar leaves bearing telia were soaked in tap water for 6 h, and then incubated at $19 \pm 1°C$, adaxial (upper) surface uppermost, on wet filter paper in large (24×24 cm) Petri dishes. After 24 h, the production of basidia and basidiospores on the poplar leaves was checked under a stereomicroscope. Leaves producing large quantities of basidiospores were placed 5 cm over larch seedlings for 1 day, in large ($30 \times 20 \times 15$ cm) transparent plastic boxes (Curver). After 1 day, the poplar leaves were withdrawn, and the larch seedlings were maintained in the plastic boxes in a growth chamber ($19 \pm 1°C$, photoperiod 16/8 h) for 4–5 weeks. The larch plants were visually inspected every day to check the formation of pycnia, i.e., the haploid stage of the rust fungus, and subsequently aecia on the larch needles. Pycnia were not manually crossed, but spermatization of pycnia occurred naturally, likely through contacts between larch needles, or through the presence of small flies (dark-winged fungus gnats, Sciaridae, Diptera) that thrive in the larch seedling substratum.

Production of urediniospores of the offspring

Larch needles bearing individual aecia were harvested every 2 days, and then aeciospores from each single aecium (i.e., S1 individual) were inoculated onto one 12-mm-diameter poplar leaf disc (*Populus deltoides* × *P. nigra* 'Robusta') as previously described (Husson et al., 2013). After 8–10 days incubation, the sporulating leaf discs were used to inoculate 4–5 entire poplar leaves as previously described (Husson et al., 2013), in order to obtain 50–100 mg of clonal urediniospores per S1 individual for genomic DNA extraction.

GENETIC ANALYSIS

Genotyping of the offspring

In order to avoid any bias in segregation and linkage disequilibrium analyses, the genetic purity of 138 of the offspring was verified twice through genotyping with a set of 25 polymorphic microsatellite loci specific to *M. larici-populina*: MLP12 (Barrès et al., 2006), MLP49, MLP50, MLP54, MLP55, MLP56, MLP57, MLP58, MLP59, MLP66, MLP68, MLP71, MLP73, MLP77, MLP82, MLP83, MLP87, MLP91, MLP92, MLP93, MLP94, MLP95, MLP96, MLP97, and MLP100 (Xhaard et al., 2009, 2011). The first verification was performed on the poplar leaf disc inoculated with aeciospores from larch, in order to check the presence of a unique genotype in each S1 individual. The second verification was performed on the urediniospores after multiplication on poplar leaves, in order to detect any contamination from a non-related rust isolate. DNA was extracted from infected poplar leaf discs and from urediniospores using the BioSprint 96 DNA plant kit used in combination with the BioSprint automated workstation (Qiagen), as previously described (Barrès et al., 2008). Microsatellite loci amplification and fragment analysis were performed as previously described (Xhaard et al., 2011).

Segregation and linkage disequilibrium analyses

Segregation and linkage disequilibrium analyses were performed on 115 genetically pure S1 individuals, using the 14 microsatellite loci (out of the 25), which are heterozygous in the parental strain 98AG31 (**Table 1**). According to the Mendelian laws, loci which are homozygous in the parental strain are not expected to segregate in the progeny, whereas a 1 (homozygous 1): 2 (heterozygous): 1 (homozygous 2) segregation is expected for loci which are heterozygous in the parental strain. Chi-squared tests with a significance level of 0.05 were performed to test whether the observed segregation deviated from the expected ratio.

Linkage disequilibrium between all pairs of loci was tested using Fisher's exact test procedure (Garnier-Gere and Dillmann, 1992) implemented in GENEPOP on the Web (http://genepop.curtin.edu.au) (Rousset, 2008) with the following parameters of the Markov chain: 2000 dememorization steps, 250 batches, and 2000 iterations per batch. To adjust the resulting *P*-value distribution for multiple tests, we used the false discovery rate (FDR) procedure (Benjamini and Yekutieli, 2001). The resulting adjusted *P*-values are called *Q*-values. This procedure is implemented in the R package *Q*-value (Storey and Tibshirani, 2003).

GENOMIC ANALYSIS

Genomic DNA extraction

Genomic DNA was extracted from a subset of 47 genetically pure S1 individuals of *M. larici-populina*, plus the parental strain 98AG31. For each individual, 10 aliquots of 5 mg of urediniospores were placed in 2 ml Eppendorf tube, with two 3-mm-diameter glass beads and 20 1-mm-diameter glass beads. The spores were homogenized for 1 min at 30 Hz using a MM 200

Table 1 | Characteristics and genome position of the 14 microsatellite loci, polymorphic in strain 98AG31, used for segregation and linkage disequilibrium analyses in *M. larici-populina*.

Locus name	Repeat motif	Scaffold no.	Position (bp)	Scaffold length (bp)
MLP49	(GAT)18	37	803,561	929,451
MLP54	(ATG)14	8	807,623	2,004,758
MLP55	(ATC)15	1	2,804,879	4,071,029
MLP56	(AAC)7	5	2,550,898	2,603,268
MLP58	(AAG)12	3	728,772	3,255,379
MLP59	(ATC)13	8	1,544,938	2,004,758
MLP73	(AC)14	38	809,210	903,405
MLP77	(AT)10	15	629,856	1,649,323
MLP82	(TAC)10	15	1,199,792	1,649,323
MLP87	(TGT)8	11	834,087	1,841,903
MLP91	(GTT)10	40	444,668	887,115
MLP92	(TTG)11	26	241,476	1,146,214
MLP94	(TTC)10	1	236,499	4,071,029
MLP12	(AAG)10	NA	NA	NA

Scaffold/sequence position is given according to the M. larici-populina genome sequence v1.0 (http://genome.jgi-psf.org/Mellp1) retrieved on April 1, 2014. NA, not available.

Mixer Mill (TissueLyser, Retsch, Qiagen) and then suspended in 1 ml of hot (65°C) CTAB buffer (CTAB 2%, Tris pH9 0.1 M, NaCl 1.4 M, EDTA 0.02 M, β-mercaptoethanol 0.2%). The content of the 10 Eppendorf tubes was pooled in a 50 ml Falcon tube. After carefully mixing by inverting, the tubes were incubated for 30 min at 65°C. One volume of phenol/chloroform/isoamyl alcohol (50:48:2) (Euromedex) was added to the 50 ml Falcon tube. The content of the tube was carefully mixed and then centrifuged at 8000 rpm for 10 min. The aqueous phase was transferred to a new tube and one volume of chloroform was added. The content of the tube was carefully mixed and then centrifuged at 8000 rpm for 10 min. The aqueous phase was transferred to a new tube. RNA was digested with 100 µl of RNaseA (Fermentas, 10 $\mu g.\mu l^{-1}$) for 30 min at 37°C. One volume of chloroform was added and the tubes were centrifuged at 8000 rpm for 10 min. The aqueous phase was recovered, distributed as aliquots of 800 µl in Eppendorf tubes, and 600 µl of isopropanol was added to each tube. Then the tubes were centrifuged at 14,000 rpm for 30 min at 4°C. The supernatant was removed by pouring liquid from the tube, and the DNA pellets were washed with 200 µl of 70% ethanol. The tubes were centrifuged at 14,000 rpm for 10 min at 4°C and washed again with 70 µl of 70% ethanol. The supernatant was removed by pipetting liquid from the tubes, the pellets were dried for 30 min under a fume hood, and then resuspended in TE 1× buffer (Tris 10 mM, EDTA 1 mM, pH 8.0). Quality and quantity of recovered high molecular weight DNA was assessed by electrophoresis on agarose gel and with a QuBit fluorometer (Life Technologies).

Whole-genome resequencing

Genome DNA re-sequencing was performed by IntegraGen (Evry, France). Forty-eight genomic DNA libraries were prepared using TruSeq DNA sample preparation kit (v3) followed by paired-end 100 bases massively parallel sequencing on Illumina HiSeq 2000. Briefly, 3 µg of each sample of genomic DNA were fragmented by sonication and purified to yield fragments of 400–500 bp. Paired-end adaptor oligonucleotides from Illumina were ligated on repaired A-tailed DNA fragments, then purified and enriched by PCR cycles. Each library was quantified by qPCR before equimolar pooling of the 48 libraries. The 48-plex pool was sequenced on one flowcell lane of Illumina HiSeq 2000 platform as paired-end 100 bp reads. Image analysis and base calling were performed using Illumina Real Time Analysis (RTA) Pipeline with default parameters.

Mapping and SNP detection

Mapping was performed with BWA version 0.6.2 (Li and Durbin, 2009). Version 1.0 of the *M. larici-populina* 98AG31 genome assembly (http://genome.jgi-psf.org/Mellp1) was used as reference in index using the IS algorithm. The reference genome contains 462 scaffolds and a total of 101,129,028 bp. Read alignment was performed using default options of the *aln* and *sampe* commands except maximum insert size (750 bp). Alignments were stored in the pileup format using SAMtools version 0.1.18 (Li et al., 2009).

We filtered sites presenting a potential single-nucleotide polymorphism (SNP) using liberal thresholds. All 48 individuals (including the parental strain) were considered jointly for each position, excluding all reads for which the base at this position was called with a quality Phred score lower than 25. We considered sites for which the minority allele (the second in frequency, if more than two) was represented by at least three copies, when merging all individuals. It was also required that each of the three genotypes was represented by two individuals each, where individuals were crudely assigned to one of the three genotypes (heterozygote if both alleles were represented, homozygote otherwise). If more than two alleles were observed, only the two most frequent were considered.

Identification of recombination blocks

In order to take into account genotype uncertainty when sequencing depth for a single individual is low, we used an *ad-hoc* scoring function to represent the amount of evidence regarding the homozygous/heterozygous status of each site. The function gives a score of 0 (undetermined) if the depth is 0 or 1, positive scores if only one allele is observed and negative scores if two alleles are observed. Overall, scores are bounded by -1 and $+1$. We fitted the logistic function $c/(1 + \exp[-r(x - b)])$, where x is the sequencing depth at a given SNP, to the observed proportion of observed heterozygotes in the 98AG31 parental strain, which is necessarily heterozygous in all true SNP positions. Minimum mean square error estimation yielded $r = 0.51714$, $b = 3.9148$, and $c = 0.98404$. This function was assumed to give the probability of observing a heterozygote if the genotype is truly heterozygous. Since, in our setting, true heterozygotes and true homozygotes are equally likely, we can use the same function as the probability that a genotype is truly homozygous if it is observed to be homozygous. If a genotype was observed to be homozygous, we used an arbitrary weighting scheme to

take into account both depth and allele frequency evenness when considering genotypes observed to be heterozygous. The score was computed as $(M/m\text{-}1)/200\text{-}1$ where M and m are the majority and minority allele absolute frequencies, respectively. In addition, the score of each SNP incorporates the scores of neighboring SNPs. Thus, the score of a given position was actually computed as the sum of the score of the focus SNP and the 15 neighboring SNPs on each side, weighted by the distance (a normal distribution with standard deviation 10,000 bp is used for weights, which gives high weights to SNPs about 20,000 kb apart).

A further step of smoothing was performed in order to identify recombination blocks. We assumed that a region of consecutive SNPs with constant heterozygous or homozygous status to be a block that was transmitted without recombination. We defined these blocks as regions of consecutive SNPs for which the score kept the same sign and at least one SNP exhibited an absolute value larger than 0.5. To avoid excessive false positive recombination points, we extended recombination blocks over SNPs with scores of opposite sign provided that they did not exceed |0.2| and up to a SNP with a score of at least |0.2| of the corresponding sign. These blocks represented overlapping regions of the genome with different putative recombination points defined in each individual. We defined a sample-wise set of blocks by dividing the genome in regions based on all putative recombination points. As a result recombination blocks were regions in which no recombination event had taken place.

Next, we generated the phased sequence of the two parental haplotypes of all blocks. We retrieved the majority allele of all SNPs for individuals that were classified as homozygous for the region under consideration. This generated the sequence of all putative homozygotes for each non-recombining region. We generated the maximum-likelihood phylogeny of those sequences using PhyML version 20120412 (Guindon and Gascuel, 2003) using the Jukes and Cantor model of substitution (all other parameters left to defaults). The two alleles should be represented by two deeply diverged clades in the resulting tree. For this reason, we excluded all recombination groups for which there was no internal branch representing at least 90% of the total tree length. This is expected to exclude recombination groups that contained undetected recombination events or a large proportion of erroneous data.

For all pairs of recombination blocks, we used P-value for Fisher's exact test of independence (based on the 3×3 matrix of the three genotype frequencies at both sites) as a measure of linkage disequilibrium. The P-value was not computed if fewer than 10 individuals had non-missing data for both sites, or if there was less than one copy of each homozygous genotype and fewer than two copies of each heterozygous genotype (over the two loci).

RESULTS

PRODUCTION OF THE *M. LARICI-POPULINA* S1 PROGENY

In order to generate the *M. larici-populina* S1 progeny, 4-month-old poplar plants were spray-inoculated with *M. larici-populina* strain 98AG31 in a greenhouse. Uredinia appeared on the abaxial surface of the inoculated poplar leaves 8–10 days after inoculation (**Figure 1A**). Pale brown telia began to appear on the adaxial surface of the leaves 3–4 weeks after inoculation (**Figure 1B**) and became dark brown and then black during the following weeks, as the leaves became senescent (**Figure 1C**). In December, almost all the poplar leaves were covered with black telia. After overwintering for 2 months outside in natural winter conditions (**Figure 1D**), poplar leaves bearing telia were tested for basidiospores production in laboratory. After water-soaking and incubation on wet filter paper for 1 day, masses of basidia and basidiospores, resulting from meiosis, were produced on the adaxial surface of the leaves (**Figures 1E,F**). Basidia were used to inoculate larch seedlings (**Figure 1G**). Pycnia appeared on the adaxial surface of larch needles 5–7 days after inoculation (**Figure 1H**). Spermatization of pycnia by pycniospores of the opposite mating type occurred naturally, resulting in plasmogamy of two haploid cells and formation of a dikaryotic mycelium. The resulting aecia appeared on the abaxial surface of larch needles 10–15 days after inoculation (**Figure 1I**), and individual aecia representing single selfed progeny were harvested and then multiplied on poplar leaf discs.

GENOTYPING OF THE OFFSPRING

The genetic purity of 138 of the offspring was verified twice (on initial inoculation onto poplar leaf discs and after 2–3 rounds of multiplication on poplar leaves) through genotyping with a set of 25 polymorphic microsatellite loci specific to *M. larici-populina*. Results of genotyping allowed the detection of two types of contamination that may occur during the offspring production process. On the one hand, 4.3% of the individuals exhibited an "external" contamination, i.e., a contamination of an offspring with a non-related *M. larici-populina* isolate, resulting in the presence of non-parental alleles at one or several microsatellite loci. On the other hand, 12.3% of the individuals exhibited an "internal" contamination, i.e., a mixture of two of the offspring, resulting in the presence of unbalanced peak heights of parental alleles at one or several heterozygous microsatellite loci. Both types of contamination were detected at the first (i.e., poplar leaf disc inoculated with aeciospores from larch) and the second (i.e., urediniospores after multiplication on poplar leaves) verification. Isolates identified as contaminated were deleted from further analyses.

SEGREGATION ANALYSIS

In order to validate the progeny's meiotic origin, a segregation analysis was performed on 115 genetically pure S1 individuals, using the 14 microsatellite loci (out of the 25), which are heterozygous in the parental strain 98AG31. Twelve out of the 14 microsatellite loci exhibited a Mendelian segregation of 2 (heterozygous): 1 (homozygous 1): 1 (homozygous 2), as expected (**Figure 2**). Two loci (MLP59 and MLP49) differed significantly from the expected ratio (P-value = 0.034 and 0.005, respectively), exhibiting an excess of heterozygotes (61.4 and 65.8% of heterozygotes, respectively).

LINKAGE DISEQUILIBRIUM ANALYSIS

In order to detect pairs of microsatellite loci, which are genetically linked, a linkage disequilibrium analysis was performed between all pairs of microsatellite loci. Two out of the 91 pairwise linkage disequilibrium tests were found significant (Q-value < 0.05) (**Table 2**). Highly significant linkage disequilibrium was found for

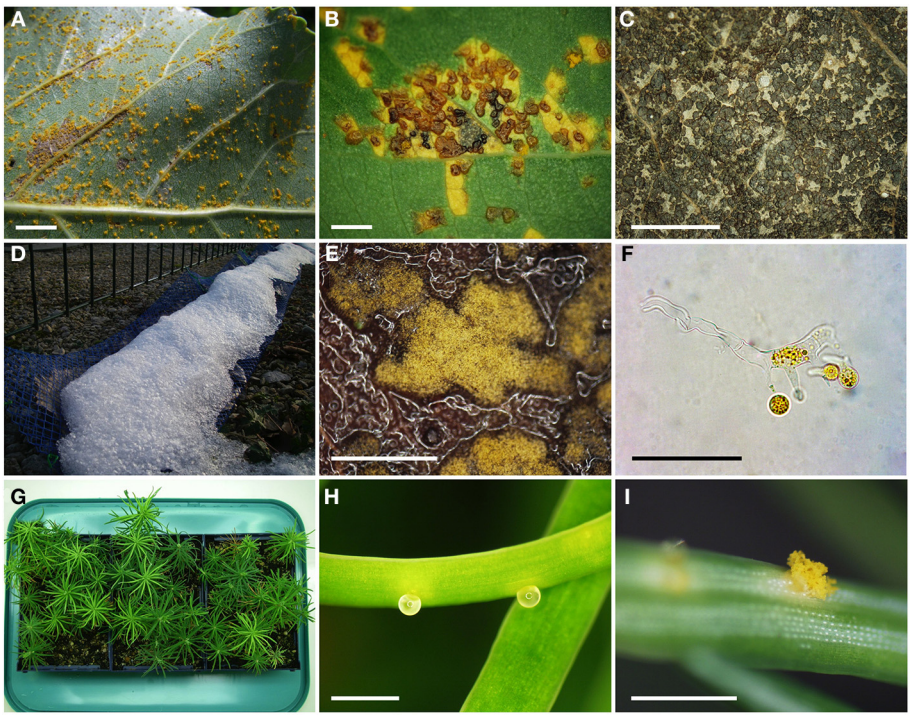


FIGURE 1 | Successive stages of the life cycle observed in the production of the *M. larici-populina* S1 progeny. (A) Inoculated poplar leaf covered with uredinia on the abaxial surface (scale bar 1 cm). **(B)** Light brown to dark brown telia forming on the adaxial surface of a poplar leaf (scale bar 1 mm). **(C)** Black mature telia obtained at leaf fall (scale bar 5 mm). **(D)** Outdoor overwintering of poplar leaves in the plastic mesh bags. **(E)** Telia producing basidia and basidiospores (scale bar 1 mm). **(F)** Basidium producing four basidiospores (scale bar 50 μm). **(G)** Young larch seedlings used for inoculation with basidiospores. **(H)** Pycnia on the adaxial surface of a larch needle (scale bar 1 mm). **(I)** Aecium producing aeciospores on the abaxial surface of a larch needle (scale bar 1 mm).

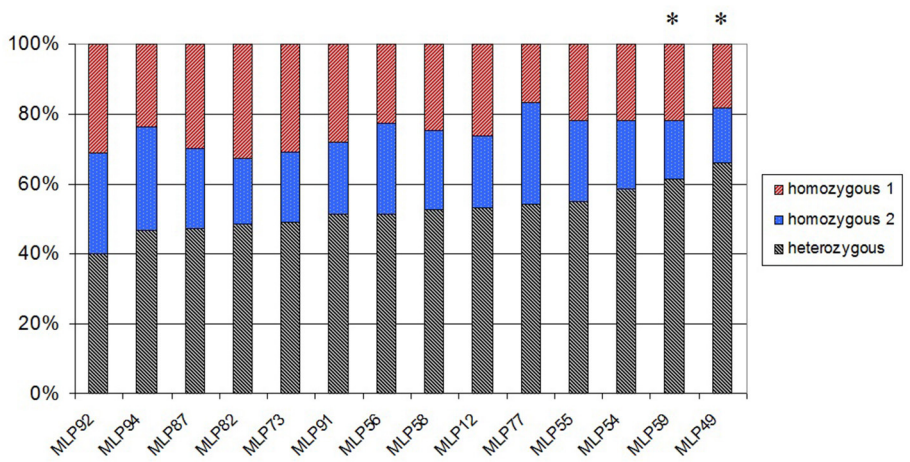


FIGURE 2 | Percentage of individuals in the *M. larici-populina* S1 progeny, which are heterozygous, homozygous 1, and homozygous 2 at each of the 14 polymorphic microsatellite loci. Stars denote loci with a significant deviation from the expected 1:2:1 Mendelian segregation for $\alpha = 0.05$.

MLP77/MLP56 and MLP77/MLP82 pairs, with Q-value $< 10^{-4}$ for both. Therefore, these two pairs of loci are genetically linked pairwise. Conversely, no significant linkage disequilibrium was found for all other pairs of loci, with Q-value ranging from 0.253 to 0.808.

GENOME SEQUENCING AND MAPPING

The whole genome of the 47 S1 individuals of *M. larici-populina*, plus the parental strain 98AG31, was resequenced using the Illumina HiSeq 2000 technology. A total of over 40 billions bp were sequenced, with a relatively uneven repartition of

Table 2 | Matrix of *Q*-values for all pairwise linkage disequilibrium tests between microsatellite loci.

	MLP12	MLP49	MLP54	MLP55	MLP56	MLP58	MLP59	MLP73	MLP77	MLP82	MLP87	MLP94	MLP92
MLP49	0.73544												
MLP54	0.69846	0.73544											
MLP55	0.73544	0.73544	0.73544										
MLP56	0.73544	0.73544	0.73544	0.73544									
MLP58	0.73544	0.34450	0.78528	0.73544	0.73544								
MLP59	0.57010	0.57010	0.73544	0.73544	0.78528	0.38876							
MLP73	0.73544	0.69500	0.73544	0.73544	0.73544	0.78528	0.34450						
MLP77	0.80770	0.73544	0.57010	0.80770	**$<10^{-4}$**	0.52127	0.52127	0.78670					
MLP82	0.78399	0.73544	0.52127	0.73544	0.73544	0.76627	0.73544	0.76627	**$<10^{-4}$**				
MLP87	0.73544	0.73544	0.73544	0.69846	0.73544	0.76627	0.78528	0.73544	0.76627	0.76627			
MLP94	0.69846	0.73544	0.39876	0.77008	0.78528	0.73544	0.73544	0.80770	0.80770	0.76627	0.76627		
MLP92	0.73544	0.73544	0.80770	0.73544	0.78527	0.57010	0.44613	0.73544	0.73544	0.34450	0.73544	0.76627	
MLP91	0.73544	0.44613	0.78399	0.58058	0.73544	0.25345	0.30801	0.73544	0.39876	0.39876	0.73544	0.73544	0.73544

Significant Q-values (Q < 0.05) are indicated in bold text.

sequencing depth per individual (Table S1). Over three quarters of reads (268 millions) mapped to the reference genome, leading to an average genome sequencing depth of 265X when considering all individuals, but ranging from 1.3X to 9.9X with most individuals presenting a final genome sequencing depth of about 5–7X. Genome sequences were deposited in GenBank under the BioProject ID PRJNA255081 and the SRA accession number SRP044324.

DETECTION OF SNPs AND DETERMINATION OF RECOMBINATION BLOCKS

Based on the criteria described in Materials and Methods, we identified 144,566 SNPs across the genome. The average density is 1.4 SNP/kb, but exhibits a marked heterogeneity across the genome (see the example in **Figure 3** for scaffold 1). Based on a smoothed scoring approach, we identified all potential recombination points and assigned regions to homozygote and heterozygote status (see the example in **Figure 4** for scaffold 1). This led to the definition of a total of 3302 recombination blocks in the genome, of which 2580 exhibited two phylogenetically distinct homozygous alleles and were considered robust enough to be analyzed. It can be noted that the recombination blocks may have different sizes, both in terms of number of SNPs included and of physical region covered. The number of SNPs per block ranges from just 1 to 1465 and the length of the corresponding physical region reaches more than half a Mbp (Table S2). In total, 140 of the 462 scaffolds contain at least one block (other scaffolds either have no SNP or the recombination block was excluded). Sixty-one have more than 10 blocks, and the three largest scaffolds have over 100 blocks. Each individual appears to be a mixture of homozygous and heterozygous regions as expected, although there appears to be a slight bias toward homozygous regions (**Table 3**). As expected, the fraction of the genome of the parental strain that is assigned was exclusively heterozygous. For individuals for which sequencing depth was too low, the smoothed scores rarely achieved threshold scores and large fractions of their genome were left unassigned.

ANALYSIS OF GENOMIC LINKAGE DISEQUILIBRIUM

The 2580 recombination blocks were treated each as a diallelic marker with three genotypic states. All 3,326,910 pairwise comparisons were considered, and 3,149,154 Fisher's exact tests were performed (excluding pairs with too much missing data), of which 61,743 involved markers belonging to the same assembly scaffold. The analysis of the decay of linkage disequilibrium against physical distance shows that highly significant (*P*-values less than 10^{-4}) linkage disequilibrium extends to over a distance of 1 Mbp (**Figure 5**).

The study of pairwise linkage disequilibrium test between all pairs of markers (both within and between assembly scaffolds) allowed us to identify signatures of statistical linkage between different scaffolds (see **Figure 6** for a focus on the first 7 scaffolds). A systematic but liberal analysis identified 158 pairs of scaffolds that may follow each other, grouping a total of 81 scaffolds into 23 linkage groups (Table S3). Although 381 scaffolds (representing together over 35 Mbp) are left unlinked, the 81 grouped scaffolds contain 66 of the 100 largest scaffolds and represent more than 65 Mbp (65% of the genome). One of these links across scaffolds can be seen between scaffolds 2 and 4 in **Figure 6** (second row, fourth column). While most pairs of sites exhibits low levels of linkage (white or blue pixels), a cluster of highly correlated pairs of sites appear at the upper-right corner of this panel, indicating linkage disequilibrium between the beginning of scaffold 2 and the end of scaffold 4.

However, this approach fails to account for signatures of linkage disequilibrium between scaffolds when they do not occur at the end of both scaffolds. A striking but not exceptional example can be seen in **Figure 6** between the beginning of scaffold 7 and a region at the beginning of the second third of scaffold 1 (in the reverse orientation). In addition, there is no signature of linkage between the first and second thirds of scaffold 1. This pattern, which suggests that scaffolds are linked internally rather than serially, was found at multiple instances among pairwise comparisons, sometimes with internal connections in both scaffolds (**Figure 6**, Figure S1).

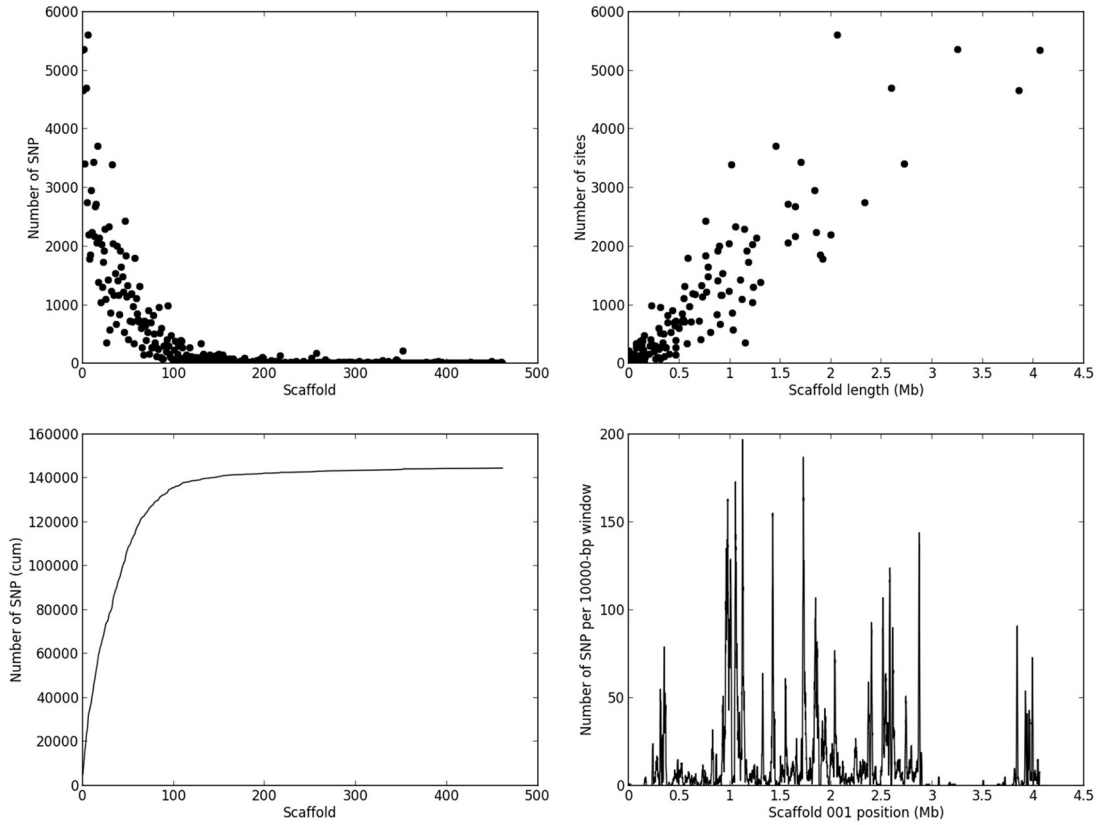


FIGURE 3 | Distribution of SNPs across the *M. larici-populina* genome. The figure presents the number of SNPs detected per scaffold (upper-left panel, the scaffold are sorted by decreasing length in the genome assembly), and the resulting cumulative number of SNPs per scaffold (lower-left panel), the number of SNPs as a function of scaffold length (upper-right panel), and the local density of SNPs in scaffold 1 taken as an example (lower-right panel).

DISCUSSION

PRODUCTION OF THE *M. LARICI-POPULINA* S1 PROGENY

In this study, we developed a segregating S1 progeny of the poplar rust fungus *M. larici-populina,* through selfing of the reference strain 98AG31 on larch plants. Pioneer work on classical genetics of rust fungi began in the 1930–1940's with the studies of Harold H. Flor on the flax/flax rust pathosystem (Flor, 1935, 1942). Since *Melampsora lini* is autoecious (i.e., completes its life cycle on a single host plant, flax), the production of F1 and F2 progenies through selfing or outcrossing was easier compared to heteroecious rusts. Nevertheless, early work on genetics of heteroecious cereal rusts also began in the 1940's (Johnson and Newton, 1946). In the present work several difficulties were overcome to manage the life cycle of the poplar rust fungus in controlled conditions. The first obstacle was the break of dormancy of teliospores. Several attempts were made to break dormancy in laboratory conditions with repeated cycles of freezing-thawing and wetting-drying for several months (Zambino et al., 2000), but remained unsuccessful. Thus, the "natural overwintering" method (Pei et al., 1999; Leonard and Szabo, 2005) was used instead.

The second difficulty stems from the fact that it is not possible to grow basidiospore-derived haploid mycelium of *M. larici-populina in vitro* in order to genotype a haploid progeny, as can be done with other Basidiomycetes, such as Agaricomycotina (Doudrick et al., 1995; Labbé et al., 2008), and with some rust fungi, such as the fusiform rust fungus (Doudrick et al., 1993). Therefore, we decided to inoculate larch plants, the alternate host, and to genotype aeciospore-derived dikaryotic individuals of *M. larici-populina*.

The third difficulty was the availability of fresh larch needles at just the time that germinating teliospores were available. Several attempts were made to infect detached larch flushing buds produced from 2-year-old larch plants (Pei et al., 1999), but with limited success because the larch needles could not be maintained alive for 2–3 weeks without rotting. The use of larch seedlings has overcome this difficulty.

The fourth difficulty was the feasibility of performing controlled outcrosses with *M. larici-populina* by matching larch needles bearing pycnia derived from different telial sources. Although some authors have obtained outcrossed progenies (Pei et al., 1999), it is technically difficult to obtain single isolated pycnia on larch needles in order to perform controlled spermatization of individual pycnia derived from one strain with spermatia derived from another strain. Therefore, we decided to study a S1 (selfed) progeny obtained through spermatization of pycnia from one telial source (strain 98AG31) with spermatia obtained from the same telial source. One drawback of this strategy is that only loci that are heterozygous in the parental strain segregate in the

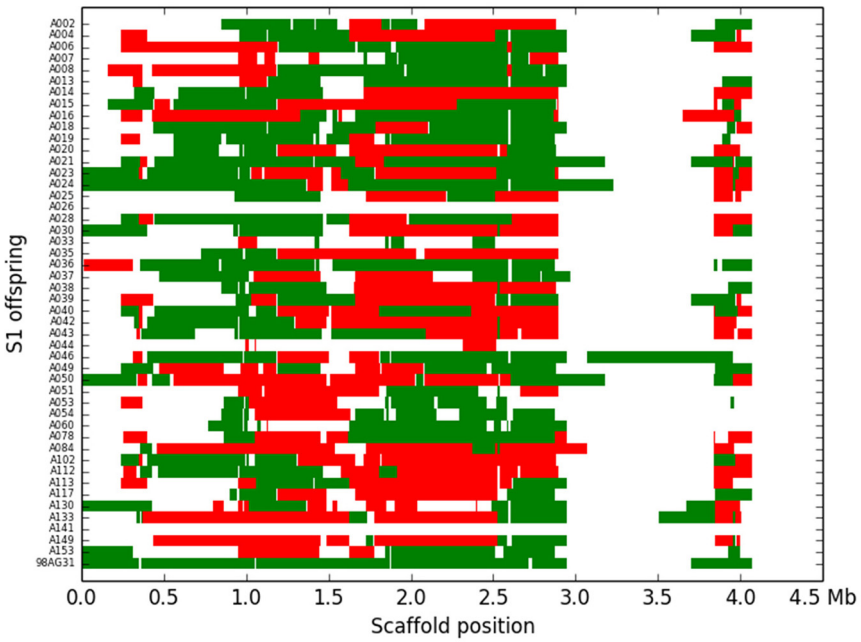


FIGURE 4 | Recombination blocks and assignment of the 48 *M. larici-populina* individuals (including the parental strain) for scaffold 1. Each line represents an S1 individual, the last being the parental strain. The blocks are represented against their physical location along scaffold 1. Green blocks represent heterozygote regions, red blocks represent homozygote regions and white areas represent unassigned regions.

S1 progeny. Nevertheless, thanks to the low level of inbreeding found in natural *M. larici-populina* populations (Barrès et al., 2008; Xhaard et al., 2011, 2012), as many as 144,566 segregating SNPs were observed in the progeny studied, which is largely sufficient for mapping purposes.

SEGREGATION ANALYSIS

Results of the segregation analysis showed that two out of the 14 polymorphic microsatellite loci did significantly depart from the expected 1:2:1 ratio, namely loci MLP49 (located on scaffold 8) and MLP59 (located on scaffold 37). This result can still be explained by chance alone under Mendelian segregation. Applying the Bonferroni correction for multiple testing leads to a significance threshold of 0.0035, so that both loci no longer depart significantly from the expected ratio. Alternatively, a significant excess of heterozygotes may be due to the proximity of these loci to mating type loci. Eleven putative pheromone precursor genes and four pheromone receptor genes, which may be involved in the mating type, were annotated in the *M. larici-populina* genome (Duplessis et al., 2011). However, none of these potential mating type genes were located on scaffolds 8 or 37, where loci MLP49 and MLP59 are located. Considering that the genomic organization of the mating type loci is still unresolved for the poplar rust fungus, we cannot conclude that the mating type loci influence the excess of heterozygotes observed for loci MLP49 and MLP59.

Other forms of selection may cause an excess of heterozygotes, such as over dominant selection. This type of selection is known to occur within host-pathogen interactions (Hughes and Nei, 1988), and may therefore affect our experiment since *M. larici-populina* is an obligate biotroph and can be cultivated only in interaction with its host. Analysis of genome-wide data could help analyze putative signatures of selection. In this first study, however, SNPs obtained through whole-genome sequencing were filtered such as all three genotypes were represented, which could have caused a bias when testing for departure from Mendelian ratios, by excluding sites more often in case of distortion of Mendelian segregation.

LINKAGE DISEQUILIBRIUM ANALYSES

Pairwise linkage disequilibrium tests on microsatellite data allowed us to bring to light two pairs of microsatellite loci in genetic linkage. Three different cases were encountered: (i) loci known to be physically linked because they are located on the same scaffold and found genetically linked; (ii) loci that are not physically linked but found genetically linked; and (iii) loci that are not found genetically linked despite being physically linked.

As expected, loci MLP77 (scaffold 15; 629,856 bp) and MLP82 (scaffold 15; 1,199,792 bp) were found in significant linkage disequilibrium (Q-value $< 10^{-4}$), since they are located on the same scaffold at 569,936 bp distance, which proved the efficiency of this method to confirm physical linkage by genetic linkage tests. In addition, locus MLP77 also was found in significant linkage disequilibrium with locus MLP56 (scaffold 5; 2,550,898 bp), with a Q-value $< 10^{-4}$, despite located on different scaffolds. Consequently, this pair of loci is genetically linked and the respective scaffolds (scaffold 5 and scaffold 15) should be grouped into a linkage group. Furthermore, since locus MLP56 is located at the end of scaffold 5 and MLP77 is located at first third of scaffold 15, the most likely physical link should be between the end of scaffold 5 and the beginning of scaffold 15. Thus, the microsatellite-based

Table 3 | Proportion of recombination blocks determined as heterozygous, homozygous, or unassigned for all *M. larici-populina* individuals.

Individual	Homozygous	Heterozygous	Unassigned
98AG31-A002	0.40	0.13	0.47
98AG31-A004	0.48	0.18	0.34
98AG31-A006	0.36	0.22	0.42
98AG31-A007	0.34	0.11	0.55
98AG31-A008	0.34	0.20	0.46
98AG31-A013	0.32	0.20	0.48
98AG31-A014	0.39	0.16	0.45
98AG31-A015	0.43	0.28	0.30
98AG31-A016	0.48	0.24	0.28
98AG31-A018	0.40	0.16	0.44
98AG31-A019	0.41	0.18	0.41
98AG31-A020	0.38	0.15	0.46
98AG31-A021	0.36	0.30	0.33
98AG31-A023	0.40	0.26	0.34
98AG31-A024	0.33	0.28	0.40
98AG31-A025	0.35	0.15	0.50
98AG31-A026	0.02	0.01	0.97
98AG31-A028	0.50	0.26	0.24
98AG31-A030	0.40	0.19	0.41
98AG31-A033	0.25	0.06	0.69
98AG31-A035	0.35	0.16	0.49
98AG31-A036	0.37	0.14	0.49
98AG31-A037	0.35	0.11	0.54
98AG31-A038	0.38	0.14	0.48
98AG31-A039	0.38	0.17	0.45
98AG31-A040	0.47	0.20	0.33
98AG31-A042	0.40	0.24	0.36
98AG31-A043	0.31	0.22	0.47
98AG31-A044	0.06	0.03	0.91
98AG31-A046	0.38	0.16	0.46
98AG31-A049	0.37	0.21	0.43
98AG31-A050	0.45	0.26	0.29
98AG31-A051	0.29	0.13	0.59
98AG31-A053	0.31	0.11	0.59
98AG31-A054	0.38	0.11	0.51
98AG31-A060	0.10	0.16	0.74
98AG31-A078	0.41	0.21	0.38
98AG31-A084	0.46	0.17	0.37
98AG31-A102	0.48	0.16	0.36
98AG31-A112	0.45	0.22	0.33
98AG31-A113	0.38	0.15	0.47
98AG31-A117	0.37	0.14	0.49
98AG31-A130	0.31	0.20	0.48
98AG31-A133	0.40	0.18	0.42
98AG31-A141	0.01	0.01	0.98
98AG31-A149	0.37	0.15	0.49
98AG31-A153	0.35	0.13	0.51
98AG31	0.00	0.27	0.73

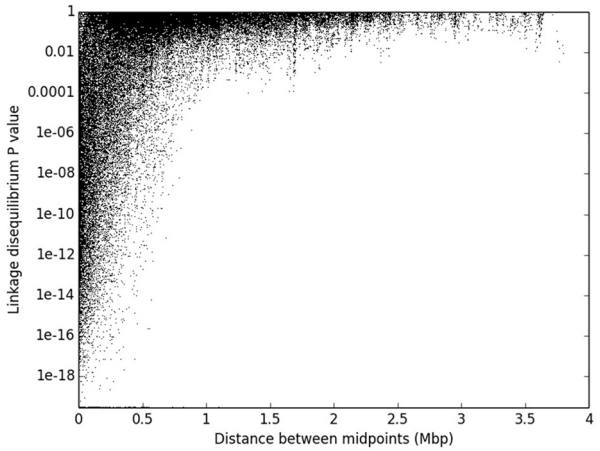


FIGURE 5 | Decay of linkage disequilibrium against physical distance. The value of Fisher's exact test of association between 61,743 pairs of recombination blocks belonging to the same scaffold is represented as a function of the physical distance between block midpoints.

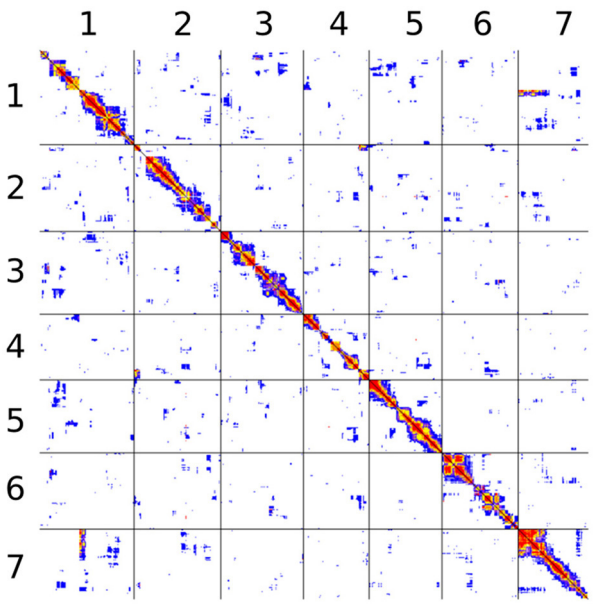


FIGURE 6 | Pairwise linkage disequilibrium between all markers belonging to assembly scaffolds 1 through 7. The figure does not account for physical distance. Black lines separate scaffolds. All pairs are represented twice, above and below the diagonal. Each pair is represented by a single pixel. The color code is based on Fisher's exact test *P*-value: red, $P < 10^{-10}$; orange, $P < 10^{-8}$; yellow, $P < 10^{-6}$; purple, $P < 10^{-4}$; blue, $P < 10^{-2}$; white, larger *P*-values or not computed.

linkage disequilibrium analysis allowed us to build one linkage group for a total length of 4,252,591 bp, which accounts for about 4.2% of the total genome length. However, these two scaffolds were not found in the same linkage group as defined from the SNP-based linkage analysis (Table S3).

Counter-intuitive results were observed for MLP94 (scaffold 1; 236,499 bp)/MLP55 (scaffold 1; 2,804,879 bp), and MLP54 (scaffold 8; 807,623 bp)/MLP59 (scaffold 8; 1,544,938 bp) pairs. Although each pair of loci is located on a single scaffold, no linkage disequilibrium was detected (Q-value = 0.77 and 0.74, respectively). Loci MLP94 and MLP55 are located on scaffold 1 at a distance of 2,568,380 bp, which could be far enough to break any

genetic linkage. This result is consistent with the SNP-based linkage analysis, which showed that linkage is unlikely to be detected at distances higher than 2 Mbp. However, the situation is different for loci MLP54 and MLP59, which are located on scaffold 8 at a distance of 737,315 bp, and for which no linkage disequilibrium was found despite this relatively small distance. The SNP-based linkage analysis showed that linkage could be still detectable at distances up to 1 Mbp. This loss of linkage disequilibrium could be due to a recombination hotspot in this specific region (Petes, 2001). Another possible explanation would be misassembly of scaffold 8, resulting in a chimeric scaffold. This latter hypothesis is supported by the linkage disequilibrium discontinuities observed along scaffold 8 (Figure S1).

In a second step, we aimed to adopt a whole-genome perspective on linkage disequilibrium patterns. We used shifts from homozygous to heterozygous genomic regions detected from whole-genome sequencing data to scale genomic data down to 2580 recombination blocks in which all SNPs were non-recombining. These blocks were then treated as markers, allowing us to integrate data from successive SNPs, and to cope with sequencing errors and missing data that could have had a strong impact due to the relatively small sample size and the unequal sequencing depth among individuals. The approach was however limited to genomic regions that were polymorphic within the parental strain, thereby generating informative segregating markers in the offspring. Thanks to the low level of inbreeding and the relatively high level of polymorphism in the *M. larici-populina* populations, the amount of diversity proved to be sufficient for our purpose (more than one SNP per kb on average).

Linkage disequilibrium could be detected between markers located at nearby locations on the same assembly scaffold, but also between markers located on different scaffolds, while many pairs of markers located on the same scaffold exhibited no linkage disequilibrium. We found that, as expected, linkage disequilibrium decays with increasing distance, as shown with microsatellite markers cited earlier, with linkage still detectable at distances up to 1 Mbp. Based on the 2580 SNP-based markers, we detected 23 putative linkage groups, including scaffolds that exhibit signals of linkage disequilibrium with at least one other scaffold of the same group. Interestingly, some of the between-scaffold signatures of linkage disequilibrium obtained with whole-genome sequencing data pointed out potential genome assembly issues. As tentatively evidenced using microsatellite-based analysis, it appears that, at several instances, contiguous genomic regions within a scaffold display discontinuities in genetic linkage. These results point to the fact that most of the largest scaffolds might have to be redefined, i.e., split and rearranged. Due to the large size and high repetitive sequence content (for a fungus), the genome of *M. larici-populina* is not easy to assemble properly, even based on high-depth genome sequencing. It is thus possible that some of the scaffolds are actually chimers.

Noteworthy, the pattern of linkage disequilibrium among these whole-genome markers (23 linkage groups) did not match the observation made using the 14 microsatellite markers. In other words, the linkage group made of scaffold 5 and 15 evidenced by the microsatellite-based linkage analysis is not supported by the SNP-based scaffold merging. This discrepancy can be further explained by a careful look at the pattern of SNP-based linkage along scaffold 15 (Figure S1) and may constitute another example of putative genome misassembly. First, we observed a clear break in statistical linkage between the very beginning of scaffold 15 and the rest of this scaffold. Second, while the very beginning of scaffold 15 is undoubtedly linked with scaffold 12 (which defined linkage group 7, Table S3), we observed that the SNP-based markers located just after the genetic break (beginning of the second part of scaffold 15) are statistically linked with markers located at the end of scaffold 5. This observation is fully consistent with the linkage group formed by the microsatellite loci MLP77 (scaffold 15) and MLP56 (scaffold 5). The microsatellite-based analysis thus enables us to extend the linkage group 4 (consisting in scaffold 73 linked to the beginning of scaffold 5) by adding the largest portion of scaffold 15 (linked with the end of scaffold 5).

The linkage disequilibrium analysis performed over a narrow number (14) of microsatellite markers allowed us to define one linkage group in the *M. larici-populina* genome. The subsequent use of 2480 markers integrating the information of almost 150,000 SNPs detected in 47 S1 individuals suggested 23 linkage groups. Both methods have strengths and weaknesses—microsatellites were typed on more offspring individuals, and SNPs are available at higher density along genomes—but they are complementary. While the SNP-based and microsatellite-based linkage analyses appeared inconsistent at first sight, we demonstrated that these two analyses converged to question current genome assembly. These encouraging results demonstrate that an accurate genetic map could be constructed with a larger number of *M. larici-populina* S1 offspring, by using the method for constructing ultra-high-density linkage maps with high-quality SNPs based on low-coverage resequencing, as already described for plants (Huang et al., 2009; Xie et al., 2010) and fungi (Au et al., 2013).

Besides constituting by itself a valuable resource for investigating the architecture of complex traits of *M. larici-populina*, such a genetic map will provide complementary data for completing the genome assembly. The assembly of the genome of *M. larici-populina* has been made difficult by its large size (compared to most fungi) and especially its high content in repetitive sequences (Duplessis et al., 2011). However, the reference genome is an essential resource for genetic approaches as genes of interest might be much more difficult to identify if they are located in repeat-rich and poorly assembled regions. Genetic mapping is an excellent complementary approach, because, unlike physical sequencing, it is much less sensitive to the presence of repetitive sequences (it still is because of the possibility of ambiguous mapping). The integration of genetic mapping data to the current version of the genome assembly of *M. larici-populina* might therefore represent a significant advance.

AUTHOR CONTRIBUTION

Michaël Pernaci, Stéphane De Mita, Fabien Halkett, and Pascal Frey designed research; Michaël Pernaci, Stéphane De Mita, Axelle Andrieux, Jérémy Pétrowski, and Pascal Frey performed research; Michaël Pernaci, Stéphane De Mita, Fabien Halkett, and Pascal Frey analyzed data; and Michaël Pernaci, Stéphane De

Mita, Fabien Halkett, Sébastien Duplessis, and Pascal Frey wrote the paper.

ACKNOWLEDGMENTS

We are grateful to Mathilde Chertier for her help in genotyping the offspring and DNA extractions, to Bénédicte Fabre for her help in the production of poplar plants, and to Katherine J. Hayden and the two reviewers for their helpful comments. This work was supported by grants from the French National Research Agency (ANR-12-ADAP-0009, GANDALF project; ANR-13-BSV7-0011, FunFit project) and from INRA (EFPA innovative project). The UMR IAM is supported by a grant from the French National Research Agency (ANR-11-LABX-0002-01, Laboratory of Excellence ARBRE). Michaël Pernaci was supported by a PhD fellowship from the French Ministry of Education and Research (MESR).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00454/abstract

REFERENCES

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., et al. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* 21, 610–617. doi: 10.1101/gr.115402.110

Andrivon, D., Pilet, F., Montarry, J., Hafidi, M., Corbière, R., Achbani, E. H., et al. (2007). Adaptation of *Phytophthora infestans* to partial resistance in potato: evidence from French and Moroccan populations. *Phytopathology* 97, 338–343. doi: 10.1094/phyto-97-3-0338

Au, C., Cheung, M., Wong, M., Chu, A. K. K., Law, P. T. W., and Kwan, H. (2013). Rapid genotyping by low-coverage resequencing to construct genetic linkage maps of fungi: a case study in *Lentinula edodes*. *BMC Res. Notes* 6:307. doi: 10.1186/1756-0500-6-307

Barrès, B., Dutech, C., Andrieux, A., Caron, H., Pinon, J., and Frey, P. (2006). Isolation and characterization of 15 microsatellite loci in the poplar rust fungus, *Melampsora larici-populina*, and cross-amplification in related species. *Mol. Ecol. Notes* 6, 60–64. doi: 10.1111/j.1471-8286.2005.01137.x

Barrès, B., Halkett, F., Dutech, C., Andrieux, A., Pinon, J., and Frey, P. (2008). Genetic structure of the poplar rust fungus *Melampsora larici-populina*: evidence for isolation by distance in Europe and recent founder effects overseas. *Infect. Genet. Evol.* 8, 577–587. doi: 10.1016/j.meegid.2008.04.005

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188. doi: 10.1186/1471-2105-9-114

Broggini, G. A. L., Bus, V. G. M., Parravicini, G., Kumar, S., Groenwold, R., and Gessler, C. (2011). Genetic mapping of 14 avirulence genes in an EU-B04 x 1639 progeny of *Venturia inaequalis*. *Fungal Genet. Biol.* 48, 166–176. doi: 10.1016/j.fgb.2010.09.001

Brun, H., Chevre, A. M., Fitt, B. D. L., Powers, S., Besnard, A. L., Ermel, M., et al. (2010). Quantitative resistance increases the durability of qualitative resistance to *Leptosphaeria maculans* in *Brassica napus*. *New Phytol.* 185, 285–299. doi: 10.1111/j.1469-8137.2009.03049.x

Covarelli, L., Beccari, G., Tosi, L., Fabre, B., and Frey, P. (2013). Three-year investigations on leaf rust of poplar cultivated for biomass production in Umbria, Central Italy. *Biomass Bioenerg.* 49, 315–322. doi: 10.1016/j.biombioe.2012.12.032

Cozijnsen, A. J., Popa, K. M., Purwantara, A., Rolls, B. D., and Howlett, B. J. (2000). Genome analysis of the plant pathogenic ascomycete *Leptosphaeria maculans*; mapping mating type and host specificity loci. *Mol. Plant Pathol.* 1, 293–302. doi: 10.1046/j.1364-3703.2000.00033.x

Cronn, R., Liston, A., Parks, M., Gernandt, D. S., Shen, R., and Mockler, T. (2008). Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* 36, e122. doi: 10.1093/nar/gkn502

Dangl, J. L., Horvath, D. M., and Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science* 341, 746–751. doi: 10.1126/science.1236011

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510. doi: 10.1038/nrg3012

Doudrick, R., Nelson, C., and Nance, W. (1993). Genetic analysis of a single urediniospore culture of *Cronartium quercuum* f. sp. *fusiforme*, using random amplified polymorphic DNA markers. *Mycologia* 85, 902–911. doi: 10.2307/3760673

Doudrick, R. L., Raffle, V. L., Nelson, C. D., and Furnier, G. R. (1995). Genetic analysis of homokaryons from a basidiome of *Laccaria bicolor* using random amplified polymorphic DNA (RAPD) markers. *Mycol. Res.* 99, 1361–1366. doi: 10.1016/S0953-7562(09)81222-7

Dowkiw, A., Voisin, E., and Bastien, C. (2010). Potential of Eurasian poplar rust to overcome a major quantitative resistance factor. *Plant Pathol.* 59, 523–534. doi: 10.1111/j.1365-3059.2010.02277.x

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). "Advancing knowledge on biology of rust fungi through genomics," in *Advances in Botanical Research*, ed F. M. Martin (Oxford: Academic Press), 173–209.

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Fabre, F., Rousseau, E., Mailleret, L., and Moury, B. (2012). Durable strategies to deploy plant resistance in agricultural landscapes. *New Phytol.* 193, 1064–1075. doi: 10.1111/j.1469-8137.2011.04019.x

Flor, H. (1942). Inheritance of pathogenicity in *Melampsora lini*. *Phytopathology* 32, 653–669.

Flor, H. H. (1935). Physiologic specialization of *Melampsora lini* on *Linum usitatissimum*. *J. Agric. Res.* 51, 819–837.

Foulongne-Oriol, M. (2012). Genetic linkage mapping in fungi: current state, applications, and future trends. *Appl. Microbiol. Biotechnol.* 95, 891–904. doi: 10.1007/s00253-012-4228-4

Foulongne-Oriol, M., Spataro, C., Cathalot, V., Monllor, S., and Savoie, J.-M. (2010). An expanded genetic linkage map of an intervarietal *Agaricus bisporus* var. *bisporus* x *A. bisporus* var. *burnettii* hybrid based on AFLP, SSR and CAPS markers sheds light on the recombination behaviour of the species. *Fungal Genet. Biol.* 47, 226–236. doi: 10.1016/j.fgb.2009.12.003

Garnier-Gere, P., and Dillmann, C. (1992). A computer program for testing pairwise linkage disequilibria in subdivided populations. *J. Hered.* 83, 239.

Gérard, P. R., Husson, C., Pinon, J., and Frey, P. (2006). Comparison of genetic and virulence diversity of *Melampsora larici-populina* populations on wild and cultivated poplar and influence of the alternate host. *Phytopathology* 96, 1027–1036. doi: 10.1094/phyto-96-1027

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* 52, 696–704. doi: 10.1080/10635150390235520

Hacquard, S., Delaruelle, C., Frey, P., Tisserant, E., Kohler, A., and Duplessis, S. (2013). Transcriptome analysis of poplar rust telia reveals overwintering adaptation and tightly coordinated karyogamy and meiosis processes. *Front. Plant Sci.* 4:456. doi: 10.3389/fpls.2013.00456

Hacquard, S., Petre, B., Frey, P., Hecker, A., Rouhier, N., and Duplessis, S. (2011). The Poplar-Poplar rust interaction: insights from genomics and transcriptomics. *J. Pathog.* 2011:716041. doi: 10.4061/2011/716041

Hahn, M. W., Zhang, S. V., and Moyle, L. C. (2014). Sequencing, assembling, and correcting draft genomes using recombinant populations. G3 (Bethesda) 4, 669–679. doi: 10.1534/g3.114.010264

Heilman, P. E. (1999). Planted forests: poplars. *New Forests* 17, 89–93. doi: 10.1023/a:1006515204167

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., et al. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res.* 19, 1068–1076. doi: 10.1101/gr.089516.108

Huang, X., Kurata, N., Wei, X., Wang, Z.-X., Wang, A., Zhao, Q., et al. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490, 497–501. doi: 10.1038/nature11532

Hughes, A. L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335, 167–170. doi: 10.1038/335167a0

Husson, C., Ioos, R., Andrieux, A., and Frey, P. (2013). Development and use of new sensitive molecular tools for diagnosis and detection of *Melampsora* rusts on cultivated poplar. *Forest Pathol.* 43, 1–11. doi: 10.1111/efp.12007

Johnson, R. (1979). The concept of durable resistance. *Phytopathology* 69, 198–199.

Johnson, T., and Newton, M. (1946). Specialization, hybridization, and mutation in the cereal rusts. *Bot. Rev.* 12, 337–392. doi: 10.1007/bf02861524

Jorge, V., Dowkiw, A., Faivre-Rampant, P., and Bastien, C. (2005). Genetic architecture of qualitative and quantitative *Melampsora larici-populina* leaf rust resistance in hybrid poplar: genetic mapping and QTL detection. *New Phytol.* 167, 113–127. doi: 10.1111/j.1469-8137.2005.01424.x

Kaye, C., Milazzo, J., Rozenfeld, S., Lebrun, M.-H., and Tharreau, D. (2003). The development of simple sequence repeat markers for *Magnaporthe grisea* and their integration into an established genetic linkage map. *Fungal Genet. Biol.* 40, 207–214. doi: 10.1016/j.fgb.2003.08.001

Kema, G. H., Goodwin, S. B., Hamza, S., Verstappen, E. C., Cavaletto, J. R., Van Der Lee, T. A., et al. (2002). A combined amplified fragment length polymorphism and randomly amplified polymorphism DNA genetic kinkage map of *Mycosphaerella graminicola*, the septoria tritici leaf blotch pathogen of wheat. *Genetics* 161, 1497–1505.

Kubisiak, T. L., Anderson, C. L., Amerson, H. V., Smith, J. A., Davis, J. M., and Nelson, C. D. (2011). A genomic map enriched for markers linked to Avr1 in *Cronartium quercuum* f.sp. *fusiforme*. *Fungal Genet. Biol.* 48, 266–274. doi: 10.1016/j.fgb.2010.09.008

Labbé, J., Zhang, X., Yin, T., Schmutz, J., Grimwood, J., Martin, F., et al. (2008). A genetic linkage map for the ectomycorrhizal fungus *Laccaria bicolor* and its alignment to the whole-genome sequence assemblies. *New Phytol.* 180, 316–328. doi: 10.1111/j.1469-8137.2008.02614.x

Lannou, C. (2012). Variation and selection of quantitative traits in plant pathogens. *Annu. Rev. Phytopathol.* 50, 319–338. doi: 10.1146/annurev-phyto-081211-173031

Larraya, L. M., Pérez, G., Ritter, E., Pisabarro, A. G., and Ramirez, L. (2000). Genetic linkage map of the edible Basidiomycete *Pleurotus ostreatus*. *Appl. Environ. Microbiol.* 66, 5290–5300. doi: 10.1128/aem.66.12.5290-5300.2000

Leonard, K. J., and Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol. Plant Pathol.* 6, 99–111. doi: 10.1111/j.1364-3703.2005.00273.x

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760. doi: 10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Manzo-Sánchez, G., Zapater, M.-F., Luna-Martínez, F., Conde-Ferráez, L., Carlier, J., James-Kay, A., et al. (2008). Construction of a genetic linkage map of the fungal pathogen of banana *Mycosphaerella fijiensis*, causal agent of black leaf streak disease. *Curr. Genet.* 53, 299–311. doi: 10.1007/s00294-008-0186-x

McDonald, B. A., and Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* 40, 349–379. doi: 10.1146/annurev.phyto.40.120501.101443

Pariaud, B., Ravigne, V., Halkett, F., Goyeau, H., Carlier, J., and Lannou, C. (2009). Aggressiveness and its role in the adaptation of plant pathogens. *Plant Pathol.* 58, 409–424. doi: 10.1111/j.1365-3059.2009.02039.x

Pedersen, C., Rasmussen, S. W., and Giese, H. (2002). A genetic map of *Blumeria graminis* based on functional genes, avirulence genes, and molecular markers. *Fungal Genet. Biol.* 35, 235–246. doi: 10.1006/fgbi.2001.1326

Pei, M. H., Royle, D. J., and Hunter, T. (1999). Hybridization in larch-alternating *Melampsora epitea* (*M. larici-epitea*). *Mycol. Res.* 103, 1440–1446. doi: 10.1017/S0953756299008655

Petes, T. D. (2001). Meiotic recombination hot spots and cold spots. *Nat. Rev. Genet.* 2, 360–369. doi: 10.1038/35072078

Pinon, J., Frey, P., and Husson, C. (2006). Wettability of poplar leaves influences dew formation and infection by *Melampsora larici-populina*. *Plant Dis.* 90, 177–184. doi: 10.1094/pd-90-0177

Rousset, F. (2008). GENEPOP ′007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol. Ecol. Resour.* 8, 103–106. doi: 10.1111/j.1471-8286.2007.01931.x

Sicard, D., Legg, E., Brown, S., Babu, N. K., Ochoa, O., Sudarshana, P., et al. (2003). A genetic map of the lettuce downy mildew pathogen, *Bremia lactucae*, constructed from molecular markers and avirulence genes. *Fungal Genet. Biol.* 39, 16–30. doi: 10.1016/S1087-1845(03)00005-7

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Terashima, K., Matsumoto, T., Hayashi, E., and Fukumasa-Nakai, Y. (2002). A genetic linkage map of *Lentinula edodes* (shiitake) based on AFLP markers. *Mycol. Res.* 106, 911–917. doi: 10.1017/S0953756202006275

Vialle, A., Frey, P., Hambleton, S., Bernier, L., and Hamelin, R. (2011). Poplar rust systematics and refinement of *Melampsora* species delineation. *Fungal Divers.* 50, 227–248. doi: 10.1007/s13225-011-0129-6

Xhaard, C., Andrieux, A., Halkett, F., and Frey, P. (2009). Characterization of 41 microsatellite loci developed from the genome sequence of the poplar rust fungus, *Melampsora larici-populina*. *Conserv. Genet. Resour.* 1, 21–25. doi: 10.1007/s12686-009-9005-z

Xhaard, C., Barrès, B., Andrieux, A., Bousset, L., Halkett, F., and Frey, P. (2012). Disentangling the genetic origins of a plant pathogen during disease spread using an original molecular epidemiology approach. *Mol. Ecol.* 21, 2383–2398. doi: 10.1111/j.1365-294X.2012.05556.x

Xhaard, C., Fabre, B., Andrieux, A., Gladieux, P., Barrès, B., Frey, P., et al. (2011). The genetic structure of the plant pathogenic fungus *Melampsora larici-populina* on its wild host is extensively impacted by host domestication. *Mol. Ecol.* 20, 2739–2755. doi: 10.1111/j.1365-294X.2011.05138.x

Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., et al. (2010). Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10578–10583. doi: 10.1073/pnas.1005931107

Xu, X. (2012). Super-races are not likely to dominate a fungal population within a life time of a perennial crop plantation of cultivar mixtures: a simulation study. *BMC Ecol.* 12:16. doi: 10.1186/1472-6785-12-16

You, X., Shu, L., Li, S., Chen, J., Luo, J., Lu, J., et al. (2013). Construction of high-density genetic linkage maps for orange-spotted grouper *Epinephelus coioides* using multiplexed shotgun genotyping. *BMC Genet.* 14:113. doi: 10.1186/1471-2156-14-113

Yu, H., Xie, W., Wang, J., Xing, Y., Xu, C., Li, X., et al. (2011). Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. *PLoS ONE* 6:e17595. doi: 10.1371/journal.pone.0017595

Zambino, P. J., Kubelik, A. R., and Szabo, L. J. (2000). Gene action and linkage of avirulence genes to DNA markers in the rust fungus *Puccinia graminis*. *Phytopathology* 90, 819–826. doi: 10.1094/phyto.2000.90.8.819

Zhou, X., Xia, Y., Ren, X., Chen, Y., Huang, L., Huang, S., et al. (2014). Construction of a SNP-based genetic linkage map in cultivated peanut based on large scale marker development using next-generation double-digest restriction-site-associated DNA sequencing (ddRADseq). *BMC Genomics* 15:351. doi: 10.1186/1471-2164-15-351

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 31 May 2014; accepted: 21 August 2014; published online: 10 September 2014.

Citation: Pernaci M, De Mita S, Andrieux A, Pétrowski J, Halkett F, Duplessis S and Frey P (2014) Genome-wide patterns of segregation and linkage disequilibrium: the construction of a linkage genetic map of the poplar rust fungus Melampsora larici-populina. Front. Plant Sci. 5:454. doi: 10.3389/fpls.2014.00454

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Pernaci, De Mita, Andrieux, Pétrowski, Halkett, Duplessis and Frey. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 15 September 2014
doi: 10.3389/fpls.2014.00450

Patterns of genomic variation in the poplar rust fungus *Melampsora larici-populina* identify pathogenesis-related factors

Antoine Persoons[1,2], Emmanuelle Morin[1,2], Christine Delaruelle[1,2], Thibaut Payen[1,2], Fabien Halkett[1,2], Pascal Frey[1,2], Stéphane De Mita[1,2] and Sébastien Duplessis[1,2]*

[1] Institut National de la Recherche Agronomique, Unité Mixte de Recherche 1136 Institut National de la Recherche Agronomique/Université de Lorraine Interactions Arbres/Microorganismes, Champenoux, France
[2] Université de Lorraine, Unité Mixte de Recherche 1136 Institut National de la Recherche Agronomique/Université de Lorraine Interactions Arbres/Microorganismes, Vandoeuvre-lès-Nancy Cedex, France

Edited by:
Francine Govers, Wageningen University, Netherlands

Reviewed by:
John P. Rathjen, The Australian National University, Australia
Dario Cantu, University of California, Davis, USA

***Correspondence:**
Sébastien Duplessis, INRA, Unité Mixte de Recherche 1136 INRA/Université de Lorraine Interactions Arbres/Microorganismes, 54280 Champenoux, France
e-mail: duplessi@nancy.inra.fr

Melampsora larici-populina is a fungal pathogen responsible for foliar rust disease on poplar trees, which causes damage to forest plantations worldwide, particularly in Northern Europe. The reference genome of the isolate 98AG31 was previously sequenced using a whole genome shotgun strategy, revealing a large genome of 101 megabases containing 16,399 predicted genes, which included secreted protein genes representing poplar rust candidate effectors. In the present study, the genomes of 15 isolates collected over the past 20 years throughout the French territory, representing distinct virulence profiles, were characterized by massively parallel sequencing to assess genetic variation in the poplar rust fungus. Comparison to the reference genome revealed striking structural variations. Analysis of coverage and sequencing depth identified large missing regions between isolates related to the mating type loci. More than 611,824 single-nucleotide polymorphism (SNP) positions were uncovered overall, indicating a remarkable level of polymorphism. Based on the accumulation of non-synonymous substitutions in coding sequences and the relative frequencies of synonymous and non-synonymous polymorphisms (i.e., P_N/P_S), we identify candidate genes that may be involved in fungal pathogenesis. Correlation between non-synonymous SNPs in genes encoding secreted proteins (SPs) and pathotypes of the studied isolates revealed candidate genes potentially related to virulences 1, 6, and 8 of the poplar rust fungus.

Keywords: effector, virulence, Pucciniales, obligate biotroph, genomics, polymorphism

INTRODUCTION

Worldwide, *Melampsora* spp. (Basidiomycota, Pucciniales) are the most devastating pathogens of poplars (Steenackers et al., 1996), and *Melampsora larici-populina* is a major threat in European poplar plantations (Pinon and Frey, 2005). The poplar rust fungus has a complex life cycle with five different types of spores that develop on two distinct host plants: *Populus*, on which it performs several asexual reproduction cycles during summer and autumn, and *Larix* spp., on which it performs a single sexual reproduction cycle once a year in spring. Poplars are particularly susceptible to *M. larici-populina* mostly because of their intensive monoclonal cultivation over several decades (Gérard et al., 2006). Until now eight qualitative resistances (*R1* to *R8*) have been deployed in plantations and each has been overcome by *M. larici-populina*. The most damaging resistance breakdown occurred in 1994 when the resistance *R7* was overcome and led to the invasion of France by virulent 7 *M. larici-populina* isolates (Xhaard et al., 2011). In accordance with the gene-for-gene relationship (Flor, 1971), *M. larici-populina* isolates which successfully infect resistant poplar possess the corresponding virulence factors (i.e., vir1 to vir8) determined at an avirulence locus. Up to now, none of the poplar R genes, nor the poplar rust virulence genes have been characterized (Hacquard et al., 2011).

Pathogenicity factors, i.e., effectors, contribute to the success of pathogen infection. Their recognition by cytoplasmic plant *R* receptors leads to a rapid and strong defense reaction through specific signaling cascades and expression of defense-related genes that stop pathogen growth, notably through the expression of a localized hypersensitive response at infection site (Dodds and Rathjen, 2010; Win et al., 2012). Most effectors described to date in rust fungi correspond to avirulence factors such as AvrL567, AvrP4, AvrP123, and AvrM of the flax rust fungus *Melampsora lini* (Ravensdale et al., 2011) and PGTAUSPE-10-1, a candidate AvrSr22 factor of the wheat stem rust *Puccinia graminis* f. sp. *tritici* (Upadhyaya et al., 2014), but their role in pathogenesis remain unknown. Another effector, the Rust Transferred Protein 1 (RTP1) from the bean rust fungus *Uromyces fabae*, forms fibrils in the extrahaustorial matrix and is transferred from haustoria into infected host cells, and may have protease inhibitory function (Kemen et al., 2005, 2013; Pretsch et al., 2013). So far, only a handful of fungal candidate effectors have been fully characterized

(Stergiopoulos and de Wit, 2009; Tyler and Rouxel, 2012; Giraldo and Valent, 2013). Fungal effectors share several features, which are not exclusive, i.e., most have a N-terminal secretion signal, enrichment in cysteine residues and a lack of functional homology in databases and present a small size. Such features have been widely used to determine sets of candidate effectors in the predicted proteome of fungal pathogens for which a reference genome has been sequenced (Lowe and Howlett, 2012; Duplessis et al., 2014a).

Host immunity escape by pathogens is frequently mediated by deletion or mutations in effector genes, which often show elevated levels of non-synonymous polymorphism as a result of their antagonistic co-evolution with the host (Stukenbrock and McDonald, 2009). The relative abundance of non-synonymous and synonymous polymorphisms (P_N and P_S) measures the direct effect of positive selection that tends to remove deleterious non-synonymous variants in coding sequences. When considered at the interspecific level, the rates of non-synonymous and synonymous substitutions (termed d*N* and d*S*, respectively) can be assessed to contrast patterns of variation between species (Stukenbrock and Bataillon, 2012). Such approaches have been applied at the genome scale to detect sets of candidate effectors in oomycetes and fungi (Raffaele and Kamoun, 2012; Cantu et al., 2013; Stergiopoulos et al., 2013; Stukenbrock, 2013). Evidence of positive selection was reported in avirulence genes of rust fungi at the intraspecific (AvrL567, Dodds et al., 2004; AvrP4 and AvrP123, Barrett et al., 2009) or interspecific levels (AvrP4, Van der Merwe et al., 2009). Genome-scale approaches were also used with sets of candidate effectors at the intraspecific level in *Puccinia striiformis* f. sp. *tritici* (Cantu et al., 2013) or by considering clusters of paralogous genes (CPG) in the genome of *M. larici-populina* (Hacquard et al., 2012).

Genomics is becoming a method of choice to identify new candidate effectors, particularly in obligate biotrophs where functional approaches are impeded. Only a handful of rust fungi genomes are available (Cantu et al., 2011, 2013; Duplessis et al., 2011a; Zheng et al., 2013; Nemri et al., 2014). In these, repertoires of candidate effectors corresponding to small secreted proteins (SSPs) have been defined (Hacquard et al., 2012; Saunders et al., 2012; Cantu et al., 2013; Zheng et al., 2013; Nemri et al., 2014). The poplar-poplar rust pathosystem is a model in forest pathology because it is one of the few pathosystems for which both the host and pathogen genomes are available (Tuskan et al., 2006; Duplessis et al., 2011a). *M. larici-populina* has a remarkably large diploid genome of 101 Mb enriched in repetitive and transposable elements (TE), a common feature of rust fungi genomes. There is a striking number of 16,399 predicted genes in the poplar rust genome, another feature shared with other rust fungi (Duplessis et al., 2014b). Among genes encoding secreted proteins (SPs), a set of 1184 SSP genes showing typical features of pathogen effectors was uncovered; most of these are cysteine-rich, belong to multigene families and are lineage specific (Duplessis et al., 2011a; Hacquard et al., 2012). In order to prioritize functional analysis of such candidates, other features were searched including specific expression during the interaction with the poplar host (Duplessis et al., 2011b), presence of conserved motifs in proteins, and gene families exhibiting evidences of positive selection by considering a classification into CPG (Joly et al., 2010; Hacquard et al., 2012). Another way to identify promising effectors is to study gene polymorphism at the intraspecific or interspecific level, as has been performed in *M. lini* (Ravensdale et al., 2011).

In the present study, we report on the genome sequencing of 15 *M. larici-populina* isolates and their comparison to the reference genome of isolate 98AG31 (Duplessis et al., 2011a) in order to identify patterns of genomic variations that may relate to fungal pathogenesis. Genes that accumulate intraspecific polymorphism in their coding sequence as well as in their non-coding upstream regions were scrutinized, thus providing a new filter to prioritize candidate effectors of interest.

MATERIALS AND METHODS

FUNGAL MATERIAL

Isolates were selected in a laboratory collection (Frey P., INRA Nancy, Champenoux, France) in order to maximize historical and geographical repartitions and virulence profiles (**Table 1**). Phenotypes of all isolates (i.e., combination of virulences) were confirmed in triplicate on eight poplar cultivars each carrying a single resistance (R1 to R8) to *M. larici-populina* (**Table 1**) and on the universal clone 'Robusta', as a positive control. To ensure their purity and to avoid potential clones within the selected isolates, genotyping was performed using 25 microsatellite markers (Xhaard et al., 2011). Urediniospores of each isolates were multiplied on 'Robusta' detached leaves to obtain enough material for genomic DNA isolation.

DNA ISOLATION

A total of 100–300 mg of urediniospores were used for DNA isolation using a CTAB method. Spores were crushed using a Retsch Tissue Lyser (Qiagen, Courtaboeuf, France) at a frequency of 30 Hz for 1 min. Broken spores were resuspended in CTAB buffer (Tris 0.1 M, NaCl 1.43 M, EDTA 0.02 M, CTAB 0.02 M) and heated at 65°C for 30 min. The suspension was subjected to centrifugation at 8000 rpm at room temperature for 5 min to pellet spore debris. Supernatant was gently mixed with an equal volume of phenol:chloroform:isoamyl alcohol (50:48:2; Euromedex, Souffelweyersheim, France) and centrifuged at 8000 rpm at room temperature for 10 min. The aqueous phase was recovered, gently mixed with an equal volume of chloroform and centrifuged at 8000 rpm at room temperature for 10 min. The aqueous phase was subjected to RNA digestion with RNaseA at 10 µM (Fermentas, Saint-Remy-lès-chevreuses, France) at 37°C for 30 min. A final extraction with an equal volume of chloroform was realized followed by centrifugation at 8000 rpm at room temperature for 10 min. The recovered aqueous phase was then subjected to isopropanol (0.75 of final volume) precipitation, followed by centrifugation at 14,000 rpm at 4°C for 30 min. DNA pellet was washed twice with 70%, then absolute ethanol, each followed by centrifugation at 14,000 rpm at 4°C for 10 min. The DNA pellet was finally dried under a hood for 20 min and resuspended in 1X Tris EDTA. Quality and quantity of recovered high molecular weight DNA was assessed by electrophoresis on agarose gel, by spectrophotometry (Nanodrop, Saint-Remy-lès-Chevreuse, France) and with the QuBit (Life Technologie, Villebon-sur-Yvette, France) fluorometric quantitation system.

Table 1 | Summary of *Melampsora larici-populina* isolates.

Isolate	Year	Location	Latitude, Longitude	Host	Pathotype
93ID6	1993	Champenoux (NE France)	N 48° 45′ 02″, E 06° 20′ 20″	*P. x euramericana* 'I45–51'	3-4
02Y5	2002	Charrey-sur-Saône (NE France)	N 47° 05′ 18″, E 05° 09′ 11″	*P. x euramericana* 'Robusta'	2-3-4-7-8
09BS12	2009	Mirabeau (SE France)	N 43° 41′ 29″, E 05° 40′ 21″	*P. nigra*	4-6
94ZZ15	1994	Saulchoy (N France)	N 50° 21′, E 01° 50′	*P. x euramericana* 'Luisa Avanzo'	3-4-5-7
94ZZ20	1994	Nogent-sur-Vernisson (Central France)	N 47° 50′, E 02° 45′	*P. x interamericana* 'Boelare'	3-4-7
08EA47	2008	Prelles (SE France)	N 44° 51′ 00″, E 06° 34′ 47″	*P. nigra*	2-4
95XD10	1995	Rogécourt (N France)	N 49° 39′, E 03° 25′	*P. x euramericana* 'Flevo'	3-4-5-7
08EA20	2008	Prelles (SE France)	N 44° 51′ 00″, E 06° 34′ 47″	*P. nigra*	4
08EA77	2008	Prelles (SE France)	N 44° 51′ 00″, E 06° 34′ 47″	*P. nigra*	4-6
97CF1	1997	Champenoux (NE France)	N 48° 45′ 02″, E 06° 20′ 20″	*P. x interamericana* 'Hoogvorst'	3-4-7
08KE26	2008	Mirabeau (SE France)	N 43° 41′ 29″, E 05° 40′ 21″	*P. nigra*	4
9683B13	1996	Orléans (Central France)	N 47° 49′ 39″, E 01° 54′ 40″	*P. x interamericana* '83B13'	1-3-4-5-6-7
98AG31	1998	Moy-de-l'Aisne (N France)	N 49° 45′, E 03° 21′	*P. x interamericana* 'Beaupré'	3-4-7
93JE3	1993	Champenoux (NE France)	N 48° 45′ 02″, E 06° 20′ 20′	*P. x euramericana* 'Blanc du Poitou'	2-4
98AR1	1998	Geraardsbergen (Flanders, Belgium)	N 50° 45′, E 03° 52′	*P. x interamericana* 'B71085/A1'	1-3-4-5-7-8

Isolate name, year, and location of sampling are indicated. Host indicates the poplar species/cultivar on which the isolate was sampled. The pathotype profile (combination of virulences) was confirmed in triplicate by inoculation on a differential set of poplar cultivars carrying the eight known resistances to M. larici-populina.

GENOME RE-SEQUENCING

For all isolates, except 98AR1, genomic DNA libraries were prepared using TruSeq DNA sample preparation kit (v3) followed by paired-end 100 nt massively parallel sequencing on Illumina HiSEQ2000 by Integragen (Evry, France). Briefly, 3 μg of each genomic DNA were fragmented by sonication and purified to yield fragments of 400–500 nt. Paired-end adapter oligonucleotides from Illumina were ligated on repaired A-tailed DNA fragments, then purified and enriched by PCR cycles. Each library was quantified by qPCR and sequenced on Illumina HiSeq2000 platform as paired-end 100 nt reads. Image analysis and base calling were performed using Illumina Real Time Analysis (RTA 1.13.48.0) pipeline with default parameters. Isolate 98AR1 genomic DNA was sequenced by a single read strategy of 75 bases on Illumina Genome Analyzer II (Beckman Coulter Genomics, Grenoble, France).

FILTERING AND MAPPING OF SHORT READS

Adapter and quality filtering was carried out using CLC Genomics Workbench 6.5 (CLC bio, QIAgen, Aarhus, Denmark). For each batch of reads, 3 and 10 low quality terminal nucleotides were trimmed at the 5′ and 3′ ends, respectively. FASTQ files of trimmed sequences were used to proceed with mapping onto the 98AG31 reference genome available at the Joint Genome Institute (JGI; http://genome.jgi.doe.gov/programs/fungi/index.jsf; Duplessis et al., 2011a). The 462 scaffolds composing the reference genome were uploaded in CLC Genomics Workbench and the annotation was superimposed onto the scaffolds using the annotation plugin. The following parameters were applied for mapping: masking mode = no masking; mismatch cost = 2; insertion cost = 3; deletion cost = 3; length fraction = 1.0; similarity fraction = 0.95; global alignment = no; auto-detect paired distances = yes; non-specific match handling = map randomly. Sequencing data and assemblies were deposited at at the National Center for Biotechnology Information (NCBI) and the Short Reads Archive (Bioproject PRJNA251864 study SRA accession SRP042998). Coverage and sequencing depth values were extracted from the CLC stand-alone read mapping files and were further used to compare scaffolds of resequenced isolates. Sequencing depth and coverage on each scaffold were visually inspected using the CLC read tracks functions used for further detection of structural variants.

SCAFFOLD DEPTH ANALYSIS AND VARIANTS DETECTION

Cross-comparison of average coverage and sequencing depth onto the 462 reference scaffolds was performed within and between isolates based on the CLC Genomics Workbench mapping outputs to detect the potential presence/absence of regions and the sequencing coverage or depth bias. In the case of missing regions or coverage bias, read mapping profiles and distribution of genes and TEs on the scaffolds were inspected manually. In these manual inspections, regions with high concentrations of ambiguous mappings were excluded from consideration, because of the possibility of artifactually divergent coverage. In parallel, the coverage analysis tool implemented in CLC Genomics Workbench (version 7.0) was used to detect regions within scaffolds showing significantly unexpected low or high coverage relative to the reference genome, according to a Poisson distribution of observed coverage in mapping positions (p-value threshold = 0.0001 and minimum length of the coverage region of 100 bp). Search for SP genes in the low-coverage regions was performed using an in-house Python script. Notably, this script was limited to detection of genes which laid entirely inside the corresponding region.

Single Nucleotide Variants (SNVs, i.e., Single Nucleotide Polymorphisms, SNPs), Multiple Nucleotide Variants (MNVs, i.e., successive SNVs), and small Insertion/Deletion variants (i.e., InDels) were detected in the genome of each isolate based on mapping outputs using the quality-based variant detection option of CLC Genomics Workbench (version 6.5.1). This option

considers minimum quality levels and minimum coverage of bases where the variant is detected and in surrounding bases. The following parameters were considered: neighborhood radius = 5; maximum gap and mismatch count = 2; minimum neighborhood quality = 15; minimum central quality = 20; ignore non-specific matches = yes; ignore broken pairs = yes; minimum coverage = 10; minimum variant frequency = 35%; maximum expected alleles = 2; advanced = no; require presence in both forward and reverse reads = no; ignore variants in non-specific regions = no; genetic code = standard. Variant tables were generated for all isolates. Selection of synonymous and non-synonymous polymorphism in genes and variants in 1 Kb upstream regions of genes was performed using in-house Python scripts.

SEQUENCE ANALYSIS

Gene and protein sequences and Gene Ontology (GO) and Eukaryotic Orthologous Group (KOG) functional annotations were retrieved from the *M. larici-populina* genome sequence on the Mycocosm website at the JGI (http://genome.jgi.doe.gov/programs/fungi/index.jsf). Homology searches were carried out using the Blastp algorithm (Altschul et al., 1997) against the non-redundant database at the NCBI (March 2014). AvrP4 sequences from Van der Merwe et al. (2009) and Barrett et al. (2009) were retrieved from the NCBI and used for multiple alignments with members of the CPG5464 gene family previously identified in the *M. larici-populina* genome (Hacquard et al., 2012). Alignment with variants of the CPG5464 gene family retrieved in the *M. larici-populina* isolates was conducted using the program ClustalW (Thompson et al., 2002 and gaps were manually inserted to strictly align sites reported under positive selection in the above-mentioned articles, before generating conservation profiles on the WebLogo server (Crooks et al., 2004).

KOG ENRICHMENT ANALYSIS

KOG (Tatusov et al., 2003) annotation of each *M. larici-populina* gene was retrieved from the JGI genome website. Each gene was classified according to the KOG functional classification using custom Perl scripts. Over-represented KOG categories in a selected gene set were calculated relative to the global gene distribution in the genome. Fisher's exact test was used to determine significant differences in the distribution of genes by KOG categories between the selected gene set and all genes ($p < 0.05$).

P_N/P_S ANALYSIS

For each gene, an alignment was generated with a custom Python script based on the reference genome and gene annotations (gff files from the *M. larici-populina* JGI website) taking into account the SNP variants generated by CLC Genomics Workbench. Alignments interrupted by an early stop codon were excluded from the computation of synonymous and non-synonymous polymorphisms. Polymorphism index was computed for each gene using Egglib version 2.1.6 (De Mita and Siol, 2012). This Python library computes from an alignment the number of synonymous or non-synonymous sites either polymorphic or non-polymorphic. P_N/P_S is computed as the ratio of the number of synonymous over non-synonymous polymorphisms corrected by the number of synonymous and non-synonymous sites, respectively.

RESULTS

SEQUENCING EFFICIENCY

Genomes of 15 *M. larici-populina* isolates, including the 98AG31 reference isolate, were sequenced at a targeted depth of ∼40X. A total of 64 billion bases were generated, corresponding to 2.5–6.2 billion reads per genome. After filtering, the average read length was 84.4 nt. A number of length and similarity parameters were tested for maping reads onto the reference genome. Loose default parameters tended to generate multiple mappings in repetitive sequences including large gene families, impinging on further call of variants in a given isolate (data not shown). Stringent parameters were retained (i.e., total length of the sequence showing a minimum of 95% similarity) for optimal mapping and subsequent variant calling. On average, 78% of the reads aligned to the 462 scaffolds of the reference genome (63–90%), and only one isolate had a lower percentage of mapped reads (isolate 9683B13, 40%). Examination of 1000 randomly selected unmapped reads from genome 9683B13 showed contamination with bacterial sequences (68%; >30% *Pseudomonas* sp. and >10% *Stenotrophomonas maltophilia*, data not shown), so these sequences were discarded. Overall, this led to a sequencing depth average of 32X per genome (22X–46X; **Table 2**). Overall coverage was between 90.7 and 96.3% for the 15 isolates. For all genomes sequenced with paired-end reads (that is, all except 98AR1), the number of broken paired reads was relatively moderate (<11% and average of 9%).

COVERAGE AND SEQUENCING DEPTH ANALYSIS

Cross-comparison of mapping outputs identified a bias of average coverage and sequencing depth among the 462 reference scaffolds within and between isolates. For instance, several scaffolds systematically showed very high (>100X) or low (<1X) depths in all sequenced isolates, and others showed discrepancies for a given scaffold between different isolates. Such situations were manually inspected and led to the survey of 151 scaffolds (representing about 10% of the genome sequence) for which the mapping depth profile and the presence of genes along the scaffolds were recorded (Supporting Table 1). Notably, scaffold 484 showed a systematic high depth >1000X. Four mitochondrial scaffolds were previously identified and removed from the poplar rust genome assembly (Duplessis et al., 2011a). Mapping of Illumina reads from the 15 isolates onto these four scaffolds showed much higher depth than the average observed for other scaffolds (178X–1211X, data not shown). Inspection of scaffold 484 indicated that it is most likely a portion of the mitochondrial genome. Indeed, this 5.4 Kb scaffold bears two genes showing high homology to two mitochondrial genes (ATP synthase F0 subunit and NADH dehydrogenase subunit).

For other scaffolds with systematic high coverage and sequencing depth biases, major differences are explained by missing regions in one or several isolates. Such scaffolds were marked by no mapping support for the entire scaffold, or for some regions of the scaffold at the same positions in a given subset of isolates (i.e., probable large deletions or highly variable loci). For

Table 2 | General mapping information for the 15 *Melampsora larici-populina* isolates.

Isolate	Total reads number	Mapped reads	% Mapped reads	Broken pairs	Average read length	Sequencing depth
93ID6	3,594,455,577	2,656,764,147	73.9	226,296,523	84.4	26.3
02Y5	3,691,995,994	3,218,997,193	87.2	269,105,383	85.4	31.8
09BS12	6,230,429,688	4,717,557,005	75.7	479,213,815	84.2	46.6
94ZZ15	3,653,741,644	3,290,238,877	90.1	278,395,986	85.3	32.5
94ZZ20	3,387,309,786	3,045,158,939	89.9	253,470,401	85.2	30.1
08EA47	4,659,300,813	3,460,505,640	74.3	352,258,523	83.3	34.2
95XD10	4,701,407,950	3,993,529,488	84.9	396,812,163	83.7	39.5
08EA20	4,829,802,826	3,034,419,164	62.8	290,972,918	83.2	30.0
08EA77	4,259,571,919	3,840,082,037	90.2	340,127,111	84.7	38.0
97CF1	3,570,560,916	3,083,826,749	86.4	270,564,864	84.5	30.5
08KE26	5,407,393,523	4,626,803,739	85.6	434,085,871	85.0	45.8
9683B13	6,378,404,736	2,537,206,558	39.8	223,243,679	83.1	25.1
98AG31	2,779,485,081	2,529,716,573	91.0	218,294,868	85.2	25.0
93JE3	4,310,048,066	2,796,054,256	64.9	258,175,513	84.1	27.6
98AR1	2,562,143,464	2,227,530,892	86.9	na	76.0	22.0

Illumina reads of each genome were mapped onto the 98AG31 JGI reference genome. na, not applicable.

instance, the 319 Kb scaffold 90 showed either a similar depth along the scaffold in reference isolate 98AG31 and five other isolates (pattern A; **Figure 1**), or the absence of regions at the same positions for two patterns, each grouping different isolates (patterns B and C; **Figure 1**). Pattern C exhibited an overall low sequencing depth ranging from 3.5X to 5.8X, that mostly corresponds to repetitive elements regions marked by peaks of high depth similar to those present in patterns A and B. This indicates that the missing regions were not related to sequencing depth (**Figure 1**). For pattern C with the longest missing regions, a total of 38 genes were not supported by reads, including 4 pheromone genes related to mating type in the poplar rust fungus. Despite a generally similar profile and sequencing depths within pattern C, isolates 08EA20 and 08EA77 showed a higher coverage (54.6 and 58.8%, respectively) than the other three isolates (20.8, 22.9 and 23.6%). This is explained by a light and continuous depth in the central region of the scaffold that was totally absent in the other isolates (**Figure 1**). In isolate 08EA20 two genes located at 16–17 Kb (hypothetical protein) and 22–23 Kb (chitinase) were present. In isolate 08EA77, only the chitinase encoding gene was present, whereas these two genes were missing in the other three isolates of pattern C. Assembling unmapped reads from isolates exhibiting pattern C onto the 38 missing genes using loose similarity parameters retrieved only highly divergent and/or partial sequences (data not shown). Because of the presence of pheromone genes on scaffold 90, we looked at previously described mating type loci in the *M. larici-populina* genome (Duplessis et al., 2011a). A missing region containing a pheromone gene and a STE3 pheromone receptor gene was also observed in scaffold 172 for the isolates with pattern C. This prompted us to examine the homeodomain locus, composed of the genes HD1 and HD2. The five isolates that exhibited missing regions in scaffolds 90 and 172 also presented a missing region at the homeodomain locus in the scaffold 35. Using the homeodomain loci and pheromone/receptor loci genes as baits, divergent alleles were identified for *M. larici-populina HD1*, *HD2* and some pheromone genes in the unmapped reads of these isolates (data not shown).

A total of 212 genes lie in the missing regions of the surveyed scaffolds, including 12 SP genes in 7 scaffolds (Supporting Table 1). We therefore conducted a systematic analysis of regions of 100 bp or more showing coverage differences using the CLC coverage analysis tool, in order to detect possible deletions or amplifications. In total, 18,564–81,325 regions with significantly high/low coverage differences relative to the 98AG31 reference genome were identified in the 14 isolates (Supporting Table 2). Search for SP genes within these regions revealed that between 12 (9683B13) and 59 (95XD10) SP genes are in low coverage regions, indicating a possible deletion compared to isolate 98AG31. However, we could not find any correlation between a probable SP gene deletion and the pathotypes of the isolates, i.e., the absence of a SP gene explaining virulences 1, 2, 5, 6, or 8 (98AG31 reference isolate being virulence 3, 4, 7).

POLYMORPHISM AND INSERTION/DELETION DETECTION

In order to assess polymorphism in the 15 isolates, variants (SNVs/SNPs, MNVs, and InDels) were recorded using the CLC Genomics Workbench program. The 98AG31 reference genome had been sequenced at a 6.9X sequencing depth from dikaryotic urediniospores by Sanger sequencing, following a whole-genome shotgun strategy. Therefore, the 462 scaffolds represent a chimeric version of the genome combining the two haplotypes (Duplessis et al., 2011a). Resequencing by Illumina at a sequencing depth of 25X identified a total of 93,189 variants including 86,877 SNPs, 1741 MNVs, 2945 insertions and 1626 deletions in isolate 98AG31 (representing 96,099 bases; **Table 3**), which is in close range with the 88,083 SNPs recorded by Sanger sequencing. However, only 40,001 SNPs from the initial assembly were confirmed, highlighting differences due to the sequencing approaches. An average of 163,477 variants (including 152,936 SNPs) representing 168,708 bases was found in the 14 other isolates mapped onto the reference genome, representing a larger number of polymorphic sites

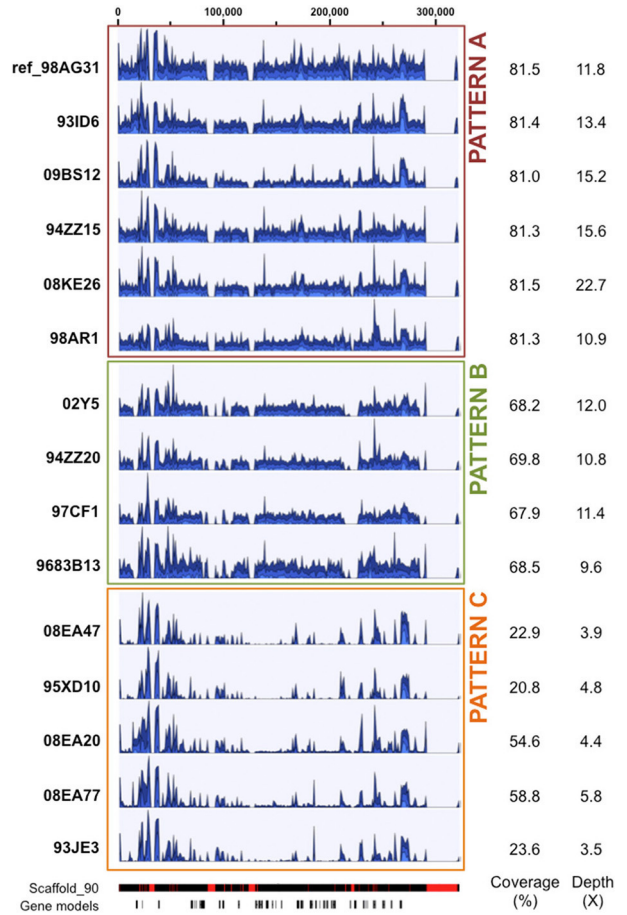


FIGURE 1 | Patterns of sequencing depth along scaffold 90 in 15 *Melampsora larici-populina* isolates. Illumina reads from 15 isolates were mapped onto the 98AG31 reference genome. Scaffold 90 is presented here to illustrate distinct patterns of sequencing depth between groups of isolates: pattern A (red box) with coverage and sequencing depth similar to 98AG31, pattern B (green box) presenting four regions of lower coverage, and pattern C (orange box) with overall reduced coverage. Graphical outputs in blue represent the local sequencing depth along scaffold 90, normalized to the maximum depth measured in each isolate. Average coverage and sequencing depth are detailed for each isolate on the right. The bars below represent scaffold 90 from JGI reference genome website (red blocks indicate gaps) and predicted gene models (38 in total). Scale in nucleotides is presented at the top. The total scaffold length is 319,043 bp.

at the inter-individual level (0.17% of the genome; 1.51 SNPs/Kb). When the 15 genomes were considered together, 11,683 SNPs were conserved, whereas in total 611,824 unique SNPs were found. The variant caller implemented in CLC allowed the determination of the zygosity of nucleotides at the polymorphic sites. The heterozygosity rate was 0.45–0.55 in 12 isolates, whereas it was lower in 09BS12 and 08KE26 (0.35 and 0.37, respectively) and higher in 98AG31 (0.85). The latter is as expected, as it was the reference genome to which reads were mapped (**Table 3**). For all genomes, the ratio of transition over transversion mutations was 2.31 ± 0.11 (**Table 4**), which is similar in range to previous observations in rust fungi (Cantu et al., 2013). Individually, all isolates except the reference 98AG31 showed similar numbers of SNPs, MNVs, and InDels (**Table 3**), indicating a homogeneous polymorphism rate at the intraspecific level. Polymorphic sites residing within coding DNA sequences (CDS) were more closely scrutinized and represented 20% of the SNPs, 17% of the MNVs, and 5% of deletions, and 5% of insertions in InDels. These proportions were rather similar in the different isolates (**Table 4**). In total, more SNPs were present in exons than in introns (average 30,077 ± 3893 SD and 14,982 ± 1871 SD, respectively; **Table 4**), but when exon and intron size were accounted for, introns tended to accumulate more SNPs than the coding sequences (data not shown).

HIGHLY VARIABLE GENES

Synonymous and non-synonymous polymorphisms within the 15 isolates were inspected in the gene complement of *M. larici-populina*, considering only SNPs that were represented in most of the observed variants (90%). Both homozygous and heterozygous SNPs were considered. For cross-comparison of SNPs between isolates, non-redundant SNPs (i.e., nucleotides in the reference isolate presenting polymorphism in at least one other isolate) were considered. Overall, a very large portion of the genes (89%) was marked at least by one SNP, and 5332 and 10 genes exhibited more than 10 and 100 SNPs, respectively (Supporting Table 3). A total of 1089 genes in the 15 isolates had more than 10 non-synonymous SNPs in CDS, the maximum number being 66 (proteinID 66139). **Table 5** presents the top 30 genes with the highest number of non-synonymous SNPs over the 15 genomes, with 20.5 SNPs/Kb and 11.8 non-synonymous SNPs/Kb on average. Homology searches by Blastp against the NCBI nr protein database indicated a putative function or presence of a conserved domain for nine of the genes, six of which are associated with predicted nuclear activity. In total, 14 genes had GO and/or KOG annotations, and the majority encode predicted proteins of unknown function. A functional KOG analysis of the 4142 genes exhibiting ≥5 non-synonymous SNPs revealed significant enrichment for gene categories related to chromatin structure and dynamics; cell cycle control, cell division and chromosome partitioning; nuclear structure; defense mechanisms and extracellular structures (**Figure 2**). SNPs were also inspected in the 1 Kb upstream regions of CDS, where they may impact transcription. Most genes also had at least one polymorphic site in their 1 Kb upstream regions (89%) and 2554 genes each had more than 10 SNPs in these regions (Supporting Table 3). Half of the 30 genes with the highest number of SNPs had an annotation in various cellular categories including two SSP genes, the other half corresponded to genes encoding predicted proteins of unknown function (Supporting Table 4).

HIGHLY VARIABLE SECRETED PROTEIN ENCODING GENES

A set of 1184 SSP-encoding genes representing candidate poplar rust effectors was previously reported (Hacquard et al., 2012). Because larger effectors were also described (e.g., flax rust AvrM; Ravensdale et al., 2011), we decided to place a particular focus on secreted protein encoding genes as possible candidate effectors (i.e., a total of 2050 SPs identified by automatic annotation, including the 1184 SSPs). We further distinguish SSPs from SPs as SSP genes were manually annotated in the *M. larici-populina*

Table 3 | Genomic variants identified in 15 *Melampsora larici-populina* isolates by mapping onto the 98AG31 JGI reference genome.

Isolate	Zygosity	Variant types	Total					
	Homozygous	Heterozygous	Deletion	Insertion	MNVs	SNVs	Variants	Nucleotides
93ID6	84,849	88,855	3534	4198	3302	162,670	173,704	179,274
02Y5	76,511	95,418	3514	4399	3348	160,668	171,929	177,658
09BS12	91,934	54,500	3170	4020	2835	136,409	146,434	151,298
94ZZ15	84,155	82,478	3485	4287	3160	155,701	166,633	172,001
94ZZ20	80,851	80,541	3385	4085	3002	150,920	161,392	166,613
08EA47	85,423	75,527	3435	4158	3026	150,331	160,950	166,117
95XD10	68,735	87,520	2909	3554	2903	146,889	156,255	160,886
08EA20	90,268	91,000	3723	4354	3469	169,722	181,268	187,146
08EA77	89,765	83,569	3599	4275	3222	162,238	173,334	178,887
97CF1	75,954	76,585	3061	3902	2958	142,618	152,539	157,525
08KE26	102,244	55,022	3578	4268	3100	146,320	157,266	162,670
9683B13	70,974	82,208	3004	3708	2866	143,604	153,182	157,967
98AG31	14,219	78,970	1626	2945	1741	86,877	93,189	96,099
93JE3	91,933	75,793	3277	3938	3182	157,329	167,726	172,951
98AR1	77,267	88,799	3315	3932	3130	155,689	166,066	170,921

MNV, Multiple Nucleotide Variant; SNV, Single Nucleotide Variant (i.e., Single Nucleotide Polymorphism).

Table 4 | Analysis of polymorphism in 15 *Melampsora larici-populina* isolates.

Isolate	SNPs					% Polymorphism in CDS			
	Tr/Tv	SNPs in exon	SNPs in intron	SNPs intergenic	Non-synonymous SNP	Deletion	Insertion	MNV	SNV
93ID6	2.30	33,428	16,489	112,753	15,950	5.0	5.7	18.3	20.5
02Y5	2.30	32,904	16,325	111,439	15,553	5.5	5.4	15.9	20.5
09BS12	2.34	26,086	13,365	96,958	12,905	5.2	5.6	16.8	19.1
94ZZ15	2.29	31,938	16,056	107,707	15,252	6.2	5.7	17.2	20.5
94ZZ20	2.29	31,035	15,461	104,424	14,859	5.2	5.2	17.7	20.6
08EA47	2.30	29,848	14,986	105,497	14,493	5.3	5.5	17.3	19.9
95XD10	2.42	29,932	14,817	101,940	15,950	4.5	4.9	14.8	20.4
08EA20	2.30	35,069	17,230	117,423	16,911	5.3	5.4	18.0	20.7
08EA77	2.35	32,152	16,383	113,703	15,653	5.4	5.4	17.1	19.8
97CF1	2.33	29,566	14,649	98,403	14,218	5.7	6.1	17.9	20.7
08KE26	2.36	27,137	13,886	105,297	13,862	5.3	5.6	16.5	18.5
9683B13	2.33	29,776	14,719	99,109	14,442	5.1	5.5	17.7	20.7
98AG31	2.27	18,749	9335	58,793	8825	6.6	5.8	19.9	21.6
93JE3	2.36	32,155	15,684	109,490	15,651	4.9	5.4	18.0	20.4
98AR1	2.21	31,389	15,352	108,948	14,441	4.7	5.2	17.0	20.2

CDS, Coding DNA sequence. Tr/Tv, rate of transition to transversion; MNV, Multiple Nucleotide Variants; SNV/SNP, Single Nucleotide Variant/Polymorphism.

genome (Hacquard et al., 2012). Overall, a very large portion of the SP genes (89%) was marked by at least one SNP and 586 exhibited 10 SNPs or more (Supporting Table 5). A total of 386 and 119 genes had more than 5 and 10 non-synonymous SNPs, respectively (maximum = 45 non-synonymous SNPs; proteinID 66458). **Table 6** presents the top 30 SP genes with the highest numbers of non-synonymous SNPs/Kb, of which 24 are SSP genes. Only six SPs showed homology to other fungal proteins, including an *M. lini* avirulence factor AvrP4, a metallopeptidase, and a pleckstrin homology-like domain involved in binding to interacting protein partners. Rates of synonymous (P_S) and non-synonymous (P_N) substitutions were calculated for all genes with the EggLib package (Supporting Table 3) and SP genes were more particularly scrutinized. The P_N/P_S rate could be measured for 14,052 genes, while 1073 genes had a mutation generating a stop codon in the sequence and were excluded. P_N/P_S showed similar distributions between SP genes and other genes (**Figure 3**) and the highest P_N/P_S (4.9) was found for a gene encoding a hypothetical protein (ProteinID_70080; Supporting Table 3). In SP genes, the highest P_N/P_S was 2.47 and corresponds to a SSP of 200 amino acids with three homologs in *Puccinia graminis* f. sp. *tritici* and no conserved domain (ProteinID_124304; Supporting Table 5). The average P_N/P_S observed in SP genes (0.20) was lower than for other genes (0.25). A total of 68 SP genes showed a $P_N/P_S >$

Table 5 | Top 30 genes accumulating non-synonymous (NS) Single Nucleotide Polymorphism (SNP).

JGI Protein ID[a]	Protein length	Transcript length	SNP	NS	Annotation	GO ID[a]	KOG ID[a]
66139	5273	15819	227	66	AAA+ ATPase	0003677	1808
84101	1325	3975	95	57	Hypothetical protein	No hit	No hit
93626	1737	5211	82	54	Hypothetical protein	No hit	No hit
62079	1821	5463	73	47	Hypothetical protein, telomere-length maintenance and DNA damage repair domain	0001584	No hit
106057	2195	6585	136	45	Hypothetical protein, NAM-like protein C-terminal domain	No hit	No hit
92944	1135	3405	71	45	Hypothetical protein, DNA breaking-rejoining enzymes, C-terminal catalytic domain	No hit	No hit
95670	893	2679	87	45	Hypothetical protein	No hit	1187
66458	929	2787	55	45	Hypothetical protein	No hit	1245
70222	1542	4626	73	44	DEAD-like helicase superfamily	No hit	0351
101154	1470	4410	79	44	Hypothetical protein	No hit	No hit
114610	948	2844	91	41	Hypothetical protein	No hit	No hit
85441	1256	3768	56	40	Hypothetical protein	No hit	0714
92226	1393	4179	59	38	Hypothetical protein	No hit	No hit
67208	1203	3609	76	37	Hypothetical protein	No hit	1015
108793	931	2793	54	37	Hypothetical protein	No hit	No hit
96388	1344	4032	63	36	Hypothetical protein	No hit	No hit
108574	2851	8553	114	35	Hypothetical protein, down-regulated in metastasis domain	No hit	No hit
91870	1131	3393	72	35	Hypothetical protein, alpha kinase domain family	0004674	3614
118268	1649	4947	108	34	Hypothetical protein, sister-chromatid cohesion C-terminus domain	0006520	No hit
68278	1507	4521	54	34	Hypothetical protein	No hit	4475
65221	568	1704	44	34	Hypothetical protein	No hit	No hit
91258	771	2313	51	33	Hypothetical protein, GCM transcription factor family motif	No hit	2992
88323	575	1725	55	33	Hypothetical protein	No hit	No hit
60895	698	2094	58	33	Hypothetical protein	No hit	2992
84177	639	1917	57	33	Hypothetical protein	No hit	No hit
92190	551	1653	52	33	Hypothetical protein	0006306	No hit
101664	1102	3306	63	32	Hypothetical protein	No hit	No hit
95815	1486	4458	46	32	Hypothetical protein	No hit	1245
107058	720	2160	51	32	Hypothetical protein	No hit	No hit
64441	1107	3321	45	31	Hypothetical protein	No hit	No hit

[a] *Protein ID number, Eukaryotic Orthologous Group (KOG) and Gene Ontology (GO) annotations were retrieved from the 98AG31 reference genome at the Joint Genome Institute Mycocosm website (http://genome.jgi.doe.gov/programs/fungi/index.jsf).*

1, whereas 668 had a $P_N/P_S > 1$ in other genes (Supporting Table 6). Among the 30 genes with the highest numbers of non-synonymous SNPs, nine have a $P_N/P_S > 1$ (**Table 6**). These genes represent particularly interesting candidates that could have evolved under the selection pressure exerted by the interaction with the host plant. No enrichment in KOG functional annotation was detected for the 736 genes presenting a $P_N/P_S > 1$.

In the panel of 15 *M. larici-populina* isolates, only two of the eight virulences described in the poplar rust fungus presented a balanced frequency: virulence 3 with six avirulent isolates and nine virulent isolates and virulence 7 with seven avirulent isolates and eight virulent isolates (**Table 1**). SP genes presenting conserved non-synonymous SNPs in avirulent isolates and not in virulent isolates (including the reference genome 98AG31 which carries virulences 3 and 7) could be strong candidates, however none of the SP genes presented such a pattern for virulence 3 and 7, suggesting that events other than non-synonymous substitutions in coding sequence may explain the emergence of the virulences 3 and 7. Four SP genes (Protein IDs 89167, 91014, 105154, and 123753) presented non-synonymous SNPs in isolates

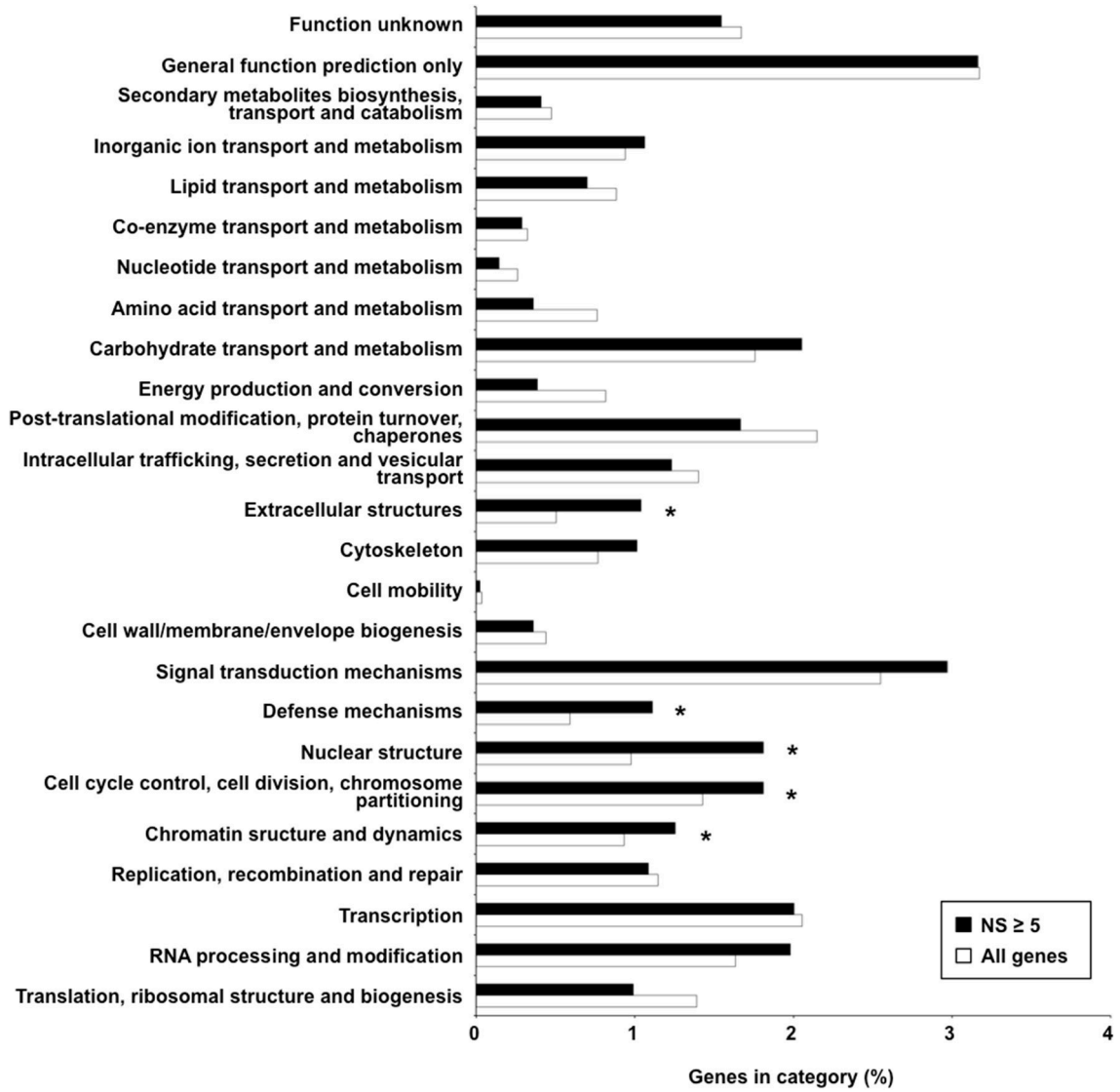


FIGURE 2 | Functional categories over-represented among genes exhibiting five non-synonymous polymorphisms or more. Percentages of genes falling in the different KOG categories among genes exhibiting five non-synonymous polymorphisms or more (NS $\geq$ 5) relative to the global gene distribution are shown. Black and white bars correspond to selected NS $\geq$ 5 genes and all genes, respectively. The category "No hits" corresponding to genes with no KOG annotation (~75% in both sets) is not represented on the graph to facilitate visualization of other categories. Significantly over-represented KOG categories are indicated by asterisks (Fisher's exact test, $p < 0.05$).

98AR1 and 02Y5 which bear the virulence 8, whereas these were absent from the other 13 avirulent isolates, suggesting these genes could be candidate effectors for virulence 8. One SP gene (Protein ID 104703) presented non-synonymous SNPs in isolates 98AR1 and 9683B13 that were absent from the other isolates, indicating that this gene could be a candidate related to virulence 1. One SP gene (Protein ID 108857) presented non-synonymous SNPs in isolates 08EA77, 9683B13, and 09BS12, whereas they were absent from the 12 other isolates, suggesting also that this gene could be a candidate for virulence 6. No correlation was found between mutations in SP genes and other virulences. Similarly, none of the genes interrupted by stop codons correlated with the pathotypes of the 15 isolates.

M. larici-populina SSP genes showing homology to *M. lini Avr* genes *AvrL567*, *AvrP123*, and *AvrP4* do not exhibit important accumulation of non-synonymous SNPs (Supporting Table 5). Interestingly, the polymorphic sites identified for the *M. lini* AvrL567 homolog in the poplar rust genome correspond to those that were previously identified by PCR-cloning in a panel of 32 *M. larici-populina* isolates (Hacquard et al., 2012), which included isolate 98AR1, validating the SNPs found in this candidate. Evidence of positive selection were previously recorded for *AvrP4* genes at the intraspecific level in *M. lini* (Barrett et al., 2009) and at the interspecific level in the Melampsoraceae family (Van der Merwe et al., 2009), as well as in a cluster of paralogous genes encoding AvrP4-homologs (multigene family CPG5464;

Table 6 | Top 30 genes encoding secreted proteins accumulating non-synonymous SNPs/Kb.

Protein ID[a]	Protein length	Transcript length	SNP	NS	NS/Kb	Annotation	KOG ID[a]	Go ID[a]
124497	77	231	5	5	21.6	hypothetical secreted protein of 8 kDa	No hit	No hit
124050	151	453	13	9	19.9	hypothetical secreted protein of 17 kDa	No hit	No hit
124361	88	264	5	5	18.9	hypothetical secreted protein of 9 kDa	No hit	No hit
109910	230	690	17	13	18.8	hypothetical secreted protein	No hit	No hit
123541	75	225	6	4	17.8	hypothetical secreted protein of 8 kDa	No hit	No hit
123852	135	405	55	7	17.3	hypothetical secreted protein of 15 kDa	No hit	No hit
104907	117	351	6	6	17.1	hypothetical secreted protein	1245	No hit
123868	139	417	15	7	16.8	hypothetical secreted protein of 15 kDa	No hit	No hit
66458	929	2787	55	45	16.1	hypothetical secreted protein	No hit	No hit
103402	151	453	15	7	15.5	hypothetical secreted protein	No hit	No hit
101262	131	393	18	6	15.3	hypothetical secreted protein	No hit	No hit
124304	200	600	10	9	15.0	hypothetical secreted protein of 22 kDa	No hit	No hit
107425	268	804	28	12	14.9	hypothetical secreted protein	No hit	No hit
124511	67	201	3	3	14.9	hypothetical secreted protein of 7 kDa	No hit	No hit
124264	90	270	5	4	14.8	hypothetical secreted protein of 10 kDa, *Melampsora lini* AvrP4 homolog	No hit	9055
107508	720	2160	51	32	14.8	hypothetical secreted protein	No hit	No hit
124351	92	276	7	4	14.5	hypothetical secreted protein of 10 kDa	No hit	No hit
95362	301	903	18	13	14.4	hypothetical secreted protein	No hit	No hit
64885	188	564	23	8	14.2	hypothetical secreted protein of 21 kDa	No hit	No hit
58423	142	426	10	6	14.1	hypothetical secreted protein of 14 kDa	No hit	No hit
124524	71	213	3	3	14.1	hypothetical secreted protein of 8 kDa	No hit	No hit
63656	315	945	22	13	13.8	hypothetical secreted protein	No hit	No hit
70838	97	291	9	4	13.7	hypothetical secreted protein of 10 kDa	No hit	No hit
123559	146	438	10	6	13.7	hypothetical secreted protein of 16 kDa	No hit	No hit
61241	392	1176	39	16	13.6	hypothetical secreted protein, PLECKSTRIN homology domain	No hit	No hit
68348	247	741	18	10	13.5	hypothetical secreted protein	No hit	No hit
123552	150	450	12	6	13.3	hypothetical secreted protein of 17 kDa	No hit	No hit
124134	125	375	14	5	13.3	hypothetical secreted protein of 14 kDa	No hit	No hit
108793	931	2793	54	37	13.2	hypothetical secreted protein	No hit	No hit
36743	179	537	8	7	13.0	hypothetical secreted protein of 21 kDa, peptidase M, neutral zinc metallopeptidase	No hit	8237

[a] *Protein ID number, Eukaryotic Orthologous Group (KOG) and Gene Ontology (GO) annotations were retrieved from the 98AG31 reference genome at the Joint Genome Institute Mycocosm website (http://genome.jgi.doe.gov/programs/fungi/index.jsf).*

Hacquard et al., 2012). The 13 members of the CPG5464 family in *M. larici-populina* were more closely examined in the 15 isolates (**Figure 4**). The 13 members of the family were rather conserved and only four had non-synonymous SNPs between isolates (CPG5464_124256, CPG5464_124262, CPG5464_124264, CPG5464_124266). In total, substitutions were noted at four different positions, two within the signal peptide and two after the conserved K/R and E/D regions. None of these substitutions corresponded to positions previously shown under positive selection at the intraspecific or interspecific level (**Figure 4**). Notably, CPG5464_124564, which includes three different substitution sites in three isolates, presented a P_N/P_S value of 1 and was among the SP genes exhibiting the highest numbers of SNPs/Kb (**Table 6**, Supporting Table 5). Among the eight homologs of *M. lini AvrM* genes, one showed 15 non-synonymous SNPs (ProteinID_124207; Supporting Table 3). Three *Uromyces fabae* RTP1 homologs have been described in *M. larici-populina* (Hacquard et al., 2012). Only one RTP1 homolog (ProteinID_123932; Supporting Table 3) that consists of a fusion between a *M. lini* HESP-327 homolog and an *U. fabae* RTP1 homolog exhibited an important number of non-synonymous SNPs (7, of which 5 reside in the C-terminal RTP1 region). No substitution occurred at the positions of the four conserved cysteine residues under purifying selection identified by Pretsch et al. (2013).

DISCUSSION

The sequencing of the *M. larici-populina* genome has opened new avenues for the study of effector genes in a model pathosystem composed of a perennial plant and an obligate biotrophic rust fungus (Duplessis et al., 2011a; Hacquard et al., 2011). A set of 1184 candidate poplar rust effectors were identified on the

basis of a combination of typical features of effectors reported in other fungal pathogens, including an initial arbitrary size filter to focus on small proteins of less than 300 amino acids (Hacquard et al., 2012). Because rust fungi effectors such as the *M. lini* AvrM avirulence factor (Ravensdale et al., 2011) can be larger, all predicted SPs were subsequently considered in the search for candidate effectors. Complementary information such as transcript profiling during host infection or the pathogen life cycle can help in reducing the set of genes likely to be *bona fide* effectors (Duplessis et al., 2011b; Hacquard et al., 2013a). Another filter commonly used to identify candidate effectors in plant pathogens is the detection of positive selection in virulence genes, indicative of the evolutionary pressure exerted by the plant-pathogen co-evolution (Alfano, 2009; Stergiopoulos and de Wit, 2009). Events such as non-synonymous substitutions, gene gain, gene loss or differential regulation of gene expression can affect avirulence genes and generate new virulences in plant pathogens; comparative genomics using new generation sequencing technologies have uncovered such types of events (Raffaele and Kamoun, 2012; Spanu, 2012). In the present study, we applied Illumina sequencing by synthesis to explore the genetic diversity of *M. larici-populina*, focusing on 15 isolates collected on poplar trees in the wild or in experimental poplar nurseries in the past 21 years in France, and with a wide range of virulence profiles. The main goal here is to provide another level of information about *M. larici-populina* genes in order to guide selection of pathogenesis-related genes, including effectors, for future functional analyses. The mapping of Illumina reads onto the 98AG31 reference genome helped in the detection of variations such as SNPs and InDels. To date, only a few reports explored genetic diversity at the genome scale in rust fungi using Illumina technology, but they provide ground for comparison within the Pucciniales order (Duplessis et al., 2014b).

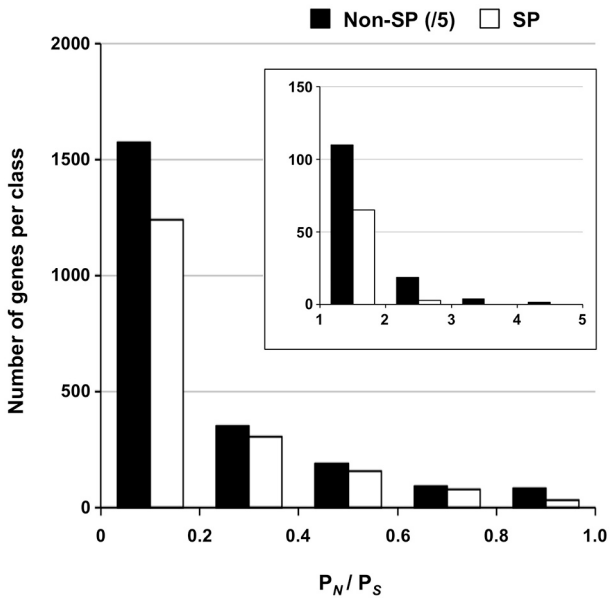


FIGURE 3 | Distribution of P_N/P_S for SP and non-SP genes. Ratios of non-synonymous to synonymous polymorphisms (P_N/P_S) between 0 and 1 are shown for SP genes and non-SP genes. The insert shows distribution of genes with a $P_N/P_S > 1$. Numbers of non-SP genes were divided by 5 for representation. Note the different scale for *y*-axes in figure and insert.

RESEQUENCING *M. LARICI-POPULINA* GENOMES REVEALS STRUCTURAL VARIATIONS

Reads were mapped onto the 98AG31 reference genome with good overall coverage and sequencing depth. Although there was

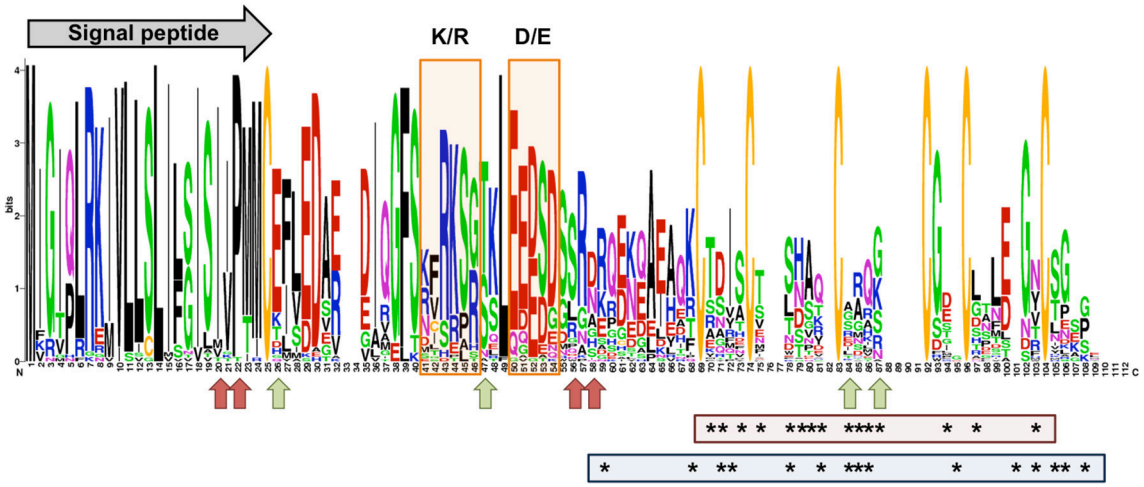


FIGURE 4 | Conservation protein profile of the *M. larici-populina* CPG5464 family and AvrP4 homologs in Melampsoraceae. The profile was designed using WebLogo with 40 sequences corresponding to the 12 members in the CPG5464 family (Hacquard et al., 2012), six variants deduced from the 15 genomes sequenced in this study, 22 AvrP4 homologs sequenced from 9 *Melampsora* spp. (Van der Merwe et al., 2009) and 16 *Melampsora lini* AvrP4 variants. The predicted signal peptide and K/R and D/E rich regions previously shown in Hacquard et al. (2012) are depicted on the profile. Green arrows point to sites under selection in Barrett et al. (2009). Red arrows point to sites of substitution observed in *M. larici-populina* variants. Asterisks in the red box indicate amino acids under positive selection in Van der Merwe et al. (2009) and asterisks in the blue box indicate amino acids under positive selection in Hacquard et al. (2012).

a narrow range in the average coverage by isolate, discrepancies were observed for given scaffolds. Particularly, the small scaffold 484 presented a strikingly high sequencing depth. Two genes encoding an ATP synthase F0 subunit and a NADH dehydrogenase subunit presenting strong similarity with resident genes of the soybean rust *Phakopsora pachyrhizi* mitochondrial genome (Stone et al., 2010) are present on this scaffold. Thus, our analysis identifies a new mitochondrial scaffold that will help in refining the genome assembly. Detailed examination of scaffolds that presented divergent coverage and sequencing depth between isolates revealed on some occasions rather large missing gene-containing regions compared to the reference genome. Although still unresolved, the poplar rust fungus seems to possess a tetrapolar mating system, as for many other basidiomycetes (Duplessis et al., 2011a). In this system, two unlinked loci govern the sexual cycle, and both loci should differ to complete mating (Fraser et al., 2007). Three distinct patterns of conserved missing regions were observed between isolates of unrelated pathotypes collected on different years at different locations (see **Table 1** for collection details). Scaffold 90 showed the most striking differences, where missing regions encompass a total of 38 genes, including four pheromone genes that were previously annotated in mating type loci of *M. larici-populina* (Duplessis et al., 2011a). Other mating type loci (i.e., the pheromone/receptor and the homeodomain loci) are also missing in these isolates suggesting that their mating type loci are highly divergent. Despite the quality of the reference genome assembly, the organization of the mating type loci is still not resolved (Duplessis et al., 2011a). This study will provide support to further explore and resolve the organization and composition of the poplar rust fungus mating loci. Other missing regions unrelated to the mating loci suggest that the poplar rusts posses a great genomic variability. In *M. oryzae*, 1.68 Mb (of a total of 38 Mb) were missing in isolate Ina168 resequenced by 454-pyrosequencing compared to the 70-15 reference genome (Yoshida et al., 2009). This has led to the discovery of many missing SSP genes including known avirulence genes between the two *M. oryzae* isolates (Yoshida et al., 2009). In *M. larici-populina*, none of the missing regions contained large numbers of SP genes (only 12 in total). By performing a wider coverage analysis in the 15 isolates, up to 59 SP genes were found in low coverage regions, representing possible deletions. However, no such deletion correlates with the poplar rust virulences. In *P. striiformis* f. sp. *tritici*, less than 1.3% of the secretome (15 SP genes) was absent between the most divergent sequenced isolates (Cantu et al., 2013), which indicates that the same set of SP genes occurs at the intraspecific level in these rust fungi.

M. LARICI-POPULINA GENOMES SHOW REMARKABLE LEVELS OF POLYMORPHISM

The reference genome 98AG31 was included in the panel of 15 isolates. This genome was previously characterized by Sanger sequencing, which provided an adequate assembly into 462 scaffolds (considering the large size of 101 Mb and a large content in TE, i.e., 45%), however at a rather low sequencing depth of 6.9X (Duplessis et al., 2011a). A total of 88,083 SNPs were previously identified in the reference genome by mapping back Sanger sequencing reads onto the assembled reference genome, with a loose criterion considering a minimum of four reads at a given position (Duplessis et al., 2011a). Illumina sequencing identified a total of 93,189 variants including 86,877 SNPs, of which only 40,001 confirmed SNPs found in the initial assembly. This finding strengthens the support for the use of resequencing at a greater depth to confidently assess SNPs. The total number of SNPs we report is slightly lower than the one found in *P. graminis* f. sp. *tritici* (129,172; Duplessis et al., 2011a). It differs, too, to the numbers reported in *P. striiformis* f. sp. *tritici*, with 81,001–108,785 depending on the isolate considered in Zheng et al. (2013) and more than 350,000 with important variations between isolates in Cantu et al. (2013). The large variation in SNPs in these studies could be explained by the wide variation in geographical origin of the isolates and the varying rates of occurrence of sexual reproduction at these sites. Population analyses of the poplar rust fungus with neutral markers indicate that the fungus frequently undergoes sexual recombination resulting in regular gene flow within natural population (Gérard et al., 2006; Barrès et al., 2008; Xhaard et al., 2011). Overall, these findings indicate a great genetic diversity in rust fungi that possess a complex life cycle with a sexual reproduction stage achieved on an alternate host (Duplessis et al., 2014b).

Because of the high TE content and the large size of the poplar rust genome, together with putatively large differences between isolates (as previously reported in *P. striiformis* f. sp. *tritici*), we did not expect *de novo* assembly to be optimal for analysis of the 14 isolates sequenced for the first time in this study. Indeed, *de novo* assembly generated large numbers of scaffolds (i.e., >30,000, data not shown). Instead, Illumina reads from the 14 isolates were directly mapped onto the 98AG31 reference genome for variants detection, similar as in Zheng et al. (2013). In *M. larici-populina*, an average of 148,532 SNPs per isolate were uncovered, which is slightly higher than in *P. striiformis* f. sp. *tritici* according to Zheng et al. (2013). The proportions of heterozygous SNPs in the two isolates 08KE26 and 09BS12 (35 and 37%, respectively), might reflect their assignment to an asexual group as described by the poplar rust population genetic analysis of Xhaard et al. (2011). A much higher proportion of heterozygous SNPs were found between *P. striiformis* f. sp. *tritici* isolates: 82–84% in Zheng et al. (2013) and 87–99% in Cantu et al. (2013). The observed differences between the two studies may reflect differences in the sequencing and analysis process used (Duplessis et al., 2014b), or could be related to a different reproduction regime, as *P. striiformis* f. sp. *tritici* is mostly asexual which fosters individual heterozygosity (Balloux et al., 2003; Halkett et al., 2005). It would be interesting to compare this with the genetic diversity in rust fungi such as *P. pachyrhizi* or *H. vastatrix* with no known sexual reproduction to date (Rodrigues et al., 1975; Goellner et al., 2010). InDel variants were also inspected and ranged from 4571 to 8077 in the 15 *M. larici-populina* isolates, which is slightly larger than in *P. striiformis* f. sp. *tritici* where 1863 on average were reported (Zheng et al., 2013), but smaller than in the yeast *Saccharomyces* sp. (Liti et al., 2009). A substantial level of polymorphism is noted in *M. larici-populina* at the intraspecific level (∼6 SNPs/Kb), which is in close accordance with those reported in the shiitake mushroom *Lentinula edodes* (4.6 SNPs/Kb, Au et al., 2013) or in the wheat stripe rust fungus

P. striiformis f. sp. *tritici* (Cantu et al., 2013). It is slightly larger than in plant pathogenic ascomycetes such as *Pyrenophora tritici-repentis* (1.9 SNPs/Kb, Manning et al., 2013), *Blumeria graminis* (less than 2 SNPs/Kb; Hacquard et al., 2013b; Wicker et al., 2013) or *Leptosphaeria maculans* (0.5 SNPs/Kb; Zander et al., 2013) but much lower than in the yeast *S. cerevisiae* (59.8 SNPs/Kb; Liti et al., 2009) or in the plant pathogen *Rhizoctonia solani* (∼15 SNPs/Kb; Hane et al., 2014). The observed differences in the levels of polymorphism could reflect evolutionary trends related to the lifestyle of these fungi. Rust fungi, exhibit a remarkable level of polymorphism, providing ground for detection of loci that may underlie the co-evolution with their associated hosts and/or their unique life cycle, which is marked by the formation of five spore types and infection of two alternate hosts (Duplessis et al., 2014b).

PATTERNS OF GENETIC VARIATIONS IN POPLAR RUST GENES UNCOVER CANDIDATE PATHOGENESIS-RELATED GENES

A large part of the variants was identified in coding sequences, similar to *P. striiformis* f. sp. *tritici* (Cantu et al., 2013; Zheng et al., 2013). In total, 89 and 74% of the 16,399 *M. larici-populina* genes were marked by at least one SNP, or one non-synonymous SNP, respectively, in one of the isolates. Such valuable information provides ground for detailed analysis of the functions that may be under selection in the poplar rust genome, particularly those evolving under the pressure of the host plant. P_N/P_S values can be informative to the detection of positive selection and the understanding of how fungi adapt to their environment (Stukenbrock and Bataillon, 2012). We examined the genes showing a $P_N/P_S > 1$ with a particular focus on candidate effectors. Strikingly, whereas other comparative genomic studies have revealed candidate effector genes under positive selection (Cooke et al., 2012; Wicker et al., 2013), we did not detect any enrichment in SP genes exhibiting a high P_N/P_S compared to all genes in the poplar rust genome. However, 68 SP genes in total showed a $P_N/P_S > 1$ and are priority candidates. Other genes falling in this category may be related to pathogenesis-related functions, but no particular enrichment in functional annotation could be detected. However, the missing regions in *M. larici-populina* isolates contain many genes encoding small proteins (i.e., less than 300 amino acids) with no predicted signal peptide. In the obligate biotroph *B. graminis*, selection analysis carried out between formae speciales identified candidate effectors with no predicted signal peptide that share other common evolutionary features with annotated effectors (Wicker et al., 2013). A total of 262 *M. larici-populina* genes encoding small proteins were found with a $P_N/P_S > 1$ (Supporting Table 6). Such small protein encoding genes are also found among *in planta* highly expressed genes of *M. larici-populina* (Duplessis et al., 2011b). Although no unconventional secretory system is known so far in rust fungi, it would be tempting to consider such proteins in future analysis as possible candidate effectors. We therefore examined the genes presenting a large proportion of non-synonymous substitutions in their sequence and detected enrichment in KOG categories related to nuclear structure and function. Interestingly, genomes of rust fungi contain significantly expanded gene families encoding helicases that may play an important role in DNA repair and maintenance, and nucleic acid and zinc-finger proteins corresponding to putative transcription factors (Duplessis et al., 2011a; Zheng et al., 2013). DNA repair systems can have a dramatic impact on genomic diversity (Seidl and Thomma, 2014) and their possible role in the evolution of the poplar rust genome is still to be determined.

In our study, variations occurring in upstream sequence of genes were also inspected, on the grounds that they may relate to regulation of expression. In total, 16% of the genes had more than 10 SNPs in their 1 Kb upstream region. Detailed transcriptome-driven analyses of conserved cis-acting regulatory elements in *P. infestans* have revealed motifs underlying specific expression of pathogenesis-related genes (Seidl et al., 2012; Roy et al., 2013a,b). The transcriptome analysis of poplar leaf infection by *M. larici-populina* has shown conserved patterns of coordinated expression of several sets of SSP genes along a time course experiment (Duplessis et al., 2011a). Several other transcriptomic studies have confirmed this trend for SP genes in rust fungi (Fernandez et al., 2012; Cantu et al., 2013; Tremblay et al., 2013; Bruce et al., 2014; Duplessis et al., 2014b). A better knowledge of cis-acting regulatory elements in the genome of *M. larici-populina* is needed to further explore the impact of mutations in upstream gene regions. Other molecular mechanisms may control regulation of expression profiles, as recently exemplified in the oilseed rape ascomycete pathogen *L. maculans* (Soyer et al., 2014). Particularly of note, a significant enrichment in genes falling in the chromatin structure and dynamics KOG category was found in genes accumulating non-synonymous SNPs, and it remains to be explored whether such a control of the chromatin structure could relate to the control of gene expression in rust fungi.

A major goal of the present study was to uncover the presence of polymorphic effectors within a set of predefined candidates that may reflect specific adaptation to the host plant in the classical scheme of the plant-pathogen arms race. A similar approach conducted in *P. striiformis* f. sp. *tritici* identified five polymorphic candidate effectors by comparing two isolates presenting distinct pathotypes (Cantu et al., 2013). Another study identified such possible avirulence genes among secreted protein transcripts showing patterns of non-synonymous mutations between different *Puccinia triticina* isolates (Bruce et al., 2014). In the panel of *M. larici-populina* isolates, virulences 1, 6, and 8 presented correlations with the presence of non-synonymous SNPs in one, one and four genes of virulent isolates compared to avirulent isolates, respectively. Such genes could be candidates underlying virulences 1, 6, and 8. No such correlation was observed for the other virulences carried by the poplar rust isolates, indicating that other events than non-synonymous substitutions in coding sequences may explain their emergence.

Sequence polymorphism has been reported in several avirulence genes of the flax rust *M. lini* (Catanzariti et al., 2006; Dodds et al., 2006; Barrett et al., 2009; Van der Merwe et al., 2009; Ravensdale et al., 2011). Homologs of flax rust avirulence genes retrieved in the *M. larici-populina* genome did not exhibit high P_N/P_S or excess of non-synonymous substitutions in the 15 isolates, except in a very few cases. Interestingly, non-synonymous substitutions observed in the CPG5464 family homologous to *M. lini* AvrP4 did not match sites previously shown under selection in *M. lini* at the intraspecific level (Barrett

et al., 2009), in Melampsoraceae at the interspecific level (Van der Merwe et al., 2009) or between members of the paralogous gene cluster CPG5464 of *M. larici-populina* (Hacquard et al., 2012). Members of this gene family are rather conserved within the Melampsoraceae, suggesting that AvrP4/CPG5464 could play an important role as an effector during the interaction with the relative host plants. A high diversity is observed at both the intraspecific and interspecific level highlighting the probable interplay with the different host plants, but to date, such an interaction in a gene-for-gene manner has only been demonstrated for the flax rust fungus (Ravensdale et al., 2011). At least one *M. larici-populina* homolog of the *M. lini AvrM* gene shows a high level of polymorphic sites (e.g., in isolate 98AR1, 30 SNPs of which 15 are non-synonymous), similar to those reported in *M. lini* (Catanzariti et al., 2006; Ravensdale et al., 2011). Some of these mutations are particularly important for the direct interaction with the corresponding *M* resistance gene in flax (Catanzariti et al., 2010; Ve et al., 2013). It will be particularly interesting to further study the potential role of AvrM homologs in the poplar-poplar rust fungus interaction.

FUTURE STEPS IN POPLAR RUST GENOMICS

Genomics is a powerful approach to identify pathogenesis-related candidates, as the present study illustrates. From the perspective of population biology, it is well-known that structure and demography can affect all loci equally. To identify loci under selection, a population genomics approach is required to take into account demographic history. A population genomics study is ongoing in collaboration with the JGI to identify loci related to virulence 7. As large portions of the genome were missing in different *M. larici-populina* isolates, it might be required to study presence/absence at a larger scale using *de novo* assembled genomes. Many mechanisms can underlie genome evolution (Raffaele and Kamoun, 2012; Seidl and Thomma, 2014) and a better knowledge of the structural rearrangements occurring in the poplar rust genome will help to determine their impact on virulence evolution. In this regard, we have initiated the genome sequencing of an avirulent 7 isolate by combining paired-end and mate-pair Illumina sequencing to compare with the virulent 7 reference genome. Together, these genomic analyses will foster functional studies by pinpointing numerous sites of sequence variation, i.e., positions that may have important implications at the structural level for the function of effectors.

AUTHOR CONTRIBUTIONS

Sébastien Duplessis and Pascal Frey designed research; Antoine Persoons, Sébastien Duplessis, Christine Delaruelle, and Pascal Frey performed research; Antoine Persoons, Sébastien Duplessis, Emmanuelle Morin, Stéphane De Mita, and Thibaut Payen analyzed data; Antoine Persoons and Sébastien Duplessis drafted the manuscript and, Antoine Persoons, Sébastien Duplessis, Pascal Frey, Fabien Halkett, Thibaut Payen, and Stéphane De Mita wrote the paper.

ACKNOWLEDGMENTS

We warmly thank Katherine Hayden for comments on the manuscript. We would like to acknowledge the help of Bénédicte Fabre (INRA Nancy) in the production of poplar plants in greenhouses and of *M. larici-populina* urediniospores. We also thank our colleagues Claude Murat and Francis Martin at INRA Nancy for fruitful discussions during the course of the study. This work was supported by the French National Research Agency through the Laboratory of Excellence ARBRE (ANR-12-LABXARBRE-01), the Young Scientist Grant POPRUST to Sébastien Duplessis (ANR-2010-JCJC-1709-01) and the GANDALF project (ANR-12-ADAP0009) and by the Région Lorraine (Researcher Award to Sébastien Duplessis). Antoine Persoons is supported by a Doctoral Scholarship from the Institut National de la Recherche Agronomique and the Region Lorraine. We thank the Joint Genome Institute for the access to the *M. larici-populina* genome sequence.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00450/abstract

REFERENCES

Alfano, J. R. (2009). Roadmap for future research on plant pathogen effectors. *Mol. Plant Pathol.* 10, 805–813. doi: 10.1111/j.1364-3703.2009.00588.x

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Au, C. H., Cheung, M. K., Wong, M. C., Chu, A. K., Law, P. T., and Kwan, H. S. (2013). Rapid genotyping by low-coverage resequencing to construct genetic linkage maps of fungi: a case study in *Lentinula edodes*. *BMC Res. Notes* 6:307. doi: 10.1186/1756-0500-6-307

Balloux, F., Lehmann, L., and De Meeus, T. (2003). The population genetics of clonal and partially clonal diploids. *Genetics* 164, 1635–1644.

Barrès, B., Halkett, F., Dutech, C., Andrieux, A., Pinon, J., and Frey, P. (2008). Genetic structure of the poplar rust fungus *Melampsora larici-populina*: evidence for isolation by distance in Europe and recent founder effects overseas. *Infect. Genet. Evol.* 8, 577–587. doi: 10.1016/j.meegid.2008.04.005

Barrett, L. G., Thrall, P. H., Dodds, P. N., van der Merwe, M., Linde, C. C., Lawrence, G. J., et al. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol. Biol. Evol.* 26, 2499–2513. doi: 10.1093/molbev/msp166

Bruce, M., Neugebauer, K. A., Joly, D. L., Migeon, P., Cuomo, C. A., Wang, S., et al. (2014). Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat. *Front. Plant Sci.* 4:520. doi: 10.3389/fpls.2013.00520

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Catanzariti, A.-M., Dodds, P. N., Ve, T., Kobe, B., Ellis, J. G., and Staskawicz, B. J. (2010). The AvrM effector from flax rust has a structured C-terminal domain and interacts directly with the M resistance protein. *Mol. Plant Microbe Interact.* 23, 49–57. doi: 10.1094/MPMI-23-1-0049

Cooke, D. E., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J., et al. (2012). Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. *PLoS Pathog.* 8:e1002940. doi: 10.1371/journal.ppat.1002940

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E., (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

De Mita, S., and Siol, M. (2012). EggLib: processing, analysis and simulation tools for population genetics and genomics. *BMC Genet.* 11, 13–27. doi: 10.1186/1471-2156-13-27

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini* AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell*. 16, 755–768. doi: 10.1105/tpc.020040

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Teh, T., Wang, C. I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8888–8893. doi: 10.1073/pnas.0602577103

Dodds, P. N., and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* 11, 539–548. doi: 10.1038/nrg2812

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014b). Advancing knowledge on biology of rust fungi through genomics. *Adv. Bot. Res.* 70, 173–209. doi: 10.1016/B978-0-12-397940-7.00006-9

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011b). *Melampsora larici-populina* transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Duplessis, S., Spanu, P. D., and Schirawski, J. (2014a). "Biotrophic fungi (powdery mildews, Rusts and Smuts)," in *Ecological Genomics of the Fungi. Plant-Interacting Fungi Section*, ed F. Martin (Oxford: Wiley-Blackwell), 149–168.

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A. S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* 9, 275–296.

Fraser, J. A., Hsueh, Y. P., Findley, K. M., and Heiman, J. (2007). "Evolution of the mating type locus: the basidiomycetes," in *Sex in Fungi: Molecular Determination and Evolutionary Implications,* eds J. Heitman, J. W. Kronstad, J. W. Taylor, and L. A. Casselton (Washington, DC: ASM Press), 19–34.

Gérard, P. R., Husson, C., Pinon, J., and Frey, P. (2006). Comparison of genetic and virulence diversity of *Melampsora larici-populina* populations on wild and cultivated poplar and influence of the alternate host. *Phytopathology* 96, 1027–1036. doi: 10.1094/PHYTO-96-1027

Giraldo, M. C., and Valent, B. (2013). Filamentous plant pathogen effectors in action. *Nat. Rev. Microbiol.* 11, 800–814. doi: 10.1038/nrmicro3119

Goellner, K., Loehrer, M., Langenbach, C., Conrath, U. W. E., Koch, E., and Schaffrath, U. (2010). *Phakopsora pachyrhizi*, the causal agent of Asian soybean rust. *Mol. Plant Pathol.* 11, 169–177. doi: 10.1111/j.1364-3703.2009.00589.x

Hacquard, S., Delaruelle, C., Frey, P., Tisserant, E., Kohler, A., and Duplessis, S. (2013a). Transcriptome analysis of poplar rust telia reveals overwintering adaptation and tightly coordinated karyogamy and meiosis processes. *Front. Plant Sci.* 4:456. doi: 10.3389/fpls.2013.00456

Hacquard, S., Joly, D. L., Lin, Y. C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A. comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (Poplar Leaf Rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hacquard, S., Kracher, B., Maekawa, T., Vernaldi, S., Schulze-Lefert, P., and Ver Loren van Themaat, E. (2013b). Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. *Proc. Natl. Acad. Sci. U.S.A.* 110, E2219–E2228. doi: 10.1073/pnas.1306807110

Hacquard, S., Petre, B., Frey, P., Hecker, A., Rouhier, N., and Duplessis, S. (2011). The poplar-poplar rust interaction: insights from genomics and transcriptomics. *J. Pathog.* 2011, 716041. doi: 10.4061/2011/716041

Halkett, F., Simon, J.-C., and Balloux, F. (2005). Tackling the population genetics of clonal and partially clonal organisms. *Trends Ecol. Evol.* 20, 194–201. doi: 10.1016/j.tree.2005.01.001

Hane, J. K., Anderson, J. P., Williams, A. H., Sperschneider, J., and Singh, K. B. (2014). Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. *PLoS Genet.* 10:e1004281. doi: 10.1371/journal.pgen.1004281

Joly, D. L., Feau, N., Tanguay, P., and Hamelin, R. C. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11:422. doi: 10.1186/1471-2164-11-422

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Kemen, E., Kemen, A., Ehlers, A., Voegele, R., and Mendgen, K. (2013). A novel structural effector from rust fungi is capable of fibril formation. *Plant J.* 75, 767–780. doi: 10.1111/tpj.12237

Liti, G., Carter, D. M., Moses, A. M., Warringer, J., Parts, L., James, S. A., et al. (2009). Population genomics of domestic and wild yeasts. *Nature* 458, 337–341. doi: 10.1038/nature07743

Lowe, R. G., and Howlett, B. J. (2012). Indifferent, affectionate, or deceitful: lifestyles and secretomes of fungi. *PLoS Pathog.* 8:e1002515. doi: 10.1371/journal.ppat.1002515

Manning, V. A., Pandelova, I., Dhillon, B., Wilhelm, L. J., Goodwin, S. B., Berlin, A. M., et al. (2013). Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3* 3, 41–63. doi: 10.1534/g3.112.004044

Nemri, A., Saunders, D. G. O., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Pinon, J., and Frey, P. (2005). "Interactions between poplar clones and *Melampsora* populations and their implications for breeding for durable resistance," in *Rust Diseases of Willow and Poplar,* eds M. H. Pei and A. R. McCracken (Wallingford: CAB International), 139–154.

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins-a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832

Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10, 417–430. doi: 10.1038/nrmicro2790

Ravensdale, M., Nemri, A., Thrall, P. H., Ellis, J. G., and Dodds, P. N. (2011). Co-evolutionary interactions between host resistance and pathogen effector genes in flax rust disease. *Mol. Plant Pathol.* 12, 93–102. doi: 10.1111/j.1364-3703.2010.00657

Rodrigues, C. J. Jr., Bettencourt, A. J., and Rijo, L. (1975). Races of the patho- gen and resistance to coffee rust. *Annu. Rev. Phytopathol.* 13, 49–70.

Roy, S., Kagda, M., and Judelson, H. S. (2013a). Genome-wide prediction and functional validation of promoter motifs regulating gene expression in spore and infection stages of *Phytophthora infestans*. *PLoS Pathog.* 9:e1003182. doi: 10.1371/journal.ppat.1003182

Roy, S., Poidevin, L., Jiang, T., and Judelson, H. S. (2013b). Novel core promoter elements in the oomycete pathogen *Phytophthora infestans* and their influence on expression detected by genome-wide analysis. *BMC Genomics* 14:106. doi: 10.1186/1471-2164-14-106

Saunders, D. G. O., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:e29847. doi: 10.1371/journal.pone.0029847

Seidl, M. F., and Thomma, B. P. H. J. (2014). Sex or no sex: evolutionary adaptation occurs regardless. *Bioessays* 36, 335–345. doi: 10.1002/bies.201300155

Seidl, M. F., Wang, R.-P., Van den Ackerveken, G., Govers, F., and Snel, B. (2012). Bioinformatic inference of specific and general transcription factor binding sites in the plant pathogen *Phytophthora infestans*. *PLoS ONE* 7:e51295. doi: 10.1371/journal.pone.0051295

Soyer, J. L., El Ghalid, M., Glaser, N., Ollivier, B., Linglin, J., Grandaubert, J., et al. (2014). Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*. *PLoS Genet.* 10:e1004227. doi: 10.1371/journal.pgen.1004227

Spanu, P. D. (2012). The genomics of obligate (and nonobligate) biotrophs. *Annu. Rev. Phytopathol*. 50, 91–109. doi: 10.1146/annurev-phyto-081211-173024

Steenackers, J., Steenackers, M., Steenackers, V., and Stevens, M. (1996). Poplar diseases, consequences on growth and wood quality. *Biomass Bioenerg.* 10, 267–274.

Stergiopoulos, I., Cordovez, V., Okmen, B., Beenen, H. G., Kema, G. H., and de Wit, P. J. (2013). Positive selection and intragenic recombination contribute to high allelic diversity in effector genes of *Mycosphaerella fijiensis*, causal agent of the black leaf streak disease of banana. *Mol. Plant Pathol*. 15, 447–460. doi: 10.1111/mpp.12104

Stergiopoulos, I., and de Wit, P. J. (2009). Fungal effector proteins. *Annu. Rev. Phytopathol.* 47, 233–263. doi: 10.1146/annurev.phyto.112408.132637

Stone, C. L., Buitrago, M. L., Boore, J. L., and Frederick, R. D. (2010). Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *Phakopsora pachyrhizi* and *P. meibomiae. Mycologia* 102, 887–897. doi: 10.3852/09-198

Stukenbrock, E. H. (2013). Evolution, selection and isolation: a genomic view of speciation in fungal plant pathogens. *New Phytol*. 199, 895–907. doi: 10.1111/nph.12374

Stukenbrock, E. H., and Bataillon, T. (2012). A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. *PLoS Pathog*. 8:e1002893. doi: 10.1371/journal.ppat.1002893

Stukenbrock, E. H., and McDonald, B. A. (2009). Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Mol. Plant Microbe Interact*. 22, 371–380. doi: 10.1094/MPMI-22-4-0371

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinform.* 4:41. doi: 10.1186/1471-2105-4-41

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr. Protoc. Bioinform. Chapter* 2:Unit 2.3. doi: 10.1002/0471250953.bi0203s00

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2013). Analysis of *Phakopsora pachyrhizi* transcript abundance in critical pathways at four time-points during infection of a susceptible soybean cultivar using deep sequencing. *BMC Genomics* 14:614. doi: 10.1186/1471-2164-14-614

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

Tyler, B. M., and Rouxel, T. (2012). "Effectors of fungi and oomycetes: their virulence and avirulence functions and translocation from pathogen to host cells," in *Molecular Plant Immunity,* ed G. Sessa (Oxford: Wiley-Blackwell), 123–167. doi: 10.1002/9781118481431.ch7

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M. A., Ellis, J. G., and Dodds, P. N. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol. Plant Microbe Interact*. 27, 255–264. doi: 10.1094/MPMI-07-13-0187-FI

Van der Merwe, M. M., Kinnear, M. W., Barrett, L. G., Dodds, P. N., Ericson, L., Thrall, P. H., et al. (2009). Positive selection in AvrP4 avirulence gene homologues across the genus *Melampsora*. *Proc. Biol. Sci*. 276, 2913–2922. doi: 10.1098/rspb.2009.0328

Ve, T., Williams, S. J., Catanzariti, A.-M., Rafiqi, M., Rahman, M., Ellis, J. G., et al. (2013). Structures of the flax-rust effector AvrM reveal insights into the molecular basis of plant-cell entry and effector-triggered immunity. *Proc. Natl. Acad. Sci. U.S.A*. 11, 17594–17599. doi: 10.1073/pnas.1307614110

Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J. P., Shatalina, M., Roffler, S., et al. (2013). The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nat. Genet*. 45, 1092–1096. doi: 10.1038/ng.2704

Win, J., Chaparro-Garcia, A., Belhaj, K., Saunders, D. G., Yoshida, K., Dong, S., et al. (2012). Effector biology of plant-associated organisms: concepts and perspectives. *Cold Spring Harb. Symp. Quant. Biol*. 77, 235–247. doi: 10.1101/sqb.2012.77.015933

Xhaard, C., Fabre, B., Andrieux, A., Gladieux, P., Barrès, B., Frey, P., et al. (2011). The genetic structure of the plant pathogenic fungus *Melampsora larici-populina* on its wild host is extensively impacted by host domestication. *Mol. Ecol.* 20, 2739–2755. doi: 10.1111/j.1365-294X.2011.05138

Yoshida, K., Saitoh, H., Fujisawa, S., Kanzaki, H., Matsumura, H., Yoshida, K., et al. (2009). Association genetics reveals three novel avirulence genes from the rice blast fungal pathogen *Magnaporthe oryzae*. *Plant Cell* 21, 1573–1591. doi: 10.1105/tpc.109.066324

Zander, M., Patel, D. A., Van de Wouw, A., Lai, K., Lorenc, M. T., Campbell, E., et al. (2013). Identifying genetic diversity of avirulence genes in *Leptosphaeria maculans* using whole genome sequencing. *Funct. Integr. Genomics* 13, 295–308. doi: 10.1007/s10142-013-0324-5

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun*. 4, 2673. doi: 10.1038/ncomms3673

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 20 August 2014; published online: 15 September 2014.

Citation: Persoons A, Morin E, Delaruelle C, Payen T, Halkett F, Frey P, De Mita S and Duplessis S (2014) Patterns of genomic variation in the poplar rust fungus Melampsora larici-populina identify pathogenesis-related factors. Front. Plant Sci. ***5****:450. doi: 10.3389/fpls.2014.00450*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Persoons, Morin, Delaruelle, Payen, Halkett, Frey, De Mita and Duplessis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 08 January 2015
doi: 10.3389/fpls.2014.00759

Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes

Narayana M. Upadhyaya[1]*, Diana P. Garnica[2], Haydar Karaoglu[3], Jana Sperschneider[1], Adnane Nemri[1], Bo Xu[1], Rohit Mago[1], Christina A. Cuomo[4], John P. Rathjen[2], Robert F. Park[3], Jeffrey G. Ellis[1] and Peter N. Dodds[1]*

[1] *Agriculture Flagship, Commonwealth Scientific and Industrial Research Organization, Canberra, ACT, Australia*
[2] *Research School of Biology, Australian National University, Canberra, ACT, Australia*
[3] *Plant Breeding Institute, Faculty of Agriculture and Environment, The University of Sydney, Narellan, NSW, Australia*
[4] *Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA*

Edited by:
David L. Joly, Université de Moncton, Canada

Reviewed by:
Nils Rostoks, University of Latvia, Latvia
Nicolas Feau, University of British Columbia, Canada

****Correspondence:***
Peter N. Dodds and Narayana M. Upadhyaya, Agriculture Flagship, Commonwealth Scientific and Industrial Research Organization, Cnr. Barry Drive and Clunies Ross Street, Black Mountain, Acton, Canberra, ACT 2601, Australia
e-mail: peter.dodds@csiro.au; Narayana.upadhyaya@csiro.au

The wheat stem rust fungus *Puccinia graminis* f. sp. *tritici* (*Pgt*) is one of the most destructive pathogens of wheat. In this study, a draft genome was built for a founder Australian *Pgt* isolate of pathotype (pt.) 21-0 (collected in 1954) by next generation DNA sequencing. A combination of reference-based assembly using the genome of the previously sequenced American *Pgt* isolate CDL 75-36-700-3 (p7a) and *de novo* assembly were performed resulting in a 92 Mbp reference genome for *Pgt* isolate 21-0. Approximately 13 Mbp of *de novo* assembled sequence in this genome is not present in the p7a reference assembly. This novel sequence is not specific to 21-0 as it is also present in three other *Pgt* rust isolates of independent origin. The new reference genome was subsequently used to build a pan-genome based on five Australian *Pgt* isolates. Transcriptomes from germinated urediniospores and haustoria were separately assembled for pt. 21-0 and comparison of gene expression profiles showed differential expression in ∼10% of the genes each in germinated spores and haustoria. A total of 1,924 secreted proteins were predicted from the 21-0 transcriptome, of which 520 were classified as haustorial secreted proteins (HSPs). Comparison of 21-0 with two presumed clonal field derivatives of this lineage (collected in 1982 and 1984) that had evolved virulence on four additional resistance genes (*Sr5*, *Sr11*, *Sr27*, *SrSatu*) identified mutations in 25 HSP effector candidates. Some of these mutations could explain their novel virulence phenotypes.

Keywords: haustoria, avirulence, resistance, secreted proteins, effectors

INTRODUCTION

Wheat stem rust, caused by *Puccinia graminis* f. sp. *tritici* (*Pgt*), is one of the most destructive diseases of wheat, barley and triticale (Leonard and Szabo, 2005; Park, 2007). In order to infect plants and cause disease, pathogens such as *Pgt* need first to overcome or evade the natural defenses of the plant. These defenses include preformed barriers, such as the waxy cuticle and inducible responses triggered by the plant innate immunity system (Jones and Takemoto, 2004). The first layer of the immune system involves recognition of pathogen associated molecular patterns (PAMPs) such as chitin or flagellin (Jones and Dangl, 2006; Dodds and Rathjen, 2010). Recognition of these factors by cell surface receptors leads to PAMP-triggered immunity (PTI), which is effective in preventing infection by non-adapted pathogens. Bacterial pathogens of plants overcome these defenses through the use of effector proteins that are delivered into host cells by a type III secretion system (Zhou and Chai, 2008), and biotrophic fungi and oomycetes also deliver effectors into host cells during infection (Giraldo and Valent, 2013). However, many of these effectors are recognized by a second layer of the plant defense system that involves intracellular receptors that are the products of the classically defined resistance (*R*) genes of the gene-for-gene system, first elucidated in the flax/flax rust pathosystem (Flor, 1971). In this context pathogen effectors are known as Avirulence (Avr) proteins and their recognition leads to rapid activation of a localized cell death termed the hypersensitive response, which is thought to limit the spread of the pathogen from the infection site (Chisholm et al., 2006). This second layer of defense has been termed effector-triggered immunity (ETI), and involves direct or indirect recognition of pathogen effector proteins by plant R proteins. Pathogens may evade this recognition by mutation of the corresponding *Avr* genes.

Many biotrophic fungi and oomycetes share a common infection process that involves the formation of haustoria, which invaginate and engage in close physical contact with the plasma membrane of host cells (Koeck et al., 2011). Haustoria play a role in nutrient acquisition and metabolism (Hahn and Mendgen, 2001; Voegele and Mendgen, 2003) and there is evidence to suggest that these structures also play a crucial role in the delivery of virulence effectors that alter defense responses and promote infection (Kemen et al., 2005; Whisson et al., 2007; Rafiqi et al., 2010). For example, all Avr genes that have been identified in the flax

rust fungus (*Melampsora lini*) encode small secreted proteins that are expressed in haustoria and are recognized inside host cells by nucleotide binding leucine-rich repeat (NB-LRR) receptors (Dodds et al., 2004; Catanzariti et al., 2006; Barrett et al., 2009; Rafiqi et al., 2010). Analyses of transcript sets from isolated haustoria of *M. lini* (Nemri et al., 2014), the stripe rust pathogen *Puccinia striiformis* f. sp. *tritici* (*Pst*; Cantu et al., 2013; Garnica et al., 2013), common bean rust *Uromyces appendiculatus* (Link et al., 2013) and soybean rust *Phakopsora pachyrhizi* (Link et al., 2013) have predicted large numbers of secreted proteins expressed in these cells, indicating that they may deliver a large set of effectors to infected host cells. In the case of the wheat stem rust pathogen, whole genome shotgun sequencing of the American *Pgt* isolate CDL 75-36-700-3 (referred to as p7a) yielded an 81.5 Mbp genome sequence (out of an estimated 88.6 Mbp scaffold assembly) predicted to contain 15,979 protein coding genes (Duplessis et al., 2011). Of these about 10% are predicted to be secreted proteins, but their expression in haustoria has not been determined.

On the host side, there are more than 50 race-specific stem rust resistance *(Sr)* genes described in wheat, either derived from this species or introgressed from its close relatives (McIntosh et al., 1995), many of which have been deployed in modern wheat cultivars to control this disease. However, resistance breakdown has occurred frequently due to mutations in existing local isolates and the emergence or migration of new isolates, such as the highly virulent *Pgt* race Ug99 (Stokstad, 2007). In some areas where the alternate host of *Pgt* (*Berberis vulgaris*) exists, sexual recombination can give rise to new virulence phenotypes. Successful control of stem rust in wheat requires constant identification of new *Sr* genes, stacking of several different *Sr* genes in cultivars, and cultural efforts to keep inoculum levels low within each geographical zone of cultivation. The two recently cloned stem rust resistance genes *Sr33* (Periyannan et al., 2013) and *Sr35* (Saintenac et al., 2013) encode classical NB-LRR type intracellular immune receptors, suggesting that, as in *M. lini*, the corresponding *Pgt* Avr proteins are likely to be effectors delivered into host cells.

In Australia, there have been at least four independent incursions of exotic stem rust isolates documented since 1925. After arrival, the four founding isolates have each evolved mainly asexually in the field through presumed stepwise mutations that overcome various *Sr* genes deployed in wheat, leading to four clonal lineages comprising many derivative mutant pathotypes (pt.) differing for virulence on various host resistance genes (Park, 2007). In this study, we have used isolates of the four founder Australian *Pgt* pathotypes of these lineages and two mutant-derivative isolates of one lineage (pt. 21-0) with additional virulence, to construct the *Pgt* pan-genome, transcriptome and secretome. Comparisons of pt. 21-0 with the two presumed clonal field mutant derivatives with virulence to four additional resistance genes (*Sr5*, *Sr11*, *Sr27*, *SrSatu*) identified alterations in 25 haustorially-expressed effector candidates, which could include the mutations that give rise to their novel virulence phenotypes.

MATERIALS AND METHODS

Puccinia graminis f. sp. *tritici* (*Pgt*) ISOLATES

Individual isolates of four Australian *Pgt* (**Table 1**) pathotypes, 21-0 (Univeristy of Sydney accession number 54129), 126-5,6,7,11 (accession number 334), 194-1,2,3,5,6 (accession number 691042), and 326-1,2,3,5,6 (accession number 690822) were used in this study. Given that each is a specific isolate of a pathotype, and for simplicity, these are referred to as isolates 21-0, 126, 194, and 326 hereafter. Each isolate represents the original detection of four separate incursions of *Pgt* into Australia isolated from the field starting from mid 1920s that have been maintained as viable cultures in liquid nitrogen at Plant Breeding Institute, Cobbitty, NSW, Australia (Park, 2007). To ensure isolate purity, a single pustule from a low density infection was isolated from each isolate and propagated on wheat cultivar Morocco in isolation prior to DNA preparation. The identity and purity of each isolate was checked by pathogenicity tests with a set of host differentials. Two additional isolates, pathotypes 34-2,12 (accession number 82246) and 34-2,12,13 (accession number 84552; referred to as isolates 34M1 and 34M2 hereafter), were also purified from single pustules by

Table 1 | Australian *Pgt* isolates used in this study and their compatibility (*Avr*/*avr* profiles) with different host *R* genes.

Pgt isolate (short name)	Incursion/ isolation year	Virulent	Avirulent	Mesothetic
126-5,6,7,11 (126)	1925	*Sr5, Sr7b,Sr8a, Sr8b, Sr15, Sr17*	*Sr6, Sr9b, Sr9e, Sr11, Sr21 Sr27, Sr30, Sr36, SrAgi, SrEM, SrSatu*	*Sr9g*
21-0	1954	*Sr7b, Sr9g*	*Sr5, Sr6, Sr8a, Sr8b,Sr9b, Sr9e, Sr11, Sr15, Sr17, Sr21 Sr27, Sr30, Sr36, SrAgi, SrEM, SrSatu*	
34-2,12 (34M1)	1982	*Sr5, Sr7b, Sr9g, Sr11, Sr27*	*Sr6, Sr8a, Sr8b, Sr9b, Sr9e, Sr15, Sr17, Sr21, Sr30, Sr36, SrAgi, SrEM, SrSatu*	
34-2,12,13 (34M2)	1984	*Sr5, Sr7b, Sr9g, Sr11, Sr27, SrSatu*	*Sr6, Sr8a, Sr8b, Sr9b, Sr9e, Sr15, Sr17, Sr21, Sr30, Sr36, SrAgi, SrEM*	
326-1,2,3,5,6 (326)	1969	*Sr6, Sr8a, Sr9b, Sr11, Sr17*	*Sr5, Sr7b, Sr8b, Sr9g, Sr9e, Sr15, Sr27, Sr30, Sr36, SrAgi, SrEM, SrSatu*	*Sr21*
194-1,2,3,5,6 (194)	1969	*Sr6, Sr7b, Sr8a, Sr9b, Sr11, Sr17*	*Sr5, Sr8b, Sr9g, Sr9e, Sr15, Sr21, Sr27, Sr30, Sr36, SrAgi, SrEM, SrSatu*	

growth on 'Coorong' (*Sr27*) and 'Satu' (*SrSatu*) triticale, respectively, and their identities and purity confirmed by pathogenicity analysis. Isolates 34M1 and 34M2 were collected from the field in 1982 and 1984, respectively. They are considered to be mutational derivatives of pt. 21-0, with added virulence for *Sr5*, *Sr11,* and *Sr27* (34M1) and *Sr5*, *Sr11*, *Sr27,* and *SrSatu* (34M2; Zwer et al., 1992). Both isolates were found to have SSR genotypes identical to isolate 21-0 when tested with 8 SSR markers (Zhang, 2008). These studies also demonstrated that isolates 126, 194, and 326 differed from each other, and from the 21-0/34M1/34M2 clade (Zhang, 2008).

For rust infection, host plants were grown at high density (∼25 seeds per 12 cm pot with compost as growth media) to the two leaf stage (∼7 days) in a growth cabinet set at 18–25°C temperature and 16 h light. Spores (–80°C stock) were first thawed and heated to 42°C for 3 min, mixed with talcum powder and dusted over the plants. Pots were placed in a moist chamber for 24 h and then transferred back to the growth cabinet. For RNA isolation, infected plant leaves with high density pustules (1 or 2 days before sporulation) were harvested, snap frozen and stored at –80°C. For DNA isolation, mature spores were collected, dried and stored at –80°C.

DNA ISOLATION FROM *Pgt* UREDINIOSPORES AND SEQUENCING

DNA was extracted from urediniospores by a CTAB extraction method (Rogers et al., 1989) with some modifications, including the use of 0.5 mm glass beads instead of fine sand and dry beating (2 × 1 min) at full speed on a dental amalgamator instead of grinding in liquid nitrogen. Extraction was carried out in several batches each with ∼50 mg of dry spores and equal volume of 0.5 mm glass beads to accumulate sufficient quantities of DNA from different isolates. After CTAB extraction, samples were treated with DNase-free RNAase, extracted with phenol/chloroform/isoamyl alcohol (25:24:1) and purified using Qiagen Genomic tips (cat No 10233, Qiagen). DNA quality was assessed using the Bioanalyzer 2100 (Agilent Technologies). Each 50 mg batch of spores yielded ∼20 ug of crude DNA, but the recovery from the Qiagen Genomic Tips was usually very low (∼15–20%) and so several batches were needed to amass sufficient genomic DNA for sequencing.

Pgt isolate 21-0 genomic DNA was sequenced by Roche GS FLX 454 technology at the Australian Genome Research Facility Ltd (AGRF – Australia). A 454 sequencing library was prepared from 5 ug of DNA using the GS General Library preparation kit (Roche Diagnostics). The library fragment size range was 300–500 bp. This library was processed using the GS emPCR and GS FLX LR70 Sequencing kits (Roche Diagnostics) and sequenced in the GS FLX machine. The sequence (fasta format.fna) and the quality score (.qual) outputs were used for further analysis as detailed later in the section.

DNA from urediniospores of isolates *Pgt* 21-0, 126, 194, and 326 were also sequenced using the Illumina GAII platform at the Broad Institute (75 bp paired-end reads). Image analysis and base calling (including quality scoring) were performed using Illumina's Pipeline Analysis Software v1.4 or later. Genomic DNA from mutant isolates 34M1 and 34M2 was sequenced on the Illumina HiSeq platform at AGRF. Libraries were prepared with Agencourt SPRIworks System1 (Beckman Coulter Genomics) using Illumina paired-end library adaptors. Fragment sizes in the library ranged 248–578 bp (including adaptors). Library clusters were generated with the automated cBot system using the Illumina TruSeq PE Cluster Synthesis v2.0 kit and sequenced (100bp paired-end reads) in HiSeq2000 using Illumina TruSeq v2.0 kits. Image intensities and quality scored base calls were performed by the built in HiSeq Control Software and fed into further analysis pipeline as detailed later in the section. Raw sequence reads generated and used in this study are being submitted to NCBI and will be associated with BioProject PRJNA253722[1].

HAUSTORIAL ISOLATION

Twenty grams of infected wheat leaves (isolate 21-0, 10-days post-infection) were sequentially washed with chilled tap water, 2% bleach, water, 70% ethanol, and Milli-Q purified water. Initial stages of haustorial isolation were performed as described previously (Catanzariti et al., 2011) using a final 20-μm pore size nylon mesh to remove the bulk of the plant cell material. Further processing was performed by Percoll gradient fractionation as described previously (Garnica and Rathjen, 2014). Briefly, the filtrate was centrifuged at 1080 *g* for 15 min and the resulting pellet was resuspended in 80 ml of suspension buffer (Percoll 30%, 0.2 M sucrose, 20 mM MOPS pH 7.2). The suspension was divided into five tubes and then centrifuged at 25,000 *g* for 30 min. The top 10 ml of each tube was recovered, diluted 10 times with isolation buffer (0.2 M Sucrose, 20 mM MOPS pH 7.2) and centrifuged at 1,080 *g* for 15 min. The pellets were resuspended in suspension buffer with Percoll 25% and taken to a second round of isolation. The final pellet was frozen in liquid nitrogen and stored at –80°C prior to RNA isolation.

RNA ISOLATION AND SEQUENCING

RNA was isolated from purified haustoria and spores germinated for 15 h on sterile distilled water (16°C in the dark). Samples were ground to a fine powder in liquid nitrogen and total RNA isolated using the RNeasy Plant Mini Kit (Qiagen). Extracted RNA was treated with RNase-free DNAse (Promega) and repurified using the RNeasy Plant Mini Kit columns. RNA quality was assessed with the Bioanalyzer 2100. About 10 μg of total RNA was processed with the mRNA-Seq Sample Preparation kit from Illumina to produce the sequencing libraries. Quality and quantity controls were run on an Agilent 2100 Bioanalyzer using a DNA 1000 chip kit and each library was diluted and used for sequencing with an Illumina Genome Analyser GX II platform (100 bp paired-end reads).

GENOME AND TRANSCRIPTOME ASSEMBLY AND ANALYSES

A consensus reference genome was built using various modules available in CLC Genomics Workbench (Version 4.5 or later, CLC bio Qiagen, Prismet) and the analysis workflow as depicted in **Figure 1A**. Combined 454 and Illumina sequencing reads from isolate 21-0 DNA were first pre-processed (quality trim 0.01, adaptor trim, minimum length 40 nt, maximum ambiguity 2 nt, terminal trim 1 nt). Read mapping was performed

[1]http://www.ncbi.nlm.nih.gov/bioproject/253722

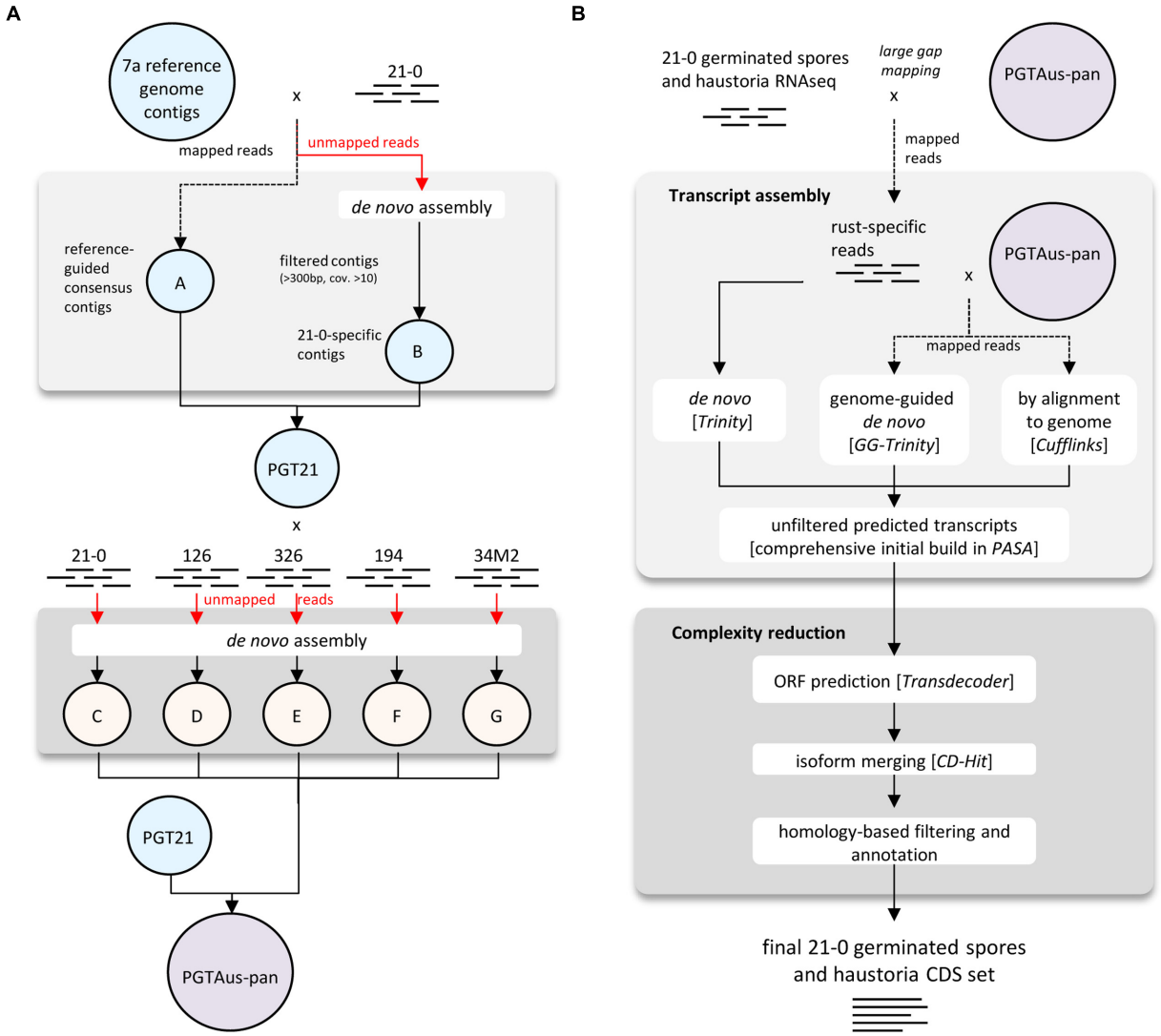


FIGURE 1 | Flow chart diagrams illustrating genome and transcriptome assembly pipelines. (A) Pipeline for assembly of PGT21 and PGTAus-pan genomes from DNA reads from Australian *Pgt* isolates. **(B)** Pipeline for assembly of isolate 21-0 transcriptome from RNA reads from isolated haustoria and germinated spores.

using the CLC module Map Reads to Reference (default parameters) and the 4557 contigs of the *Pgt* isolate CDL 75-36-700-3 (Duplessis et al., 2011)[2] as a reference for assembly. Consensus sequences derived from this mapping were taken as part A of our PGT21 reference build. The unmapped reads were assembled *de novo* and contigs of length >300 nucleotides and average coverage >10 were *de novo* assembled in a second round using 'simple assembly,' and added as part B of the PGT21 reference. DNA reads from isolates 21-0, 126, 194, 326, and 34M2 were mapped to PGT21 and the unmapped reads were *de novo* assembled separately and contigs >300 nucleotides and average coverage >10 (parts 'C,' 'D,' 'E,' 'F,' 'G' respectively) were added to the PGT21 assembly to generate the pan-genome assembly, PGTAus-pan.

For transcriptome assembly, quality trimmed (0.01 quality trim, minimum length 50) RNA reads from isolated haustoria and germinated spores from isolate 21-0 were first aligned to PGTAus-pan genome by using the CLC module large gap read mapping (default parameters) and mapped reads were extracted as fungal specific reads. Transcript models were built separately using genome-guided and *de novo* assembly with the Trinity pipeline (Grabherr et al., 2011) and a genome reference based assembly using Tophat/Cufflinks (Trapnell et al., 2012). These transcript models were used as inputs to the PASA (Program to Assemble Spliced Alignments) pipeline[3] to build a comprehensive

[2]http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html

[3]http://pasa.sourceforge.net/

transcriptome database. Open reading frame (ORF) and protein predictions (>50 amino acids) were performed using Transdecoder[4]. A further complexity reduction was then performed on a non-redundant protein set with CD-hit (Li and Godzik, 2006) for isoform/allele merging[5] (95% identity cut-off), yielding 27,150 proteins, which were reduced to 22,391 after manual curation to exclude likely spurious ORFs. Predicted proteins were analyzed for homology to known proteins (e-value 1e-20 cut-off) by PFAM domain searching (Punta et al., 2012) and by Blastp analysis (e-value 1e-05 cut-off) against a custom database of predicted proteins from *Pgt*, *Pst* (Cantu et al., 2013; Garnica et al., 2013) and *P. triticina*[6]. The presence of signal peptides was predicted using SignalP v4.0 using the SignalP-TM network function (Petersen et al., 2011). Transmembrane domains were then predicted using TmHMM (Krogh et al., 2001) and those proteins containing one or more transmembrane domains that did not overlap with the signal peptide (minimum five amino acids) were excluded from the secreted protein set. The 21-0 haustorial and germinated spore transcript models (coding sequences only) and the p7a reference transcript set (coding sequences only; 15,979 entries downloaded on 1-5-2014 from *Puccinia* Group Sequencing Project, Broad Institute of Harvard and MIT[7]) were mapped onto PGTAus-pan using the PASA pipeline.

SNP DETECTION AND INTER-ISOLATE COMPARISONS

Unless otherwise mentioned, analysis was performed using programs and plug-ins available in CLC Genomics Workbench (V. 6.5.1 or later). Quality trimmed DNA reads (quality 0.01, minimum length 50 nt, adapter trimmed, and overlapping paired-end read merging) were mapped (default settings) to the annotated PGTAus-pan reference genome. Local realignments were performed before making variant calls using Probabilistic Variant Detection, ignoring non-specific matches and broken pairs and with default parameters including minimum coverage 10, variant probability 90% and minimum variant count 2. Variant comparison tables were produced and exported as VCF or CSV files for further processing. For assigning variants to coding and non-coding sequences, we used the combined p7a and 21-0 transcript annotation and chose the longest predicted coding sequence at each locus. To infer phylogenetic relationships between the sequenced isolates, variant calls were first filtered using custom Python scripts for homozygous SNPs (indels were ignored) and then merged and converted to tabular format using VCFtools (Danecek et al., 2011). From this, SNP alignments were concatenated and used as input to FastTree (Price et al., 2010), with the –pseudo and –nt options. Phylogenetic trees were drawn and midpoint rooted using MEGA6 (Tamura et al., 2013).

GO ANNOTATION OF THE PREDICTED PROTEOME

For the gene ontology (GO) classification the set of 22,391 predicted genes was analyzed using the BLAST2GO PRO plugin in CLC genomics 6.5. Briefly, a Blastp search of predicted protein sequences against the non-redundant protein database (nr) of NCBI (Database downloaded on August-2013) was performed with a maximum expectation value of 1.0e-25, maximum number of alignments to report = 50 and highest scoring pair length = 33 amino acids. The GO terms associated with each BLAST hit were retrieved and GO annotation assignment to the query sequences was carried out using default parameters. BLAST2GO was also used for GO functional enrichment analysis of the genes differentially expressed in both germinated spores and haustoria, by performing Fisher's exact test with false discovery rate (FDR) correction to obtain an adjusted *p*-value (0.05).

RESULTS AND DISCUSSION

GENOME ASSEMBLY OF AUSTRALIAN *Pgt* ISOLATES

To investigate genetic variation amongst Australian stem rust isolates, four isolates (21-0, 126, 194, and 326) with different virulence/avirulence phenotypes on the *Sr* resistance genes represented in standard differential genotypes (**Table 1**) and representing the four independent incursions of stem rust into Australia (Park, 2007) were each analyzed by next generation sequencing. Illumina sequencing (75 bp paired ends) data from genomic DNA of isolates 21-0, 126, 194, 326 yielded 41-178 million reads after quality-based filtering (**Table 2**) that were mapped to the 81.5 Mbp reference genome (4,557 contigs, 81,521,292bp) of the American *Pgt* isolate CDL 75-36-700-3 (p7a; Duplessis et al., 2011). Between 61 and 73% of the sequence reads for each isolate could be mapped to the p7a reference genome, covering between 94.8 and 97.6 of the reference at depths of 23- to 108-fold (**Table 2**). Mapped regions in isolates 21-0, 126, and 194 showed >98% sequence identities to the p7a reference, while isolate 326 was more divergent with only 93% identity.

For each isolate more than 25% of the reads did not map to the p7a reference, suggesting that these genomes contained substantial amounts of DNA sequence not present in the p7a reference genome. Therefore, we built a new reference genome based on the sequence of isolate 21-0 (**Figure 1A**; Table S1). We obtained additional 454 sequence data for this isolate (3 million reads, 1.2 Gbp, 12X coverage, average read length 400 bp). This sequence was combined with the Illumina sequence data and first assembled against the p7a reference genome. The consensus sequences for the 4,557 contigs in this assembly were then taken as part A of our PGT21 reference build (79.2 Mbp). The remaining unmapped reads were assembled *de novo* and contigs of length >300 nucleotides and average coverage >10X (19,662) were retained and again *de novo* assembled, resulting in a total of 16,960 contigs (part B, 13.3 Mbp), which were then added to the PGT21 reference build. The complete PGT21 genome assembly then comprised 21,517 contigs and ~92.5 Mbp, about 11 Mbp larger than the p7a reference genome sequence. Much of this could represent sequence missing (gaps in the scaffold) from the p7a reference assembly, rather than isolate-specific sequence, because the p7a scaffold assembly size is 89 Mbp including gaps (Duplessis et al., 2011). The remainder of the additional sequence may represent highly variable regions between the two isolates that failed to map to the original reference sequence. The *de novo* assembled sequence region contained a similar density of heterozygous SNPs (see below) to the reference assembled

[4]http://transdecoder.sourceforge.net

[5]http://weizhong-lab.ucsd.edu/

[6]http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html

[7]http://www.broadinstitute.org/

Table 2 | Mapping of Illumina DNA reads from Australian *Pgt* isolates against p7a reference.

	Pgt Isolates			
	21-0	**194**	**326**	**126**
Total reads (quality trimmed)*	178,487,947	124,005,114	41,202,425	134,392,144
Reads mapped to reference	131,084,929	84,503,934	25,300,892	88,653,681
Percentage mapped reads	73.44	68.15	61.41	65.97
Total bases mapped to reference	8,556,866,766	5,558,168,397	1,823,251,337	5,533,033,972
Assembly length (bp)	78,726,070	78,918,599	77,273,144	79,579,366
Average times coverage	108.69	70.43	23.59	69.53
Unmapped reads	47,403,018	39,501,180	15,901,533	45,738,463
Percentage unmapped reads	26.56	31.85	38.59	34.03
Percentage coverage of reference	96.57	96.81	94.79	97.62
Percentage bases identical to reference**	98.26	98.20	93.30	98.50
Percentage mismatched bases**	1.11	1.14	1.15	0.91
Percentage reference gap bases**	3.71	3.45	5.10	2.77
Percentage assembly gap bases**	0.59	0.57	0.50	0.56

**CLC genomics workbench 4.9 or above was used for assembly (parameter settings: quality clip 0.05; conflict resolution by vote; random mapping of non-specific reads, two ambiguities allowed, read length 40–75 bp).*
***Based on BLAT analysis between p7a contigs and respective assembled sequences.*

region, indicating that it is present it both haploid nuclei, and does not represent a divergent sequence present in just one nucleus of this dikaryotic organism. Nearly 96% of the isolate 21-0 DNA reads could be remapped back to the PGT21 reference covering 99.35% of the assembly with an average nucleotide identity of 99.7% (**Table 3**).

Alignment of the DNA reads from the other Australian *Pgt* isolates (126, 194, 326), as well as from two additional isolates

Table 3 | Mapping of Illumina DNA reads from Australian *Pgt* isolates to PGT21 reference genome.

	Pgt Isolates/mutants					
	21	**34M1**	**34M2**	**194**	**326**	**126**
Total reads (quality trimmed)*	155,272,002	312,359,971	165,949,995	106,502,558	24,201,578	106,119,468
Reads mapped to reference	148,738,543	292,860,382	160,656,060	98,176,567	23,470,521	93,092,615
Percentage mapped reads	95.79	93.76	96.81	92.18	96.98	87.72
Total bases mapped	9,324,322,642	28,576,013,167	15,032,224,237	5,981,339,426	1,166,889,391	5,398,916,769
Assembly length (bp)	91,842,155	90,688,020	90,159,598	91,123,290	89,660,526	88,928,858
Average times coverage	95	303	160	61	12	57
Unmapped reads	6,533,459	19,499,589	5,293,935	8,325,991	731,057	13,026,853
Percentage unmatched reads	4.05	6.24	3.19	7.82	3.02	12.28
Assembled contigs	21,517	21,517	21,192	21,513	21,509	21,513
Percentage coverage of reference	99.25	98.00	97.59	98.48	96.90	96.10
Percentage coverage of reference part B	∼100	99.23	96.98	98.26	96.34	86.57
Percentage bases identical to reference**	99.7	99.4	99.36	99.41	99.38	98.12
Percentage mismatched bases**	0.20	0.46	0.49	0.41	0.49	1.07
Percentage Reference gap bases**	0.72	1.83	2.1	1.46	2.79	3.79
Percentage Assembly gap bases**	0.10	0.13	0.13	0.16	0.12	0.65

**CLC genomics workbench 4.9 or above was used for assembly (parameter settings: quality limit 0.01, ambiguity allowed 1 nt from each end; length 25–77 nt except with 34M2 where length 50–100 nt; conflict resolution by vote; random mapping of non-specific reads).*
***Based on BLAT analysis between PGT21 and respective assembled sequences.*

derived from 21-0 (34M1 and 34M2), to the PGT21 assembly showed that 88–97% of reads mapped to the PGT21 reference and covered 96–98% of the sequence and were at least 98.12% identical (**Table 3**). Sequence reads from the independent isolates 126, 194, and 326 covered between 87 and 98% of the additional 13.3 Mbp *de novo* assembled region of the PGT21 (part B), indicating that most of this region is not specific to isolate 21-0. To capture possible isolate-specific sequences from other Australian isolates, additional unmapped DNA reads from 21-0, 126, 194, 326, and 34M2 were *de novo* assembled independently and contigs >300 bp and >10x coverage were added to the PGT21 reference (Parts C to G respectively) to obtain the pan-genome PGTAus-pan (**Figure 1A**; Table S1). Isolate 126 showed the highest level of unique sequence (∼1%, Table S1), as well as the greatest number of mismatches, gaps and unmapped reads (**Table 3**), suggesting a greater evolutionary divergence of isolate 126 from 21-0 compared to the other isolates. This is consistent with previous studies of genetic diversity among Australian isolates of *Pgt* using other DNA marker systems (Keiper et al., 2003; Zhang, 2008). An analysis searching for the CEGMA set of 248 conserved eukaryotic genes (Parra et al., 2007) found that 237 (95.5%) were present in full or in part in the PGTAus-pan assembly, compared to 232 (93.5%) for the p7A reference genome, indicating an improvement in gene coverage in the PGTAus-pan genome compared with the p7a reference. Altogether, we have assembled a 92 Mbp *Pgt* pan-genome, which contains a significant amount of novel sequence not included in the p7a assembly. Most of this sequence is nevertheless common amongst several stem rust isolates, thus resulting in higher genome coverage for the wheat stem rust pathogen.

ANNOTATION OF TRANSCRIPTS ON THE PGTAus-PAN GENOME

As a first step toward annotating the Australian *Pgt* pan-genome, the 15,979 transcripts (ORFs only) previously predicted for p7a were mapped against PGTAus-pan using the PASA pipeline (Haas et al., 2003). Under the set stringency (alignment length >75% and identity >95%) 14,843 transcripts aligned to the genome (Table S3) and as expected, almost all (14,828) mapped to part A of PGT21 (p7a reference assembled), while 15 transcripts mapped to other parts of the pan-genome. In total, 13,554 valid ORFs (>50 amino acids) could be predicted from the mapped transcripts.

For a more comprehensive annotation of the PGTAus-pan genome, we generated a new transcript set from RNA isolated from purified haustoria and germinated spores of isolate 21-0 as outlined in **Figure 1B**. Three biological replicate samples were sequenced by Illumina HiSeq 2000 (100 bp paired ends) to also allow subsequent differential expression analysis (see below). Initial raw reads (∼25 million pairs in each replicate sample) yielded 17-23 million quality-trimmed pairs per replication (Table S2). Large-gap read mapping (CLC Genomics Workbench) to PGTAus-pan was used to extract *Pgt*-specific reads. Transcripts were built independently by three methods using pooled reads: trinity assembly using both *de novo* and genome-guided approaches and TopHat/Cufflinks assembly against the PGTAus-pan reference. Transcripts from these independent assemblies were combined and assembled using the PASA pipeline to give a comprehensive initial transcriptome set of 61,451 transcript models. Of these, 59,783 could be aligned to the genome with 55,386 correctly mapping to predicted exon boundaries (Table S3). Most of these (85.6%) mapped to the p7a reference-assembled region (part A) of the PGT21 genome, while 13.4% mapped to the *de novo* assembled region (Part B). A small number (587, ∼1%) of transcripts mapped to other parts of the pan-genome (C to G).

A total of 22,391 non-redundant protein sequences were predicted from the transcript models after complexity reduction and filtering as described in the methods (Table S4). Approximately 90% (20,242) of these proteins mapped to PGT21 part A (i.e., common to PGT21 and p7a) and 9.3% (2,091) mapped to PGT21 part B. In a Blastp search against a custom database of predicted proteins from *Pgt*, *Pst*, and *P. triticina*, a previously annotated homolog in one of these *Puccinia* species was detected for 19,311 proteins (e-value 1e-05 cut-off). Interestingly, only 15,923 showed best Blastp hits to *Pgt* proteins while the remainder returned best hits to *Pst* (2,026) or *P. triticina* (1,381) proteins. These include genes that either were not present in the Pgt7a reference genome assembly (1,041 mapped to part B of PGT21) or were not annotated in the sequence (2,062 mapping to part A). A further 3,061 ORFs/proteins had no significant Blastp hit to the *Puccinia* group protein set but did align to PGT21 and may represent novel rust genes not previously detected. Only five transcripts failed to align to the PGTAus-pan sequence, and these showed significant hits to wheat cDNA sequences suggesting they are derived from host RNA contaminants. Another three transcripts with poor alignment to PGTAus-pan sequence had better matches to wheat transcripts. None of the other sequences appeared to be derived from wheat genes. We have also flagged 26 transcripts of possible *Pgt* mitochondrial origin. The PGTAus-pan genome sequence was annotated with the aligned transcripts from both p7a and the 21-0 transcriptome build[8]. A total of 21,874 gene loci are predicted in this annotation, which is similar to the gene numbers predicted for other rust fungal genomes such as *Melampsora larici-populina* (16,399, Duplessis et al., 2011), *M. lini* (16,271, Nemri et al., 2014), and *Pst* [20-25,000, (Cantu et al., 2013; Zheng et al., 2013)].

COMPARISON OF HAUSTORIAL AND GERMINATED SPORE TRANSCRIPTOMES

We also used the RNA-Seq data from isolated haustoria and germinated urediniospores to compare gene expression between these cell types. The data were each obtained from three independent biological replicates allowing statistically robust quantitative expression analysis. The RNA-Seq tools from CLC genomics were used to align the raw Illumina reads against the reference transcript set and expression levels were quantified as reads per kilobase per million mapped reads (RPKM) for comparison of transcript levels. A total of 4,524 genes were differentially expressed between these cell types, with approximately half upregulated in haustoria and half in germinated spores (**Figure 2**).

The 22,391 predicted gene set was annotated using BLAST2GO software (Conesa et al., 2005; Figure S1A). Among all Blastp results,

[8]http://webapollo.bioinformatics.csiro.au/puccinia_graminis_tritici_PGTAus-pan/index.html

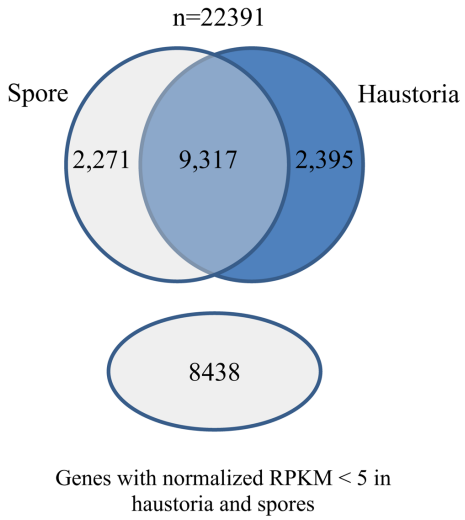


FIGURE 2 | Differential gene expression in *Pgt* germinated spores and haustoria. Venn diagram of predicted gene set showing the number of genes that had statistically significant changes in expression between germinated spores and haustoria (Baggerley's test FDR corrected *p*-value < 0.05 and expression difference > 20RPKM), genes that did not show differential expression between the two tissues, and those that showed very low expression (RPKM < 5) in both tissues.

P. graminis, *M. larici-populina, Cryptococcus neoformans*, *Agaricus bisporus,* and *Serpula lacrymans* were the top five species in terms of the total number of hits to the NCBI-nr protein database (Figure S1B). In total 7,469 (33.4%) genes could be unambiguously annotated with predicted functions and were categorized into functional classes to identify those that encode proteins with known roles in cellular processes. Direct GO count graphs were created to categorize the sequences to several groups based on their biological process ontologies (Table S5), the major functional categories are shown in **Figure 3**. Processes upregulated in germinated spores were representative of cell proliferation, such as cell cycle, DNA replication and cell wall biogenesis, whereas haustoria were committed to energy production and biosynthetic processes. Similar observations were recently made for the stripe rust pathogen *Pst* (Garnica et al., 2013). Other similarities with *Pst* included the upregulation of genes involved in the production of ATP through glycolysis, TCA cycle and oxidative phosphorylation in haustoria of *Pgt,* and upregulation of genes involved in releasing energy from stored lipid reserves and processing them via the glyoxylate/gluconeogenesis pathways in spores (Table S6). This suggests that the primary metabolism of haustoria and germinated spores of these two rust pathogens is largely the same. Recent transcriptomic studies on isolated haustoria from other rust fungi (Link et al., 2013) revealed important metabolic similarities to *Pgt* and *Pst*, supporting the idea that both the structure and the physiology of the haustorium are hallmarks of biotrophy in rust fungi.

To determine broader similarities in the gene expression profiles between *Pgt* and *Pst*, the whole set of *Pgt* predicted genes was compared to transcriptomic data for *Pst* (Garnica et al., 2013). The 12,282 transcripts from *Pst* were BLAST searched against the predicted gene set of *Pgt* and then matched accordingly to their tissue expression profile (**Figure 4**). A total of 9,962 transcripts from *Pst* (81%) showed similarity (e-value 1e-5 cut-off) to at least one predicted gene from *Pgt*. Although only 56% of the matching genes had the same expression profile in both species, most of these differences were genes showing differential expression in one species but either not deferentially expressed or expressed at a low level in the other, probably mainly reflecting differences in the sensitivity of the statistical tests applied. Despite this, there was a broad similarity in the expression data for both species. *Pgt* homologs of *Pst* genes upregulated in haustoria were enriched for haustorial-specific genes, while *Pgt* homologs of *Pst* genes upregulated in spores were enriched for spore-expressed genes. Furthermore, most of the genes belonging to the metabolic categories mentioned above showed the same expression trends in both pathogens (**Figure 3**; Table S6).

PREDICTION OF EFFECTOR CANDIDATES

To identify potential effectors in the PGT21 genome, we searched for proteins containing a predicted signal peptide (SP) in the haustorial and germinated spore transcript sets. Proteins containing one or more transmembrane domains (not overlapping with SP domain) were excluded, leaving a total of 1,924 predicted secreted proteins (Table S7). Of these, 1,590 had best Blastp hits to *Pgt* (p7a), while 103 and 81 had best hits in the *Pst* and *P. triticina* protein sets respectively, with the remaining 150 showing no hits. Of the 1,924 predicted secreted proteins, 1,824 were encoded in part A of PGT21, and 100 in part B. Over half (1,022) of these proteins have 4 or more cysteine (cys) residues while 212 have 10 or more cys residues, a common feature of many predicted and known effector proteins (Templeton et al., 1994).

Gene expression analysis detected 689 predicted secreted protein transcripts that were upregulated in haustoria (FDR corrected *p*-value < 0.05, >2 fold change, >5 normalized RPKM) while 460 were upregulated in germinated spores. Eliminating those with the lowest expression levels (<20 RPKM) left a set of 430 upregulated in haustoria and 329 in germinated spores. We considered the 430 haustorially upregulated secreted proteins as primary candidates for stem rust effectors. However, some rust effectors could also be expressed in germinated spores, as is the case for AvrM in flax rust (Catanzariti et al., 2006). Therefore we also considered those that showed high expression in haustoria (>100 RPKM) as good candidates regardless of their expression in germinated spores. This added an additional 90 genes to make a total set of 520 haustorial secreted proteins (HSPs). These included 299 proteins containing four or more cys residues and 85 with 10 or more. Only 41 of these could be annotated with putative function (PFAM hit with e-value < 1e-20), including seven carbohydrate-active enzymes, two heat shock proteins, two thaumatin-like proteins and three thioredoxin proteins (Table S7). Similar numbers of HSPs have been predicted from haustorial transcriptomes of *Pst* (437, Garnica et al., 2013) *U. appendiculatus* and *P. pachyrhizi* (395 and 149 respectively, Link et al., 2013).

GENOME DIVERSITY BETWEEN ISOLATES

We examined genome-wide sequence variation both within and between the six Australian *Pgt* isolates by aligning the sequence

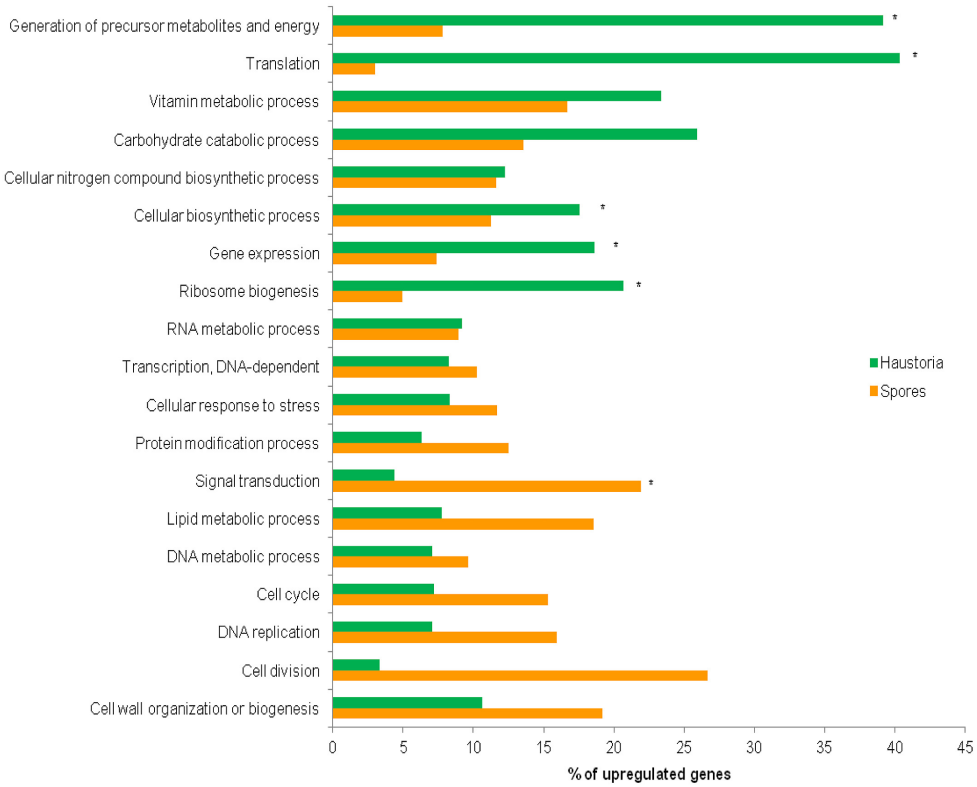


FIGURE 3 | Comparative ontology analysis of genes with statistically significant changes in expression between haustoria and germinated spores. Of the original predicted gene set, 34.6% (794) of the 2,365 haustorial-enriched genes and 46.4% (1,037) of the 2,271 genes up-regulated in germinated spores were annotated with B2G. Relevant biological process GO terms are shown on the Y-axis. Percentages of genes differentially expressed in each tissue belonging to the nominated categories are shown on the X-axis. Asterisks indicate the categories over-represented in either developmental stage found after applying Fisher's exact test, FDR < 0.05.

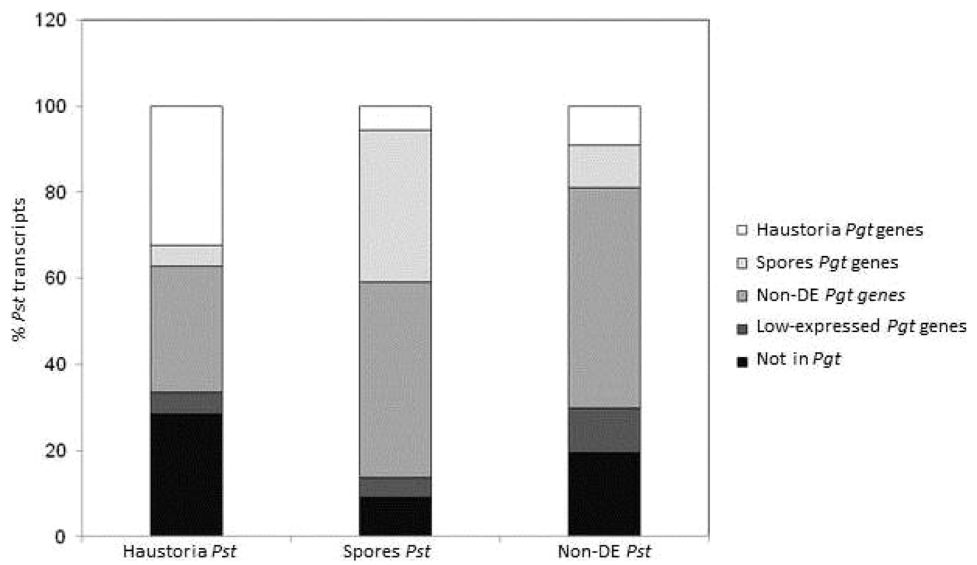


FIGURE 4 | Comparison of gene expression profiles in *Pgt* and *Pst* in haustoria and germinated spores stages. A total of 12,282 transcripts from *Pst* previously classified accordingly to their statistically significant over expression in germinated spores (2,357) and haustoria (1,989), or those that did not show differential expression between the two tissues (Non-DE, 7,936) were BLAST searched (max E-val 1.0e-6) against the *Pgt* 21-0 predicted gene set. The percentage of transcripts matching genes in the same (or different) expression profile category in *Pgt* are shown in colors.

reads from each isolate to the PGTAus-pan genome reference. For isolate 21-0, we found over 1.3 million variants, including single nucleotide variants (SNVs), multiple nucleotide variants (MNVs), and insertion/deletions (indels; Table S8A). These occurred at an overall frequency of 14.2/kb of mapped consensus sequence, with base changes (SNVs and MNVs) representing about 86% of this variation (12.3/kb). The vast majority of these variants (∼92%) occurred in a heterozygous condition, reflecting a high level of divergence between the two haploid nuclei in this dikaryotic organism. The frequency of variants was broadly similar in intergenic regions (13.37/kb), gene-coding regions (including introns, 15.88/kb) and coding sequences (12.79/kb), but there was a difference in the distribution of indels between these locations, being much more frequent in intergenic regions (14.3% of variation, 1.92/kb) than in coding sequences (5.4% of variation, 0.69/kb). A total of 153,946 (6.6/kb) variants could give rise to altered protein sequences including 136,516 non-synonymous SNPs (SNV+MNV) as well as 16,304 indels and these were distributed in 17,960 genes.

The frequency of DNA variation in the other isolates was similar to 21-0 (about 13-15/kb), except in isolate 326 where the low variant discovery (4.3/kb) may be attributed to the lower coverage of reads used in the initial mapping. As with 21-0, these isolates were heterozygous for the majority of variants (92–96%), except isolate 126, which contained a high proportion of variants that were in the homozygous state (43%). Thus, substantial variation between heterokaryons seems to be a common feature of *Pgt* isolates. Similarly, Cantu et al. (2013) found substantial polymorphism between heterokaryons in five isolates of *Pst*, with heterozygous SNPs occurring at a frequency of ∼6 per kb and representing over 90% of the total (homozygous and heterozygous) variation. Zheng et al. (2013) found a much lower rate of heterozygosity (∼1.0 SNP/kb) in *Pst* isolates, possibly because their genome assembly from fosmid clones resulted in separate assembly of allelic regions from the two haplotypes. Since *Pgt* reproduces asexually in Australia the heterozygosity present in these isolates, derived from their most recent sexual ancestor before incursion of these isolates into Australia, has been fixed. This is clearly observed in the case of the 34M1 and 34M2 which are clonally derived from 21-0 but isolated around 30 years after its incursion, and share almost all of the >1 million heterozygous SNPs that are present in 21-0. The high proportion of variant homozygosity in isolate 126 may reflect a level of inbreeding in the most recent sexual background of isolate 126, while 21-0, 194, and 326 may have arisen from more diverse populations.

To determine relationships between the *Pgt* isolates, we compared genome-wide variation between the isolates (**Figure 5A**). Variation between the four founder isolates was substantial: for instance only 7.0% of variation in 21-0 was shared with the other four isolates. In isolates 21-0, 194 and 326, unique variants were 8.4%, 4.7% and 1.0%, respectively, while isolate 126 was much more divergent with 30.4% unique variants. A phylogenetic tree constructed using the homozygous SNP data for the six isolates (**Figure 5B**), showed that isolates 34M1 and 34M2 fell into a clade derived from 21-0, consistent with the prediction that these isolates represent field-evolved mutational derivatives of 21-0 based on virulence phenotypes (Park, 2007). As noted above, isolate 126 showed greater divergence from the other isolates in this group.

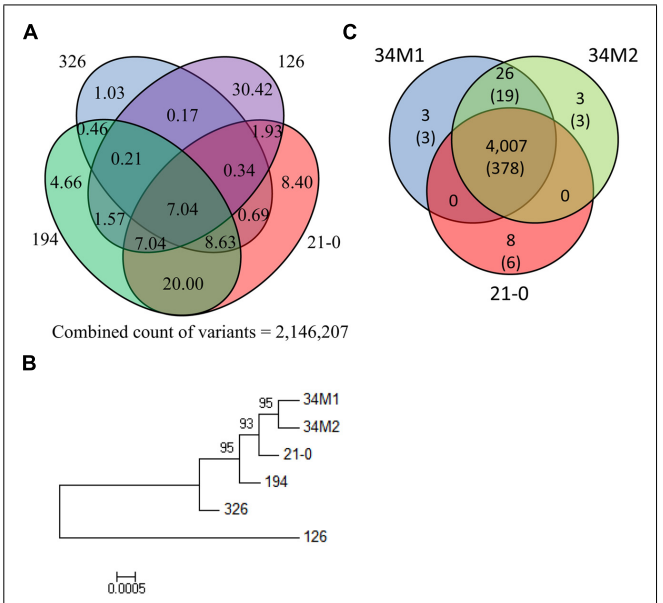


FIGURE 5 | Comparison of Australian *Pgt* isolate genome sequences. **(A)** Venn diagram showing the percentage of the total number of polymorphisms detected in the four Australian founder *Pgt* isolates (21-0, 126, 194, and 326) that are shared between different isolates. **(B)** Phylogenetic tree of Australian *Pgt* isolates. This phylogeny was produced by FastTree from homozygous SNPs in the six stem rust isolates. Local bootstrap support values are shown. This tree is midpoint rooted. **(C)** Venn diagram showing numbers of shared non-synonymous polymorphisms in HSP genes between 21-0, 34M1 and 34M2. Figures in the parenthesis refer to the number of genes in which the polymorphisms occur.

VARIATION IN EFFECTOR CANDIDATES

We examined variation in the set of 520 HSPs as these are most likely to include genes controlling virulence/avirulence differences between isolates with respect to infection on host differentials carrying different *Sr* genes. In 21-0, 402 (77%) of these genes contained sequence variants in their coding sequences (17.72/kb, almost all heterozygous), while 52 (∼10%) were not polymorphic and the remainder (∼13%) could not be scored due to incomplete mapping to the genome (Table S8B). In total, 3,843 variants (9.60/kb) occurring in 374 HSPs, gave rise to amino acid changes in the encoded proteins (including indels and frameshifts). Among the four Australian isolates, 16,322 variants were detected in 427 HSPs and showed a similar pattern of shared and unique polymorphisms as for the genome-wide variants (Figure S2). These included 5,245 non-synonymous variants in 406 genes. In a similar analysis of two UK *Pst* isolates that differ in only 2 virulence phenotypes, Cantu et al. (2013) found polymorphisms in 60 HSPs. However, this analysis only considered homozygous SNPs between the strains and heterozygous differences may account for significantly more differences. Bruce et al. (2014) observed much lower levels of diversity in effector candidates from *P. triticina*, with only 15 of 532 secreted proteins expressed *in planta*, showing amino acid differences among six isolates. However, this analysis was conducted using protein

sequences translated from consensus-derived RNAseq transcripts and thus also does not consider heterozygous variation. The true extent of variation between these strains may be significantly higher.

As indicated previously, isolates 34M1 and 34M2 represent field-derived mutants of isolate 21-0 that have gained virulence for resistance genes *Sr5*, *Sr11,* and *Sr27* and in the case of 34M2 one additional *R* gene, *SrSatu*. We therefore examined nucleotide variants that give rise to altered amino acid sequences among the HSP set in these isolates. There were a total of 4,048 such nucleotide variants, of which the vast majority (3,712) were common to all three isolates. We manually examined the remaining 336 putative SNPs that distinguished the strains to eliminate any incorrect calls. In most cases reads representing each polymorphic variant were present in all three strains, although the SNP failed to be called in one or more strains. In only one case there was a false positive call. After manual curation, 4,007 SNPs were common to all three strains, while only 40 SNPs distinguished the strains (**Figure 5C**). Of these, 26 were common to 34M1 and 34M2 and absent in 21-0, and therefore represent novel mutations in these isolates that could explain their virulence on *Sr5*, *Sr11,* or *Sr27*. These occurred in a total of 19 HSP genes. The three variants (occurring in three genes) that were unique to 34M2 could explain virulence on *SrSatu*, giving a total of 22 candidates for these four *Avr* genes. We do not know whether the progenitor pathotype 21-0, is functionally homozygous at these *Avr* loci, in which case mutation of both alleles would be required for virulence, or heterozygous in which case a single mutation would be sufficient. In addition, eight variants (in six genes) were unique to 21-0. Loss of 21-0 variants could result from a deletion of one allele, but in all of these cases other heterozygous variants are retained in the HSP gene, ruling out this possibility. Alternatively, mutation of one variant site to the opposite allelic version could lead to virulence if the pathotype was heterozygous for this character. Thus these are also possible virulence mutations, giving a further three unique candidates for these *Avr* genes, for a total of 25 (Table S9). Clearly there are more HSP genes showing variation than the four documented *Avr* changes separating 34M2 from its progenitor 21-0. Mutations in other HSP genes that altered virulence-avirulence on uncharacterized *Sr* genes in wild host species may have been selected between the 1954 and 1984 isolations of 21-0 and 34M2. Furthermore, based on the strong assumption that effectors in *Pgt* play a virulence role, selection may occur in these genes for improved adaptation to host virulence targets in wheat or wild hosts. There may also be selection for changes in 'background' effector genes that compensate for loss of function in effectors associated with virulence-avirulence toward *Sr5*, *Sr11*, *Sr27,* and *SrSatu*.

CONCLUSION

To summarize, we have generated an extended pan-genome for the wheat stem rust fungus that extends the previous reference assembly based on the p7a isolate by including about 13 Mbp of novel sequence. We carefully considered whether this additional sequence was specific to different strains, as substantial genome divergence has been observed for some other fungal plant pathogens. For instance, *Magnaporthe oryzae* strains contain up to 5% unique sequence that is dispersed throughout the genome (Yoshida et al., 2009), while *Fusarium oxysporum* contains several dispensable chromosomes that can vary in presence between strains infecting different hosts (Ma et al., 2010). Divergence between the haploid nuclei in the dikaryotic *Pgt* could also be a source of diversity in genome content. However, the vast majority of this additional sequence was represented in four unrelated isolates that each arrived in Australia at different times over the past century. The presence of heterozygous SNPs in this region also indicates that it is not derived from a single nucleus due to genome divergence between the haploid nuclei. Hence we suggest that most of this region is not strain specific, but more likely represents sequence that was simply not assembled in the p7a reference. Thus, the genome assembly presented here increases the sequenced genome coverage of this organism, improving the representation of core eukaryotic genes, and allowing the annotation of about 2000 transcripts in this region. Transcriptome assembly from germinated urediniospores and haustoria also identified a further ∼3500 transcripts not previously annotated in the p7a reference genome as well as a large number of potential alternative transcripts. Analysis of putative secreted proteins identified 520 HSPs as effector candidates, and a subset of 25 of these represent candidates for four *Avr* genes that differ between the pt 21-0 isolate and two derived isolates. We are currently performing functional analyses of these candidates by bacterial delivery to resistant host lines (Upadhyaya et al., 2014) to determine whether they encode these *Avr* recognition specificities. We are also selecting *de novo* mutants of 21-0 that acquire virulence toward *Sr5*, *Sr11*, *Sr27,* and *SrSatu* in glasshouse experiments so that sequence comparisons can be made between the candidate genes in 21-0 and the same genes in the new mutants.

AUTHOR CONTRIBUTIONS

Narayana M. Upadhyaya, Diana P. Garnica, Adnane Nemri, Jana Sperschneider, Christina A. Cuomo, Haydar Karaoglu, Bo Xu, Rohit Mago performed experiments and analyzed data. Narayana M. Upadhyaya, John P. Rathjen, Robert F. Park, Jeffrey G. Ellis, Peter N. Dodds provided scientific direction. All contributed to the preparation of the manuscript.

ACKNOWLEDGMENTS

Authors wish to thank the Two Blades Foundation for financial support, Robyn East, Dhara Bhat and Lina Ma for excellent technical assistance, Andrew Spriggs for providing custom scripts, Sharadha Sakthikumar for initial QC and variant analysis of the Illumina data generated at the Broad and Dr. Cristobal Uauy for providing the *Pst* proteome set for inclusion as a component of the *Puccinia* group protein database used in this study. Part of this work was supported through access to facilities managed by Bioplatforms Australia and funded by the Australian Government National Collaborative Research Infrastructure Strategy and Education Investment Fund Super Science Initiative. We are grateful for the assistance of the CSIRO Bioinformatics Core unit for hosting our genome browser (http://webapollo.bioinformatics.csiro.au/puccinia_graminis_tritici_PGTAus-pan/index.html) in the WebApollo server.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00759/abstract

REFERENCES

Barrett, L. G., Thrall, P. H., Dodds, P. N., van der Merwe, M., Linde, C. C., Lawrence, G. J., et al. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol. Biol. Evol.* 26, 2499–2513. doi: 10.1093/molbev/msp166

Bruce, M., Neugebauer, K. A., Joly, D. L., Migeon, P., Cuomo, C. A., Wang, S., et al. (2014). Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat. *Front. Plant Sci.* 4:520. doi: 10.3389/fpls.2013.00520

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. *sp. tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Catanzariti, A. M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Catanzariti, A. M., Mago, R., Ellis, J., and Dodds, P. (2011). Constructing haustorium-specific cDNA libraries from rust fungi. *Methods Mol. Biol.* 712, 79–87. doi: 10.1007/978-1-61737-998-7_8

Chisholm, S. T., Coaker, G., Day, B., and Staskawicz, B. J. (2006). Host-microbe interactions: shaping the evolution of the plant immune response. *Cell* 124, 803–814. doi: 10.1016/j.cell.2006.02.008

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi: 10.1093/bioinformatics/bti610

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., et al. (2011). The variant call format and VCFtools. *Bioinformatics* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini* AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16, 755–768. doi: 10.1105/tpc.020040

Dodds, P. N., and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* 11, 539–548. doi: 10.1038/nrg2812

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annu. Rev. Phytopathol.* 9, 275–296. doi: 10.1146/annurev.py.09.090171.001423

Garnica, D. P., and Rathjen, J. P. (2014). Purification of fungal haustoria from infected plant tissue by flow cytometry. *Methods Mol. Biol.* 1127, 103–110. doi: 10.1007/978-1-62703-986-4_8

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PLoS ONE* 8:e67150. doi: 10.1371/journal.pone.0067150

Giraldo, M. C., and Valent, B. (2013). Filamentous plant pathogen effectors in action. *Nat. Rev. Microbiol.* 11, 800–814. doi: 10.1038/nrmicro3119

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Hahn, M., and Mendgen, K. (2001). Signal and nutrient exchange at biotrophic plant-fungus interfaces. *Curr. Opin. Plant. Biol.* 4, 322–327. doi: 10.1016/S1369-5266(00)00180-1

Jones, D. A., and Takemoto, D. (2004). Plant innate immunity – direct and indirect recognition of general and specific pathogen-associated molecules. *Curr. Opin. Immunol.* 16, 48–62. doi: 10.1016/j.coi.2003.11.016

Jones, J. D., and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323–329. doi: 10.1038/nature05286

Keiper, F. J., Hayden, M. J., Park, R. F., and Wellings, C. R. (2003). Molecular genetic variability of Australian isolates of five cereal rust pathogens. *Mycol. Res.* 107, 545–556. doi: 10.1017/S0953756203007809

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Koeck, M., Hardham, A. R., and Dodds, P. N. (2011). The role of effectors of biotrophic and hemibiotrophic fungi in infection. *Cell. Microbiol.* 13, 1849–1857. doi: 10.1111/j.1462-5822.2011.01665.x

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Leonard, K. J., and Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol. Plant. Pathol.* 6, 99–111. doi: 10.1111/j.1364-3703.2005.00273.x

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658–1659. doi: 10.1093/bioinformatics/btl158

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2013). The haustorial transcriptomes of *Uromyces appendiculatus* and *Phakopsora pachyrhizi* and their candidate effector families. *Mol. Plant. Pathol.* 15, 379–393. doi: 10.1111/mpp.12099

Ma, L. J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M. J., Di Pietro, A., et al. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* 464, 367–373. doi: 10.1038/nature08850

McIntosh, R. A., Wellings, C. R., and Park, R. F. (1995). *Wheat Rusts: An Atlas of Resistance Genes*. Collingwood, VIC: CSIRO Melbourne.

Nemri, A., Saunders, D. G., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Park, R. F. (2007). Stem rust of wheat in Australia. *Aust. J. Agric. Res.* 58, 558–566. doi: 10.1071/AR07117

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Periyannan, S., Moore, J., Ayliffe, M., Bansal, U., Wang, X., Huang, L., et al. (2013). The gene Sr33, an ortholog of barley Mla genes, encodes resistance to wheat stem rust race Ug99. *Science* 341, 786–788. doi: 10.1126/science.1239028

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., et al. (2012). The Pfam protein families database. *Nucleic Acids Res.* 40, D290–D301. doi: 10.1093/nar/gkr1065

Rafiqi, M., Gan, P. H., Ravensdale, M., Lawrence, G. J., Ellis, J. G., Jones, D. A., et al. (2010). Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. *Plant Cell* 22, 2017–2032. doi: 10.1105/tpc.109.072983

Rogers, O., Renher, S., Bledsoe, C., Mueller, G., and Ammirati, J. (1989). Extraction of DNA from Basidiomycetes for ribosomal DNA hybridization. *Can. J. Bot.* 67, 1235–1243.

Saintenac, C., Zhang, W., Salcedo, A., Rouse, M. N., Trick, H. N., Akhunov, E., et al. (2013). Identification of wheat gene Sr35 that confers resistance to Ug99 stem rust race group. *Science* 341, 783–786. doi: 10.1126/science.1239022

Stokstad, E. (2007). Plant pathology. Deadly wheat fungus threatens world's breadbaskets. *Science* 315, 1786–1787. doi: 10.1126/science.315.5820.1786

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725–2729. doi: 10.1093/molbev/mst197

Templeton, M. D., Rikkerink, E. H. A., and Beever, R. E. (1994). Small cysteine-rich proteins and recognition in fungal-plant interactions. *Mol. Plant Microbe Interact.* 7, 320–325. doi: 10.1094/MPMI-7-0320

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578. doi: 10.1038/nprot.2012.016

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M. A., Ellis, J. G., and Dodds, P. N. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol. Plant Microbe Interact.* 27, 255–264. doi: 10.1094/MPMI-07-13-0187-FI

Voegele, R. T., and Mendgen, K. (2003). Rust haustoria: nutrient uptake and beyond. *New Phytol.* 159, 93–100. doi: 10.1046/j.1469-8137.2003.00761.x

Whisson, S. C., Boevink, P. C., Moleleki, L., Avrova, A. O., Morales, J. G., Gilroy, E. M., et al. (2007). A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450, 115–118. doi: 10.1038/nature06203

Yoshida, K., Saitoh, H., Fujisawa, S., Kanzaki, H., Matsumura, H., Yoshida, K., et al. (2009). Association genetics reveals three novel avirulence genes from the rice blast fungal pathogen *Magnaporthe oryzae*. *Plant Cell* 21, 1573–1591. doi: 10.1105/tpc.109.066324

Zhang, J. (2008). *Studies of the Triticale: Stem Rust Pathosystem at Classical and Molecular Levels*. M.Sc. thesis, The University of Sydney, Sydney.

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4, 2673. doi: 10.1038/ncomms3673

Zhou, J. M., and Chai, J. (2008). Plant pathogenic bacterial type III effectors subdue host responses. *Curr. Opin. Microbiol.* 11, 179–185. doi: 10.1016/j.mib.2008.02.004

Zwer, P. K., Park, R. F., and McIntosh, R. A. (1992). Wheat stem rust in Australia 1969-1985. *Aust. J. Agric. Res.* 43, 399–431. doi: 10.1071/AR9920399

Conflict of Interest Statement: The Guest Associate Dr. David L. Joly declares that, despite having collaborated with author Christina A. Cuomo, the review process was handled objectively. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 09 December 2014; published online: 08 January 2015.

*Citation: Upadhyaya NM, Garnica DP, Karaoglu H, Sperschneider J, Nemri A, Xu B, Mago R, Cuomo CA, Rathjen JP, Park RF, Ellis JG and Dodds PN (2015) Comparative genomics of Australian isolates of the wheat stem rust pathogen Puccinia graminis f. sp. tritici reveals extensive polymorphism in candidate effector genes. Front. Plant Sci. **5**:759. doi: 10.3389/fpls.2014.00759*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2015 Upadhyaya, Garnica, Karaoglu, Sperschneider, Nemri, Xu, Mago, Cuomo, Rathjen, Park, Ellis and Dodds. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 21 November 2013
doi: 10.3389/fpls.2013.00456

Transcriptome analysis of poplar rust telia reveals overwintering adaptation and tightly coordinated karyogamy and meiosis processes

Stéphane Hacquard[1,2†], Christine Delaruelle[1,2], Pascal Frey[1,2], Emilie Tisserant[1,2], Annegret Kohler[1,2] and Sébastien Duplessis[1,2]*

[1] *INRA, UMR 1136, Interactions Arbres-Microorganismes, Champenoux, France*
[2] *UMR 1136, Université de Lorraine, Interactions Arbres-Microorganismes, Vandoeuvre-lès-Nancy, France*

Edited by:
Guus Bakkeren, Agriculture and Agri-Food Canada, Canada

Reviewed by:
Barry Saville, Trent University, Canada
John Fellers, Agricultural Research Service - US Department of Agriculture, USA

***Correspondence:**
Sébastien Duplessis, INRA, UMR 1136, Interactions Arbres-Microorganismes, Route d'Amance, F-54280 Champenoux, France
e-mail: duplessi@nancy.inra.fr

†Present address:
Stéphane Hacquard, Department of Plant-Microbe Interactions, Max Planck Institute for Plant Breeding Research, Cologne, Germany

Most rust fungi have a complex life cycle involving up to five different spore-producing stages. The telial stage that produces melanized overwintering teliospores is one of these and plays a fundamental role for generating genetic diversity as karyogamy and meiosis occur at that stage. Despite the importance of telia for the rust life cycle, almost nothing is known about the fungal genetic programs that are activated in this overwintering structure. In the present study, the transcriptome of telia produced by the poplar rust fungus *Melampsora larici-populina* has been investigated using whole genome exon oligoarrays and RT-qPCR. Comparative expression profiling at the telial and uredinial stages identifies genes specifically expressed or up-regulated in telia including osmotins/thaumatin-like proteins (TLPs) and aquaporins that may reflect specific adaptation to overwintering as well numerous lytic enzymes acting on plant cell wall, reflecting extensive cell wall remodeling at that stage. The temporal dynamics of karyogamy was followed using combined RT-qPCR and DAPI-staining approaches. This reveals that fusion of nuclei and induction of karyogamy-related genes occur simultaneously between the 25 and 39 days post inoculation time frame. Transcript profiling of conserved meiosis genes indicates a preferential induction right after karyogamy and corroborates that meiosis begins prior to overwintering and is interrupted in Meiosis I (prophase I, diplonema stage) until teliospore germination in early spring.

Keywords: *Melampsora larici-populina*, obligate biotrophic fungus, rust lifecycle, teliospores, gene expression, microarray

INTRODUCTION

Rust fungi (basidiomycetes, pucciniales) are obligate biotrophs that belong to a monophyletic group containing ∼7000 species (Maier et al., 2003). These are one of the most widespread and devastating groups of plant pathogens that can infect monocotyledonous and dicotyledonous plants, including major crop species (Alexopoulos et al., 1996). For example, the emergence of the Ug99 race of the wheat stem rust *Puccinia graminis* f. sp. *tritici* is a serious threat to wheat production worldwide, causing massive crop losses (Singh et al., 2011; Fisher et al., 2012). The genomes of the rust fungi *P. graminis* f. sp *tritici*, infecting wheat and barberry and *Melampsora larici-populina*, infecting poplar and larch, have been sequenced and genome signatures related to their extreme parasitic lifestyle were unraveled (Duplessis et al., 2011a). *M. larici-populina* causes devastating damage on poplar plantations that are used for wood production, carbon sequestration, biofuel production, and phytoremediation (Polle et al., 2013).

Like many other rust fungi, *M. larici-populina* exhibits a complex heteroecious macrocyclic lifecycle completed on two different hosts (poplar, the telial host and larch, the aecial host) and involves five spore-producing stages (Hacquard et al., 2011a). On poplar, the fungus successively differentiates three distinct sporulation structures. The first one, produced throughout spring and summer is called the uredinium and corresponds to a yellow-orange pustule that is differentiated within 7 days on the abaxial surface of poplar leaves. This structure, which releases large amounts of dikaryotic urediniospores, is responsible for massive epidemics in poplar plantations in Europe and worldwide since successive cycles of uredinia formation occur throughout summer (Barrès et al., 2012). On senescent leaves in autumn, the fungus differentiates highly-melanized pustules called telia that produce the overwintering spore form, the dikaryotic teliospores. In spring, teliospores that have undergone karyogamy and meiosis in decaying poplar leaves germinate and produce a new structure called basidium that releases four haploid basidiospores. The basidiospores infect larch needles to form pycnia, which produce haploid pycniospores. After cross-fertilization of pycnia by pycniospores, the fungus forms aecia, which produce dikaryotic aeciospores that infect again poplar leaves. Interestingly, although most of *M. larici-populina* populations undergo host alternation on larch under temperate climates, asexual lineages that overwintered asexually on poplar were recently reported (Xhaard et al., 2011).

Teliospore ontogeny has been described in several rust fungi genera including *Cronartium*, *Chrysomyxa*, *Puccinia*, and *Melampsora* (Longo et al., 1979; Moriondo et al., 1989; Mims et al., 1996; Berndt, 1999; Driessen et al., 2005). Maturation of the teliospores is marked by an increase of the cytoplasmic density, an accumulation of lipid droplets and glycogen-like structures, and disappearance of vacuoles (Harder, 1984; Mendgen, 1984). These features may reflect a specific adaptation contributing to teliospore survival during winter. In addition, the presence of chitin in the spore wall was demonstrated using wheat germ agglutinin gold labeling (Mims and Richardson, 2005). Once produced in telia, teliospores undergo karyogamy and meiosis implicating that these spores are an important source of genetic diversity (Schumann and Leonard, 2000). In the rust fungus *P. graminis*, it has been shown that all teliospores have fusion nuclei 42 days post inoculation (dpi) and that meiosis is blocked in prophase I at the diplonema stage when spores enters dormancy (Boehm et al., 1992). Consistent with this, ultrastructure analysis of teliospores revealed that meiotic chromosome pairing (synaptonemal complexes, prophase I) is initiated shortly after karyogamy in *Gymnosporangium* (Mims, 1977, 1981) and that *Puccinia malvacearum* teliospores are in late diplonema stage when they differentiate metabasidia (O'Donnell and McLaughlin, 1981). Taken together, these data suggest that for rust species with overwintering telia, meiosis begins prior to overwintering and is interrupted in Meiosis I (prophase I, diplonema stage) until teliospore germination in early spring.

Despite the importance of the telial stage for the rust life cycle, almost nothing is known about the fungal genetic programs that are activated in this overwintering structure. Indeed, most of the recent molecular approaches to understand rust biology have focused on the analysis of gene expression in urediniospores and during host infection at the uredinial stage using Sanger EST sequencing (for a complete list, see recent review by Duplessis et al., 2012), microarrays (Jakupović et al., 2006; Hacquard et al., 2010; Duplessis et al., 2011a,b) or RNA-Seq (Fernandez et al., 2012; Petre et al., 2012; Cantu et al., 2013; Garnica et al., 2013). Recently however, ESTs libraries generated from different spore types of the wheat leaf rust fungus *Puccinia triticina* revealed a high proportion of EST sequences (87% of 697 ESTs) uniquely detected in teliospores compared to all other sampled stages (Xu et al., 2011).

In the present study, we used whole-genome custom oligoarrays to monitor fungal gene expression profiles in telia of *M. larici-populina* collected on senescent poplar leaves before overwintering. Comparative expression profiling at the telial and uredinial stages identifies genes that are only or preferentially expressed in telia, suggesting their contribution to a specific genetic program. We further investigated some candidate genes that might be involved in the teliospore differentiation process.

MATERIALS AND METHODS

PLANT GROWTH CONDITIONS AND INOCULATION PROCEDURES

Samples corresponding to resting urediniospores (USP), infected poplar leaves (INF; 96 h post inoculation, hpi), and uredinia (URE; 168 hpi) were previously described (Duplessis et al., 2011b). For microarray analysis, senescent leaves of the "Beaupré" poplar cultivar naturally infected by *M. larici-populina* and presenting dark telial pustules (telia, TEL) were harvested in October 2010 at a poplar nursery (Centre INRA Nancy Lorraine, Champenoux, 54, France). For RT-qPCR and karyogamy process analyses, development of *M. larici-populina* (strain 98AG31) telia was monitored in the susceptible poplar cultivar *Populus trichocarpa* × *Populus deltoides* "Beaupré" (compatible interaction). Resting urediniospores were collected on leaves of susceptible *P. deltoides* × *Populus nigra* "Robusta" and plant inoculation procedures were performed using the same inoculum dose of 100,000 urediniospores/ml and strictly identical culture conditions as those previously described (Rinaldi et al., 2007). Samples were harvested at intervals corresponding to the biotrophic growth (4 days post inoculation, dpi), the formation of uredinia (11 dpi), and the formation and the maturation of telia (18, 25, 32, 39, and 46 dpi). Infected leaves were incubated at 20°C until uredinia formation (11 dpi) and then transferred and maintained at 10°C to induce telia formation. At each time-point, harvested samples were immediately fixed in 4% (wt/vol) paraformaldehyde (PFA) for microscopy analyses or snap frozen in liquid nitrogen and kept at −80°C for further nucleic acid isolation. The time course was performed in triplicate.

MICROSCOPY

After fixation (3 h, 4°C) in 4% PFA (wt/vol) prepared in phosphate buffer saline (PBS), samples were washed twice with PBS and then embedded in 6% agarose (wt/vol). Transversal sections (15 μm) of INF were cut using a vibratome VT1000S (Leica, Nanterre, France) and directly transferred onto a microscopic slide. Sections were mounted in an antifade reagent with DAPI (Molecular Probes) and observed using the Palm Laser Micro dissection Microscope (Zeiss, Bernried, Germany) using the 40× objective. The number of fused and non-fused nuclei were analyzed for each sample in ∼100 teliospores from a single biological replicate.

RNA ISOLATION AND cDNA SYNTHESIS

Total RNA were isolated with the RNeasy Plant Mini kit (Qiagen, Courtaboeuf, France) from 1 mg of resting spores (USP) and from 100 mg of infected leaf tissues (INF, URE) as previously described (Duplessis et al., 2011b), including a DNase I (Qiagen) treatment according to the manufacturer's instructions to eliminate traces of genomic DNA. Total RNA from the telial stage (TEL) were isolated from 100 mg of leaf tissue using the same protocol used for the USP, INF, and URE samples. Electrophoretic RNA profiles were assessed with an Experion analyzer using the Experion RNA Standard-sens analysis kit (Bio-Rad, Marnes la Coquette, France) (**Figure S1**). For oligoarrays experiment, total RNA from the telial stage (TEL) were subjected to a single round of amplification using the MessageAmp™ II aRNA amplification kit (Ambion, Austin, TX, USA) as previously described for the USP, INF et URE samples (Duplessis et al., 2011b). RNA amplification generated more than 50 μg of amplified RNA (aRNA) and aRNA profiles were verified using the Experion analyzer and Experion RNA Standard-Sens analysis kit (Bio-Rad). Double-stranded cDNA were synthesized from 2.5 μg of aRNA using the Superscript™ Double-Stranded cDNA Synthesis Kit

(Invitrogen, Cergy Pontoise, France) according to the NimbleGen user protocol. Single dye labeling of samples, hybridization procedures, and data acquisition were performed at the NimbleGen facilities (NimbleGen Systems, Reykjavik, Iceland) following their standard protocol. For the RT-qPCR analysis, isolation of total RNA was performed using the RNeasy Plant Mini kit (Qiagen, Courtaboeuf, France) from 50 mg of infected leaf tissues (4–46 dpi) and a DNase I treatment was included to eliminate traces of genomic DNA (Qiagen). Electrophoretic RNA profiles were assessed with an Experion Analyzer using the Experion RNA Standard-sens analysis kit (Bio-Rad, Marnes la Coquette, France) (**Figure S1**).

CONSTRUCTION OF *M. larici-populina* EXON OLIGOARRAY

The *M. larici-populina* custom-exon expression oligoarray (4 × 72 K) manufactured by Roche NimbleGen Systems Limited (Madison, WI) (http://www.nimblegen.com/products/exp/index.html) contained four independent, non-identical, 60-*mer* probes per gene model coding sequence (NCBI Gene Expression Omnibus, GEO platform GPL10350). Included in the oligoarray were 17,556 coding sequences, 1063 random 60-*mer* control probes and labeling controls (Duplessis et al., 2011a). The 17,556 coding sequences correspond to the initial *M. larici-populina* gene annotation set at the Joint Genome Institute (JGI). The current annotation of the *M. larici-populina* genome is of 16,400 genes (12/06/2013), of which 13,093 (80%) were represented on the array used in the study. Oligonucleotide probes that presented a risk of cross-hybridization with poplar transcripts (i.e., fluorescence signal over the background level when arrays were hybridized with non-inoculated poplar leaf cDNA) or between transcript species expressed by different genes from a same gene family (i.e., probes with more than 90% homology between two transcripts) were not considered in our analysis.

MICROARRAY DATA ANALYSIS

Microarray probe intensities were quantile normalized across chips. Average expression levels were calculated for each gene from the independent probes on the array and were used for further analysis. Raw array data were normalized using the ARRAYSTAR software (DNASTAR, Inc. Madison, WI, USA). A transcript was deemed expressed when its signal intensity was three-fold higher than the mean signal-to-noise threshold (cut-off value) of 1063 random oligonucleotide probes present on the array. All expression assays were conducted on three independent biological replicates. A Student *t*-test with false discovery rate (FDR) (Benjamini-Hochberg) multiple testing correction was applied to the data (ARRAYSTAR software). Transcripts with a significant *p*-value (<0.05) and more than a 2-fold change in transcript level were considered as differentially expressed. The expression datasets are available at the NCBI GEO as serie #GSE49099.

HEATMAPS OF GENE EXPRESSION PROFILES

Heatmaps of *M. larici-populina* gene expression profiles were generated using the Genesis expression analysis package (Sturn et al., 2002). To derive expression patterns of genes in the different fungal developmental stages (USP, INF, URE, TEL), log2 expression ratios (Relative Expression Indexes, REI) were calculated between the normalized expression level for a given gene at a given fungal developmental stage and the geometrical mean expression level calculated across all 4 fungal developmental stages (Duplessis et al., 2011b). Functional gene annotation was based on Blastp search against the Swissprot database (Bairoch and Apweiler, 2000).

KOG ENRICHMENT ANALYSIS

We obtained KOG (eukaryotic orthologous groups) (Tatusov et al., 2003) annotation of each *M. larici-populina* gene by using RPSBLAST against the KOG database (*e*-value < 1e-5). Each gene was classified according to KOG functional classification using custom perl scripts. Over-represented KOG categories among telia- or uredinia-induced genes were calculated relative to the global gene distribution. The significance of over-represented functional KOG categories were evaluated using the Fisher's Exact Test ($p < 0.05$).

RT-qPCR

To monitor karyogamy- and meiosis-related transcript expression profiles during telia formation, 16 genes were selected for RT-qPCR assays (**Table S1**). Specific primers amplifying fragments ranging from 152 to 247 were designed for each gene using Primer 3 (Rozen and Skaletsky, 2000). Absence of cross annealing was checked in the *M. larici-populina* (http://genome.jgi-psf.org/Mellp1/Mellp1.home.html) and *P. trichocarpa* (http://www.phytozome.net/) genome sequences using the blastn algorithm. First-strand cDNA synthesis was performed using 500 ng total RNA and cDNA were amplified strictly following procedures described in Hacquard et al. (2012). Transcript expression levels were normalized with the *M. larici-populina* reference genes a-tubulin (*Mlp-aTUB*) and elongation factor (*Mlp-ELF1*a) as previously described (Hacquard et al., 2011b).

RESULTS

DEVELOPMENT OF UREDINIA AND TELIA WITHIN POPLAR LEAVES

Uredinia and telia that are formed by the rust fungus *M. larici-populina* on the susceptible poplar cultivar "Beaupré" are represented in **Figure 1**. During summer, yellow-orange uredinia pustules are formed on the abaxial surface of poplar leaves about 1 week after urediniospore landing on poplar leaf epidermis (**Figure 1A**). Early in autumn, the asexual uredinial cycle stops and black telia pustules (**Figure 1B**) start to differentiate on the adaxial surface of poplar leaves. Whereas uredinia continuously release important amounts of urediniospores that are dispersed over large distances by wind to cyclically infect poplar throughout summer (**Figure 1C**), teliospores are produced only once a year in telia and those are tightly encapsulated between the plant epidermis and the palisade mesophyll of poplar leaves (**Figure 1D**). This structure provides adequate conditions for teliospore overwintering.

M. larici-populina GENE EXPRESSION PROFILING IN TELIA AND COMPARISON WITH THE UREDINIAL STAGE

To identify *M. larici-populina* genes significantly regulated (*p*-value <0.05; $-2 <$ fold change > 2) or specifically expressed

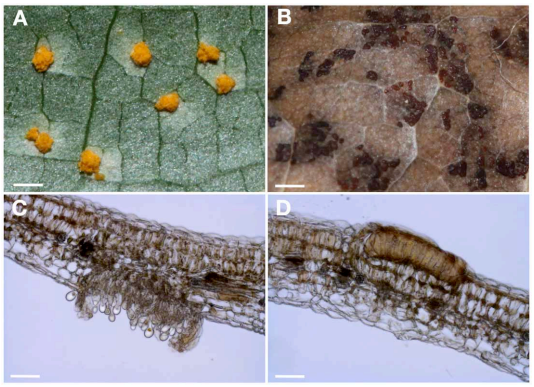


FIGURE 1 | *M. larici-populina* uredinia and telia on poplar leaves. (A) Macroscopic picture of uredinia formed 7 days post inoculation by the virulent *M. larici-populina* 98AG31 strain on the abaxial surface of leaves of the "Beaupré" poplar cultivar. **(B)** Macroscopic picture of telia observed early in autumn on the adaxial surface of senescent "Beaupré" leaves naturally infected by *M. larici-populina*. Scale bar = 1 mm. **(C)** Transversal section of a "Beaupré" leaf infected by the virulent *M. larici-populina* 98AG31 strain showing a mature uredinium releasing urediniospores on the abaxial surface. **(D)** Transversal section of a "Beaupré" leaf naturally infected by *M. larici-populina* showing a mature telia on the adaxial surface into which teliospores are tightly encapsulated. Scale bar = 50 μm.

at the telial stage, we used a whole-genome custom oligoarray onto which oligonucleotides matching to 13,093 genes of *M. larici-populina* were spotted (Duplessis et al., 2011a). We compared transcript levels measured in telia (TEL) with those detected in three previously described samples related to the uredinial stage (urediniospores: USP, infection process: INF, uredinia: URE) (Duplessis et al., 2011b). Principal component analysis of transcript expression levels measured in USP, INF, URE, and TEL showed very good consistency between the three biological replicates of collected telia (**Figure 2A**). Furthermore, it appears clearly that distinct genetic programs are expressed by the rust fungus at the four developmental stages surveyed. USP, INF, and TEL samples are distributed apart by the PCA and URE samples have an intermediate position, which could suggest overlapping expression patterns between the different types of samples (**Figure 2A**).

GENES SPECIFICALLY EXPRESSED IN TELIA

Among the 13,093 genes present on the oligoarray, 10,412 were expressed in at least one of the four considered developmental stages (**Figure 2B**, **Table S2**). Among those, 6569 were expressed in all four fungal developmental stages and 9588 were expressed in telia including 395 telia-specific genes. Interestingly, this is the largest number of specific transcripts found for a given stage since only 200 were detected for USP, 181 for INF, and 147 for URE suggesting that specific functions are activated in the *M. larici-populina* overwintering spore-producing structure. **Table 1** summarizes the set of telia-specific genes showing high level of transcript accumulation at that stage (gene expression level >800). Interestingly, a gene encoding a saccharopine dehydrogenase, involved in the biosynthesis of the amino

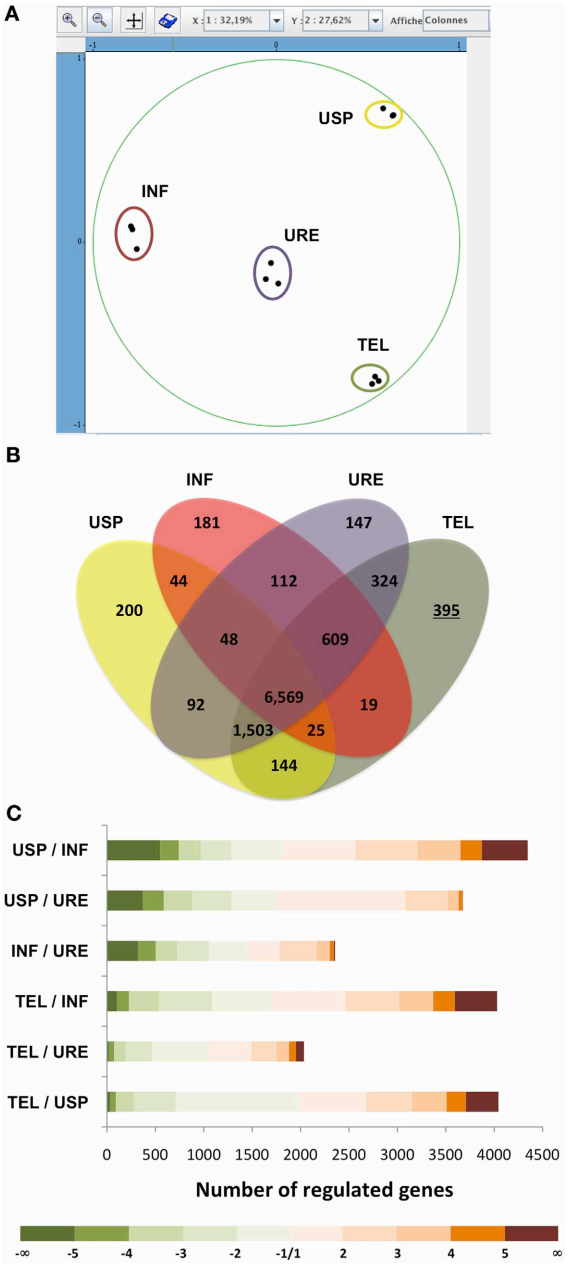


FIGURE 2 | Microarray expression analysis of telia and comparison with expression data of other fungal developmental stages. (A) Principal component analysis (PCA) of *M. larici-populina* transcript levels measured in urediniospores (USP), during infection (INF), in uredinia (URE), and in telia (TEL) using custom oligoarrays (three biological replicates per stage were used for the PCA). Expression level of each gene assessed in a given biological situation and a given biological replicate was reported to the mean expression level calculated for the 12 hybridizations (six conditions × three replicates) and was $\log_{10}$-normalized before proceeding with PCA. The PCA plot places biological conditions along the two axes (X and Y) explaining 32.19 and 27.62% of the variance observed within samples. **(B)** Venn diagram showing the number of *M. larici-populina* genes expressed in each condition. The underlined number corresponds to the number of genes specifically expressed in telia. **(C)** Number of genes significantly regulated between the analyzed fungal developmental stages [$\log_2$ Fold-Change (FC) >1, $p < 0.05$]. Color coding corresponds to $\log_2$FC.

acid L-Lysine through the α-aminoadipate pathway (Xu et al., 2006), ranged among the most highly expressed telia-specific gene set (Expression level >10,000, **Table 1**). The high proportion of genes encoding unknown proteins (47/68), including 11 Small Secreted Proteins (SSPs), suggests that telia formation and teliospores production is mostly driven by complex and largely unknown mechanisms. Nevertheless, genes encoding several carbohydrate Active Enzymes (CAZymes, Cantarel et al., 2009) including a pectinesterase and a cutinase (*Mlp-93655*, *Mlp-123715*) (Carbohydrate esterase families 8 and 5, respectively), an alpha-1,6-mannosyltransferase (*Mlp-27594*, glycosyl transferase family 32) and an endoglucanase (*Mlp-95634*, glycolsyl hydrolase family 12) could be identified (**Table 1**). Consistent with the fact that karyogamy and meiosis occur during teliospore maturation, several meiotic and karyogamy related genes were also specifically expressed in the telial structure (**Table 1**) such as those encoding the meiotic recombination protein rec8, the cell division control protein 15, the meiotic nuclear division protein 1 as well as the nuclear fusion protein Kar5 and Kar9 (*Mlp-93153*, *Mlp-94329*, *Mlp-106571*, *Mlp-112713*, *Mlp-94206*).

GENES REGULATED IN TELIA COMPARED WITH OTHER *M. larici-populina* DEVELOPMENTAL STAGES

By comparing transcript expression levels in TEL and URE, we identified 2035 genes significantly regulated ($-2 <$ fold-change >2, p-value < 0.05) between the two spore-producing structures (**Figure 2C**, **Table S2**). A relatively similar set of regulated genes was detected between the INF and URE conditions (<2500) but larger numbers were detected between USP and INF, USP, and URE, TEL and INF, or TEL and USP (>3600) (**Figure 2C**). This result is in accordance with the principal component analysis (**Figure 2A**) and supports the idea that similar sets of genes are shared by the genetic programs triggered in the rust fungus *M. larici-populina* for the production of both telia and uredinia. Despite potential common features, the substantial number of genes identified as significantly regulated in TEL compared with URE (1003 up- and 1032 down-regulated genes) indicates that specific pathways may be activated and could explain the structural and functional differences characterizing the two structures (**Figure 2C**).

GENES UP-REGULATED IN TELIA COMPARED WITH UREDINIA

A functional KOG analysis of the 1003 significantly up-regulated genes in TEL compared with URE reveals significant enrichment for gene categories related to defense mechanism, inorganic ion transport and metabolism, secondary metabolites biosynthesis transport and catabolism as well as general function (**Figure 3A**). Global expression profiling of these up-regulated genes across all stages revealed that most transcripts significantly accumulated in TEL compared with URE are preferentially expressed in telia except for a cluster of genes that also show high transcripts accumulation in USP (**Figure 3B**). Interestingly, genes belonging to this cluster are not or barely detected during infection (INF) and several encode transporters (**Figure 3B**), including a calcium-transporting ATPase (*Mlp-86276*), an aquaporin (*Mlp-26257*), an MFS efflux pump (*Mlp-72481*), an MFS general substrate transporter (*Mlp-42763*), a pleiotropic drug resistance transporter (*Mlp-50834*), a quinate permease (*Mlp-47943*), and a sulfate permease (*Mlp-39732*) (**Figure 3C**). A total of 113 genes encoding SSPs of unknown function previously categorized as putative candidate effectors (Hacquard et al., 2012) are induced in TEL compared with URE (**Figure 3B**). Considering the development of telia on senescent leaf tissues, these are most likely not effectors engaged in the manipulation of host cell immunity (Win et al., 2012). Alternatively, they may also have dual functions during the rust lifecycle such as recently reported for the rust transferred protein 1. In addition to its ability to be transferred in the host cytoplasm (Kemen et al., 2005), RTP1 is also capable of fibril formation (Kemen et al., 2013). Thus, some candidate effectors may also have a structural role during teliospores production and maturation. Investigation of the set of significantly regulated genes (fold-change >4) by a functional annotation based on Blastp search against the Swissprot database highlights major biological processes induced during telia formation (**Figure 3C**). Several genes potentially related to the overwintering process were identified including 5 aquaporins (*Mlp-106246*, *Mlp-79395*, *Mlp-84885*, *Mlp-26257*, *Mlp-117123*), 3 osmotin/thaumatin-like proteins (TLPs; *Mlp-76068*, *Mlp-79324, Mlp-85787*), a trehalose-like protein (*Mlp-67317*), a calcineurin temperature suppressor (*Mlp-71212*), and a calcium-transporting ATPase (*Mlp-48992*). Consistent with the functional KOG analysis (**Figure 3A**), secondary metabolites transport appears to be active in telia since transcripts encoding 3 MFS toxin efflux pumps (*Mlp-106478*, *Mlp-108871*, *Mlp-72481*) and a pleiotropic drug resistance transporter (*Mlp-50835*) were strongly accumulated. In addition to the four afore-mentioned telia-specific karyogamy and meiosis-related genes (**Table 1**), the meiosis-specific protein HOP1, a kinase-like protein (related to cdc15), and the meiotic recombination protein SPO11 are also regulated between TEL and URE, with higher transcript levels detected in TEL. A significant number of genes (8 in total) encoding plant cell wall degrading enzymes dominate among the most highly up-regulated genes in telia supporting an extensive plant cell wall remodeling during telia development. These include CAZymes targeting cellulose (GH12 and GH61 families), hemicellulose (GH10 and GH27 families), pectin (CE8 and GH28 families), and hemicellulose/pectin (GH43 family). Genes encoding two multi-copper oxidase laccase-like proteins, previously detected in a *P. triticina* teliospores EST library (Xu et al., 2011), were also strongly up-regulated in the overwintering telial structure.

GENES DOWN-REGULATED IN TELIA COMPARED WITH UREDINIA

A functional KOG analysis of the 1032 significantly down-regulated genes in TEL compared with URE reveals a significant enrichment for gene categories related to carbohydrate transport and metabolism and unknown proteins (**Figure 4A**). Transcript profiling of these down-regulated genes revealed that most are preferentially expressed in uredinia but a subset also shows a higher transcript accumulation in USP or during the infection process (INF) (**Figure 4B**). Importantly, the cluster of genes showing high transcript accumulation during the infection process is particularly enriched with SSPs suggesting they may encode biotrophy-associated effectors involved in poplar

Table 1 | *M. larici-populina* genes highly and specifically expressed in telia.

Protein_ID[a]	Expression level[b]				Definition	Cat[c]	Length[d]	SP[e]	Blastp *Pgt*[f]
	USP	INF	URE	TEL					
86396	87	47	76	14,466	Hypothetical protein		184		PGTG_08052
101624	28	27	27	12,711	Saccharopine dehydrogenase	A	434		PGTG_03759
87547	38	22	25	11,939	Hypothetical SSP, PR-1-like protein	B	280	Y	PGTG_07743
101938	60	41	33	11,876	Hypothetical SSP	B	133	Y	–
93655	26	77	45	11,548	Pectinesterase (CE8)	C	316	Y	PGTG_08012
36325	102	28	51	10,214	Homogentisate 1,2-dioxygenase	A	414		–
107118	22	22	23	9702	Hypothetical SSP, PR-1-like protein	B	236	Y	PGTG_16765
27934	34	28	28	8747	Glycosyl transferase (GT8)	C	188		–
88840	65	157	66	7067	Hypothetical protein		519		PGTG_11713
27594	59	53	44	7029	Alpha-1,6-mannosyltransferase (GT32)	C	249		–
58538	32	31	90	6606	Short-chain dehydrogenase	A	248		PGTG_15283
104079	26	23	37	6478	Hypothetical protein		337		PGTG_03998
110542	34	97	49	5369	Hypothetical SSP	B	133		–
95634	36	38	29	5193	Endoglucanase (GH12)	C	179		PGTG_03891
108357	24	61	69	4899	Hypothetical protein		508		PGTG_16954
110660	69	38	41	4587	Hypothetical protein		513		PGTG_15396
92814	26	22	24	4576	Hypothetical protein		768		PGTG_06712
93153	34	34	44	4305	Meiotic recombination protein rec8	D	774		PGTG_02404
87054	71	66	80	4229	Hypothetical protein		279		PGTG_13349
59662	30	21	23	3898	Hypothetical protein		326		–
93477	25	38	27	3199	Hypothetical protein		481		PGTG_02434
112713	67	38	76	3023	Nuclear fusion protein KAR5	D	736		PGTG_16428
101708	31	146	68	2622	Aldehyde dehydrogenase	A	496		PGTG_15008
104617	29	34	39	2541	Hypothetical protein		308		–
86447	34	30	23	2401	Hypothetical protein		573		PGTG_18083
92775	28	32	28	2392	Hypothetical protein		269		–
109764	76	102	62	2269	Hypothetical protein		482		–
85892	48	31	30	2239	Hypothetical protein		402		PGTG_13349
84888	30	34	51	2146	Hypothetical protein		626		PGTG_00821
92678	60	74	30	2140	Hypothetical protein		419		PGTG_01636
123561	32	31	76	2077	Hypothetical SSP	B	133	Y	–
103910	51	71	37	2072	Hypothetical protein		455		PGTG_10254
62289	45	45	56	1924	Hypothetical SSP	B	265	Y	PGTG_06969
61331	63	26	41	1883	Hypothetical protein		197		–
26257	102	38	48	1842	Aquaporin (MIP)	E	263		PGTG_02867
89049	45	54	47	1767	Hypothetical protein		447		PGTG_03343
60216	34	26	49	1702	Hypothetical SSP	B	214	Y	–
94329	26	37	51	1573	Cell division control protein 15	D	451		PGTG_16937
59440	70	33	31	1553	Hypothetical protein		336		PGTG_09936
113347	44	31	59	1536	Hypothetical SSP	B	136	Y	–
123715	44	81	80	1520	Cutinase (CE5)	C	351	Y	PGTG_01091
109924	29	24	18	1449	Hypothetical protein		297		–
110784	69	56	33	1442	Hypothetical protein		518		PGTG_15984
66126	29	28	22	1426	Carbohydrate esterase (CE16)	C	262		PGTG_18191
106798	23	39	31	1419	Hypothetical secreted protein		325		–
101151	27	19	19	1362	Choline dehydrogenase	A	594		PGTG_18542
110516	42	41	67	1336	Hypothetical protein		687	Y	–
90260	83	41	42	1279	Homeobox protein		116		PGTG_07066
64744	108	218	83	1274	Hypothetical protein		290		–
91904	102	227	69	1269	Hypothetical protein		363		–
93339	31	26	32	1249	Hypothetical protein		435		PGTG_01619

(Continued)

Table 1 | Continued

Protein_ID[a]	Expression level[b]				Definition	Cat[c]	Length[d]	SP[e]	Blastp *Pgt*[f]
	USP	INF	URE	TEL					
90367	49	54	41	1216	Hypothetical protein		589		PGTG_04717
103627	40	18	26	1192	Hypothetical protein		167		–
102200	36	29	83	1174	Hypothetical SSP	B	153	Y	–
61074	29	43	66	1124	Hypothetical protein		454		PGTG_03998
63861	31	27	28	1105	Hypothetical protein		485		–
65250	20	22	19	1071	Hypothetical SSP	B	195	Y	PGTG_06052
70334	44	35	26	1061	Hypothetical protein		109		–
113400	45	46	31	1011	Hypothetical protein		498		PGTG_19950
63565	78	201	99	989	Hypothetical protein		176		PGTG_11018
106571	84	33	42	870	Meiotic nuclear division protein 1	D	206		PGTG_18915
25325	33	57	33	869	L-gulonolactone oxidase	A	442		PGTG_07192
101135	106	59	50	866	Hypothetical protein		220		–
94206	108	95	53	850	hypothetical protein (distantly related to KAR9)	D	656	Y	–
107936	45	55	75	821	Hypothetical SSP	B	160	Y	–
105800	55	45	73	810	Hypothetical protein		342		PGTG_1213

[a] *Protein ID number of corresponding best gene model in the M. larici-populina JGI genome sequence.*

[b] *Normalized transcript levels are presented. Transcripts levels in urediniospores (USP), infection process (INF), and uredinia (URE) samples are below our arbitrary cut-off (less than three-fold higher than the mean signal-to-noise threshold) and were thus considered as not expressed. TEL, Telia.*

[c] *Cat, Category: A, metabolism enzymes; B, small secreted proteins (SSPs); C, carbohydrate-active enzymes; D, meiosis and karyogamy-related genes; E, transporters.*

[d] *Protein length (amino acids).*

[e] *SP, Signal Peptide. SPs were predicted according to SignalP v3.0.*

[f] *Pgt, Puccinia graminis f. sp. tritici (http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiHome.html).*

immunity manipulation (**Figure 4B**). By looking at the functional annotation of significantly regulated genes (fold-change < -4), we could identify processes that are activated in uredinia and switched-off in telia (**Figure 4C**). Among these, we identified 4 genes encoding mating-related proteins including two mating-type STE3 pheromone receptors, a pheromone-regulated multi-spanning membrane protein (Prm1) and a putative b mating type protein. Consistent with the high content of carotenoid in the urediniospores, two genes encoding carotenoid ester lipase precursors were also up-regulated during urediniospore production and release. Among the genes encoding transporters that are down-regulated in telia compared to uredinia (fold change < -4), we identified six MFS general substrate transporters, four oligopeptides transporters, two auxin efflux carriers, a monosaccharide transporter (related to *Uromyces fabae* HXT1, Voegele et al., 2001) and an aquaporin (**Figure 4C**). In addition, 14 genes encoding CAZymes and targeting the plant cell wall (GH7, CE8), the fungal cell wall (GH18, GH71), or both (GH2, GH5) showed altered transcript accumulation in TEL compared with URE suggesting that these genes are involved in urediniospore production, maturation processes, or release from host tissues (**Figure 4C**).

EXPRESSION PROFILING OF KARYOGAMY AND MEIOSIS-RELATED GENES DURING TELIA FORMATION

To determine the temporal dynamics of karyogamy and meiosis-related gene expression profiles during telia formation, a time-course interaction survey has been carried out between the *M. larici-populina* virulent strain 98AG31 and detached leaves of the susceptible poplar cultivar "Beaupré." Samples were harvested at intervals corresponding to the biotrophic growth phase (4 dpi), uredinia (11 dpi), and the formation and the maturation of telia (18, 25, 32, 39, and 46 dpi). Transversal sections of infected leaf tissues followed by DAPI-staining revealed that the first aggregates of fungal cells, corresponding to telia initials, are formed between 14 and 18 dpi. At 18 dpi all observed non-mature teliospores were dikaryotic (**Figures 5A,B**). The number of fused nuclei increase dramatically from 20% at 25 dpi to 80% at 32 dpi, indicating that karyogamy mainly takes place during this time frame (**Figures 5A,B**). Consistent with this, RT-qPCR expression profiling of the karyogamy-related genes *Kar5* and *Kar9* revealed that both genes are induced as soon as 25 dpi and their transcripts strongly accumulate at 32 dpi for *Kar9* and 39 dpi for *Kar5* (**Figure 5C**). All observed teliospores have fused nuclei at 46 dpi (**Figures 5A,B**). Importantly, all the conserved eukaryotic meiotic genes analyzed in this study (*Rec8*, *Mre11*, *Rad50*, *Rad51*, *MutS4*, *MutS5*, *Spo11*, *Mnd1,* and *Mlh1;* Malik et al., 2008) are induced in differentiated telia and their transcripts predominantly accumulate at late stages (i.e., 39 and 46 dpi, **Figure 5C**). This result suggests that with our experimental conditions, meiosis is initiated soon after karyogamy since at 39 dpi, more than 90% of the observed teliospores already contain fused nuclei (**Figure 5B**). Rad51, essential for double-strand break meiotic repair and Rec8, involved in sister chromatin cohesion, show a particular expression pattern with sustained transcript accumulation throughout the telia differentiation process, except at 46dpi where the transcripts were barely detected. This result may indicate that Rad51 and Rec8 are already produced during karyogamy, before meiosis takes place.

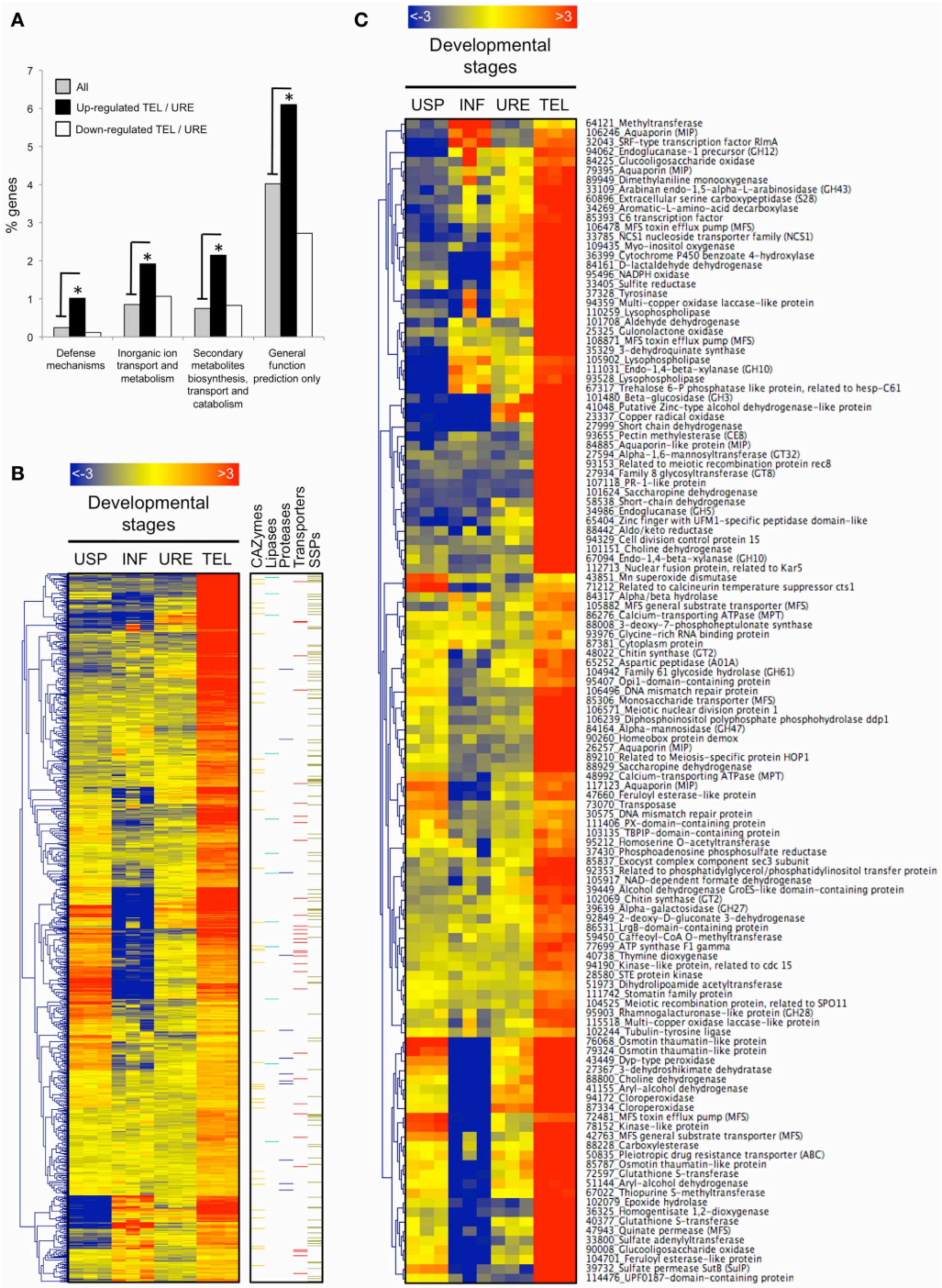


FIGURE 3 | Genes significantly up-regulated in telia compared with uredinia. (A) Over-represented KOG categories among telia-induced genes relative to the global gene distribution. Black and white bars correspond to the distribution of genes significantly up and down regulated in telia (TEL) compared with uredinia (URE) ($\log_2$FC >1, $p < 0.05$), respectively, into functional KOG categories. Gray bars correspond to the global gene distribution. Only the significantly over-represented functional KOG categories are presented. *indicate statistically significant differences (Fisher's Exact Test, $p < 0.05$) **(B)** Heatmap of transcript expression levels in all four fungal developmental stages for genes significantly up-regulated in telia compared with uredinia ($\log_2$FC >1, $p < 0.05$). Over-represented (red) or under-represented (blue) transcripts are depicted as $\log_2$ fold-changes relative to the mean expression level measured across all four stages. USP, urediniospores; INF, poplar infected leaves; URE, uredinia; TEL, telia. On the right side, genes belonging to five pathogenicity-related categories (carbohydrate active-enzymes, lipases, proteases, transporters, and small secreted proteins) are highlighted with color bars. **(C)** Among the genes presented in the panel **(B)**, only those showing a higher transcript induction in TEL compared with URE ($\log_2$FC >2, $p < 0.05$) and having a functional annotation (based on the swissprot database) are highlighted. JGI protein identification number and the associated function are indicated.

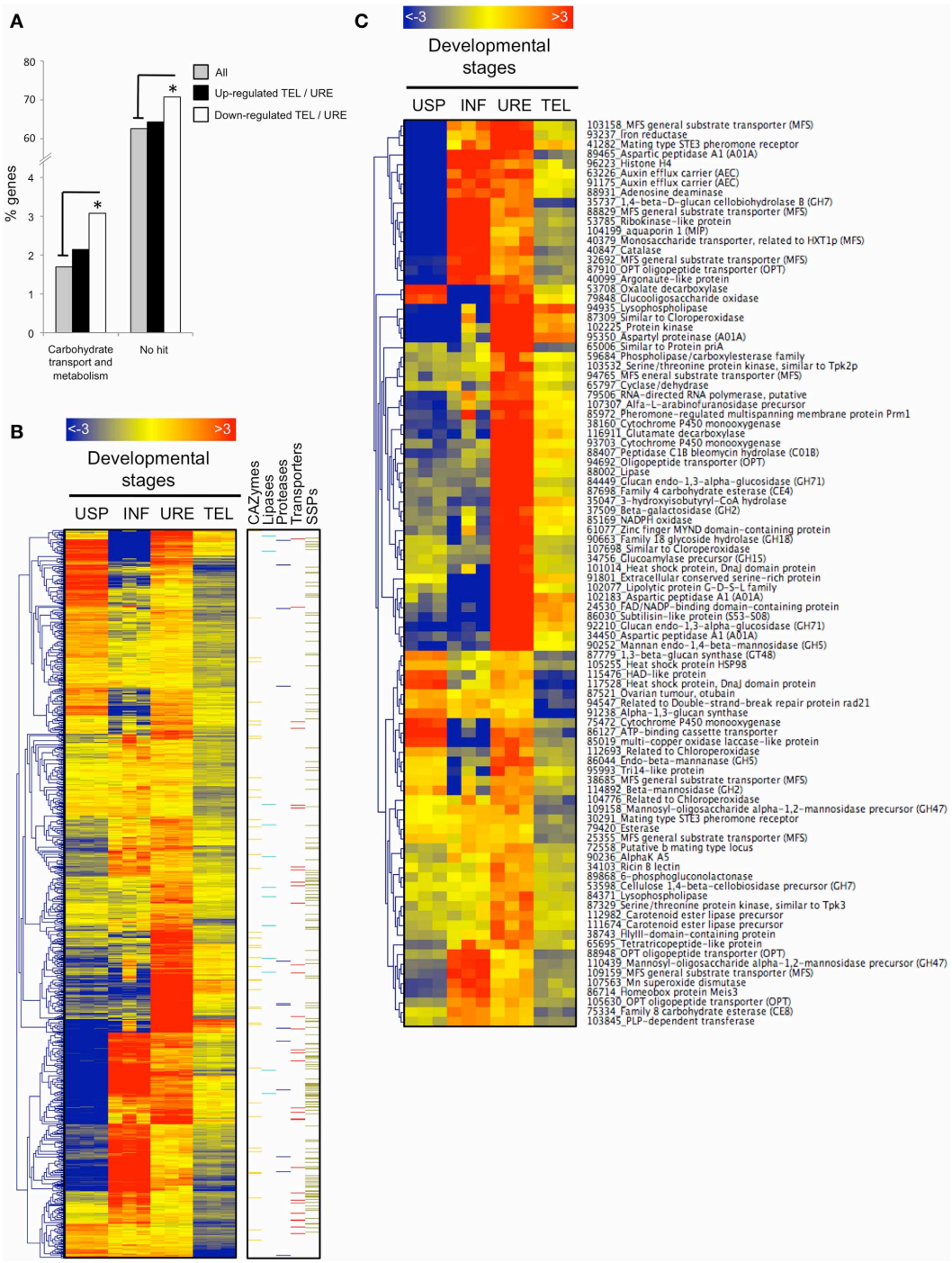


FIGURE 4 | Genes significantly down-regulated in telia compared with uredinia. (A) Over-represented KOG categories among telia-repressed genes relative to the global gene distribution. Black and white bars correspond to the distribution of genes significantly up and down regulated in telia (TEL) compared with uredinia (URE) ($\log_2$FC >1, $p < 0.05$), respectively, into functional KOG categories. Gray bars correspond to the global gene distribution. Only the significantly over-represented functional KOG categories are presented. *indicate statistically significant differences (Fisher's Exact Test, $p < 0.05$). Heatmap of transcript expression levels in all four fungal developmental stages for genes significantly down-regulated in telia compared with uredinia ($\log_2$FC < -1, $p < 0.05$). Over-represented (red) or under-represented (blue) transcripts are depicted as $\log_2$ fold-changes relative to the mean expression level measured across all four stages. USP, urediniospores; INF, poplar infected leaves; URE, uredinia; TEL, telia. On the right side, genes belonging to five pathogenicity-related categories (carbohydrate active-enzymes, lipases, proteases, transporters, and small secreted proteins) are highlighted with color bars. **(C)** Among the genes presented in the panel **(B)**, only those showing high transcript repression in telia compared with uredinia ($\log_2$FC < -2, $p < 0.05$) and having a functional annotation (based on the swissprot database) are highlighted. JGI protein identification number and the associated function are indicated.

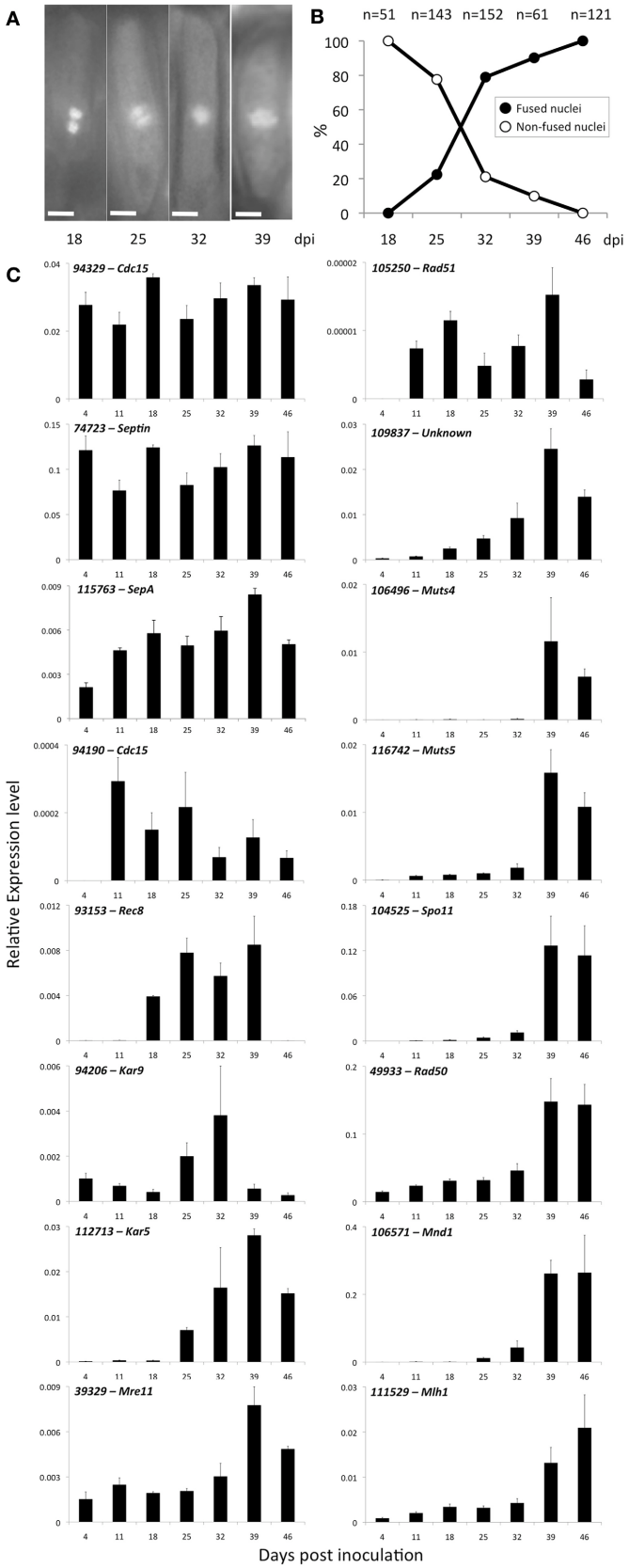


FIGURE 5 | Continued

FIGURE 5 | Karyogamy dynamics and meiotic-related gene expression profiles during telia formation. (A) Representative pictures of DAPI-stained teliospores nuclei during telia formation and maturation. From left to right, 18, 25, 32, and 39 days post inoculation. Scale bar = 5 μm. **(B)** Dynamics of karyogamy during telia formation and maturation. Percentage of fused and non-fused nuclei is indicated for each stage and the total numbers of counted teliospores are indicated above the graph. **(C)** Karyogamy, meiosis, and cytokinesis-related gene expression profiles monitored by RT-qPCR during telia formation and maturation. For each gene, expression levels were normalized with a-tubulin (*Mlp-aTUB*) and elongation factor (*Mlp-ELF1a*) reference genes.

DISCUSSION

The telial stage of rust fungi plays a crucial role in the fungal life cycle as it produces overwintering spores (i.e., teliospores) in which karyogamy and meiosis take place. Numerous ultrastructural studies have been conducted on telia and teliospores (Longo et al., 1979; Mendgen, 1984; Moriondo et al., 1989; Mims et al., 1996; Berndt, 1999; Driessen et al., 2005; Mims and Richardson, 2005), however, the functions activated by rust fungi during teliospores production and maturation remain poorly described. In the present study, we report in the poplar-poplar rust model pathosystem (Duplessis et al., 2009; Hacquard et al., 2011a) the transcriptome of telia produced by the rust fungus *M. larici-populina* using whole genome exon oligoarrays and RT-qPCR.

Our data show that the genetic program expressed in telia is more similar to the genetic program activated in uredinia than those observed in isolated resting urediniospores and during the biotrophic growth during poplar leaf infection, suggesting that overlapping sets of transcripts are important for both sporulation structures. Notably, we found at least two times more telia-specific genes compared to other investigated stages and most of them (69%) encode unknown proteins. Consistent with this, EST sequencing of pycniospores, aeciospores, teliospores, and urediniospores of the rust fungus *P. triticina* revealed that pycniospores and teliospores yield the largest sets of unique gene sequences (837 and 605, respectively), the majority of them (81 and 86%, respectively) having no functional annotation (Xu et al., 2011). Taken together, these results suggest that teliospore production involves largely unknown biological processes.

The telia structure plays a key role for spore survival over winter (Mendgen, 1984). Many transcripts are specifically expressed in telia or differentially up-regulated between telia and uredinia in *M. larici-populina* and they may be related to adaptation to cold temperatures and adverse winter conditions. Among these genes, several encode aquaporin water channels that may prevent osmotic damage of cells due to freezing. Previous studies have shown that aquaporins have desiccation and freeze tolerance functions in microorganisms, including bacteria, yeast, and fungi (Tanghe et al., 2006). Interestingly, one of the above-mentioned *M. larici-populina* aquaporin gene specifically expressed in telia (*Mlp-26257*) is orthologous to *AQY1*, a gene previously characterized together with *AQY2* in *Saccharomyces cerevisiae* and involved in a rapid, osmotically driven efflux of water during the freezing process that reduce intracellular ice crystal formation and resulting cell damage (Tanghe et al., 2002). Moreover, it has been also shown that AQY1 may also play a role in spore maturation in *S. cerevisiae* by allowing water outflow (Sidoux-Walter et al., 2004). We also identified three genes encoding osmotin/TLPs with a higher expression in telia than uredinia. In plants, osmotins belong the pathogenesis-related 5 family and have high sequence similarity with thaumatins that are sweet-tasting proteins (AnŽlovar and Dermastia, 2003). Genes encoding osmotins/TLPs are induced in plants in response to pathogens (Petre et al., 2011), cold (Kuwabara et al., 2002), drought (Jung et al., 2005), and osmotic stress (Singh et al., 1987). Induction of osmotins/TLPs during abiotic stress is often associated with osmotic adaptation in plant cells (Singh et al., 1987; Liu et al., 2010). Two osmotins/TLPs identify as highly expressed in *M. larici-populina* telia (*Mlp-76068*, *Mlp-79324)* correspond to small TLPs recently reported in basidiomycetes (Petre et al., 2011). These small TLPs belong to a monophyletic group with *Puccinia* TLPs indicating they may have evolved independently in pucciniales and plants (Petre et al., 2011). The role of TLPs in fungi has not been elucidated yet but our data suggest they might serve a possible role as an osmoprotectant in response to damaging effects of desiccation that can occur in teliospores during winter. Teliospores are highly melanized structures, and melanin is thought to provide protection against adverse environmental conditions. Genes encoding multi-copper oxidase laccase-like proteins, also identified in an EST library of *P. triticina* teliospores, are induced in telia and could be implicated in the biosynthesis of the melanin pigment (Xu et al., 2011). In basidiomycetes, recognition of mating partners is achieved through a pheromone/pheromone receptor system encoded by mating loci (Kronstad and Staben, 1997). In the smut fungus *Ustilago maydis*, the binding of pheromone to the receptor induces signaling cascades through specific mitogen-activated protein kinases pathways, and it is also marked in mating partners by the formation of conjugation tubes as well as G2 cell cycle arrest which ensure a synchronous stage of the cell cycle prior further developmental stages (Brefort et al., 2009). In the present case, we noticed that transcript levels from *M. larici-populina* mating loci genes, including pheromone receptor STE3 genes and a putative b mating loci gene are higher at late stages of plant colonization (i.e., formation of new urediniospores in the plant mesophyll) than in resting urediniospores or telia, although higher in telia than in resting urediniospores. A similar transcript profile is observed for the pheromone-regulated multispanning membrane protein Prm1 gene which is involved in plasma membrane fusion events during mating (Heiman and Walter, 2000). It is tempting to speculate that these mating-related genes could play a role in signaling during the rust fungus spore development and/or the control of cell cycle progression and cell fusion during formation of sporogenous hyphae and urediniospores.

Karyogamy and meiosis are crucial cellular processes that take place in teliospores. They play a fundamental role in generating genetic diversity by promoting recombination between chromosome homologs. In fungi, meiosis can drive genome plasticity and facilitates rapid adaptation to changing environments

(Wittenberg et al., 2009; Goodwin et al., 2011) and it is a crucial process for pathogenic rust fungi to overcome R-gene mediated host disease resistance by diversification of virulence effectors. Karyogamy monitoring during teliospore formation and maturation revealed that teliospore initials are formed between poplar epidermal and palisade parenchyma cells around 16 dpi. DAPI-staining of teliospore nuclei also indicates that karyogamy is a dynamic process that mainly takes place between 25 and 39 dpi. Consistent with this, a previous study has shown that when *Populus tremula* leaves begin to wither, marginal teliospores of telia formed by *Melampsora pinitorqua* are in a dikaryotic stage whereas the more central ones are already in the diploid stage (Longo et al., 1979). *M. larici-populina* karyogamy-related genes *Kar5* and *Kar9* are both transiently induced during telial development from 25 dpi and their transcripts accumulate at 32 dpi for *Kar9* and 39 for *Kar5*, corroborating their implication in the karyogamy process. From 90 to 100% of the analyzed teliospores have fused nuclei at 39 and 46 dpi, respectively. As the microscopic observation of nuclei during karyogamy has been carried out only on a single biological replicate, we cannot exclude that slights variations may occur when analysing additional replicates. However, our results are consistent with previous results in *P. graminis* showing that all DAPI-stained teliospore protoplasts have fused nuclei at 42 dpi (Boehm et al., 1992). Several transcripts encoding meiosis-related genes are induced in *M. larici-populina* telia. RT-qPCR expression profiles of conserved meiosis-related genes during telia differentiation revealed transcripts accumulation between 39 and 46 dpi, indicating that meiosis occurs soon after karyogamy in the experimental conditions used in the study. This observation may differ under natural conditions with decreasing temperature during autumn. Consistent with the fact that meiosis is already initiated at 39 dpi, a spotty DAPI-staining was observed for most nuclei at that stage (data not shown). Among the conserved meiosis genes analyzed, Spo11 is a transesterase that creates DNA double strand breaks in homologous chromosomes (meiotic prophase 1, leptonema stage) (Keeney et al., 1997), Hop1 is a protein is required for synaptonemal complex formation (meiotic prophase 1, zygonema stage) (Aravind and Koonin, 1998), Mnd1 is a protein that ensure accurate and efficient meiotic interhomolog repair (meiotic prophase 1, pachynema stage) (Gerton and DeRisi, 2002) and REC8 is involved in sister chromatin cohesion (prophase 1) (Klein et al., 1999). These genes were identified as specifically expressed or differentially regulated in telia using oligoarrays and they are all involved in the early meiotic prophase stages (leptonema, zygonema, pachynema), supporting that meiosis is blocked in prophase I at the diplonema stage when teliospores enter dormancy (Boehm et al., 1992).

The accumulation of other transcripts may also reflect telia-specific features. For instance, a transcript encoding a saccharopine dehydrogenase in particular is specifically and highly expressed in telia. This gene belongs to the α-aminoadipate pathway that leads to the biosynthesis of the amino acid L-Lysine (Xu et al., 2006) and may play a crucial role in the overwintering structure. Several fungal alkaloids or peptides have lysine as a structural element or biosynthetic precursor and may accumulate in telia. Contrary to the uredinium that breaks through the epidermis to release large amounts of urediniospores (Hacquard et al., 2010), the telium is a structure that is encapsulated between poplar epidermis and palisade mesophyll cells that remains stable over the winter season. Expression of a cocktail of lytic enzymes that target the plant cell wall at early stages of telia development such as cellulases, hemicellulases, and pectinases might reflect accommodation of the telial structure to the decaying plant tissue.

To conclude, our transcriptomic analysis gives a first overview of the genetic program activated by rust fungi during telia formation. Particularly, we identified several genes encoding osmotins/thaumatin, aquaporin, and multi-copper oxidase laccase-like proteins that may reflect specific adaptation to cold environment and overwintering. Furthermore, our time course experiment study revealed the precise temporal dynamics of karyogamy and meiosis processes and suggests these are tightly regulated during teliospore formation and maturation.

AUTHOR CONTRIBUTION

Stéphane Hacquard, Pascal Frey, and Sébastien Duplessis designed research; Stéphane Hacquard and Christine Delaruelle performed research; Stéphane Hacquard, Emilie Tisserant, and Annegret Kohler analyzed data; and Stéphane Hacquard, Pascal Frey, and Sébastien Duplessis wrote the paper.

ACKNOWLEDGMENTS

We would like to acknowledge the help of Miss Leila Parizadeh for telia sampling and fruitful discussions and continuous support from our colleague Francis Martin (INRA Nancy). We also acknowledge Prof. Salvatore Moricca for helpful discussions. This work was supported by public grants overseen by the French National Research Agency (ANR) as part of the "Investissements d'Avenir" program (ANR-11-LABX-0002-01, Lab of Excellence ARBRE) and the Young Scientist Grant POPRUST to Sébastien Duplessis (ANR-2010-JCJC-1709-01) and by the Région Lorraine (Researcher Award to Sébastien Duplessis).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2013.00456/abstract

Figure S1 | Electrophoretic profiles of total RNA collected in the study.

Table S1 | Summary of *M. larici-populina* genes selected for RT-qPCR analysis.

Table S2 | Genome-wide expression analysis of *M. larici-populina* genes measured at the telial and uredinial stages using microarray.

REFERENCES

Alexopoulos, C. J., Mims, C. W., and Blackwell, M. (1996). *Introductory. Mycology, 4th Edn*. New York, NY: Wiley.

AnŽlovar, S., and Dermastia, M. (2003). The comparative analysis of osmotins and osmotin-like PR-5 proteins. *Plant Biol.* 5, 116–124. doi: 10.1055/s-2003-40723

Aravind, L., and Koonin, E. V. (1998). The HORMA domain: a common structural denominator in mitotic checkpoints, chromosome synapsis and DNA repair. *Trends Biochem. Sci.* 23, 284–286. doi: 10.1016/S0968-0004(98)01257-2

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi: 10.1093/nar/28.1.45

Barrès, B., Dutech, C., Andrieux, A., Halkett, F., and Frey, P. (2012). Exploring the role of asexual multiplication in poplar rust epidemics: impact on diversity and genetic structure. *Mol. Ecol.* 21, 4996–5008. doi: 10.1111/mec.12008

Berndt, R. (1999). *Chrysomyxa* rust: morphology and ultrastructure of D-haustoria, uredinia, and telia. *Can. J. Bot.* 77, 1469–1484. doi: 10.1139/cjb-77-10-1469

Boehm, E. W. A., Wenstrom, J. C., McLaughlin, D. J., Szabo, L. J., Roelfs, A. P., and Bushnell, W. R. (1992). An ultrastructural pachytene karyotype for *Puccinia graminis* f.sp. *tritici*. *Can. J. Bot.* 70, 401–413. doi: 10.1139/b92-054

Brefort, T., Doehlemann, G., Mendoza-Mendoza, A., Reissmann, S., Djamei, A., and Kahmann, R. (2009). *Ustilago maydis* as a pathogen. *Annu. Rev. Phytopathol.* 47, 423–445. doi: 10.1146/annurev-phyto-080508-081923

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37, D233–D238. doi: 10.1093/nar/gkn663

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. tritici reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Driessen, S. A., O'Brien, P. A., and Hardy, G. E. (2005). Morphology of the rust fungus *Puccinia boroniae* revisited. *Mycologia* 97, 1330–1334. doi: 10.3852/mycologia.97.6.1330

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011b). *Melampsora larici-populin*a transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Duplessis, S., Joly, D. L., and Doods, P. N. (2012). "Rust effectors," in *Effectors in Plant-Microbe Interactions, 1st Edn.,* eds F. Martin and S. Kamoun (Chichester: John Wiley and Sons, Ltd.), 155–193.

Duplessis, S., Major, I. Martin, F., and Séguin, A. (2009). Poplar and pathogen interactions: insights from populus genome-wide analyses of resistance and defense gene families and gene expression profiling. *Crit. Rev. Plant Sci.* 28, 309–334. doi: 10.1080/07352680903241063

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A. S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., et al. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature* 484, 186–194. doi: 10.1038/nature10947

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PLoS ONE* 8:e67150. doi: 10.1371/journal.pone.0067150

Gerton, J. L., and DeRisi, J. L. (2002). Mnd1p: an evolutionarily conserved protein required for meiotic recombination. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6895–6900. doi: 10.1073/pnas.102167899

Goodwin, S. B., M'barek, S. B., Dhillon, B., Wittenberg, A. H., Crane, C. F., Hane, J. K., et al. (2011). Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* 7:e1002070. doi: 10.1371/journal.pgen.1002070

Hacquard, S., Delaruelle, C., Legué, V., Tisserant, E., Kohler, A., Frey, P., et al. (2010). Laser capture microdissection of uredinia formed by *Melampsora larici-populina* revealed a transcriptional switch between biotrophy and sporulation. *Mol. Plant Microbe Interact.* 23, 1275–1286. doi: 10.1094/MPMI-05-10-0111

Hacquard, S., Joly, D. L., Lin, Y. C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hacquard, S., Petre, B., Frey, P., Hecker, A., Rouhier, N., and Duplessis, S. (2011a). The poplar-poplar rust interaction: insights from genomics and transcriptomics. *J. Pathog.* 2011:716041. doi: 10.4061/2011/716041

Hacquard, S., Veneault-Fourrey, C., Delaruelle, C., Frey, P., Martin, F., and Duplessis, S. (2011b). Validation of *Melampsora larici-populina* reference genes for in planta RT-quantitative PCR expression profiling during time-course infection of poplar leaves. *Physiol. Mol. Plant Pathol.* 75,106–112. doi: 10.1016/j.pmpp.2010.10.003

Harder, D. E. (1984). "Developmental ultrastructure of hyphae and spores," in *The Cereal Rusts*, eds W. R. Bushnell and A. P. Roelfs (London: Academic Press Inc.), 333–373.

Heiman, M. G., and Walter, P. (2000). Prm1p, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating. *J. Cell Biol.* 151, 719–730. doi: 10.1083/jcb.151.3.719

Jakupović, M., Heintz, M., Reichmann, P., Mendgen, K., and Hahn, M. (2006). Microarray analysis of expressed sequence tags from haustoria of the rust fungus *Uromyces fabae*. *Fungal Genet. Biol.* 43, 8–19. doi: 10.1016/j.fgb.2005.09.001

Jung, Y. C., Lee, H. J., Yum, S. S., Soh, W. Y., Cho, D. Y., Auh, C. K., et al. (2005). Drought-inducible-but ABA-independent-thaumatin-like protein from carrot (*Daucus carota L.*). *Plant Cell Rep.* 24, 366–373. doi: 10.1007/s00299-005-0944-x

Keeney, S., Giroux, C. N., and Kleckner, N. (1997). Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell* 88, 375–384. doi: 10.1016/S0092-8674(00)81876-0

Kemen, E., Kemen, A., Ehlers, A., Voegele, R., and Mendgen, K. (2013). A novel structural effector from rust fungi is capable of fibril formation. *Plant J.* 75, 767–780. doi: 10.1111/tpj.12237

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Klein, F., Mahr, P., Galova, M., Buonomo, S. B., Michaelis, C., Nairz, K., et al. (1999). A central role for cohesins in sister chromatid cohesion, formation of axial elements, and recombination during yeast meiosis. *Cell* 98, 91–103. doi: 10.1016/S0092-8674(00)80609-1

Kronstad, J. W., and Staben, C. (1997). Mating type in filamentous fungi. *Annu. Rev. Genet.* 31, 245–276. doi: 10.1146/annurev.genet.31.1.245

Kuwabara, C., Takezawa, D., Shimada, T., Hamada, T., Fujikawa, S., and Arakawa, K. (2002). Abscisic acid- and cold-induced thaumatin-like protein in winter wheat has an antifungal activity against snow mould, *Microdochium nivale*. *Physiol. Plant.* 115, 101–110. doi: 10.1034/j.1399-3054.2002.1150112.x

Liu, J. J., Sturrock, R., and Ekramoddoullah, A. K. (2010). The superfamily of thaumatin-like proteins: its origin, evolution, and expression towards biological function. *Plant Cell Rep.* 29, 419–436. doi: 10.1007/s00299-010-0826-8

Longo, N., Moriondo, F., and Naldini Longo, B. (1979). Ultrastructural observations on teliospores of *Melampsora pinitorqua* Rostr. *Caryologia* 32, 223–240.

Maier, W., Begerow, D., Weiß, M., and Oberwinkler, F. (2003). Phylogeny of the rust fungi: an approach using nuclear large subunit ribosomal DNA sequences. *Can. J. Bot.* 81, 12–23. doi: 10.1139/b02-113

Malik, S. B., Pightling, A. W., Stefaniak, L. M., Schurko, A. M., and Logsdon, J. M. Jr. (2008). An expanded inventory of conserved meiotic genes provides evidence for sex in *Trichomonas vaginalis*. *PLoS ONE* 3:e2879. doi: 10.1371/journal.pone.0002879

Mendgen, K. (1984). "Development and physiology of teliospores," in *The Cereal Rusts*, eds W. R. Bushnell and A. P. Roelfs (London: Academic Press Inc.), 375–398.

Mims, C. W. (1977). Ultrastructure of teliospore formation in the cedar-apple rust fungus *Gymnosporangium juniperi-virginianae*. *Can. J. Bot.* 55, 2319–2329. doi: 10.1139/b77-263

Mims, C. W. (1981). Ultrastructure of teliospore germination and basidiospore formation in the rust fungus *Gymnosporangium clavipes*. *Can. J. Bot*. 59, 1041–1049. doi: 10.1139/b81-142

Mims, C. W., Liljebjelke, K. A., and Covert, S. F. (1996). Ultrastructure of telia and teliospores of the rust fungus *Cronartium quercuum* f.sp. *fusiforme*. *Mycologia* 88, 47–56. doi: 10.2307/3760783

Mims, C. W., and Richardson, E. A. (2005). Light and electron microscopy of teliospores and teliospore germination in the rust fungus *Coleosporium ipomoeae*. *Can. J. Bot*. 83, 451–458. doi: 10.1139/b05-020

Moriondo, F., Naldini Longo, B., Longo, N., Drovandi, F., and Gonnelli, T. (1989). Some observations on the life-cycle of *Melampsora pulcherrima* (Bub.) Maire. *Phytopathol. Mediterr.* 28, 46–52.

O'Donnell, K. L., and McLaughlin, D. J. (1981). Ultrastructure of meiosis in the hollyhock rust fungus, *Puccinia malvacearum*. I. Prophase I - prometaphase I. *Protoplasma* 108, 225–244. doi: 10.1007/BF02224421

Petre, B., Major, I., Rouhier, N., and Duplessis, S. (2011). Genome-wide analysis of eukaryote thaumatin-like proteins (TLPs) with an emphasis on poplar. *BMC Plant Biol.* 11:33. doi: 10.1186/1471-2229-11-33

Petre, B., Morin, E., Tisserant, E., Hacquard, S., Da Silva, C., Poulain, J., et al. (2012). RNA-Seq of early-infected poplar leaves by the rust pathogen *Melampsora larici-populina* uncovers PtSultr3;5, a fungal-induced host sulfate transporter. *PLoS ONE* 7:e44408. doi: 10.1371/journal.pone.0044408

Polle, A., Janz, D., Teichmann, T., and Lipka, V. (2013). Poplar genetic engineering: promoting desirable wood characteristics and pest resistance. *Appl. Microbiol. Biotechnol.* 97, 5669–5679. doi: 10.1007/s00253-013-4940-8

Rinaldi, C., Kohler, A., Frey, P., Duchaussoy, F., Ningre, N., Couloux, A., et al. (2007). Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust *Melampsora larici-populina*. *Plant Physiol.* 144, 347–366. doi: 10.1104/pp.106.094987

Rozen, S., and Skaletsky, H. J. (2000). "Primer3 on the WWW for general users and for biologist programmers," in *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, eds S. Krawetz and S. Misener (Totowa, NJ: Humana Press), 365–386.

Schumann, G. L., and Leonard, K. J. (2000). Stem rust of wheat (black rust). *Plant Health Instr*. doi: 10.1094/PHI-I-2000-0721-01

Sidoux-Walter, F., Pettersson, N., and Hohmann, S. (2004). The *Saccharomyces cerevisiae* aquaporin Aqy1 is involved in sporulation. *Proc. Natl. Acad. Sci. U.S.A.* 101, 17422–17427. doi: 10.1073/pnas.0404337101

Singh, N. K., Bracker, C. A., Hasegawa, P. M., Handa, A. K., Buckel, S., Hermodson, M. A., et al. (1987). Characterization of osmotin: a thaumatin-like protein associated with osmotic adaptation in plant cells. *Plant Physiol.* 85, 529–536. doi: 10.1104/pp.85.2.529

Singh, R. P., Hodson, D. P., Huerta-Espino, J., Jin, Y., Bhavani, S., Njau, P., et al. (2011). The emergence of Ug99 races of the stem rust fungus is a threat to world wheat production. *Annu. Rev. Phytopathol.* 49, 465–481. doi: 10.1146/annurev-phyto-072910-095423

Sturn, A., Quackenbush, J., and Trajanoski, Z. (2002). Genesis: cluster analysis of microarray data. *Bioinformatics* 18, 207–208. doi: 10.1093/bioinformatics/18.1.207

Tanghe, A., Van Dijck, P., Dumortier, F., Teunissen, A., Hohmann, S., and Thevelein, J. M. (2002). Aquaporin expression correlates with freeze tolerance in baker's yeast, and overexpression improves freeze tolerance in industrial strains. *Appl. Environ. Microbiol.* 68, 5981–5989. doi: 10.1128/AEM.68.12.5981-5989.2002

Tanghe, A., Van Dijck, P., and Thevelein, J. M. (2006). Why do microorganisms have aquaporins? *Trends Microbiol.* 14, 78–85. doi: 10.1016/j.tim.2005.12.001

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41. doi: 10.1186/1471-2105-4-41

Voegele, R. T., Struck, C., Hahn, M., and Mendgen, K. (2001). The role of haustoria in sugar supply during infection of broad bean by the rust fungus *Uromyces fabae*. *Proc. Natl. Acad. Sci. U.S.A.* 98, 8133–8138. doi: 10.1073/pnas.131186798

Win, J., Chaparro-Garcia, A., Belhaj, K., Saunders, D. G., Yoshida, K., Dong, S., et al. (2012). Effector biology of plant-associated organisms: concepts and perspectives. *Cold Spring Harb. Symp. Quant. Biol.* 77, 235–247. doi: 10.1101/sqb.2012.77.015933

Wittenberg, A. H., van der Lee, T. A., Ben M'barek, S., Ware, S. B., Goodwin, S. B., Kilian, A., et al. (2009). Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. *PLoS ONE* 4:e5863. doi: 10.1371/journal.pone.0005863

Xhaard, C., Fabre, B., Andrieux, A., Gladieux, P., Barrès, B., Frey, P., et al. (2011). The genetic structure of the plant pathogenic fungus *Melampsora larici-populina* on its wild host is extensively impacted by host domestication. *Mol. Ecol.* 20, 2739–2755. doi: 10.1111/j.1365-294X.2011.05138.x

Xu, H., Andi, B., Qian, J., West, A. H., and Cook, P. F. (2006). The alpha-aminoadipate pathway for lysine biosynthesis in fungi. *Cell. Biochem. Biophys.* 46, 43–64. doi: 10.1385/CBB:46:1:43

Xu, J., Linning, R., Fellers, J., Dickinson, M., Zhu, W., Antonov, I., et al. (2011). Gene discovery in EST sequences from the wheat leaf rust fungus *Puccinia triticina* sexual spores, asexual spores and haustoria, compared to other rust and corn smut fungi. *BMC Genomics* 12:161. doi: 10.1186/1471-2164-12-161

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 July 2013; accepted: 23 October 2013; published online: 21 November 2013.

*Citation: Hacquard S, Delaruelle C, Frey P, Tisserant E, Kohler A and Duplessis S (2013) Transcriptome analysis of poplar rust telia reveals overwintering adaptation and tightly coordinated karyogamy and meiosis processes. Front. Plant Sci. **4**:456. doi: 10.3389/fpls.2013.00456*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2013 Hacquard, Delaruelle, Frey, Tisserant, Kohler and Duplessis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 14 March 2014
doi: 10.3389/fpls.2014.00088

Overview of the functional virulent genome of the coffee leaf rust pathogen *Hemileia vastatrix* with an emphasis on early stages of infection

Pedro Talhinhas[1], Helena G. Azinheira[1]*, Bruno Vieira[2], Andreia Loureiro[1], Sílvia Tavares[1], Dora Batista[1], Emmanuelle Morin[3,4], Anne-Sophie Petitot[5], Octávio S. Paulo[2], Julie Poulain[6], Corinne Da Silva[6], Sébastien Duplessis[3,4], Maria do Céu Silva[1] and Diana Fernandez[5]

[1] *Centro de Investigação das Ferrugens do Cafeeiro/BioTrop/Instituto de Investigação Científica Tropical, Oeiras, Portugal*
[2] *Computational Biology and Population Genomics Group, Centro de Biologia Ambiental, Faculdade de Ciências da Universidade de Lisboa, Lisboa, Portugal*
[3] *Institut National de la Recherche Agronomique, Centre INRA Nancy Lorraine, UMR 1136 INRA/Université de Lorraine Interactions Arbres/Micro-organismes, Champenoux, France*
[4] *Université de Lorraine, UMR 1136 INRA/Université de Lorraine Interactions Arbres/Micro-organismes, Faculté des Sciences et Technologies, Vandoeuvre-lès-Nancy, France*
[5] *Institut de Recherche pour le Développement, UMR 186 IRD-Cirad-UM2 Résistance des Plantes aux Bioagresseurs, Montpellier, France*
[6] *Genoscope, Centre National de Séquençage, Commissariat à l'Energie Atomique, Institut de Génomique, Evry, France*

Edited by:
David L. Joly, Université de Moncton, Canada

Reviewed by:
Hossein Borhan, Agriculture and Agri-Food Canada, Canada
Ralf Thomas Voegele, Universität Hohenheim, Germany

***Correspondence:**
Helena G. Azinheira, Centro de Investigação das Ferrugens do Cafeeiro, BioTrop, Instituto de Investigação Científica Tropical, Quinta do Marquês, 2784-505 Oeiras, Portugal
e-mail: hgazinheira@gmail.com

Hemileia vastatrix is the causal agent of coffee leaf rust, the most important disease of coffee *Arabica*. In this work, a 454-pyrosequencing transcriptome analysis of *H. vastatrix* germinating urediniospores (gU) and appressoria (Ap) was performed and compared to previously published *in planta* haustoria-rich (H) data. A total of 9234 transcripts were identified and annotated. Ca. 50% of these transcripts showed no significant homology to international databases. Only 784 sequences were shared by the three conditions, and 75% were exclusive of either gU (2146), Ap (1479) or H (3270). Relative transcript abundance and RT-qPCR analyses for a selection of genes indicated a particularly active metabolism, translational activity and production of new structures in the appressoria and intense signaling, transport, secretory activity and cellular multiplication in the germinating urediniospores, suggesting the onset of a plant-fungus dialogue as early as at the germ tube stage. Gene expression related to the production of carbohydrate-active enzymes and accumulation of glycerol in germinating urediniospores and appressoria suggests that combined lytic and physical mechanisms are involved in appressoria-mediated penetration. Besides contributing to the characterization of molecular processes leading to appressoria-mediated infection by rust fungi, these results point toward the identification of new *H. vastatrix* candidate virulence factors, with 516 genes predicted to encode secreted proteins.

Keywords: appressorium, coffee leaf rust, germinating urediniospore, haustorium, pyrosequencing, transcriptome

INTRODUCTION

Rust diseases have been a long standing threat for centuries and have reshaped cultivation of crops and breeding strategies. Coffee leaf rust caused by *Hemileia vastatrix* Berk & Broome is the major disease of Arabica cultivated coffees (*Coffea arabica* L.) (Silva et al., 2006). *H. vastatrix* is considered as one of the most primitive phylogenetic lineages of the Pucciniales (Aime, 2006; Silva et al., 2012) and has no alternate host known so far. Since the 19th century, when it caused suppression of the coffee cultivation in Sri Lanka, the disease gained a worldwide distribution, reaching nearly all regions of the world where coffee is grown with severe economical damages. Breeding and selection of coffee resistant genotypes to different fungal races from several parts of the world has been successful (Silva et al., 2006), but as a consequence of the high adaptive potential of the pathogen, the emergence of new rust pathotypes and the corresponding breakdown of resistance has been observed in many improved coffee varieties in several countries (Várzea and Marques, 2005; Diniz et al., 2012; Cressey, 2013). Thus, currently coffee leaf rust still stands as the major constraint to Arabica coffee production.

During infection of their hosts, rust fungi differentiate several specialized infection structures such as germtubes, appressoria, stomatal vesicles, infection hyphae, haustoria, and spore-forming cells. Until recently, most of the biological knowledge gained at the molecular level on rust fungi was derived from EST sequencing, mainly from ungerminated and germinating urediniospores, rust-infected tissues, isolated haustoria and some spore types at other stages of their complex life cycle (for a review see Duplessis et al., 2012; Fernandez et al., 2013). However, some differentiation stages are not sufficiently covered yet and lack description and information, such as appressoria formation (Fernandez et al., 2013). For instance, two studies conducted in *Phakospora pachyrhizi* and *Puccinia triticina* reported that a high proportion of genes of unknown functions were expressed at the appressorial stage (Hu et al., 2007; Stone et al., 2012). Appressoria may be also differentiated by other pathogenic fungi, enabling

host cuticle penetration through physical and/or chemical mechanisms. Key features of these specialized structures include the production of an extracellular matrix for adhesion to the surface, the accumulation of molar concentrations of glycerol for generating turgor pressure and the differentiation of a penetration hypha (Deising et al., 2000). Appressoria differentiated from urediniospores typically form over host stomata, and a penetration hypha is subsequently formed at the base of the appressorium to invade the substomatic chamber. There are evidences that mechanical pressure (about 0.35 MPa) is exerted by the penetration hypha when penetrating the stoma (Terhune et al., 1993). This is considerably less than the pressure exerted by some fungi that penetrate directly through the cuticle, such as *Magnaporthe oryzae* or *Colletotrichum* spp. (Howard et al., 1991; Chen et al., 2004), but enough to distort stomatal guard cell lips (Terhune et al., 1993). Nevertheless, rust fungi must also possess machinery for lytic penetration of host cuticle and cell wall, since hyphae produced by germinating basidiospores are capable of direct penetration into host epidermal cells (Voegele et al., 2009).

Until very recently, no genomic resources were available for *H. vastatrix*. After several years of lagging behind other rust fungi on genomic research, Fernandez et al. (2012) reported on the 454-transcriptome sequencing of rust-infected coffee leaves. This study generated 22,774 contigs of which 30% were assigned to *H. vastatrix*. Analysis of these *in planta* expressed sequence tags (ESTs) revealed that the majority (60%) had no homology in public genomic databases, representing potential coffee rust-specific genes. Nevertheless, *H. vastatrix* candidate effectors likely related to host infection and orthologous to other rust fungi, were identified among 382 predicted secreted proteins (Fernandez et al., 2012). Still, there is no knowledge of transcripts expressed at early stages of infection that could provide a more integrative scenario on the molecular mechanisms governing this pathosystem.

Complementing the knowledge gained into the *in planta* transcriptome of coffee rust, here we report on the in-depth transcriptome analysis of *H. vastatrix* by 454-based RNA-Seq during urediniospore germination and appressorium formation, two early and key stages of infection. Comparison of these specific stages with infected leaves allows an integrative characterization of transcript expression profiles during the course of biotrophic growth and infection. In particular, the identification of genes related with appressorium formation leads to novel insights into a stage that has been poorly described at the molecular level.

MATERIALS AND METHODS

BIOLOGICAL MATERIAL, RNA ISOLATION AND cDNA SYNTHESIS

Hemileia vastatrix isolate CIFC 178a (race XIV: genotype $v_2v_3v_4v_5$) was multiplied on its differential host plant (*C. arabica* accession CIFC H147/1, carrying the resistance factors S_H2, S_H3, S_H4 and S_H5). An *in vitro* method was used to produce germinating urediniospores and appressoria to ensure the generation of cDNA libraries with no contaminating plant sequences (Azinheira et al., 2001; Vieira et al., 2012). For the germinating urediniospores sample (gU), 19 mg of spores were spread in sterile distilled water in Petri dishes and incubated for 18 h at 24°C under darkness. For the appressoria sample (Ap), 15 mg of spores were spread over oil-collodion membranes (Vieira et al., 2012) in Petri dishes, sprayed with water and incubated for 24 h at 24°C and 100% relative humidity, under darkness. For an accurate sample characterization, urediniospore germination and appressoria formation were quantified, showing that the germinating urediniospores sample (gU) comprised over 50% of germinating urediniospores. The appressoria sample (Ap) comprised over 60% of germinating urediniospores with appressoria. These are considered rather fair rates for *H. vastatrix* (Azinheira et al., 2001).

Samples gU and Ap were harvested, immediately frozen in liquid nitrogen and the RNA was isolated from each sample with the RNeasy Plant minikit (Qiagen, Hilden, Germany), including an in-solution DNase treatment following the manufacturer's instructions. RNA concentration and integrity were evaluated by spectrometry (Lambda EZ201, Perkin-Elmer, Waltham-MA, USA) and capillary electrophoresis (Bioanalyzer 2100, Agilent, Santa Clara-CA, USA) respectively. Following the SMARTer Pico PCR cDNA Synthesis Kit (Clontech, Saint-Germain-en-Laye, France) protocol, cDNA were synthesized from 1 µg total RNA using SMARTScribe Reverse Transcriptase (Clontech) and amplified using Advantage 2 Polymerase (Clontech). cDNA fragments, which ranged between 500 and 3000 bp, were purified with the NucleoSpin Extract II kit (Macherey-Nagel, Düren, Germany) and their quality and concentration were evaluated by electrophoresis.

PYROSEQUENCING AND ASSEMBLY OF 454 READS

For each sample, 20 µg cDNA was used for 454-pyrosequencing run on half of a picotitre plate on a Genome Sequencer FLX System using long-read GS FLX Titanium chemistry (Roche; www.454.com) at the Genoscope (Centre National de Séquençage, Evry, France; www.genoscope.cns.fr) following standard procedures recommended by Roche.

Raw sequences obtained for gU and Ap samples were assembled into contigs using Newbler 2.5 (Roche) with default parameters. For comparative purposes, the MIRA 3.2 assembler (http://sourceforge.net/apps/mediawiki/mira-assembler) was also employed. The relative abundance (Ra) of transcripts was calculated as the ratio between the number of 454 reads per contig and the length of the assembled contig (Vega-Arreguín et al., 2009).

BIOINFORMATIC ANALYSIS OF TRANSCRIPTS

As previously described (Fernandez et al., 2012), sequence homology searches were performed against several databases: the NCBI non-redundant (nr) nucleotide and protein databases (www.ncbi.nlm.nih.gov), the genome sequences of *Melampsora larici-populina* and *Puccinia* spp. (Cantu et al., 2011; Duplessis et al., 2011a; www.jgi.doe.gov and www.broadinstitute.org, respectively); the euKaryotic Orthologous Group (KOG) database (Tatusov, 2003); the Pathogen-Host Interaction (PHI-base v3.2) reference database (Winnenburg et al., 2007; www.phi-base.org); the Phytopathogenic Fungi and Oomycete EST Database (COGEME v1.6; Soanes and Talbot, 2006); and a Pucciniales EST database (168,199 ESTs retrieved from GenBank in November 2012—unchanged number as of December 2013). Besides these, 16,831 transcripts from the *M. larici-populina*

frozen gene catalog (http://genome.jgi-psf.org/Mellp1/Mellp1.download.ftp.html) and 20,567 (*P. graminis* f. sp. *tritici*) and 11,638 (*P. triticina*) from the *Puccinia* spp. transcript catalogue (http://www.broadinstitute.org/annotation/genome/puccinia_group/MultiDownloads.html) were also considered. Homology searches were performed using BLAST algorithms (Altschul et al., 1997) with a cut-off criterion (*e*-value $< 10^{-5}$). For each search against a given database, only the best hit was considered. The assignment of 454-contig sequences into KOG functional categories was obtained using Reverse psi-BLAST (RPSBLAST; Altschul et al., 1997) against the KOG database.

Open reading frames (ORFs) were predicted with the translation tool getorf from the European Molecular Biology Open Software Suite (EMBOSS; http://emboss.bioinformatics.nl/cgi-bin/emboss/getorf) using default parameters. ORFs below 18 amino acids were not considered. A secretome bioinformatics pipeline was employed to define a tentative set of secreted proteins encoded by *H. vastatrix* transcripts, using SignalP v4.0 (Petersen et al., 2011), TargetP v1.1 (Emanuelsson et al., 2000) and TMHMM v2.0 (Krogh et al., 2001).

The catalytic and carbohydrate-binding modules (or functional domains) of enzymes that degrade, modify, or create glycosidic bonds (carbohydrate-active enzymes—CAZymes) was investigated by blastp comparison of predicted polypeptides to the CAZymes database (www.cazy.org; Cantarel et al., 2009) and to the CAZymes from *M. larici-populina* and *P. graminis* f. sp. *tritici* (Duplessis et al., 2011a). Similarly, proteins involved in membrane transport were investigated by blastp searches against predicted polypeptides in the Transporter Classification Database (www.tcdb.org; Saier et al., 2006, 2009).

Contigs from gU and Ap samples, as well as those predicted as fungal from a 21 day infected-coffee leaf sample (sample H; Fernandez et al., 2012), were compared using a best reciprocal BLAST hit approach with BioEdit 7.0.4.1 (Hall, 1999). Pairs of contigs with an *e*-value lower than 10^{-30} were considered as representing the same transcript and assembled. Ra values were calculated for each transcript present in more than one library, and these values were compared across the libraries in order to evaluate variations in expression levels. For such, τ values were calculated for each gene based on the normalized Ra values, in order to account for differences in library sizes. The expression specificity index (τ) is defined as $\tau = \frac{\sum_{i=1}^{n}(1-x_i)}{n-1}$, where n is the number of tissues and x_i is the expression profile component normalized by the maximal component value (Yanai et al., 2005). The genes with the most stable expression across the three libraries were selected (105 genes with τ values below 0.25). Average Ra values were calculated among these genes for each library (0.02684 for gU, 0.03158 for Ap and 0.03039 for H) and Ra values for each contig in each library were normalized to the gU sample, following the strategy described by Ekblom et al. (2010) based on the guidelines provided by Mank et al. (2008).

RT-qPCR

Germinating urediniospores (gU) and appressoria (Ap) samples for *H. vastatrix* isolate 178a were obtained as described above. *In planta* time course samples were collected at 18 h (mostly containing appressoria) and 1, 2, 3, 7, 14, and 21 days after inoculation for the compatible interaction between isolate 178a and the *C. arabica* genotype H147/1, as previously described (Diniz et al., 2012; Vieira et al., 2012). Fungal germination, appressoria formation and the differentiation of infection structures *in planta* were monitored by light microscopy as previously described (Vieira et al., 2012). RNA extraction, cDNA synthesis and RT-qPCR experiments were performed as previously described (Vieira et al., 2012), using Hv00099, 40 S ribosomal protein and glyceraldehyde-3-phosphate dehydrogenase as reference genes (Vieira et al., 2011) and ungerminated urediniospores as the control sample. A set of 43 genes was selected for RT-qPCR analysis based on RNA-Seq expression profiles and assigned functions. Primers (Supplementary Data 1) were designed as previously described (Vieira et al., 2012).

RESULTS AND DISCUSSION

454-PYROSEQUENCING DATA FOR GERMINATING UREDINIOSPORES AND APPRESSORIA SAMPLES

Given that different 454-pyrosequencing data assemblers are available and are known to generate diverse results (Kumar and Blaxter, 2010), MIRA 3.2 and Newbler 2.5 were compared in this study. In general, MIRA produced shorter and more numerous contigs. The overall homology scores of contigs to a Pucciniales EST database was better for the Newbler assembly (data not shown), suggesting a better quality of the assemblage which led us to use Newbler assembly in this study.

For samples gU and Ap (**Table 1**), a total of 455,807 sequence reads (113,404,366 nucleotides) was generated and assembled into 9108 contigs (4267 for gU and 3626 for Ap), with ca. 24% sequences remaining as either too short/low quality sequences (7%) or singletons (17%). Among those, 1214 contigs (13%) were <100 bp and not further considered in the analysis. The remaining 7894 contigs (Supplementary Data 2) had a mean length of 656 bp (**Table 1**), with 16% contigs larger than 1 kb (3.7% larger than 2 kb). Mean number of reads per contig was 41.0, with 11% contigs over 50 reads. Mean relative abundance (Ra) was 0.1153, with 18% contigs (1424) representing transcripts with a medium to high rate of expression (Ra > 0.05).

In the absence of genomic information for *H. vastatrix*, contigs were compared to sequences deposited in databases (summary in **Table 2** and results by contig listed in Supplementary Data 2), and 54% contigs had homology (*e*-value $< 10^{-5}$) to the NCBI nr nucleotide database using blastn (Supplementary Data 2, columns G–I).

A total of 13,951 sequences obtained from the 21-days *H. vastatrix* infected-coffee leaf samples (H library) and previously predicted as of plant origin (Fernandez et al., 2012) were compared to the gU+Ap sequences, from which 22 showed an homology *e*-value below 10^{-60} (19 of which had *e*-value of 0.0; Supplementary Data 3). This analysis showed that only 0.1% of the sequences predicted as of plant origin (Fernandez et al., 2012) were wrongly assigned to this class. Similarly, among 2060 contigs previously classified as "not attributed/not resolved," 28 had homology to gU+Ap sequences with an *e*-value below 10^{-60} (21 of which had *e*-value of 0.0). These 50 contigs (*e*-value $< 10^{-60}$) were incorporated in the present study, together with the 4415 fungal contigs initially identified in the H library,

summing a total of 4465 H contigs (Fernandez et al., 2012) to our dataset.

COMPARISON TO THE *IN PLANTA* EXPRESSED FUNGAL SEQUENCES

A best reciprocal BLAST strategy was used to compare the contigs from both gU and Ap libraries, as well as the fungal contigs from the H library (Fernandez et al., 2012). This enabled the identification and re-assembly of 784 sequences shared by the three libraries, 1145 shared only by gU and Ap, 219 by Ap and H and 192 by gU and H (**Figure 1** and Supplementary Data 4, columns B–D). The remaining 6894 sequences (75%) are exclusive of each library. Altogether, 9234 unique *H. vastatrix* sequences were identified, which represents >50% of the total number of genes predicted from the genomes of *M. larici-populina* (16,399 genes) and *P. graminis* f. sp. *tritici* (17,773 genes) (Duplessis et al., 2011a). In order to further ascertain a measure of the genome coverage obtained in this study, we compared each of the three libraries (gU, Ap, and H) separately, along with the total set of 9234 sequences, against the FUNYBASE database containing 246 families of single-copy orthologs obtained from 21 genomes (core

Table 1 | Descriptive statistics for *Hemileia vastatrix* 454 pyrosequenced cDNA libraries of germinated urediniospores (gU) and appressoria (Ap).

Library	gU	Ap
Number of bases	67773266	45631100
Number of sequences	269199	186608
Mean size of reads (bp)	251.8	244.5
Number of contigs	4267	3626
Mean size of contigs (bp)	676	632
Size of contigs (bp)*	188/546/1293/3754	192/530/1139/4860
Mean number of reads per contig	48.5	32.2
Reads per contig*	4/10/63/3326	4/9/52/2077
Mean relative abundance (Ra)	0.1456	0.0797
Relative abundance (Ra)*	0.0087/0.0190/0.0967/26.72	0.0085/0.0187/0.0777/14.41

**Values correspond to 10/50/90/100 percentiles.*

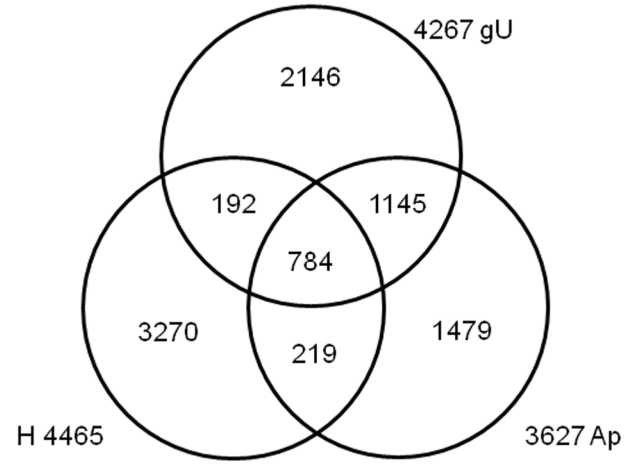


FIGURE 1 | Venn diagram depicting the comparison of *Hemileia vastatrix* EST libraries, using a best reciprocal BLAST hit strategy. Number of contigs is indicated for the germinating urediniospores (gU), appressoria (Ap) and infected leaves 21 days after inoculation (H) libraries.

Table 2 | *Hemileia vastatrix* transcript homology in databases and KOG functional categories classification.

Library	gU	Ap	Total	% of all gU contigs	% of all Ap contigs	% of all contigs
NCBI nr_blastn	2119	2126	4245	49.66	58.62	53.78
Mlp genome_tblastx	2356	2159	4515	55.21	59.53	57.20
Pgt_genome_tblastx	2294	2099	4393	53.76	57.87	55.65
Pt_genome_tblastx	2410	2171	4581	56.48	59.86	58.03
Pst_genome_tblastx	2408	2159	4567	56.43	59.53	57.85
EST_Pucciniales_tblastx	2507	2334	4841	58.75	64.35	61.33
SwissProt_blastx	1301	1362	2663	30.49	37.55	33.73
PHIbase_tblastx	482	444	926	11.30	12.24	11.73
COGEME_tblastx	2016	2042	4058	47.25	56.30	51.41
KOG	1691	1690	3381	39.63	46.59	42.83
Posttranslational modification, protein turnover, chaperones				12.5	13.6	
Translation, ribosomal structure and biogenesis				12.4	19.7	
Intracellular trafficking, secretion, and vesicular transport				8.2	5.7	
Energy production and conversion				7.9	9.1	
Signal transduction mechanisms				7.5	6.3	
Lipid transport and metabolism				5.7	6.4	

Summary of the number and percentage of hits in homology searches of the H. vastatrix germinated urediniospores (gU) and appressoria (Ap) contig libraries against the NCBI nr database (NCBI nr_blastn), rust genomic and transcriptomic databases (Mlp_genome_tblastx, Pgt_genome_tblastx, Pt_genome_tblastx, Pst_genome_tblastx, EST_Pucciniales_tblastx), the SwissProt database (SwissProt_blastx) and functional databases (PHIbase_tblastx, COGEME_tblastx, KOG).

fungal genes) (Marthey et al., 2008). Each individual library contained only about half of the 246 genes (39% in H and 55% both in Ap and gU; data not shown), but the gU+Ap+H library included 174 (71%) of those core genes (Supplementary Data 4, columns BN-BP). These results indicate that the 9234 unique *H. vastatrix* transcripts provide a significant coverage of the *H. vastatrix* functional genome.

According to their RNA-Seq expression values and assigned functions, the expression profiles of a set of 43 genes was analyzed by RT-qPCR along the time course of a compatible interaction (**Table 3**).

GENE FUNCTION

Over 72% of the 9234 *H. vastatrix* transcripts had no specific KOG category assigned (No hits, "Function unknown" or "General function prediction only"; **Table 4** and Supplementary Data 4, columns BK-BM). Within the remaining transcripts, the most represented KOG categories are "Translation, ribosomal structure and biogenesis" (14%) and "Post-translational modification, protein turnover, chaperones" (12%), while other nine categories represent 5–8% each.

A total of 4040 transcripts (44% of total) presented homologies against the NCBI non-redundant nucleotide database (Supplementary Data 4, columns V–X). A similar value (45%) was obtained by comparison (tblastx) with the Pucciniales EST database (Supplementary Data 4, columns AY-BA), with the most frequent organisms being *M. larici-populina*, *P. triticina*, *P. graminis* f. sp. *tritici* and *P. pachyrhizi* in similar proportions to those reported for the gU and Ap libraries (Supplementary Data 4, columns Y–AN). A total of 2992 transcripts (32%) have homology to all available rust genome sequences, suggesting that the corresponding genes are conserved among the Pucciniales. Only 16 transcripts showed homology to *P. graminis* f. sp. *tritici* or *P. triticina* mitochondrial sequences (Supplementary Data 4, columns AS-AX). A total of 141 and 148 transcripts did not show homology to *M. larici-populina* and *P. striiformis* gene models respectively, although showing significant homology to their genome sequences, which could indicate actual genes that were not predicted by automatic annotations in the corresponding genomes (Supplementary Data 4, columns AB–AE and AO–AR). Interestingly, 4707 transcripts (51% of total) showed no homology to the rust genes identified in genome sequences or EST databases (at a cut off *e*-value of 10^{-10}), suggesting they may correspond to highly divergent or specific *H. vastatrix* genes. In fact, among these, only 3.2% have a specific KOG category assigned, with an overrepresentation of categories involved in cellular structure, nucleic acid activity and signaling ("Cytoskeleton," "RNA processing and modification," "Transcription" and "Signal transduction mechanisms").

A total of 3573 transcripts showed homology to annotated fungal genes listed in the COGEME database (Supplementary Data 4, columns BH-BJ), *Ustilago maydis* (21%), *Giberella* spp. (13%), and *M. oryzae* (13%) being the most represented species. Further, 588 transcripts showed homology to fungal pathogenicity-related genes listed in the PHI database (Supplementary Data 4, columns BE-BG; 94% of which also have homologues in the COGEME database), mostly from *M. oryzae* (19%), *Candida albicans* (18%), and *Cryptococcus neoformans* (14%). About 70% of these 588 transcripts had specific KOG categories assigned, with categories such as "Cell cycle control, cell division, chromosome partitioning," "Secondary metabolites biosynthesis, transport and catabolism" and "Signal transduction mechanisms" overrepresented as compared to KOG categories assigned to all genes (**Table 4**).

For each transcript, the size of the predicted polypeptide was compared to the size of the corresponding polypeptide in the *P. graminis* f. sp. *tritici* genome. Among the 3686 hits to the *P. graminis* f. sp. *tritici* genome, 24% were >90% the length of their orthologs (46% proteins were >50% size; Supplementary Data 4, columns BQ-BS).

As *H. vastatrix* 454 cDNA libraries were not normalized, the number of reads contained in each contig can be considered a relative expression level of each gene. For each contig, the number of reads was divided by the length of the contig, resulting in a Relative Abundance (Ra) index (Supplementary Data 4, columns G-O). Comparison among contigs from different libraries required a normalization step in order to account for differences in library sizes (Supplementary Data 4, columns P–S). Based on the comparative analysis of the expression levels identified in each library, nine different expression profiles were defined (Supplementary Data 4, column U). The most frequent profiles observed reflect the fact that 75% of contigs are exclusive of a single library, with 87% of the 9234 predicted transcripts presenting profiles 1, 2, or 3 (**Table 5**, row 4).

The analysis of relative abundance values according to the KOG category of each gene (**Table 5**) suggests a particularly active metabolism, translational activity and production of new structures in the Ap sample and both intense signaling and secretory activity and cellular multiplication in germinating urediniospores. In the H sample, over-represented KOG categories suggest intense signaling and nutrient uptake from the host to the fungus, as previously pointed out (Fernandez et al., 2012).

SECRETED PROTEINS

A total of 467 putative secreted proteins were predicted with a secretion prediction pipeline composed of the SignalP, TargetP and TMHMM programmes (Supplementary Data 4, columns BT to CI for SignalP, CJ to CO for TargetP, CP to CS for TMHMM, CT for TMHMM vs. SignalP comparison and CU for final secretion prediction score). Besides these, other transcripts showing high homology (*e*-value $< 10^{-30}$) to the *M. larici-populina* or the *P. graminis* f. sp. *tritici* predicted secreted proteins (Duplessis et al., 2011a) were also selected (Supplementary Data 4, column CV) even if not detected by the pipeline. Since these sequences were shorter than their orthologs, the signal peptide may be lacking from sequence. From this list of 516 transcript encoding putative secreted proteins, 87 and 70% entries presented less than 300 amino acids, and 200 amino acids, respectively (Supplementary Data 4, column BR). Also, 82 of these translated gene sequences are highly enriched in cysteine residues (5–15% of all amino acids; Supplementary Data 4, column CW), the vast majority of which (78) is less than 200 aa, similar to what was reported for *M. larici-populina* small secreted proteins (Hacquard et al., 2012). Nearly 60% of these 82 sequences contain

Table 3 | Heatmap of *Hemileia vastatrix* genes expression profiles in germinating urediniospores (gU), *in vitro*-obtained appressoria (Ap) and *in planta* samples (from 18h to 21d after inoculation) obtained by RT-qPCR by comparison with 454 pyrosequencing-derived relative abundance transcript levels.

Gene	gU	Ap	18h	1d	2d	3d	7d	14d	21d	KOG function	Annotation (according to Supplementary Data 4)
Hv00125	1.16367	19.3825	2.11073	0.40276	0.26266	0.0869	0.21031	0.0255	0.00246	Signaling	Fungal conserved hypothetical protein similar to gas1
Hv00156	0.70207	19.7096	9.06186	9.12287	2.74469	1.95129	4.4071	0.3216	0.02838	Metabolism	Hexokinase
Hv00175	0.45556	0.29447	3.12743	4.31055	9411.5	2563.85	10548	175.507	26.569	Metabolism	Long-chain fatty acid CoA ligase
Hv00191	0.03	56.1	36.3	19.5	8.2	39.8	1.02	1.05	3.12	Metabolism	Acetyl-CoA C-acyltransferase
Hv00297	0.51073	2.35799	0.06277	0.01619	0.52742	0.08316	0.13036	0.02135	0.00623	Oxidative stress	Manganese superoxide dismutase
Hv00334	0.66625	1.31543	0.24244	0.13565	0.33702	0.06125	0.11578	0.03311	0.01143	Signaling	GTP-binding protein ypt1
Hv00373	2.04393	2.14363	1.9811	1.89263	1.972	1.81027	2.05546	1.59249	1.60879	Signaling	Mitogen-activated protein kinase (MAPK)
Hv00409	35.2036	216.796	1.14688	0.81432	2.465	0.72411	1.46818	0.00581	0.00044	Signaling	Fungal conserved hypothetical protein similar to Gas2
Hv00489	6.33369	48.0067	1.78482	1.64577	0.00442	12.5255	0.67139	0.01991	0.01893	Signaling	Fungal conserved hypothetical protein similar to Gas1
Hv00491	0.74366	19.0622	0.90118	0.66227	70.355	0.32046	1.68915	0.85444	0.01163	Metabolism	Malate dehydrogenase
Hv00616	1.11902	1.31567	0.99129	1.26598	0.986	0.92218	0.94056	0.89716	1.10502	Signaling	Mitogen-activated protein kinase, Hog-1
Hv00622	1.24335	12.1162	3.22935	6.22324	5.49718	0.76849	0.93931	0.2717	0.08024	Metabolism	Acyl-CoA oxidase
Hv00643	2.24832	24.0975	1.78299	0.86742	0.16458	0.17574	0.03718	0.01411	0.00131	Cell structure	Chitinase/extracellular matrix protein
Hv00704	1.02253	12.4865	0.01606	0.01829	0.03554	0.00998	0.016	0.21356	0.19162	Metabolism	Family 37 glycoside hydrolase (neutral trehalase)
Hv00717	0.4797	16.8843	1.80606	1.85386	0.02887	0.00811	0.52011	0.00046	8E-05	Metabolism	Fungal conserved hypothetical protein similar to malate synthase A
Hv00898	0.63376	3.81303	2.08204	2.04629	2.14348	0.25311	8.06796	0.06247	0.01499	Signaling	Fungal conserved hypothetical protein similar to catalytic subunit of cyclic AMP-dependent protein kinase
Hv01055	0.17642	67.4803	0.17342	0.19745	0.38368	0.10773	2.01277	0.27596	0.00106	Metabolism	Enoyl-CoA hydratase
Hv01133	1.16988	3.04568	0.9911	0.80173	0.34996	0.08809	0.19209	0.10159	0.02631	Metabolism	Citrate synthase (putative peroxysomal)
Hv01266	1.6161	9.49352	1.90843	0.73394	0.69469	0.079	0.11407	0.00913	0.00477	Metabolism	Isocitrate lyase
Hv01268	6.62721	6.03687	3.0546	1.30407	1.81178	0.39949	0.42311	0.02548	0.00541	Oxidative stress	Manganese superoxide dismutase
Hv01400	21.7861	32.1225	0.39251	0.44691	0.86842	0.24384	0.39086	0.01384	0.39722	Metabolism	Fungal conserved hypothetical protein similar to glycerol-3-phosphatase 1
Hv01431	0.58351	23.194	3.98369	0.01802	0.03502	1.14611	0.01576	0.60223	0.00794	Signaling	Catalytic subunit of cyclic AMP-dependent protein kinase
Hv01534	2.43435	54.6528	99.8917	39.3294	0.03511	4.09924	0.0158	7.29429	0.07934	Transport	Copper-transporting atpase 1
Hv01548	9.07846	12.0714	12.7266	5.93327	3.30629	2.60159	0.74004	1.02832	0.58709	Signaling	Adenylyl cyclase
Hv01594	0.63282	0.05241	0.3587	0.40842	2805.94	0.81017	766.871	18.6638	8.90275	Metabolism	Carnitine palmitoyltransferase II protein
Hv01628	3.76138	38.715	0.02326	0.02649	0.05147	26.5421	7.63213	0.00082	1.58465	Metabolism	Acyl-CoA dehydrogenase
Hv01629	1.71268	5.68132	0.47669	0.6158	0.21829	0.0037	0.06997	0.06322	0.01645	Metabolism	Short-chain specific acyl-coa dehydrogenase, mitochondrial precursor
Hv01805	0.60063	3.41873	0.69978	0.79676	1334.16	6916.29	171696	4467.17	303.319	Metabolism	Fungal conserved hypothetical protein similar to enoyl-CoA hydratase
Hv01932	0.61077	2.80827	3.4634	1.02717	1.43976	0.34943	1.16013	0.43472	0.16774	Signaling	Mitogen activated protein kinase kinase, map2k-ste7
Hv01988	0.58754	5.48309	68.5928	69.5714	2221.03	2952.97	133758	2386.12	256.303	Metabolism	Carnitine palmitoyltransferase II protein
Hv02022	0.3602	44.8358	10.3906	3.76459	0.04084	0.93193	2.73293	0.62952	0.04667	Signaling	G-protein beta subunit
Hv04402	0.69534	12.3966	0.01395	0.01589	0.03087	0.00867	0.01389	0.00049	0.08973	Metabolism	Family 20 Glycosyltransferase (Trehalose-phosphate synthase TPS, UDP-forming)
Hv06436	0.55549	31.6094	2.61774	1.81984	0.01325	0.56496	0.00596	0.23063	0.11324	Signaling	cAMP-dependent protein kinase type 3
Hv06448	1.1389	6.48226	3.70593	4.01523	0.04881	0.01371	0.02197	0.00078	0.0225	Metabolism	NAD-dependent glycerol-3-phosphate dehydrogenase
Hv06788	0.30129	7.70897	1.99418	1.18127	0.92582	0.0721	0.19353	0.17328	0.02286	Metabolism	Glycerol kinase

(Continued)

Table 3 | Continued

Gene	gU	Ap	18h	1d	2d	3d	7d	14d	21d	KOG function	Annotation (according to Supplementary Data 4)
Hv06883	3.62681	34.4684	2.06459	0.81996	1.73846	0.50857	1.0788	0.41016	0.71987	Signaling	Fungal conserved hypothetical protein similar to serine/threonine-protein kinase PRP4m
Hv07140	1	195.751	246.896	4713.12	1426.44	446.111	786.522	33.3738	0.07913	Signaling	Mitogen-activated protein kinase HOG1
Hv07174	1.79233	37.7162	2.69853	3.49281	0.02431	0.00683	0.01094	0.33983	0.00929	Signaling	G-protein beta subunit
Hv07400	1.24777	41.1467	11.1695	7.05488	4.73257	3.38405	13.2912	0.32332	0.02379	Metabolism	Carnitine palmitoyltransferase II protein
Hv08812	0.33774	114.518	4.75008	37.0662	113.31	39.9515	22.0792	0.16746	0.02904	Metabolism	Glycerol-3-phosphate dehydrogenase [NAD+]
Hv08827	1.03811	20.2036	1.9725	1.07745	1.87259	0.37207	1.24266	0.34945	0.54076	Signaling	Protein phosphatase 1 regulatory subunit 7
Hv08833	1.14907	1.23147	0.80187	0.41262	0.00392	0.0011	0.27852	0.40565	0.18488	Metabolism	Citrate synthase
Hv09049	4.05405	113.847	5.59448	2.96287	2.48302	0.10502	0.44865	0.35815	0.73447	Metabolism	Glycerol kinase

Values represent "fold change" related to the levels of expression in resting urediniospores (except for transcript 07140, where germinating urediniospores were used because of no amplification was obtained in resting urediniospores); Color scale: green to red denote lowest to highest expression levels across time points for each gene.

a [YFW]×C motif (Supplementary Data 4, column CX), while only 20% of the remaining 434 sequences (<5% cysteines) possess that motif. An overrepresentation of this particular motif was similarly observed for the small secreted proteins of the poplar rust fungus (Hacquard et al., 2012). Secreted proteins transcripts tend to present high relative expression values: while they represent 5.6% of all 9234 genes in this study, they represent 12–14% of genes with Ra > 1 (**Table 5**). Moreover, the sum of all Ra values for predicted secreted proteins is higher in gU than in Ap or H, although more genes were identified in H (**Figure 2**). In addition, 46 of these genes encoding predicted secreted proteins present homology to genes in the PHI database (Supplementary Data 5) whose mutants in various fungi exhibit either loss of pathogenicity or reduced virulence phenotype, 21 of which were up-regulated in the gU library, 13 in Ap and 11 in H according to Ra values.

Four transcripts (00303, 00357, 01043, and 04304) encoding predicted secreted proteins are orthologous of the rust transferred protein (RTP1) genes (Pretsch et al., 2013) from *U. fabae*, *M. occidentalis* or *M. medusae* f. sp. *deltoidis* (Supplementary Data 6). Orthologs of these four genes were also identified in *M. larici-populina* and *Puccinia* spp. genomes (Supplementary Data 7) and the overall similarity among genes is quite low (Supplementary Data 6, columns Z–AD). Three of these transcripts were previously identified in *H. vastatrix* transcripts (Fernandez et al., 2012; see Supplementary Data 6). Transcript 04304 was exclusively detected in the H library (Supplementary Data 6), corroborating the observations by Vieira et al. (2012) and the expression profile of RTP1 in *U. fabae* (Kemen et al., 2005). Transcript 01043 was detected in Ap and H, and transcripts 00303 and 00357 were identified in the three libraries. Different expression profiles could be observed, transcript 00303 being highly expressed in Ap, transcript 01043 more expressed in Ap and H and transcript 00357 showing similar Ra values across the three libraries. *H. vastatrix* RTP1 orthologs show distinct expression profiles for the different members of this single gene family. Such an observation confirms the very dynamic and specific transcriptional process at the gene family level that was reported for gene families encoding small secreted proteins of *M. larici-populina* during time course infection of poplar leaves (Duplessis et al., 2011b).

Some *H. vastatrix* transcripts are orthologous of haustorially expressed secreted proteins (HESP) identified in *Melampsora lini* (Dodds et al., 2006; Barrett et al., 2009) (Supplementary Data 7). Among these, HESP-178 is orthologous to the transcripts 01506 and 04456, detected respectively in gU and Ap, and in H libraries (Supplementary Data 8). HESP-379 is orthologous to transcript 00258, which was identified in the three libraries at decreasing levels of expression along the differentiation stages, confirming previous observations (Fernandez et al., 2012). HESP-767 is orthologous to transcript 09298 only identified in library Ap. No homology was detected to *Melampsora* spp. Avr genes (Dodds et al., 2004), presumably because of their poor conservation across the Pucciniales (Catanzariti et al., 2006; Duplessis et al., 2011a).

Several transcripts present homology to genes involved in the alleviation of oxidative stress caused by ROS. By instance, two transcripts (00297 and 01268) with homology to Mn-type

Table 4 | Distribution of *Hemileia vastatrix* genes by KOG categories considering all genes and different groups of genes according to their homology or predicted function.

Total	All genes		Gene with homology to the available rust genomes (Mlp, Pgt, Pst)		Genes with no homologies		Secreted proteins		PHI hits	
	9234	%	2992	%	4707	%	516	%	588	%
Amino acid transport and metabolism	155	5.30	119	5.68	1	0.534759	1	0.77	25	5.26
Carbohydrate transport and metabolism	153	5.23	114	5.44	5	2.673797	7	5.38	24	5.05
Cell cycle control, cell division, chromosome partitioning	77	2.63	61	2.91	6	3.208556	9	6.92	27	5.68
Cell wall/membrane/envelope biogenesis	58	1.98	35	1.67	1	0.534759	7	5.38	9	1.89
Chromatin structure and dynamics	73	2.50	54	2.58	8	4.278075	2	1.54	7	1.47
Coenzyme transport and metabolism	29	0.99	25	1.19	0	0	3	2.31	5	1.05
Cytoskeleton	141	4.82	63	3.01	33	17.64706	9	6.92	18	3.79
Defense mechanisms	19	0.65	8	0.38	0	0	0	0.00	4	0.84
Energy production and conversion	233	7.97	182	8.68	4	2.139037	6	4.62	40	8.42
Extracellular structures	20	0.68	7	0.33	6	3.208556	2	1.54	0	0.00
Inorganic ion transport and metabolism	101	3.45	68	3.24	2	1.069519	5	3.85	24	5.05
Intracellular trafficking, secretion, and vesicular transport	187	6.40	135	6.44	11	5.882353	7	5.38	22	4.63
Lipid transport and metabolism	175	5.98	136	6.49	3	1.604278	6	4.62	37	7.79
Nuclear structure	25	0.85	15	0.72	3	1.604278	5	3.85	5	1.05
Nucleotide transport and metabolism	31	1.06	22	1.05	1	0.534759	0	0.00	2	0.42
Posttranslational modification, protein turnover, chaperones	351	12.00	290	13.84	8	4.278075	22	16.92	65	13.68
Replication, recombination and repair	48	1.64	33	1.57	2	1.069519	3	2.31	5	1.05
RNA processing and modification	147	5.03	105	5.01	20	10.69519	2	1.54	12	2.53
Secondary metabolites biosynthesis, transport and catabolism	103	3.52	70	3.34	2	1.069519	5	3.85	45	9.47
Signal transduction mechanisms	238	8.14	148	7.06	32	17.1123	11	8.46	77	16.21
Transcription	162	5.54	91	4.34	18	9.625668	9	6.92	15	3.16
Translation, ribosomal structure and biogenesis	398	13.61	315	15.03	21	11.22995	9	6.92	7	1.47
General function prediction only	340	3.68	234	7.82	21	0.446144	12	2.33	78	13.27
Function unknown	168	1.82	92	3.07	9	0.191205	11	2.14	9	1.53
No hits	6159	66.70	796	26.60	4526	96.15466	382	74.03	91	15.48

For the non-specific categories ("No hits," "Function unknown" or "General function prediction only"), % refers to the total number of genes. For the remaining categories, % refers to the total number of genes with specific KOG categories assigned; highlighted cells correspond to information that is referred to in the article.

Table 5 | Distribution (%) of *Hemileia vastatrix* genes by expression profile across the three differentiation stages, considering all genes and different groups of genes according to their predicted function and relative abundance.

Expression profile	1 (gU= Ap<H)	2 (gU< Ap>H)	3 (gU> Ap=H)	4 (gU= Ap>H)	5 (gU> Ap<H)	6 (gU< Ap=H)	7 (gU= Ap=H)	8 (gU< Ap<H)	9 (gU> Ap>H)	Total
All genes	35.70	27.93	23.47	1.70	2.66	0.83	0.11	2.19	5.41	9234
Genes present in the three libraries	3.44	46.43	2.68	6.63	6.89	6.12	1.28	13.78	12.76	784
KOG CATEGORIES										
Amino acid transport and metabolism	29.29	35.00	15.00	3.57	5.00	2.86	0.00	5.71	3.57	155
Carbohydrate transport and metabolism	28.10	43.79	13.07	2.61	2.61	1.96	0.00	4.58	3.27	153
Cell cycle control, cell division, chromosome partitioning	12.70	31.75	28.57	3.17	6.35	3.17	0.00	4.76	9.52	77
Cell wall/membrane/envelope biogenesis	16.67	38.89	27.78	1.85	3.70	0.00	0.00	1.85	9.26	58
Chromatin structure and dynamics	25.37	35.82	20.90	1.49	1.49	0.00	1.49	5.97	7.46	73
Coenzyme transport and metabolism	11.54	53.85	23.08	0.00	3.85	0.00	0.00	7.69	0.00	29
Cytoskeleton	20.45	36.36	28.41	2.27	2.27	2.27	0.00	2.27	5.68	141
Defense mechanisms	18.75	56.25	18.75	0.00	0.00	6.25	0.00	0.00	0.00	19
Energy production and conversion	13.84	49.11	16.52	4.46	4.02	0.45	0.00	5.36	6.25	233
Extracellular structures	16.67	38.89	16.67	0.00	5.56	0.00	0.00	0.00	22.22	20
Inorganic ion transport and metabolism	20.00	33.00	30.00	3.00	1.00	2.00	1.00	0.00	10.00	101
Intracellular trafficking, secretion, and vesicular transport	18.06	28.39	26.45	5.16	5.16	0.65	0.65	2.58	12.90	187
Lipid transport and metabolism	15.38	51.92	18.59	2.56	2.56	1.92	0.00	1.92	5.13	175
Nuclear structure	29.41	17.65	35.29	0.00	5.88	0.00	0.00	5.88	5.88	25
Nucleotide transport and metabolism	12.90	45.16	25.81	6.45	0.00	0.00	0.00	3.23	6.45	31
Posttranslational modification, protein turnover, chaperones	16.56	44.70	13.91	2.32	5.63	3.64	0.66	4.64	7.95	351
Replication, recombination and repair	14.63	26.83	46.34	0.00	4.88	0.00	0.00	0.00	7.32	48
RNA processing and modification	27.59	28.28	24.83	4.14	4.83	3.45	0.00	3.45	3.45	147
Secondary metabolites biosynthesis, transport and catabolism	40.00	28.42	15.79	3.16	1.05	1.05	0.00	2.11	8.42	103
Signal transduction mechanisms	30.10	28.57	21.43	3.06	4.59	0.00	0.00	3.06	9.18	238
Transcription	26.95	30.50	20.57	0.71	2.84	2.13	0.00	2.84	13.48	162
Translation, ribosomal structure and biogenesis	6.79	63.97	8.88	1.57	4.18	2.35	0.78	7.57	3.92	398
All genes except No hits, Function unknown or General function prediction only	19.69	41.59	18.88	2.72	3.87	1.84	0.31	4.14	6.97	2924
General function prediction only	17.57	36.49	29.39	3.04	2.03	0.68	0.00	2.70	8.11	340
Function unknown	25.60	32.14	24.40	2.98	5.95	0.00	0.00	1.79	7.14	168
Ra > 0.05 (%)	37.52	24.81	14.07	2.12	2.72	0.45	0.00	4.54	13.77	660
Ra > 0.1 (%)	38.86	22.83	18.75	1.90	1.36	0.54	0.00	2.45	13.32	368
Ra > 1 (%)	28.57	23.21	33.93	3.57	0.00	0.00	0.00	1.79	8.93	56
SECRETED PROTEINS										
All (%)	34.50	27.13	17.25	1.55	3.49	0.97	0.39	6.20	8.53	516
Ra > 0.05 (%)	12.50	26.25	12.50	3.75	2.50	2.50	0.00	13.75	26.25	80
Ra > 0.1 (%)	13.95	23.81	21.43	2.38	0.00	2.38	0.00	7.14	28.57	43
Ra > 1 (%)	12.50	25.00	37.50	12.50	0.00	0.00	0.00	12.50	0.00	8

Highlighted cells correspond to information that is referred to in the article.

	Number of genes			Relative abundance		
	gU	Ap	H	gU	Ap	H
Secreted proteins	249	242	301	78.22	29.53	25.56
<200aa	79	61	129	31.8625	11.2911	14.0697
>5% Cysteines	27	29	48	8.1476	1.7498	3.7998
[YFW]xC motif containing	54	61	78	6.3661	11.7250	1.5273
CAZymes	89	90	57	6.809	3.358	1.225
Carbohydrate Esterase	9	12	11	0.4511	0.3405	0.2845
Carbohydrate-Binding Module	2	5	2	0.0578	0.1237	0.0663
Glycoside Hydrolase	43	40	32	2.8917	1.8615	0.6192
Glycosyltransferase	32	30	12	0.3682	0.4281	0.1109
Transporters	138	127	98	6.590	2.994	2.539
Metabolic pathways	349	395	305	14.320	8.995	22.990
Pathogenicity genes	286	292	186	6.6649	5.4458	4.9834
Signal transduction mechanisms	38	40	27	0.551	0.5401	0.6832
Posttranslational modification, protein turnover, chaperones	22	28	19	0.2415	0.4502	0.5434
Energy production and conversion	18	26	18	0.2759	0.4666	0.3882
Lipid transport and metabolism	19	24	13	0.2290	0.3573	0.1802
Secondary metabolites biosynthesis, transport and catabolism	7	16	10	0.0575	0.2008	1.0602
KOG categories						
Amino acid transport and metabolism	67	85	82	0.6158	1.0915	7.0027
Carbohydrate transport and metabolism	65	88	74	0.6730	1.3595	1.3705
Cell cycle control, cell division, chromosome partitioning	51	40	39	0.9816	0.6809	0.5840
Cell wall/membrane/envelope biogenesis	33	31	21	0.8085	0.8442	1.2180
Chromatin structure and dynamics	43	38	39	1.7490	7.0806	3.3650
Coenzyme transport and metabolism	17	18	11	0.1554	0.3872	0.1487
Cytoskeleton	78	67	63	2.8448	7.7092	5.9115
Defense mechanisms	8	12	4	0.0569	0.1171	0.1062
Energy production and conversion	130	152	98	1.9564	3.0311	1.8325
Extracellular structures	12	12	5	4.2132	2.1092	0.0562
Inorganic ion transport and metabolism	61	50	33	2.9121	1.6024	1.0367
Intracellular trafficking, secretion, and vesicular transport	139	97	75	1.8922	1.6718	1.2310
Lipid transport and metabolism	98	112	56	1.1721	2.0621	0.7688
Nuclear structure	17	10	13	0.3586	1.9497	0.1668
Nucleotide transport and metabolism	19	19	9	0.1962	0.3500	0.2885
Posttranslational modification, protein turnover, chaperones	204	228	159	3.0086	5.1990	3.0526
Replication, recombination and repair	35	18	18	0.3704	0.2931	0.2156
RNA processing and modification	84	64	74	1.0546	1.1414	0.2150
Secondary metabolites biosynthesis, transport and catabolism	41	46	48	2.5150	1.0394	9.7508
Signal transduction mechanisms	132	112	111	4.0205	8.0394	6.7880
Transcription	98	77	68	1.9524	1.7127	1.2772
Translation, ribosomal structure and biogenesis	212	325	151	4.4119	11.4528	4.9655
General function prediction only	201	171	131	3.1023	5.5329	10.8636
Function unknown	101	77	75	2.2658	1.4814	1.5117
No hits	2524	1846	3166	311.775	186.672	304.552
All genes	4267	3626	4465	350.37	245.69	355.71

FIGURE 2 | Heatmaps of the number of genes and sum of their relative abundance values (=number of transcripts/transcript length) in the three libraries (germinating urediniospores, gU; appressoria, Ap; infected leaves 21 days after inoculation, H) for transcripts according to the main categories under analysis. Color scale: green to red denote lowest to highest expression values for each gene.

superoxide dismutase, orthologs of the *Cryptococcus gattii* pathogenicity-required manganese superoxide dismutase gene (*sod2*) (Narasipura et al., 2005) were identified. Transcript 00297 is ortholog of *U. fabae* gene *Uf058*, *P. graminis* f. sp. *tritici* gene PGTG_04728 and *M. larici-populina* gene 107563. These *H. vastatrix* genes are up-regulated in germinating urediniospores and their proteins are predicted to be secreted, suggesting an early role in response to plant defense responses (**Table 3**).

CAZymes

The comparison of *H. vastatrix* transcripts to the carbohydrate-active enzymes (CAZymes) database (www.cazy.org; Cantarel et al., 2009) and to the predicted *M. larici-populina* and *P. graminis* f. sp. *tritici* CAZymes (Duplessis et al., 2011a) enabled the identification of 148 putative CAZymes in the coffee rust fungus. This number represents ca. 45 % of the CAZymes in the poplar and the wheat stem rust fungi genomes (Supplementary Data 9 and Supplementary Data 4, column DB), similar to those arising from the comparison of the total number of transcripts predicted in this study to the number of genes in those two genomes. However, the number of *H. vastatrix* CAZymes transcripts varies according to the type and family of enzymes. For instance, 13 and 14 genes belonging to the Glycoside Hydrolase family 47 were identified in *M. larici-populina* and in *P. graminis* f. sp. *tritici* respectively, while only two transcripts were detected in *H. vastatrix*. Several other gene families are found in comparable numbers in the three fungal species, even so in the most abundant families (e.g., Carbohydrate Esterase family 4, Glycoside Hydrolase families 5 and 16, Glycosyltransferase family 2). On the contrary, some transcript families found in *H. vastatrix* are absent from the genomes of *M. larici-populina* (e.g., Glycosyltransferase families 25 and 43) or *P. graminis* f. sp. *tritici* (Glycoside Hydrolase families 51 and 92). Additionally, eight Glycoside Hydrolase family 7 genes were identified both in *M. larici-populina* and in *P. graminis* f. sp. *tritici*, but none in *H. vastatrix* transcripts. CAZyme transcripts were more frequently expressed in the gU library and less in H (**Figure 2**). Among CAZymes transcripts identified in this study, 31 presented homology to genes in the PHI database.

TRANSPORTERS

A comparison to the Transporter Classification Database (www.tcdb.org; Saier et al., 2006, 2009) and to transporters from *M. larici-populina* and *P. graminis* f. sp. *tritici* (Duplessis et al., 2011a) enabled the identification of 215 transcripts encoding putative transporters. This represents ca. 60% of the number of transporters inferred from the *M. larici-populina* and *P. graminis* f. sp. *tritici* genome sequences (Duplessis et al., 2011a), again a similar proportion to that reported for other *H. vastatrix* transcript categories. However, deviations to this proportion occur in different transporter families. By instance, a family expansion is apparent in *H. vastatrix* for the F-ATPase family (H+- or Na+-translocating F-type, V-type and A-type ATPase) with 25 different transcripts predicted in *H. vastatrix*, against 20 in *M. larici-populina*, 22 in *P. graminis* f. sp. *tritici*, and 19–25 in a selection of basidiomycetes (Duplessis et al., 2011a). Similarly, variations in the Ra values can be related to the transporter type (Supplementary Data 10 and 11). In general, these results suggest that the transport capacity is at least as high in gU or Ap as in H.

Among the 215 *H. vastatrix* transcripts encoding putative transporters, 60 show homology to the PHI database. Both Ra and gene number are higher in gU and lower in H (**Figure 2**). Fourty are ATP-dependent transporters, including members of the ATP-binding cassette superfamily, (transcripts 01804, 07317, and 09267) and members of the P-type ATPase superfamily (transcripts 00176, 00302, 00402, 01534, and 07365), which are mostly expressed in the gU and Ap libraries. These results are corroborated by RT-qPCR for transcript 01534, with induction of expression both in *in vitro* and *in planta* appressorial samples (**Table 3**). The transcript 00302 is an ortholog of the *M. oryzae* P-type ATPase gene (*pde1*), required for the development of penetration hyphae and the proliferation of the fungus (Balhadère and Talbot, 2001) and was detected in the three *H. vastatrix* libraries at relatively constant expression levels. Orthologs of this gene were also identified in other rusts species (Broeker et al., 2006; Jakupović et al., 2006; Yin et al., 2009; Duplessis et al., 2011a), with elevated expression values in *P. pachyrhizi* appressoria (Stone et al., 2012). Among members of the Voltage-gated K^+ Channel β subunit family are two transcripts (00184 and 00427) identified in the three *H. vastatrix* libraries, and one (04218) only identified in the H library but at higher Ra values.

METABOLIC PATHWAYS

The availability of nutrients for the fungus is very scarce at the early stages of the infection process and energy must be obtained from urediniospore contents. Carbohydrate metabolism by glycolysis/tricarboxylic acid cycle (TCA)/glyoxylate shuttle and lipids metabolism seems to be crucial to the success of the penetration process (Solomon et al., 2004). In the present study, orthologs of genes coding several key enzymes of glycolysis and TCA pathways were identified that presented higher Ra values in gU and Ap datasets (Supplementary Data 12, panels A and E). Polyols and trehalose are among the sugars mobilized during germination (D'Enfert et al., 1999; Voegele and Mendgen, 2003). One of the major roles of trehalose seems to be the regulation of glycolysis. In the trehalose biosynthetic pathway, the intermediate trehalose 6-phosphate plays an important metabolic regulatory role by controlling glycolysis through hexokinase. In *H. vastatrix*, transcript 00156, orthologous of a hexokinase, is upregulated in Ap according to RT-qPCR results. Two transcripts (04402 and 04553) were identified orthologous of trehalose-6-phosphate synthase genes in *M. larici-populina* (gene 33497) and *P. graminis* f. sp. *tritici* (PGT_06208), and RT-qPCR results showed the accumulation of transcript 04402 in appressoria. *H. vastatrix* transcript 00704, an ortholog of a neutral trehalase (*M. larici-populina* gene 116200), was detected in the three libraries, RT-qPCR showing a peak of expression in the appressoria and at 21 days after inoculation, suggesting a close control of trehalose/trehalose-6-P levels at these stages.

The glycolysis pathway leads to the production of pyruvate after convertion into acetyl-CoA. This pathway is fundamental for cell survival since it provides intermediate metabolites and other important small molecules, such as ATP and NADH. In the present dataset, all enzymes involved in this pathway were

detected (Supplementary Data 12, panel A). A close connection between glycolysis and other pathways such as pentose phosphates and β-oxidation suggests the existence of a tight control of carbohydrate mobilization and utilization. Dihydroxy acetone phosphate, produced by aldolase by the glycerol-3-phosphate shuttle, can lead to the formation of glycerol (Supplementary Data 12, panel C) (Cronwright et al., 2002). In *H. vastatrix*, transcript 08812, ortholog of a glycerol 3-phosphate dehydrogenase, was identified in the Ap library. RT-qPCR analysis (**Table 3**) further revealed its expression during pre- and post-penetration events, strongly decreasing at late colonization stages. Similarly, transcript 01400 (glycerol 3-phosphatase gene ortholog), was accumulated in gU and Ap samples (**Table 3**). In *Saccharomyces cerevisiae*, the role of glycerol has been described in the maintenance of the cytosolic redox state (Cronwright et al., 2002). Besides, in fungi such as *Magnaporthe* or *Colletotrichum*, the important turgor pressure built in appressoria is mediated by the accumulation of very large amounts of glycerol in the cell (de Jong et al., 1997; Soanes et al., 2012). In *H. vastatrix*, transcripts with homology to glycerol 3-phosphatase and NAD^+-dependent glycerol 3-phosphate dehydrogenase (transcripts 01400 and 06448, respectively) showed higher expression during germination and appressoria formation according both to 454 pyrosequencing and RT-qPCR results. Increased levels of these enzymes were also described in *P. pachyrhizi* at the appressorial stage (Stone et al., 2012). The glycerol formed is metabolized by the action of a glycerol kinase (transcripts 06788 and 09049) the expression of which is also observed during appressorial formation according both to 454 pyrosequencing and RT-qPCR results, suggesting the importance of the maintenance of glycerol contents. While the sum of Ra values suggests higher expression of genes related to metabolism in the H library, a higher number of genes was identified in the Ap library (**Figure 2**).

Beyond the glycerol-3-phosphate shuttle, glycerol generation may also be achieved by the mobilization of storage lipids through degradation of triacylglycerol by triacylglycerol lipases (EC 3.1.1.3) (Thines et al., 2000). In fact, flexibility in lipid metabolism and ability to divert intermediates from glycolysis identified in *M. oryzae* was suggested to be important for rapid glycerol accumulation during appressorium development (Dean et al., 2005). In this study, the results suggested a high rate of lipid metabolism during germination and appressoria formation. Among the 16 putative lipases (transcripts 00223, 00443, 00530, 00606, 01163, 01746, 01917, 02201, 04308, 06521, 06529, 07167, 07216, 07621, and 09011), 12 were found in the gU library and nine in the Ap library, while only two transcripts were expressed in the H library (Supplementary Data 4). Lipid metabolism is important for ATP generation and as a source of intermediates to secondary metabolic pathways. Fatty acids are oxidized by β-oxidation, a pathway that has been referred crucial for appressorium formation, in addition to the glyoxylate cycle, to enable utilization of acetyl-CoA for central carbohydrate metabolism (Kretschmer et al., 2012; Soanes et al., 2012). The present study enabled the identification of orthologs of all genes involved in β-oxidation pathways in *M. larici-populina* and *P. graminis* f. sp. *tritici* (Supplementary Data 12, panel D). The comparison among the three *H. vastatrix* libraries revealed that fatty acid degradation increased in Ap as indicated by the increased expression of transcripts coding for several β-oxidation enzymes such as long-chain fatty acid CoA ligase (transcript 00175), acyl-CoA dehydrogenase (transcript 01629), enoyl CoA hydratase (transcript 01055), 3-hydroxyacyl-CoA dehydrogenase (transcript 01628), and 3-ketoacyl-CoA thiolase (transcript 00191). A similar profile was detected for acyl-CoA oxidase (transcript 00622). RT-qPCR profiles for these transcripts further revealed a second peak of expression at 2 days for transcripts 00191 and 01628, and at 7 days for transcript 00175. Transcripts 01594, 01988 and 07400 are orthologs of *M. oryzae* carnitine acetyl-transferase gene (*crat1*), involved in transport of peroxisomal acetyl-CoA. *M. oryzae* deletion mutants for this gene show reduced appressoria melanisation, and are not able to elaborate penetration pegs or infection hyphae (Ramos-Pamplona and Naqvi, 2006). Interestingly, in *H. vastatrix* these transcripts were only identified in gU and Ap libraries, further suggesting their potential involvement in appressorium-mediated infection. RT-qPCR analyses illustrate different expression profiles for these three transcripts: while 07400 is induced during appressoria formation both *in vitro* and *in planta*, with a second peak of induction at 7 days, transcripts 01594 and 01988 are mostly over-expressed during hyphal colonization of host tissues, after 2 days for transcript 01594 and as early as appressoria differentiation for transcript 01988 (**Table 3**).

The glyoxylate cycle provides means for cells to assimilate two-carbon compounds into the TCA cycle and channel these via gluconeogenesis to the biosynthesis of glucose, (Supplementary Data 12, panel E). Generally, induction of the glyoxylate cycle indicates that a cell is using lipid metabolism as its predominant source for ATP generation, involving β-oxidation of fatty acids and the production of acetyl CoA. In *H. vastatrix*, results showed the presence of transcripts coding for all enzymes of the glyoxylate cycle. Ra values, as well as RT-qPCR analysis for transcripts 00491, 00717, 01133, 01266, and 08833, suggest an increasing level of expression during appressoria formation. The fact that glyoxylate cycle allows the connection between lipid and carbon metabolism may be particularly important for foliar pathogenic fungi that need to germinate and develop specific infection structures before having access to plant nutrients (Wang et al., 2003).

SIGNALING

A total of 25 *H. vastatrix* transcripts presented homology to genes involved in signaling, whose mutants in various fungi exhibit either loss of pathogenicity or reduced virulence phenotype (recorded in the PHI database) (Supplementary Data 5). In the cAMP pathway, transcripts 01548 and 08827 are orthologs of pathogenicity-required adenylyl cyclase (*cdc35*) and adenylate cyclase (*cac1*) genes, respectively from *Candida albicans* (Rocha et al., 2001) and *Colletotrichum lagenarium* (Yamauchi et al., 2004), necessary for filamentous growth. Matching the relevance of this gene for spore germination and differentiation of infection structures from appressoria, RT-qPCR analysis showed the accumulation of transcript 08827 at the appressorial stage and of transcript 01548 at early infection stages, from urediniospore germination until 3 days (**Table 3**). Also, transcripts 00898 and 01431 are orthologs of the *Colletotrichum trifolii* pathogenicity-required catalytic subunit of cyclic AMP-dependent protein kinase gene

(*pkac*), necessary for penetration and sporulation (Yang and Dickman, 1999). RT-qPCR profiles showed induction of their expression in appressorial samples both obtained *in vitro* and *in planta* (**Table 3**), compatible with the involvement of these genes in penetration. Another protein kinase involved in the cAMP pathway is the *M. oryzae*/*C. trifolii* pathogenicity-required *cpkA* gene, ortholog of *H. vastatrix* transcript 06436, required for appressorium formation and pathogenesis (Mitchell and Dean, 1995; Yang and Dickman, 1999). RT-qPCR results further suggested an activation of the expression of this transcript in the appressorial stage (**Table 3**).

Several MAP kinases and serine/threonine kinases were identified, and RT-qPCR results further corroborated induction of their expression in germinating urediniospores and/or in appressoria. By instance, orthologs of the *Ustilago maydis kpp6* and *ubc3* genes (Mayorga and Gold, 1999; Brachmann et al., 2003), the *Cryphonectria parasitica cpmk1* gene (Park et al., 2004), the *Claviceps purpurea cpmk1* gene (Mey et al., 2002) and the *Cryptococcus neoformans* var. *grubii hog1* gene (Bahn et al., 2005) were identified. *H. vastatrix* transcripts 06883 and 07140, orthologs respectively of the *C. purpurea* and of *C. parasitica cpmk1* genes, were both identified only in the gU library. RT-qPCR analysis showed distinct expression profiles, with transcript 06883 induced at pre-penetration stages only, and transcript 07140 expressed at all infection stages except late colonization and resting urediniospores (**Table 3**). On the contrary, both the *ubc3* type transcript 00373 and the *hog1* type transcript 00616 were identified in the three libraries showing stable expression profiles, as corroborated by RT-qPCR results (**Table 3**). In *M. oryzae*, the Pmk1 MAP-kinase pathway has a major role in controlling appressorium morphogenesis (Soanes et al., 2012). Also, two MAPK kinases (MAP2K) were identified, corresponding to transcripts 01932 and 01813, of the Ste7 and the Mkk1 types respectively (Hamel et al., 2012), both of them identified in the gU and Ap libraries. RT-qPCR results further showed that the expression of transcript 01932 is observed at early infection stages (**Table 3**). The *H. vastatrix* transcripts 00125 and 00489, and 00409, orthologs respectively of the of the *M. oryzae* MAP kinase-regulated *gas1* and *gas2* genes (Xue et al., 2002), were identified in the three libraries, with high expression levels (Ra > 1) in the gU and Ap libraries. RT-qPCR results corroborate the induction of expression of these transcripts in germinating urediniospores and in appressoria, with a second peak of expression recorded for transcripts 00409 and 00489 respectively at 2 and 3 days (**Table 3**). Interestingly, transcript accumulation for *gas* orthologs in *P. pachyrhizi* and *Uromyces appendiculatus* purified haustoria were also reported (Link et al., 2013). The present results indicated that *gas* expression is not solely related to the rust haustorial infection structure, but also to earlier stages such as spore germination. A group of *H. vastatrix* transcripts show homology to G protein subunits genes from the PHI base. Heterotrimeric G-proteins transduce extracellular signals to various downstream effectors (e.g., MAP kinases) in eukaryotic cells. Transcript 06565 shows homology to the *Cryptococcus neoformans* virulence-related *gpa1* gene (Alspaugh et al., 1997) and to the *M. larici-populina* heterotrimeric G-protein α subunit 3 gene (*gpa3*) (Duplessis et al., 2011a). In *H. vastatrix*, transcript 06565 was only detected in the gU library, suggesting an involvement in pre-penetration events, in agreement with the profile of its *M. larici-populina* ortholog (gene 47478) (Duplessis et al., 2012). Orthologs of G-protein β subunit genes involved in appressorium formation, including genes *mgb1* from *M. oryzae* (Nishimura et al., 2003), *cgb1* from *Cochliobolus heterostrophus* (Ganem et al., 2004) and *Bpp1* from *U. maydis* (Müller et al., 2004), were identified in *H. vastatrix* (transcripts 00968 and 02022 for *mgb1* and 07174 for *cgb1*/*Bpp1*), all of them in the gU library, indicative of a possible role in appressoria formation in *H. vastatrix*. RT-qPCR expression profiling further showed induction of transcripts 02022 and 07174 during early pre-penetration events (**Table 3**).

OTHER GENES IDENTIFIED IN RUST TRANSCRIPTOMIC/GENOMIC STUDIES

Several orthologs of *U. fabae in planta*-induced genes (PIGs) were identified in *H. vastatrix* (Supplementary Data 13). The *U. fabae* PIGs genes showed induced expression in *Vicia faba* infected leaves as compared to germinating urediniospores (Jakupović et al., 2006). The majority of *U. fabae* transcripts with a predicted function have orthologs in *H. vastatrix* (Supplementary Data 13).

A comparison of *H. vastatrix* genes to *P. pachyrhizi* genes expressed in germinating urediniospores (Posada-Buitrago and Frederick, 2005) reveals that the two most expressed *P. pachyrhizi* genes, *Pp0104* and *Pp0417*, have no significant homologies in *H. vastatrix* (Supplementary Data 14). In the same way, the comparison of the 9234 *H. vastatrix* transcript to a collection of ESTs and proteins differentially expressed in *P. pachyrhizi* appressoria (Stone et al., 2012) shows a very limited number of genes in common between both studies (Supplementary Data 14). Interestingly however, the comparison of 4492 *P. pachyrhizi* haustorial ESTs (Link et al., 2013) to the 9234 *H. vastatrix* transcripts identified 1668 hits to 1132 unique *H. vastatrix* transcripts. Half of them corresponded to *H. vastatrix* transcripts not detected in the H library (Supplementary Data 14). A similar situation was observed when comparing the 7561 *U. appendiculatus* haustorial ESTs (Link et al., 2013) to the *H. vastatrix* transcripts (data not shown).

Among the 156 *M. larici-populina* annotated genes that are >10-fold up-regulated in infected leaves as compared to urediniospores (Duplessis et al., 2011a), only 22% have orthologs in the present *H. vastatrix* dataset, including 12 transporters (mostly sugar and ion transporters), 10 secreted proteins and six glycoside hydrolases (Supplementary Data 4, columns Y-AA). Among the 235 *P. graminis* f. sp. *tritici* annotated genes that are the >10-fold up-regulated in infected leaves as compared to urediniospores (Duplessis et al., 2011a), 49% have orthologs in the present *H. vastatrix* dataset, although half of these are predicted ribosomal genes. Unlike for *M. larici-populina*, none of these *P. graminis* f. sp. *tritici* genes include glycoside hydrolases or secreted proteins and only two transporters were identified.

An expanded number of multigene families have been reported in *M. larici-populina* and *P. graminis* f. sp. *tritici* as compared to other Basidiomycetes (Duplessis et al., 2011a). Among those expanded families, the number of *H. vastatrix* transcripts-based predicted genes is higher than those for *M. larici-populina* or *P. graminis* f. sp. *tritici* for the major facilitator superfamily,

helicase or chitinase, and under-represented for families such as serine/threonine protein kinase and sugar transporter (**Table 6**). While the current study does not cover all differentiation stages of the *H. vastatrix* life cycle and transcripts expressed at low level may not be represented, it is interesting to note that some gene families are over-represented in comparison to annotated genome sequences.

CONCLUSIONS

In this study, 7894 contigs were obtained by 454 pyrosequencing of cDNA from *H. vastatrix* germinating urediniospores and appressoria. These transcripts, along with 4465 *in planta* expressed contigs (Fernandez et al., 2012), were assembled into 9234 annotated transcripts. This number represents an important fraction (>50%) of the genes predicted in rust sequenced genomes so far (Duplessis et al., 2012). In addition, this elevated gene number for *H. vastatrix* is corroborated by other database comparisons, such as the core fungal genes database (FUNYBASE), the carbohydrate-active enzymes (CAZy) database or the Transporter Classification Database (TCDB). Database comparisons further indicate that half of these transcripts (4707) present no significant homology to genomic or transcriptomic data from other rusts, potentially representing novel or very divergent *H. vastatrix* genes.

Annotation of *H. vastatrix* transcripts and comparison of their relative abundance in each of the three sampling stages suggest a particularly active metabolism, translational activity, production of new structures and signaling in appressoria and intense transport, secretory activity and cellular multiplication in germinating urediniospores (**Figure 2**). Transcripts encoding putative carbohydrate-active enzymes and different types of transporters are more expressed in germinating urediniospores and appressoria, and lesser at late infection stages. Among transcripts involved in metabolic pathways, an active lipid metabolism was observed at pre-penetration stages compared to late infection stages, while aminoacid and carbohydrate metabolism was more active in post-penetration samples. Moreover, the homology of *H. vastatrix* transcripts to genes known to be involved and/or required for pathogenicity in other fungal plant pathogens, namely in appressoria-mediated infection, enabled the identification of an array of putative pathogenicity factors, a large proportion of which are expressed as early as during germ-tube elongation. Also, while melanized cuticle-breaching appressoria have been thoroughly investigated over the last few decades, namely in *M. oryzae* and *Colletotrichum* spp. (Deising et al., 2000; Kleemann et al., 2012), the present study represents an important insight into genes expressed in non-melanized stomata-penetrating appressoria. To this end, induction of expression of genes related to the production of carbohydrate-active enzymes and to the accumulation of glycerol in germinating urediniospores and appressoria suggests that combined lytic and physical mechanisms are involved in appressoria-mediated penetration of coffee leaf stomata.

This early activation of signaling, transport and secretory pathways suggests a precocious plant-fungus dialogue, which is corroborated by the possible induction of an hypersensitive reaction in stomatal cells of some resistant coffee varieties as early as at the appressorial stage (Silva et al., 2002; Ganesh et al., 2006; Diniz et al., 2012), thus prompting further studies targeting the identification of virulence/avirulence factors (and their resistance/susceptibility conterparts) expressed at these early stages of the plant-fungus interaction.

Table 6 | Comparison of the number of *Hemileia vastatrix* (Hv) genes to the number of members of gene families in *Melampsora larici-populina* (Mlp) and *Puccinia graminis* f. sp. *tritici* (Pgt) reported (Duplessis et al., 2011a) as expanded in relation to other Basidiomycetes.

Gene family	Hv	Mlp	Pgt
Amino acid transporter	12	11	12
Carbohydrate deacetylase	11	11	10
Cell division/GTP binding protein	13	6	13
Chitinase	33	14	9
Helicase	19	14	13
Histone H3	9	11	8
Major facilitator superfamily	50	29	17
Serine/threonine protein kinase	36	87	73
Sugar transporter	7	19	15
Superoxide dismutase	6	7	20
Zinc finger protein	29	4	67

Colour scale: green to red denote lowest to highest number of genes for each gene family.

AUTHOR CONTRIBUTIONS

This study was conceived and directed by Pedro Talhinhas, Helena G. Azinheira, Sébastien Duplessis, Maria do Céu Silva, and Diana Fernandez. The laboratorial experiments were conducted by Pedro Talhinhas, Helena G. Azinheira, Andreia Loureiro, Sílvia Tavares, and Anne-Sophie Petitot. 454-pyrosequencing was conducted by Julie Poulain and Corinne Da Silva. Bioinformatic analyses were conducted by Bruno Vieira, Emmanuelle Morin, and Octávio S. Paulo. Biological interpretation of bioinformatics analyses were conducted by Pedro Talhinhas, Helena G. Azinheira, Andreia Loureiro, Sílvia Tavares, Sébastien Duplessis, and Diana Fernandez. Pedro Talhinhas, Helena G. Azinheira, Andreia Loureiro, Sílvia Tavares, Dora Batista, Octávio S. Paulo, Sébastien Duplessis, Maria do Céu Silva, and Diana Fernandez wrote the paper. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was undertaken through a French-Portuguese collaborative project (Partenariat Hubert Curien PHC-Pessoa 22583XM) funded by the Ministère des Affaires Étrangères et Européennes of France. The work was funded by Portuguese national funds through Fundação para a Ciência e a Tecnologia (projects PTDC/AGR-AAM/71866/2006 and PTDC/AGR-GPL/114949/2009 and grants SFRH/BPD/47008/2008, SFRH/BPD/65965/2009 and SFRH/BPD/88994/2012), and by CEA/Genoscope-INRA-IRD Collaborative project (http://

www.genoscope.cns.fr/spip/Identification-of-virulence.html) (France), whose funding is gratefully acknowledged.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00088/abstract

REFERENCES

Aime, M. C. (2006). Toward resolving family-level relationships in rust fungi (Uredinales). *Mycoscience* 47, 112–122. doi: 10.1007/s10267-006-0281-0

Alspaugh, J. A., Perfect, J. R., and Heitman, J. (1997). *Cryptococcus neoformans* mating and virulence are regulated by the G-protein alpha subunit GPA1 and cAMP. *Genes Dev.* 11, 3206–3217. doi: 10.1101/gad.11.23.3206

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Azinheira, H. G., Silva, M. C., Guerra-Guimarães, L., Mendgen, K., Rodrigues, Jr., and C., Pinto Ricardo, C. (2001). "Development of infection structures of *Hemileia vastatrix* on artificial membranes," in *11th Conference of the Mediterranean Phytopathological Union, 17–20 September 2001* (Évora: Andalus Academic Publishing), 353–355.

Bahn, Y. S., Kojima, K., Cox, G. M., and Heitman, J. (2005). Specialization of the HOG pathway and its impact on differentiation and virulence of *Cryptococcus neoformans*. *Mol. Biol. Cell* 16, 2285–2300. doi: 10.1091/mbc.E04-11-0987

Balhadère, P. V., and Talbot, N. J. (2001). PDE1 encodes a P-type ATPase involved in appressorium-mediated plant infection by the rice blast fungus *Magnaporthe grisea*. *Plant Cell* 13, 1987–2004. doi: 10.1105/TPC.010056

Barrett, L. G., Thrall, P. H., Dodds, P. N., van der Merwe, M., Linde, C. C., Lawrence, G. J., et al. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol. Biol. Evol.* 26, 2499–2513. doi: 10.1093/molbev/msp166

Brachmann, A., Schirawski, J., Müller, P., and Kahmann, R. (2003). An unusual MAP kinase is required for efficient penetration of the plant surface by *Ustilago maydis*. *EMBO J.* 22, 2199–2210. doi: 10.1093/emboj/cdg198

Broeker, K., Bernard, F., and Moerschbacher, B. M. (2006). An EST library from *Puccinia graminis* f. *sp. tritici* reveals genes potentially involved in fungal differentiation. *FEMS Microbiol. Lett.* 256, 273–281. doi: 10.1111/j.1574-6968.2006.00127.x

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombrad, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res.* 37, D233–D238. doi: 10.1093/nar/gkn663

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. *sp. tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially-expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Chen, Z. J., Nunes, M. A., Silva, M. C., and Rodrigues, C. J. (2004). Appressorium turgor pressure of *Colletotrichum kahawae* might have a role in coffee cuticle penetration. *Mycologia* 96, 1199–1208. doi: 10.2307/3762135

Cressey, D. (2013). Coffee rust regains foothold. *Nature* 493, 587. doi: 10.1038/493587a

Cronwright, G. R., Rohwer, J. R., and Prior, B. A. (2002). Metabolic control analysis of glycerol synthesis in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* 68, 4448–4456. doi: 10.1128/AEM.68.9.4448-4456.2002

D'Enfert, C., Bonini, B. M., Zapella, P. D. A., Fontaine, T., da Silva, A. M., and Terenzi, H. F. (1999). Neutral trehalases catalyse intracellular trehalose breakdown in the filamentous fungi *Aspergillus nidulans* and *Neurospora crassa*. *Mol. Microbiol.* 32, 471–484. doi: 10.1046/j.1365-2958.1999.01327.x

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., et al. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* 434, 980–986. doi: 10.1038/nature03449

Deising, H. B., Werner, S., and Wernitz, M. (2000). The role of fungal appressoria in plant infection. *Microbes Infect.* 2, 1631–1641. doi: 10.1016/S1286-4579(00)01319-8

de Jong, J. C., McCormack, B. J., Smirnoff, N., and Talbot, N. J. (1997). Glycerol generates turgor in rice blast. *Nature* 389, 244–245. doi: 10.1038/38418

Diniz, I., Talhinhas, P., Azinheira, H. G., Várzea, V., Medeira, C., Maia, I., et al. (2012). Cellular and molecular analyses of coffee resistance to *Hemileia vastatrix* and nonhost resistance to *Uromyces vignae* in the resistance-donor genotype HDT832/2. *Eur. J. Plant Pathol.* 133, 141–157. doi: 10.1007/s10658-011-9925-9

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini* AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16, 755–768. doi: 10.1105/tpc.020040

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Teh, T., Wang, C. I. A., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8888–8893. doi: 10.1073/pnas.0602577103

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011b) *Melampsora larici-populina* transcript profiling during germination and time-course infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant-Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Duplessis, S., Joly, D. L., and Dodds, P. N. (2012). "Rust effectors," in *Effectors in Plant-Microbe Interactions*, eds F. Martin and S. Kamoun (New York, NY: Wiley), 155–193.

Ekblom, R., Balakrishnan, C. N., Burke, T., and Slate, J. (2010). Digital gene expression analysis of the zebra finch genome. *BMC Genomics* 11:219. doi: 10.1186/1471-2164-11-219

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016. doi: 10.1006/jmbi.2000.3903

Fernandez, D., Talhinhas, P., and Duplessis, S. (2013). "Rust fungi: achievements and future challenges on genomics and host-parasite interactions," in *The Mycota XI, Agricultural Applications 2nd Edn.*, ed F. Kempken (Berlin: Springer-Verlag), 315–341.

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H. G., Vieira, A., Loureiro, A., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals *in planta* expressed pathogen secreted proteins and plant functions expressed in a late compatible plant-rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Ganem, S., Lu, S. W., Lee, B. N., Chou, D. Y., Hadar, R., Turgeon, B. G., et al. (2004). G-protein beta subunit of *Cochliobolus heterostrophus* involved in virulence, asexual and sexual reproductive ability, and morphogenesis. *Eukaryot. Cell* 3, 1653–1663. doi: 10.1128/EC.3.6.1653-1663.2004

Ganesh, D., Petitot, A., Silva, M. C., Alary, R., Lecouls, A. C., and Fernandez, D. (2006). Monitoring of the early molecular resistance responses of coffee (*Coffea arabica* L.) to the rust fungus (*Hemileia vastatrix*) using real-time quantitative RT-PCR. *Plant Sci.* 170, 1045–1051. doi: 10.1016/j.plantsci.2005.12.009

Hacquard, S., Joly, D. L., Lin, Y.-C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hall, T. A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.* 41, 95–98.

Hamel, L.-P., Nicole, M.-C., Duplessis, S., and Ellis, B. E. (2012). Mitogen-activated protein kinase signaling in plant-interacting fungi: distinct messages from conserved messengers. *Plant Cell* 24, 1327–1351. doi: 10.1105/tpc.112.096156

Howard, R. J., Ferrari, M. A., Roach, D. H., and Money, N. P. (1991). Penetration of hard substrates by a fungus employing enormous turgor pressures. *Proc. Natl. Acad. Sci. U.S.A.* 88, 11281–11284. doi: 10.1073/pnas.88.24.11281

Hu, G. G., Linning, R., McCallum, B., Banks, T., Cloutier, S., Butterfield, Y., et al. (2007). Generation of a wheat leaf rust, *Puccinia triticina*, EST database from stage-specific cDNA libraries. *Mol. Plant Pathol.* 8, 451–467. doi: 10.1111/j.1364-3703.2007.00406.x

Jakupović, M., Heintz, M., Reichmann, P., Mendgen, K., and Hahn, M. (2006). Microarray analysis of expressed sequence tags from haustoria of the rust fungus *Uromyces fabae*. *Fungal Genet. Biol.* 43, 8–19. doi: 10.1016/j.fgb.2005.09.001

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Kleemann, J., Rincon-Rivera, L. J., Takahara, H., Neumann, U., Ver Loren van Themaat, E., van der Does, H. C., et al. (2012). Sequential delivery of host-induced virulence effectors by appressoria and intracellular hyphae of the phytopathogen *Colletotrichum higginsianum*. *PLoS Pathog.* 8:e1002643. doi: 10.1371/journal.ppat.1002643

Kretschmer, M., Klose, J., and Kronstad, J. W. (2012). Defects in mitochondrial and peroxisomal β-oxidation influence virulence in the maize pathogen *Ustilago maydis*. *Eukaryot. Cell* 11, 1055–1066. doi: 10.1128/EC.00129-12

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Kumar, S., and Blaxter, M. L. (2010). Comparing de novo assemblers for 454 transcriptome data. *BMC Genomics* 11:571. doi: 10.1186/1471-2164-11-571

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2013). The haustorial transcriptomes of *Uromyces appendiculatus* and *Phakopsora pachyrhizi* and their candidate effector families. *Mol. Plant Pathol.* doi: 10.1111/mpp.12099. [Epub ahead of print].

Mank, J. E., Hultin−Rosenberg, L., Zwahlen, M., and Ellegren, H. (2008). Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. *Am. Nat.* 171, 35–43. doi: 10.1086/523954

Marthey, S., Aguileta, G., Rodolphe, F., Gendrault, A., Giraud, T., Fournier, E., et al. (2008). FUNYBASE: a FUNgal phYlogenomic dataBASE. *BMC Bioinformatics* 9:456. doi: 10.1186/1471-2105-9-456

Mayorga, M. E., and Gold, S. E. (1999). A MAP kinase encoded by the ubc3 gene of *Ustilago maydis* is required for filamentous growth and full virulence. *Mol. Microbiol.* 34, 485–497. doi: 10.1046/j.1365-2958.1999.01610.x

Mey, G., Oeser, B., Lebrun, M. H., and Tudzynski, P. (2002). The biotrophic, non-appressorium-forming grass pathogen *Claviceps purpurea* needs a Fus3/Pmk1 homologous mitogen-activated protein kinase for colonization of rye ovarian tissue. *Mol. Plant Microbe Interact.* 15, 303–312. doi: 10.1094/MPMI.2002.15.4.303

Mitchell, T. K., and Dean, R. A. (1995). The cAMP-dependent protein kinase catalytic subunit is required for appressorium formation and pathogenesis by the rice blast pathogen *Magnaporthe grisea*. *Plant Cell* 7, 1869–1878.

Müller, P., Leibbrandt, A., Teunissen, H., Cubasch, S., Aichinger, C., and Kahmann, R. (2004). The Gβ-subunit-encoding gene bpp1 controls cyclic-AMP signaling in *Ustilago maydis*. *Eukaryot. Cell* 3, 806–814. doi: 10.1128/EC.3.3.806-814.2004

Narasipura, S. D., Chaturvedi, V., and Chaturvedi, S. (2005). Characterization of *Cryptococcus neoformans* variety *gattii* SOD2 reveals distinct roles of the two superoxide dismutases in fungal biology and virulence. *Mol. Microbiol.* 55, 1782–1800. doi: 10.1111/j.1365-2958.2005.04503.x

Nishimura, M., Park, G., and Xu, J. R. (2003). The G-beta subunit MGB1 is involved in regulating multiple steps of infection-related morphogenesis in *Magnaporthe grisea*. *Mol. Microbiol.* 50, 231–243. doi: 10.1046/j.1365-2958.2003.03676.x

Park, S. M., Choi, E. S., Kim, M. J., Cha, B. J., Yang, M. S., and Kim, D. H. (2004). Characterization of HOG1 homologue, CpMK1, from *Cryphonectria parasitica* and evidence for hypovirus-mediated perturbation of its phosphorylation in response to hypertonic stress. *Mol. Microbiol.* 51, 1267–1277. doi: 10.1111/j.1365-2958.2004.03919.x

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Posada-Buitrago, M. L., and Frederick, R. D. (2005). Expressed sequence tag analysis of the soybean rust pathogen *Phakopsora pachyrhizi*. *Fungal Genet. Biol.* 42, 949–962. doi: 10.1016/j.fgb.2005.06.004

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins—a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Ramos-Pamplona, M., and Naqvi, N. I. (2006). Host invasion during rice-blast disease requires carnitine-dependent transport of peroxisomal acetyl-CoA. *Mol. Microbiol.* 61, 61–75. doi: 10.1111/j.1365-2958.2006.05194.x

Rocha, C. R., Schröppel, K., Harcus, D., Marcil, A., Dignard, D., Taylor, B. N., et al. (2001). Signaling through adenylyl cyclase is essential for hyphal growth and virulence in the pathogenic fungus *Candida albicans*. *Mol. Biol. Cell* 12, 3631–3643. doi: 10.1091/mbc.12.11.3631

Saier, M. H. Jr., Tran, C. V., and Barabote, R. D. (2006). TCDB: the transporter classification database for membrane transport protein analyses and information. *Nucleic Acids Res.* 34, D181–D186. doi: 10.1093/nar/gkj001

Saier, M. H. Jr., Yen, M. R., Noto, K., Tamang, D. G., and Elkan, C. (2009). The Transporter Classification Database: recent advances. *Nucleic Acids Res.* 37, D274–D278. doi: 10.1093/nar/gkn862

Silva, D. N., Vieira, A., Talhinhas, P., Azinheira, H. G., Silva, M. C., Fernandez, D., et al. (2012). "Phylogenetic analysis of *Hemileia vastatrix* and related taxa using a genome-scale approach," in *Proceedings of the 24th International Conference on Coffee Science, 11–16 November 2012, San José*, ed Association for Science and Information on Coffee (Paris), 1404–1408.

Silva, M. C., Nicole, M., Guerra-Guimarães, L., and Rodrigues, C. J. Jr. (2002). Hypersensitive cell death and post-haustorial defence responses arrest the orange rust (*Hemileia vastatrix*) growth in resistant coffee leaves. *Physiol. Mol. Plant Pathol.* 60, 169–183. doi: 10.1006/pmpp.2002.0389

Silva, M. C., Várzea, V., Guimarães, L. G., Azinheira, H. G., Fernandez, D., Petitot, A., et al. (2006). Coffee resistance to the main diseases: leaf rust and coffee berry disease. *Braz. J. Plant Physiol.* 18, 119–147. doi: 10.1590/S1677-04202006000100010

Soanes, D. M., Chakrabarti, A., Paszkiewicz, K. H., Dawe, A. L., and Talbot, N. J. (2012). Genome-wide transcriptional profiling of appressorium development by the rice blast fungus *Magnaporthe oryzae*. *PLoS Pathog.* 8:e1002514. doi: 10.1371/journal.ppat.1002514

Soanes, D. M., and Talbot, N. J. (2006). Comparative genomic analysis of phytopathogenic fungi using expressed sequence tag (EST) collections. *Mol. Plant Pathol.* 7, 61–70. doi: 10.1111/j.1364-3703.2005.00317.x

Solomon, P. S., Lee, R. C., Wilson, T. J. G., and Oliver, R. P. (2004). Pathogenicity of *Stagonospora nodorum* requires malate synthase. *Mol. Microbiol.* 53, 1065–1073. doi: 10.1111/j.1365-2958.2004.04178.x

Stone, C. L., McMahon, M. B., Fortis, L. L., Nuñez, A., Smythers, G. W., Luster, D. G., et al. (2012). Gene expression and proteomic analysis of the formation of *Phakopsora pachyrhizi* appressoria. *BMC Genomics* 13:269. doi: 10.1186/1471-2164-13-269

Tatusov, R. L. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4:41. doi: 10.1186/1471-2105-4-41

Terhune, B. T., Bojko, R. J., and Hoch, H. C. (1993). Deformation of stomatal guard cell lips and microfabricated artificial topographies during appressorium formation. *Exp. Mycol.* 17, 70–78. doi: 10.1006/emyc.1993.1006

Thines, E., Weber, R. W., and Talbot, N. J. (2000). MAP kinase and protein kinase A-dependent mobilization of triacylglycerol and glycogen during appressorium turgor generation by *Magnaporthe grisea*. *Plant Cell* 12, 1703–1718. doi: 10.1105/tpc.12.9.1703

Várzea, V. M. P., and Marques, D. V. (2005). "Population variability of *Hemileia vastatrix* vs coffee durable resistance," in *Durable Resistance to Coffee Leaf Rust*, eds L. Zambolim, E. Zambolim, and V. M. P. Várzea (Viçosa: Universidade Federal de Viçosa), 53–74.

Vega-Arreguín, J. C., Ibarra-Laclette, E., Jiménez-Moraila, B., Martínez, O., Vielle-Calzada, J. P., Herrera-Estrella, L., et al. (2009). Deep sampling of the Palomero maize transcriptome by a high throughput strategy of pyrosequencing. *BMC Genomics* 10:299. doi: 10.1186/1471-2164-10-299

Vieira, A., Talhinhas, P., Loureiro, A., Duplessis, S., Fernandez, D., Silva, M. C., et al. (2011). Validation of RT-qPCR reference genes for *in planta* expression studies in *Hemileia vastatrix*, the causal agent of coffee leaf rust. *Fungal Biol.* 115, 891–901. doi: 10.1016/j.funbio.2011.07.002

Vieira, A., Talhinhas, P., Loureiro, A., Thürich, J., Duplessis, S., Fernandez, D., et al. (2012). Expression profiling of genes involved in the biotrophic colonisation of *Coffea arabica* leaves by *Hemileia vastatrix*. *Eur. J. Plant Pathol.* 133, 261–277. doi: 10.1007/s10658-011-9864-5

Voegele, R. T., Hahn, M., and Mendgen, K. (2009). "The uredinales: cytology, biochemistry, and molecular biology," in *The Mycota, 5. Plant Relationships.* Vol. 2, ed H. Deising (Berlin: Springer), 69–98.

Voegele, R. T., and Mendgen, K. (2003). Rust haustoria: nutrient uptake and beyond. *New Phytol.* 159, 93–100. doi: 10.1046/j.1469-8137.2003.00761.x

Wang, Z. Y., Thornton, C. R., Kershaw, M. J., Debao, L., and Talbot, N. J. (2003). The glyoxylate cycle is required for temporal regulation of virulence by the plant pathogenic fungus *Magnaporthe grisea*. *Mol. Microbiol.* 47, 1601–1612. doi: 10.1046/j.1365-2958.2003.03412.x

Winnenburg, R., Baldwin, T. K., Urban, M., Rawlings, C., Köhler, J., and Hammond-Kosack, K. E. (2007). PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res.* 34, D459–D464. doi: 10.1093/nar/gkm858

Xue, C., Park, G., Choi, W., Zheng, L., Dean, R. A., and Xu, J. R. (2002). Two novel fungal virulence genes specifically expressed in appressoria of the rice blast fungus. *Plant Cell* 14, 2107–2119. doi: 10.1105/tpc.003426

Yamauchi, J., Takayanagi, N., Komeda, K., Takano, Y., and Okuno, T. (2004). cAMP-pKA signaling regulates multiple steps of fungal infection cooperatively with Cmk1 MAP kinase in *Colletotrichum lagenarium*. *Mol. Plant Microbe Interact.* 17, 1355–1365. doi: 10.1094/MPMI.2004.17.12.1355

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* 21, 650–659. doi: 10.1093/bioinformatics/bti042

Yang, Z., and Dickman, M. B. (1999). *Colletotrichum trifolii* mutants disrupted in the catalytic subunit of cAMP-dependent protein kinase are nonpathogenic. *Mol. Plant Microbe Interact.* 12, 430–439. doi: 10.1094/MPMI.1999.12.5.430

Yin, C., Chen, X., Wang, X., Han, Q., Kang, Z., and Hulbert, S. H. (2009). Generation and analysis of expression sequence tags from haustoriaof the wheat stripe rust fungus *Puccinia striiformis* f. *sp. tritici*. *BMC Genomics* 10:626. doi: 10.1186/1471-2164-10-626

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 22 January 2014; paper pending published: 06 February 2014; accepted: 24 February 2014; published online: 14 March 2014.

Citation: Talhinhas P, Azinheira HG, Vieira B, Loureiro A, Tavares S, Batista D, Morin E, Petitot A-S, Paulo OS, Poulain J, Da Silva C, Duplessis S, Silva MC and Fernandez D (2014) Overview of the functional virulent genome of the coffee leaf rust pathogen Hemileia vastatrix with an emphasis on early stages of infection. Front. Plant Sci. ***5****:88. doi: 10.3389/fpls.2014.00088*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Talhinhas, Azinheira, Vieira, Loureiro, Tavares, Batista, Morin, Petitot, Paulo, Poulain, Da Silva, Duplessis, Silva and Fernandez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

MINI REVIEW ARTICLE
published: 20 August 2014
doi: 10.3389/fpls.2014.00416

Effector proteins of rust fungi

Benjamin Petre[1,2,3], David L. Joly[4] and Sébastien Duplessis[1,2]*

[1] INRA, UMR 1136 Interactions Arbres/Microorganismes, Centre INRA Nancy Lorraine, Champenoux, France
[2] UMR 1136 Interactions Arbres/Microorganismes, Faculté des Sciences et Technologies, Université de Lorraine, Vandoeuvre-lès-Nancy, France
[3] The Sainsbury Laboratory, Norwich Research Park, Norwich, UK
[4] Département de Biologie, Université de Moncton, Moncton, NB, Canada

Edited by:
Ken Shirasu, University of California, Davis, USA

Reviewed by:
Ralf Thomas Voegele, Universität Hohenheim, Germany
Jeffrey Ellis, Commonwealth Scientific and Industrial Research Organisation, Australia

***Correspondence:**
Sébastien Duplessis, INRA, UMR 1136 Interactions Arbres/Microorganismes, Centre INRA Nancy Lorraine, Champenoux 54280, France
e-mail: duplessi@nancy.inra.fr

Rust fungi include many species that are devastating crop pathogens. To develop resistant plants, a better understanding of rust virulence factors, or effector proteins, is needed. Thus far, only six rust effector proteins have been described: AvrP123, AvrP4, AvrL567, AvrM, RTP1, and PGTAUSPE-10-1. Although some are well established model proteins used to investigate mechanisms of immune receptor activation (avirulence activities) or entry into plant cells, how they work inside host tissues to promote fungal growth remains unknown. The genome sequences of four rust fungi (two Melampsoraceae and two Pucciniaceae) have been analyzed so far. Genome-wide analyses of these species, as well as transcriptomics performed on a broader range of rust fungi, revealed hundreds of small secreted proteins considered as rust candidate secreted effector proteins (CSEPs). The rust community now needs high-throughput approaches (effectoromics) to accelerate effector discovery/characterization and to better understand how they function *in planta*. However, this task is challenging due to the non-amenability of rust pathosystems (obligate biotrophs infecting crop plants) to traditional molecular genetic approaches mainly due to difficulties in culturing these species *in vitro*. The use of heterologous approaches should be promoted in the future.

Keywords: Pucciniales, rust fungi, genomics, transcriptomics, effectoromics

THE KNOWN RUST FUNGAL EFFECTOR PROTEINS

Plant pathogens secrete effector proteins into host tissues to promote infection through the manipulation of host processes (Win et al., 2012). During host colonization, rust fungi form haustoria that invaginate the host plasma membrane within the host cell cavity. These structures mediate the molecular traffic between the parasite and its host, and notably the delivery of effector proteins into host cells (Rafiqi et al., 2012), although other structures such as infection hyphae are also likely to be involved in this molecular traffic (Rafiqi et al., 2010). Until now, six effector proteins have been identified in three different rust species: AvrM, AvrL567, AvrP123, and AvrP4 in the flax rust fungus *Melampsora lini*, the Rust Transferred Protein RTP1 in the bean rust fungus *Uromyces fabae*, and PGTAUSPE-10-1 in the wheat stem rust fungus *Puccinia graminis* f. sp. *tritici* (**Table 1**; Kemen et al., 2005; Ellis et al., 2007; Upadhyaya et al., 2014). They are all secreted proteins expressed in haustoria, with no clearly identified biochemical function. How they promote fungal growth inside host tissues remains unknown (**Table 1**). In contrast, their avirulence (Avr) properties (i.e., the ability to trigger specific immune responses) and/or their trafficking mechanisms (i.e., how they enter plant cells) are better understood.

The four *M. lini* effector proteins were first identified as effectors due to their Avr properties (Ellis et al., 2007). More recently, a screen with a bacterial protein delivery system in wheat revealed the *P. graminis* f. sp. *tritici* protein PGTAUSPE-10-1 which causes cell death in the host line carrying the resistance gene Sr22; PGTAUSPE-10-1 was thus considered as a candidate AvrRs22 effector (Upadhyaya et al., 2014). *M. lini* AvrL567 and AvrM are model Avrs for the study of effector recognition by immune receptors. Both proteins are recognized inside plant cells by specific immune receptors following a direct physical interaction (**Table 1**; Dodds et al., 2004, 2006; Catanzariti et al., 2006, 2010). For both effectors, 3D structure-driven amino acid substitutions revealed multiple contact points mediating the interaction with their cognate receptor (Wang et al., 2007; Ravensdale et al., 2011; Ve et al., 2013). Amino acid residues within these contact points are highly variable, suggesting that an arms race is taking place between these effectors and their corresponding receptors. Such knowledge of Avr-receptor interactions is valuable for engineering improved immune receptors with expanded effector recognition (Harris et al., 2013; Segretin et al., 2014), which may ultimately help to develop broad-spectrum resistance in plants (Dangl et al., 2013).

All six rust effector proteins are thought to be translocated from haustoria into host cells (**Table 1**). RTP1 and AvrM have been directly shown to traffic from haustoria to plant cells during infection (Kemen et al., 2005, 2013; Rafiqi et al., 2010), whereas the direct recognition of AvrM and AvrL567 by cytosolic plant immune receptors indirectly demonstrates their internalization in the plant cell (Ellis et al., 2007). Current mechanistic models based on pathogen-free assays suggest that AvrP4, AvrM, and AvrL567 proteins can enter plant cells autonomously (Catanzariti et al., 2006; Kale et al., 2010; Rafiqi et al., 2010). Rafiqi et al. (2010) further showed that AvrL567 and AvrM cell entry is mediated by divergent N-terminal uptake domains, carrying hydrophobic residues that are critical for cell entry in the case of

Table 1 | Rust effector proteins.

Effector protein	aa residues (mature)	Signal peptide	Expression	Localization in infected tissues	Avr property (immune receptor)	Biochemical function	Role in virulence
AvrM	284–347	Yes	Haustorium[a]	Haustorium, EHMx, plant cytosol[a]	Yes (M)	nd	nd
AvrL567	127	Yes	Haustorium	Plant cytosol	Yes (L5, L6, L7)	nd	nd
AvrP123	94	Yes	Haustorium	Plant nucleus	Yes (P, P1, P2, P3)	nd	nd
AvrP4	65	Yes	Haustorium	Plant cytosol	Yes (P4)	nd	nd
RTP1	201	Yes	Haustorium[a]	Haustorium/ EHMx/plant cytosol/ plant nucleus[a]	nd	Protease inhibitor/filament-forming	nd
PGTAUSPE-10-1	np	np	Haustorium	nd	yes[b]	nd	nd

The table details the rust fungi effector proteins reported so far.
Avr, Avirulence; aa, amino acid; EHMx, extra-haustorial matrix; nd, not determined; ND, not detected; np, not published.
[a] *Direct evidence of the presence of the protein acquired by immunolocalization.*
[b] *a host-specific toxic effect was detected.*

AvrM (Ve et al., 2013). This model and the assays used to build it are currently debated, and the need to study effector trafficking during the infection has been stressed (Petre and Kamoun, 2014).

Effector proteins are anticipated to be key molecules for pathogenicity, although very little is known about how they function within host tissues. Among the six characterized rust effectors, none possess a clearly identified biochemical function or a detected virulence activity (**Table 1**). Indeed, *M. lini* transgenic lines silencing AvrL567 did not show any reduced growth on flax, suggesting that this effector is not required for full virulence (Lawrence et al., 2010). As discussed by the authors, this could be explained by a high functional redundancy in the *M. lini* effector repertoire (Lawrence et al., 2010). Such redundancy was also observed in the effector repertoires of bacterial plant pathogens (Kvitko et al., 2009), and represents an obstacle for the functional characterization of virulence effector functions through genetic approaches. However, recent progresses have been made regarding RTP1, a conserved rust effector that seems to work as a protease inhibitor (Pretsch et al., 2013). On the other hand, Kemen et al. (2013) reported that RTP1 accumulates within the host-parasite interface and forms filaments. The authors proposed a role as a structural effector, possibly stabilizing fungal structures during infection. A model that integrates the different RTP1 localizations and proposed functions remains to be drawn. Several methods for the genetic transformation of *M. lini* and *U. fabae*, as well as for host-induced gene silencing (HIGS) of *Puccinia triticina* have been reported (Lawrence et al., 2010; Djulic et al., 2011; Panwar et al., 2013). Such methods, although they are still at various stages of development, represent valuable tools to investigate the contribution of individual effectors to virulence during infection.

POST-GENOMIC APPROACHES IDENTIFY A PLETHORA OF RUST SECRETED PROTEINS CONSIDERED AS CANDIDATE EFFECTORS

In the past few years, a typical profile has emerged for plant pathogen effectors. Fungal proteins are usually considered candidate secreted effector proteins (CSEPs) if they possess a signal peptide for secretion, a small size and no other targeting sequence or transmembrane domains (Stergiopoulos and de Wit, 2009; Rouxel and Tyler, 2012; Saunders et al., 2012). Such CSEPs attract more attention when they are expressed during infection or when they present signatures of rapid evolution. Besides, expression in specific infection structures such as haustoria, often considered as a major site of effector delivery, provides another level of information. Some authors also take advantage of conserved amino acid motifs or predicted protein structures to establish large CSEP classes (Godfrey et al., 2010; Pedersen et al., 2012). Homology to known rust effectors and organization in gene families or in physical clusters have also been considered to refine these sets of CSEPs (Hacquard et al., 2012; Saunders et al., 2012). In rust fungi, such criteria have been applied in the frame of effector mining pipelines that combined genome-wide analyses and transcriptomics to reveal amazingly rich catalogs of rust CSEPs (Cantu et al., 2011, 2013; Duplessis et al., 2011a; Fernandez et al., 2012; Hacquard et al., 2012; Saunders et al., 2012; Garnica et al., 2013; Zheng et al., 2013; Bruce et al., 2014; Link et al., 2014; Nemri et al., 2014; **Table 2**).

GENOME-WIDE ANALYSES OF CSEPs

The genome sequences of four rust species have been published so far: *Melampsora larici-populina* (poplar leaf rust fungus; Duplessis et al., 2011a), *M. lini* (flax rust fungus; Nemri et al., 2014), *P. graminis* f. sp. *tritici* (wheat stem rust fungus;

Table 2 | Secreted proteins considered as rust effector candidates in transcriptome studies.

Species	Interaction, biological stage	Transcriptome approach	Number of transcripts detected	Detailed analysis of CSEPs	Publication
Hemileia vastatrix	Infected leaves	454-pyrosequencing GS-FLX titanium	6,763 fungal transcripts	382 predicted CSEPs	Fernandez et al. (2012)
H. vastatrix	Urediniospores and appressoria	454-pyrosequencing GS-FLX titanium	9,234 unique fungal transcripts	516 predicted CSEPs; abondant among the most highly expressed genes, particularly *in planta*	Talhinhas et al. (2014)
Melampsora larici-populina	Laser capture microdissection of infected leaves	Custom whole-genome oligoarrays	7,288 to 8,145 transcripts expressed in uredinia or in mesophyll tissues	19 CSEPs in the 25 most highly up-regulated transcripts in palisade mesophyll (haustoria) compared to uredinia	Hacquard et al. (2010)
M. larici-populina	Infected leaves, urediniospores	Custom whole-genome oligoarrays	>7,500 transcripts expressed in each biological condition tested	509 of 1,184 predicted CSEP genes expressed *in planta*; 50 CSEP among the top 100 genes up-regulated *in planta*	Duplessis et al. (2011a)
M. larici-populina	Time-course infection of leaves	Custom whole-genome oligoarrays	<500 early expressed transcripts; up to 8 326 transcripts *in planta*	270 CSEP genes specifically expressed *in planta*; distinct sets of >500 CSEP genes coordinately expressed along the time course	Duplessis et al. (2011b)
M. larici-populina	Early infected leaves	454-pyrosequencing GS-FLX titanium	90,398 contigs; 649 reads aligned to 361 fungal genes	19 early expressed CSEP genes among 40 fungal genes supported by more than 3 reads	Petre et al. (2012)
M. larici-populina	Telia (autumn)	Custom whole-genome oligoarrays	9,588 transcripts expressed in telia	11 SSP genes specifically expressed in telia; 113 SSP genes up-regulated in telia *vs.* uredinia	Hacquard et al. (2013)
Phakopsora pachyrhizi	Purified haustoria	454-pyrosequencing GS-FLX titanium	4,483 *P. pachyrhizi* unique contigs	156 contigs encoding CSEPs	Link et al. (2014)
P. pachyrhizi	Infected leaves	Illumina GA II	32,940 *P. pachyrhizi* contigs	176 predicted CSEP genes	Tremblay et al. (2012)
P. pachyrhizi	Time-course infection of leaves	Illumina GA II	Up to 12,284 *P. pachyrhizi* transcripts expressed	Not mentioned	Tremblay et al. (2013)
Puccinia graminis f. sp. *tritici*	Infected leaves, urediniospores	Custom whole-genome oligoarrays	9,818 transcripts expressed in total	442 of 1,106 predicted CSEP genes expressed *in planta*; 29 CSEPs in top-100 *in planta* up-regulated genes	Duplessis et al. (2011a)

(Continued)

Table 2 | Continued

Species	Interaction, biological stage	Transcriptome approach	Number of transcripts detected	Detailed analysis of CSEPs	Publication
Puccinia striiformis f. sp. *tritici* (5 isolates)	Infected leaves and purified haustoria	Illumina Genome Analyzer II	12–28.8 Millions reads from infected leaves and purified haustoria	933 CSEPs; 57 and 31 CSEP genes induced or repressed in haustoria *vs. in planta*, respectively	Cantu et al. (2013)
P. striiformis f. sp. *tritici*	Purified haustoria and urediniospores	454-pyrosequencing GS-FLX titanium and Illumina GA II	12,282 transcripts from combined transcriptomes	437 Haustoria Secreted Proteins (HSP); expression confirmed for 71 HSP genes by RT-qPCR	Garnica et al. (2013)
Puccinia triticina (6 isolates)	Infected leaves	Illumina RNA-Seq	222,571 fungal reads	543 CSEP transcripts (445 shared by the 6 isolates)	Bruce et al. (2014)
Uromyces appendiculatus	Purified haustoria	454-pyrosequencing GS-FLX Titanium	7,582 *U. appendiculatus* contigs	413 contigs encoding CSEPs	Link et al. (2014)

This table compiles the most recent genome-scale transcriptome studies in rust fungi (i.e., custom genome oligarrays and 454/Illumina-based RNA-Seq). Identification of expressed CSEPs is detailed. See Duplessis et al. (2012) for a detailed analysis of previous transcriptome studies in rust fungi based on Sanger expressed sequence tags or cDNA-arrays.

Duplessis et al., 2011a) and *Puccinia striiformis* f. sp. *tritici* (wheat stripe rust fungus; Cantu et al., 2011, 2013; Zheng et al., 2013). Genome-wide effector mining in these four species revealed hundreds of genes encoding CSEPs. In *M. larici-populina*, 1,184 CSEPs have been identified from 1,898 genes encoding predicted secreted proteins (Duplessis et al., 2011a). In *M. lini*, 762 priority CSEPs were selected from 1,085 genes encoding predicted secreted proteins (Nemri et al., 2014). In *P. graminis* f. sp. *tritici*, 1,106 CSEP genes were selected from 1,934 genes encoding predicted secreted proteins (Duplessis et al., 2011a). In *P. striiformis* f. sp. *tritici*, different reports of selected sets of CSEPs have been published. In this rust fungus, a total of 2,092 CSEP coding genes were considered in isolate CY-32 (Zheng et al., 2013) while the draft genome of isolate PST-130 led to 1,088 filtered CSEPs out of 1,188 genes coding predicted secreted protein (Cantu et al., 2011). However, genome re-sequencing of four other isolates and cross-comparison with PST-130 has led to a revision of gene numbers and to a larger set of 2,999 predicted CSEPs (Cantu et al., 2013).

All rust fungi genomes are marked by expansions of gene families, particularly those encoding secreted proteins. For instance, the largest CSEP gene family in *M. larici-populina* includes 111 members (Duplessis et al., 2011a). Noteworthy, a part of these genes were not predicted by algorithms but rather found by manual curation, highlighting the importance of expert annotation of these atypical gene families of small proteins (Duplessis et al., 2011a; Hacquard et al., 2012). This last observation is important to consider when performing cross-comparison between genomes showing different degrees of annotation. Since RXLR or LXLFLAK conserved motifs found in oomycetes helped defining large effector families (Win et al., 2007), a particular focus on motif search was given in rust CSEPs. The motif [YFW]xC has been reported in the genomes of obligate biotrophic pathogens of cereals, including *P. graminis* f. sp. *tritici* (Godfrey et al., 2010). In *M. larici-populina*, this motif is common, eventually with positional constraints, but with no restriction to the N-terminus of CSEPs (Hacquard et al., 2012). Nonetheless, functional and structural characterization for the [YFW]xC motif is lacking at the moment, and no evidence for a role in translocation has been provided so far.

Another common trend observed in rust candidate effector repertoires is the large proportion of species-, family- or order-specific CSEPs (Duplessis et al., 2014a). A large majority of species-specific CSEP genes (nearly 70%) were first observed in *M. larici-populina*. With the sequencing of the flax rust genome this number has reduced, as only 4% of the *M. lini* CSEP genes were found to be species-specific and more than half had a homolog in one of the three other sequenced rust genomes (Nemri et al., 2014). Interestingly, *M. lini Avr* genes homologs are only found in *M. larici-populina* and thus could be considered family-specific effectors, whereas other genes such as *Uromyces* spp. *RTP1* or some Haustorially Expressed Secreted Proteins (HESPs) identified in *M. lini* are conserved across rust fungi (Fernandez et al., 2012). Sequencing more genomes among Pucciniales, particularly in uncovered taxonomic families, will definitely help defining the common set of core rust effectors and those that may be related to host adaptation (Duplessis et al., 2014b).

TRANSCRIPTOMICS IDENTIFY CSEPS IN MANY RUST SPECIES

Rust fungi have rather large genomes (89–190 Mb) and an important content in repetitive elements (>43% of total genomes), which impedes the systematic sequencing and assembly of targeted species (Duplessis et al., 2014b). Indeed, genome size estimates for certain rust species go beyond the numbers given above (Leonard and Szabo, 2005; Tavares et al., 2014). Whole-genome oligoarrays or RNA-Seq has thus proven to be useful in gathering relevant information about the transcriptomes of rust fungi. A strong stage specific regulation of protein secretion has been demonstrated in *U. fabae* (Link and Voegele, 2008), and novel high-throughput approaches confirmed a coordinated expression of CSEPs during host infection, in a temporal (expression at specific time-points) or spatial (expression in specific structures) manner (**Table 2**). For instance, transcripts profiling during time-course infection of poplar leaves by *M. larici-populina* revealed waves of expression for more than 500 CSEP transcripts (Hacquard et al., 2010; Duplessis et al., 2011b; Petre et al., 2012). Moreover, such temporal succession of expression patterns has been confirmed in other rust species such as *Hemileia vastatrix* (Fernandez et al., 2012), *P. striiformis* f. sp. *tritici* (Cantu et al., 2013), and *Puccinia triticina* (Bruce et al., 2014). This highlights the need for a better understanding of expression regulation in rust fungi, whether by transcription factors or via epigenetic control, such as reported in *Phytophthora infestans* or in *Leptosphaeria maculans* (Judelson, 2012; Soyer et al., 2014).

Interestingly, different reports showed that *U. fabae* RTP1 homologs may have different localizations (Kemen et al., 2005; Hacquard et al., 2012). RTP1 also exhibits a dynamic pattern of localization in the extra-haustorial matrix and within host cells during the infection process (Kemen et al., 2013), illustrating once more that rust effectors deployment is probably finely regulated in time and space. In this regard, a major issue with *in planta* expression study is the occurrence of different fungal cell types (germ tubes, appressoria, substomatal vesicles, infection hyphae, haustoria, sporogenous hyphae, and newly formed spores), which implies that the observed expression levels are often a mixture of different cell types at different stages. After the seminal paper that described a method to purify haustoria from the bean rust fungus (Hahn and Mendgen, 1997) and the one reporting on *M. lini* HESPs that included several Avr genes (Catanzariti et al., 2006), haustoria purification has been combined with RNA-Seq studies to prioritize CSEPs likely delivered by these infection structures (Cantu et al., 2013; Garnica et al., 2013, Link et al., 2014). Laser capture microdissection has also been coupled to transcriptomics to distinguish between biotrophic and sporogenous areas in poplar leaves infected by *M. larici-populina* (Hacquard et al., 2010). This study demonstrated that CSEPs are predominantly and highly expressed in the area containing infection hyphae and haustoria.

In order to complete their life cycle, heterecious rust fungi infect two unrelated host species. To do so, it is likely that they express host-specific effector sets. However, except for the wheat leaf rust *P. triticina* (Xu et al., 2011), only a small portion of the life cycle has been surveyed in most rust species. Recently, in order to expand our understanding of the transcriptome of *M. larici-populina*, gene expression analyses were conducted on rust telia collected from decaying leaves (Hacquard et al., 2013). This study revealed that CSEP-encoding genes were expressed in these tissues, suggesting that CSEPs might have additional roles unrelated to the interaction with the living host plant (Hacquard et al., 2013). Ongoing transcriptome profiling studies in different rust species will help to determine the sets of CSEP genes expressed along the life cycle. Such studies may reveal CSEPs with a host-specific expression, which represent host-adapted effectors (Duplessis et al., 2014b).

TOWARDS UNIFIED EFFECTOR MINING AND EFFECTOROMICS PIPELINES

Various studies combined genome sequencing and transcriptomics to provide sets of CSEPs. Automated pipelines for effector mining should be unified and systematically applied to forthcoming rust fungi genomes to provide a solid foundation for future comparative analyses in Pucciniales. However, an important point to consider is the need for an accurate curation of CSEP-encoding genes in these genomes and the screening of additional time points in time-course studies and/or spore stages. Some early genome-wide surveys of CSEPs in plant interacting fungi arbitrarily focused on small proteins because of the commonly observed small size of effectors and in order to reduce manual gene curation efforts (Stergiopoulos and de Wit, 2009, Duplessis et al., 2011a). Considering that rust fungi effectors can exhibit greater size (e.g., *M. lini* AvrM), such an arbitrary cut-off should not be considered in future analyses of rust CSEPs.

To face the growing number of CSEPs made available by effector mining studies, and to better understand their functions in plant cells, we need tools to study them directly *in planta*. This relies on the ability to genetically transform the plant to perform high-throughput functional analyses (also referred to as "effectoromics"). Rust fungi hosts (e.g., wheat, soybean, flax, or poplar), are not easily amenable to molecular genetic approaches. However, non-host model plants can be used to characterize and screen CSEPs. For instance, the *Agrobacterium*-mediated transient genetic transformation of *Nicotiana benthamiana* has proven useful to rapidly express effector proteins into plant cell, but has been largely ignored in rust effector biology. This system allows combining many different approaches (cell-biology, protein biochemistry, hypersensitive response and infection assays) all in one. Thus, such approaches may help in (1) determining the sub-cellular localization of candidate effector proteins using effector-fluorescent protein fusions, (2) identifying interacting partners within protein complexes, (3) detecting candidate effector capacity to enhance susceptibility during infection with selected *N. benthamiana* pathogens (thus validating a role in virulence), and (4) testing their recognition by specific immune receptors.

AUTHOR CONTRIBUTIONS

Benjamin Petre and Sébastien Duplessis compiled data from the literature and drafted the manuscript. All the authors wrote and revised the article.

ACKNOWLEDGMENTS

The authors would like to thank the reviewers for their helpful comments to set the final version of this article. Benjamin Petre

thanks Diane Saunders (JIC, UK), Sophien Kamoun (TSL, UK), and Kofyorty Stactafyzal (NRP, UK) for great discussions. Benjamin Petre is supported by INRA, in the framework of a Contrat Jeune Scientifique, by the European Union, in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills' fellowship (under grant agreement no. 267196) and by the LABEX Arbre, through the award of a mobility grant. Research in The Sainsbury Lab is supported by the Gatsby Charitable Foundation, the European Research Council, and the Biotechnology and Biological Sciences Research Council (BBSRC). Research in David L. Joly lab is supported by the New Brunswick Innovation Foundation. Sébastien Duplessis acknowledges the support of the French ANR for a grant part of the "Investissements d'Avenir" program (ANR-11-LABX-0002-01, Lab of Excellence ARBRE) and the Young Scientist Grant POPRUST to Sébastien Duplessis (ANR-2010-JCJC-1709-01), and the Région Lorraine.

REFERENCES

Bruce, M., Neugebauer, K. A., Joly, D. L., Migeon, P., Cuomo, C. A., Wang, S., et al. (2014). Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat. *Front. Plant Sci.* 4:520 doi: 10.3389/fpls.2013.00520

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Catanzariti, A-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Catanzariti, A-M., Dodds, P. N., Ve, T., Kobe, B., Ellis, J. G., and Staskawicz, B. J. (2010). The AvrM effector from flax rust has a structured C-terminal domain and interacts directly with the M resistance protein. *Mol. Plant Microbe Interact.* 23, 49–57. doi: 10.1094/MPMI-23-1-0049

Dangl, J. L., Horvath, D. M., and Staskawicz, B. J. (2013). Pivoting the plant immune system from dissection to deployment. *Science* 341, 746–751. doi: 10.1126/science.1236011

Djulic, A., Schmid, A., Lenz, H., Sharma, P., Koch, C., Wirsel, S. G., et al. (2011). Transient transformation of the obligate biotrophic rust fungus *Uromyces fabae* using biolistics. *Fungal Biol.* 115, 633–642. doi: 10.1016/j.funbio.2011.03.007

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16, 755–768. doi: 10.1105/tpc.020040

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Teh, T., Wang, C. I., Ayliffe, MA., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8888–8893. doi: 10.1073/pnas.0602577103

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011b). *Melampsora larici-populina* transcript profiling during germination and time-course infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Duplessis, S., Joly, D. J., and Dodds, P. N. (2012). "Rust effectors," in *Effectors in Plant-Microbes Interactions*, eds F. Martin and S. Kamoun (Oxford: Wiley-Blackwell), 155–193.

Duplessis, S., Spanu, P. D., and Schirawski, J. (2014a). "Biotrophic fungi (powdery mildews, Rusts and Smuts)," in *Ecological Genomics of the Fungi. Plant-Interacting Fungi Section,* ed F. Martin (Hoboken, NJ: Wiley-Blackwell), 149–168.

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014b). Advancing knowledge on biology of rust fungi through genomics. *Adv Bot. Res.* 70, 173–209. doi: 10.1016/B978-0-12-397940-7.00006-9

Ellis, J. G., Dodds, P. N., and Lawrence, G. J. (2007). Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. *Annu. Rev. Phytopathol.* 45, 289–306. doi: 10.1146/annurev.phyto.45.062806.094331

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A-S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals *in planta*-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PLoS ONE* 8:e67150. doi: 10.1371/journal.pone.0067150

Godfrey, D., Böhlenius, H., Pedersen, C., Zhang, Z., Emmersen, J., and Thordal-Christensen, H. (2010). Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genomics* 11:317. doi: 10.1186/1471-2164-11-317

Hacquard, S., Delaruelle, C., Legué, V., Tisserant, E., Kohler, A., Frey, P., et al. (2010). Laser capture microdissection of uredinia formed by *Melampsora larici-populina* revealed a transcriptional switch between biotrophy and sporulation. *Mol. Plant Microbe Interact.* 23, 1275–1286. doi: 10.1094/MPMI-05-10-0111

Hacquard, S., Delaruelle, C., Frey, P., Tisserant, E., Kohler, A., and Duplessis, S. (2013). Transcriptome analysis of poplar rust telia reveals overwintering adaptation and tightly coordinated karyogamy and meiosis processes. *Front. Plant Sci.* 4:456. doi: 10.3389/fpls.2013.00456

Hacquard, S., Joly, D. L., Lin, Y.-C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hahn, M., and Mendgen, K. (1997). Characterization of in planta-induced rust genes isolated from a haustorium-specific cDNA library. *Mol. Plant Microbe Interact.* 10, 427–437. doi: 10.1094/MPMI.1997.10.4.427

Harris, C. J., Slootweg, E. J., Goverse, A., and Baulcombe, D. C. (2013). Stepwise artificial evolution of a plant disease resistance gene. *Proc. Natl. Acad. Sci. U.S.A.* 110, 21189–21194. doi: 10.1073/pnas.1311134110

Judelson, H. (2012). Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. *Eukaryot. Cell* 11, 1304–1312. doi: 10.1128/ec.00155-12

Kale, S. D., Gu, B., Capelluto, D. G., Dou, D., Feldman, E., Rumore, A., et al. (2010). External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell* 142, 284–295. doi: 10.1016/j.cell.2010.06.008

Kemen, E., Kemen, A., Ehlers, A., Voegele, R., and Mendgen, K. (2013). A novel structural effector from rust fungi is capable of fibril formation. *Plant J.* 75, 767–780. doi: 10.1111/tpj.12237

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Kvitko, B. H., Park, D. H., Velasquez, A. C., Wei, C-F., Russel, A. B., Martin, G. B., et al. (2009). Deletions in the repertoire of *Pseudomonas syringae* pv. tomato DC3000 type III secretion effector genes reveal functional overlap among effectors. *PLoS Pathog.* 5:e1000388. doi: 10.1371/journal.ppat.1000388

Lawrence, G. J., Dodds, P. N., and Ellis, J. G. (2010). Transformation of the flax rust fungus, *Melampsora lini*: selection via silencing of an avirulence gene. *Plant J.* 61, 364–369. doi: 10.1111/j.1365-313X.2009.04052.x

Leonard, K. J., and Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol. Plant Pathol.* 6, 99–111. doi: 10.1111/j.1364-3703.2005.00273.x

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2014). The haustorial transcriptomes of *Uromyces* appendiculatus and Phakopsora pachyrhizi and their candidate effector families. *Mol. Plant Pathol.* 15, 379–393. doi: 10.1111/mpp.12099

Link, T. I., and Voegele, R. T. (2008). Secreted proteins of *Uromyces fabae*: similarities and stage specificity. *Mol. Plant Pathol.* 9, 59–66. doi: 10.1111/j.1364-3703.2007.00448.x

Nemri, A., Saunders, D. G., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Panwar, V., McCallum, B., and Bakkeren, G. (2013). Endogenous silencing of *Puccinia triticina* pathogenicity genes through in planta-expressed sequences leads to the suppression of rust diseases on wheat. *Plant J.* 73, 521–532. doi: 10.1111/tpj.12047

Pedersen, C., Ver Loren van Themaat, E., McGuffin, L. J., Abbott, J. C., Burgis, T. A., Barton, G., et al. (2012). Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics* 13:694. doi: 10.1186/1471-2164-13-694

Petre, B., and Kamoun, S. (2014). How do filamentous pathogens deliver effector proteins into plant cells? *PLoS Biol.* 12:e1001801. doi: 10.1371/journal.pbio.1001801

Petre, B., Morin, E., Tisserant, E., Hacquard, S., Da Silva, C., Poulain, J., et al. (2012). RNA-Seq of early-infected poplar leaves by the rust pathogen *Melampsora larici-populina* uncovers PtSultr3;5, a fungal-induced host sulfate transporter. *PLoS ONE* 7:e44408. doi: 10.1371/journal.pone.0044408

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins-a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Rafiqi, M., Gan, P. H., Ravensdale, M., Lawrence, G. J., Ellis, J. G., Jones, D. A., et al. (2010). Internalization of flax rust avirulence proteins into flax and tobacco cells can occur in the absence of the pathogen. *Plant Cell* 22, 2017–2032. doi: 10.1105/tpc.109.072983

Rafiqi, M., Ellis, J. G., Ludowici, V. A., Hardham, A. R., and Dodds, P. N. (2012). Challenges and progress towards understanding the role of effectors in plant-fungal interactions. *Curr. Opin. Plant Biol* 15, 477–82. doi: 10.1016/j.pbi.2012.05.003

Ravensdale, M., Bernoux, M., Ve, T., Kobe, B., Thrall, P. H., Ellis, J. G., et al. (2011). Intramolecular interaction influences binding of the Flax L5 and L6 resistance proteins to their AvrL567 ligands. *PLoS Pathog.* 8:e1003004. doi: 10.1371/journal.ppat.1003004

Rouxel, T., and Tyler, B. M. (2012). "Effector of fungi and Oomycetes: their virulence and avirulence functions and translocation from pathogen to host cells," in *Molecular Plant Immunity*, ed. G. Sessa (Oxford: John Wiley & Sons), 123–167.

Saunders, D. G., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:e29847. doi: 10.1371/journal.pone.0029847

Segretin, M. E., Pais, M., Franceschetti, M., Chaparro-Garcia, A., Bos, J. I., Banfield, M. J., et al. (2014). Single amino acid mutations in the potato immune receptor R3a expand response to *Phytophthora* effectors. *Mol. Plant Microbe Interact.* 27, 624–637. doi: 10.1094/MPMI-02-14-0040-R

Soyer, J. L., El Ghalid, M., Glaser, N., Ollivier, B., Linglin, J., Grandaubert, J., et al. (2014). Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans*. *PLoS Genetics* 10:e1004227. doi: 10.1371/journal.pgen.1004227

Stergiopoulos, I., and de Wit, P. J. (2009). Fungal effector proteins. *Annu. Rev. Phytopathol.* 47, 233–263. doi: 10.1146/annurev.phyto.112408.132637

Talhinhas, P., Azinheira, H. G., Vieira, B., Loureiro, A., Tavares, S., Batista, D., et al. (2014). Overview of the functional virulent genome of the coffee leaf rust pathogen *Hemileia vastatrix* with an emphasis on early stages of infection. *Front. Plant Sci.* 5:88. doi: 10.3389/fpls.2014.00088

Tavares, S., Ramos, A. P., Pires, A. S., Azinheira, H. G., Caldeirinha, P., Link, T., et al. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. *Front. Plant Sci.* 5:422. doi: 10.3389/fpls.2014.00422

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2012). Identification of genes expressed by *Phakopsora pachyrhizi*, the pathogen causing soybean rust, at a late stage of infection of susceptible soybean leaves. *Plant Pathol.* 61, 773–786. doi: 10.1111/j.1365-3059.2011.02550.x

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2013). Analysis of *Phakopsora pachyrhizi* transcript abundance in critical pathways at four time-points during infection of a susceptible soybean cultivar using deep sequencing. *BMC Genomics* 11:614. doi: 10.1186/1471-2164-14-614

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M. A., Ellis, J. G., and Dodds, P. N. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol. Plant Microbe Interact.* 27, 255–264. doi: 10.1094/MPMI-07-13-0187-FI

Ve, T., Williams, S. J., Catanzariti, A-M., Rafiqi, M., Rahman, M., Ellis, J. G., et al. (2013). Structures of the flax-rust effector AvrM reveal insights into the molecular basis of plant-cell entry and effector-triggered immunity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 17594–17599. doi: 10.1073/pnas.1307614110

Wang, C. I., Guncar, G., Forwood, J. K., Teh, T., Catanzariti, A. M., Lawrence, G. J., et al. (2007). Crystal structures of flax rust avirulence proteins AvrL567-A and -D reveal details of the structural basis for flax disease resistance specificity. *Plant Cell* 19, 2898–2912. doi: 10.1105/tpc.107.053611

Win, J., Chaparro-Garcia, A., Belhaj, K., Saunders, D. G., Yoshida, K., Dong, S., et al. (2012). Effector biology of plant-associated organisms: concepts and perspectives. *Cold. Spring Harb. Symp. Quant. Biol.* 77, 235–247 doi: 10.1101/sqb.2012.77.015933

Win, J., Morgan, W., Bos, J., Krasileva, K. V., Cano, L. M., Chaparro-Garcia, A., et al. (2007). Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *Plant Cell* 19, 2349–2369. doi: 10.1105/tpc.107.051037

Xu, J., Linning, R., Fellers, J., Dickinson, M., Zhu, W., Antonov, I., et al. (2011). Gene discovery in EST sequences from the wheat leaf rust fungus *Puccinia triticina* sexual spores, asexual spores and haustoria, compared to other rust and corn smut fungi. *BMC Genomics* 12:161. doi: 10.1186/1471-2164-12-161

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4, 2673 doi: 10.1038/ncomms3673

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 21 June 2014; accepted: 04 August 2014; published online: 20 August 2014.

Citation: Petre B, Joly DL and Duplessis S (2014) Effector proteins of rust fungi. Front. Plant Sci. 5:416. doi: 10.3389/fpls.2014.00416

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Petre, Joly and Duplessis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 13 January 2014
doi: 10.3389/fpls.2013.00520

Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat

Myron Bruce[1], Kerri A. Neugebauer[2], David L. Joly[3], Pierre Migeon[2], Christina A. Cuomo[4], Shichen Wang[2], Eduard Akhunov[2], Guus Bakkeren[5], James A. Kolmer[6] and John P. Fellers[1]*

[1] *USDA-ARS Hard Winter Wheat Genetics Research Unit, Department of Plant Pathology, Manhattan, KS, USA*
[2] *Department of Plant Pathology, Kansas State University, Manhattan, KS, USA*
[3] *Département de biologie, Université de Moncton, Moncton, NB, Canada*
[4] *Broad Institute of MIT and Harvard, Cambridge, MA, USA*
[5] *Pacific Agri-Food Research Centre, Agriculture and Agri-Food Canada, Summerland, BC, Canada*
[6] *USDA–ARS Cereal Disease Laboratory, Department of Plant Pathology, University of Minnesota, St. Paul, MN, USA*

Edited by:
Sébastien Duplessis, INRA, France

Reviewed by:
Mahmut Tör, University of Worcester, UK
Scot Hulbert, Washington State University, USA

****Correspondence:***
John P. Fellers, USDA-ARS, Department of Plant Pathology, 4008 Throckmorton Hall, Manhattan, KS 66506, USA
e-mail: john.fellers@ars.usda.gov

Wheat leaf rust, caused by the basidiomycete *Puccinia triticina*, can cause yield losses of up to 20% in wheat producing regions. During infection, the fungus forms haustoria that secrete proteins into the plant cell and effect changes in plant transcription, metabolism, and defense. It is hypothesized that new races emerge as a result of overcoming plant resistance via changes in the secreted effector proteins. To understand gene expression during infection and find genetic differences associated with races, RNA from wheat leaves infected with six different rust races, at 6 days post inoculation, was sequenced using Illumina. As *P. triticina* is an obligate biotroph, RNA from both the host and fungi were present and separated by alignment to the *P. triticina* genome and a wheat EST reference. A total of 222,571 rust contigs were assembled from 165 million reads. An examination of the resulting contigs revealed 532 predicted secreted proteins among the transcripts. Of these, 456 were found in all races. Fifteen genes were found with amino acid changes, corresponding to putative avirulence effectors potentially recognized by 11 different leaf rust resistance (*Lr*) genes. Twelve of the potential avirulence effectors have no homology to known genes. One gene had significant similarity to cerato-platanin, a known fungal elicitor, and another showed similarity to fungal tyrosinase, an enzyme involved in melanin synthesis. Temporal expression profiles were developed for these genes by qRT-PCR and show that the genes expression patterns were consistent between races from infection initiation to just prior to spore eruption.

Keywords: *Puccinia triticina*, secreted peptides, effectors, leaf rust, RNA sequencing

INTRODUCTION

Wheat leaf rust can cause extensive economic impact on the wheat producing areas of the world, with significant yield losses reported during epidemics. The causative agent of wheat leaf rust is *Puccinia triticina* (*Pt*) which is an obligate biotrophic basidiomycete (Bolton et al., 2008). In the Great Plains of the US, the pathogen spreads by asexual urediniospores that land on the leaf surface and germinate under appropriate conditions. Using thigmotrophic interactions with the surface of the leaf, the germ tube orients itself perpendicular to the leaf veins and grows until it reaches a stomata. The germ tube generates an appressorium over the stomata and penetrates the interior of the leaf. Directly beneath the site of penetration, the haustorial mother cell is formed and infection structures penetrate the plant cell wall to form feeding structures called haustoria. While the haustorium penetrates the plant cell wall, it does not rupture the plant cell membrane. Communication between the growing fungus and the plant occurs across the haustorial membrane-plasma membrane interface in the form of proteins and other small molecules secreted by the fungus. The secreted proteins, some of which may be effectors, perform a variety of functions including host transcriptional reprogramming to benefit pathogen growth and mitigation of host defenses (Bolton et al., 2008).

Plants have several systems of defense to guard against infection of pathogens. Effector triggered immunity is the most distinct and stems from the host's ability to recognize the presence or activity of pathogen effectors. The host cell responds by inducing a localized response to isolate the infection. Originally, this was described at the genetic level by Flor as the "gene for gene" hypothesis (Flor, 1942). Flor's model posits that if a pathogen protein (coded by an avirulence gene) is recognized by a plant protein (coded by a resistance gene, or *R* gene), then resistance is triggered in the plant. This is usually characterized by hypersensitive cell death, or HR (Mur et al., 2008). While the genetics of avirulence in the flax rust (*Melampsora lini*)—flax (*Linum usitatissimum*) pathosystem had been described by Flor (1971), the molecular mechanism underlying the interaction between pathogen avirulence gene and host resistance gene has only become more clear as technologies advance. Dodds et al. validated the theory in flax rust by showing that a member of the gene family, *AvrL567*, is expressed in haustoria and secreted into the plant cell cytoplasm where its activity is recognized (Dodds et al., 2004).

As research has progressed, the activities of various "avirulence" genes have been shown to have a beneficial effect on the pathogen's fitness and/or ability to cause disease. Thus, the corresponding gene products were renamed "effectors." Effectors from plant pathogenic fungi include protease inhibitors, chitin binding proteins, metalloproteases, and many genes of unknown function (Stergiopoulos and de Wit, 2009). Several effectors from rust pathogens have been described. Rust transferred proteins (RTP) have a demonstrated protease inhibitor function (Pretsch et al., 2013) and RTP1 was recently shown to be a structural effector involved in filament formation in the extra-haustorial matrix (Kemen et al., 2013). RTPs were first described in *Uromyces fabae* (bean rust pathogen) and shown to translocate from the haustorium to the plant cytoplasm (Kemen et al., 2005). *Pt* encodes three proteins from this family, though their function have not been determined (Pretsch et al., 2013).

There are many difficulties encountered with *Pt* as a study system. The fungus is an obligate biotroph which cannot be cultured outside of its host. Additionally, the alternate host, *Thalictrum speciosissium*, is required for the sexual stage of the organism but is not known to be widely present in North America (Bolton et al., 2008). While controlled crosses have been useful in examining heritability of avirulence factors (Samborski and Dyck, 1968; Statler, 2000), the crossing and purification process is time-consuming and can take up to two years to develop a mapping population. Therefore, genetic studies of race structure and pathogenicity require novel approaches and are usually performed on the asexual urediniospore stage of the pathogen's life cycle.

Genomic resources for fungal pathogens provide valuable research tools in understanding the dynamics of plant-pathogen interactions (Dean et al., 2005; Kamper et al., 2006; Cantu et al., 2011; Duplessis et al., 2011a; Fernandez et al., 2012). Hacquard et al. recently published a genome-wide analysis of the poplar leaf rust pathogen (*Melampsora larici-populina*) small secreted proteins, detailing 29 different secreted cysteine repeat-containing protein families with a total of 228 proteins represented (Hacquard et al., 2012). With the availability of *Pt* and *P. graminis* f. sp. *tritici* (wheat stem rust fungus, *Pgt*) reference genomes (http://www.broadinstitute.org/scientific-community/data) a search for the same patterns in the predicted proteomes of these species revealed only four and eleven proteins, respectively. Of these, only two were predicted to be secreted in *Pt* and nine in *Pgt. graminis* (Bruce et al, unpublished data). This indicates that the *Puccinia* group may not use the same effector set as the *Melampsora* group, which may be related to differences between their host plants. To date, only one avirulence effector from cereal rusts has been verified (Nirmala et al., 2011). The presence of two rust proteins from *Pgt* is recognized by the barley resistance protein, RPG1. The rust proteins directly interact with RPG1 in yeast two hybrid experiments and activate an RPG1-mediated hypersensitive response (Nirmala et al., 2011).

With the rapidly decreasing cost of genome and transcriptome sequencing, understanding the mechanisms of pathogenesis and virulence in these organisms is becoming less difficult. In this study, six *Pt* races were inoculated on a susceptible host. Six days after inoculation (DAI), leaves were harvested and RNA extracted. Transcript-enriched RNA was sequenced using Illumina next generation sequencing and the resulting reads assembled. To identify potential fungal effectors, amino acid changes found within secreted peptides were identified in the assembly and correlated to the virulence patterns observed for the races. Using this approach, we have identified 15 candidate avirulence effectors and characterized their expression during the infection process.

RESULTS

The six *Pt* targeted races were all found in North America and their avirulence/virulence combinations are listed in **Table 1**. MHDS and MLDS belong to North American lineage 3 (NA3; Tremblay et al., 2013) and were collected in 2004 in Kansas and Ohio, respectively. These two races only differ in their reaction to *Lr 9* and *Lr16* using the standard differential set. MJBJ, THBJ, TDBG, and TNRJ belong to lineage NA5 and were collected in 1997 in Nebraska, 2005 in Texas, 2004 in Texas and 2004 in Kansas, respectively (Tremblay et al., 2013). Each are much more varied in their reactions to the differential wheat lines. TNRJ was the most virulent *Pt* race at the time this study was started. Wheat plants from the susceptible cultivar Thatcher (Tc) were inoculated separately, with each of the six rust races. The inoculations were heavy with a majority of the leaf area showing a significant infection reaction at 6 days post inoculation. Pustule formation was apparent, but urediniospores had not erupted (**Figure 1**).

Following inoculation and incubation, leaves with heavy infections were harvested and RNA was extracted. Total RNA was treated to remove ribosomal RNA, without requiring a polyA selection. Transcriptome RNA representing the wheat and leaf rust transcriptomes was fragmented and used to generate first and second strand cDNA. The second strand cDNA was size-fractionated and amplified in gel prior to Illumina sequencing. A total of 165 million raw reads were generated by Illumina sequencing using standard parameters with paired end 60 bp reads. The sequenced transcriptomes were assembled using Trinity (Grabherr et al., 2011). The transcriptomes were separated

Table 1 | Listing of *P. triticina* races used in the experiment.

Race 1	*Lr2a, Lr2c, Lr16, Lr24, Lr26, Lr3ka, Lr11, Lr30, Lr18, Lr1, Lr3a, Lr9, Lr17, LrB, Lr10/Lr14a*
MLDS	*Lr2a, Lr2c, Lr16, Lr24, Lr26, Lr3ka, Lr11, Lr30, Lr18/Lr1, Lr3a, Lr9, Lr17, LrB, Lr10, Lr14a*
MHDS	*Lr2a, Lr2c, Lr9, Lr24, Lr3ka, Lr11, Lr30, Lr18/Lr1, Lr3a, Lr16, Lr26, Lr17, LrB, Lr10, Lr14a*
MJBJ	*Lr2a, Lr2c, Lr9, Lr26, Lr3ka, Lr11, Lr17, Lr30, LrB, Lr18/Lr1, Lr3a, Lr16, Lr24, Lr10, Lr14a*
TDBG	*Lr9, Lr16, Lr26, Lr3ka, Lr11, Lr17, Lr30, LrB, Lr14a, Lr18/Lr1, Lr2a, Lr2c, Lr3a, Lr10, Lr24*
THBJ	*Lr9, Lr24, Lr3ka, Lr11, Lr17, Lr30, LrB, Lr18/Lr1, Lr2a, Lr2c, Lr3a, Lr16, Lr26, Lr10, Lr14a*
TNRJ	*Lr16, Lr26, Lr17, LrB, Lr18/Lr1, Lr2a, Lr2c, Lr3a, Lr9, Lr24, Lr3ka, Lr11, Lr30, Lr10, Lr14a*

Races are named based on their reaction to leaf rust resistance gene (Lr) differential Thatcher isolines as described by McIntosh et al. (1995). Lr genes with a low infection type (IT = 0; resistant) are highlighted in blue and Lr genes with a high infection type (IT = 3–4) susceptible are in red. Race1 has the pathotype BBBD.

into wheat-associated and leaf rust-associated files by aligning the resulting assembled contigs to the TIGR wheat EST database (available at http://www.jcvi.org) or to the leaf rust draft genome V2 of Race1 (pathotype BBBD, http://www.broadinstitute.org/scientific-community/data). A total of 222,571 leaf rust contigs were identified from the assembled contigs. Statistics for the individual races are found in **Table 2**.

The bioinformatic workflow for identification of secreted proteins in RNA-Seq transcripts is represented in **Figure 2**. SignalP, TargetP, and TMHMM prediction algorithms were used to determine which proteins among the predicted genes in the race 1 reference genome may be secreted (Petersen et al., 2011). Six frame translations of the assembled transcripts from the six races were produced *in silico* to make a set of amino acid sequences from the assembly data. There were 1450 predicted Race1 (BBBD) secreted amino acid sequences. These were used as a BLASTP query against the six race translated transcriptome (Altschul et al., 1990). The BLAST results were parsed using a custom Python script [http://www.python.org, (Cock et al., 2009)] to extract the Race1 (BBBD) identifier, the six race protein identifier, number of positive matches, number of identity matches and the alignment length. There were a total of 4726 alignments from the six race proteins with an identity/alignment length ratio greater than 0.95, of these, 543 unique secreted proteins were found across the six races, and 445 were shared by all six races (**Supplementary Table 1**).

To determine whether or not the virulence patterns observed could be correlated to non-synonymous protein changes, the BLASTP query results were parsed and filtered using a Python script [http://www.python.org, (Cock et al., 2009)] to extract gene identifiers, unique race identifiers and match quality. This data was filtered by race to generate lists of gene names that matched the reference sequence exactly, or had at least one SNP resulting in a non-synonymous amino acid substitution. As the reference Race1 (BBBD) is avirulent on 15 out of 16 of the resistance genes in the differential set, proteins that matched the reference exactly were assigned to a list matching races with a low infection type, incompatible interaction. Proteins with SNPs resulting in non-synonymous codon changes were associated with a high infection type compatible interaction. For

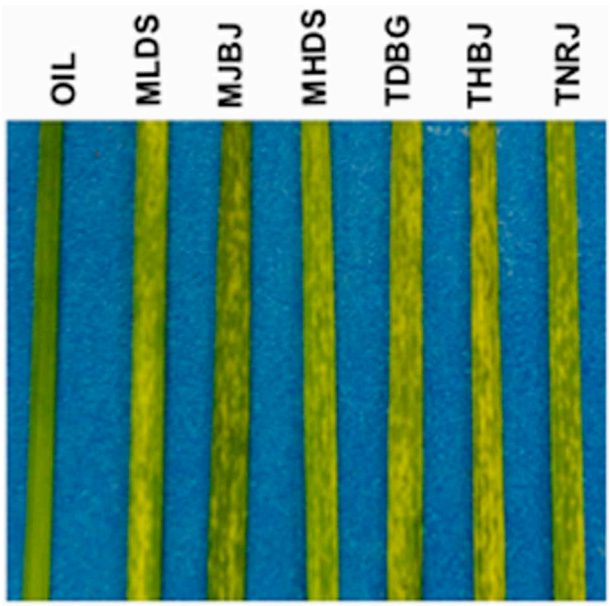


FIGURE 1 | Infection phenotypes of six *P. triticina* races on the susceptible spring wheat cultivar Thatcher at six days post inoculation, before sporulating pustules appear. Races are listed at the top. Oil represents the oil only control.

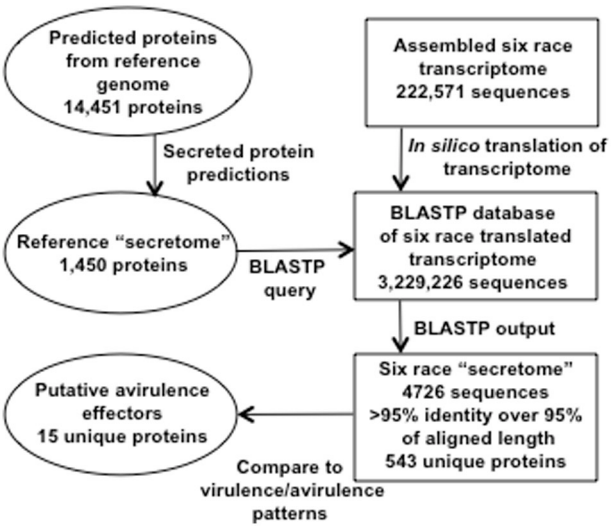


FIGURE 2 | Schematic of bioinformatic pipeline used to identify 15 secreted peptides that are candidates for avirulence effectors.

Table 2 | Illumina RNAseq data and assembly statistics.

Race	Raw reads	Mb reads	Contigs	N25 (bp)	N50 (bp)	N75 (bp)	Max (bp)
MHDS	26419162	3.170 Mb	38192	1492	817	471	12254
MLDS	25556420	3.066 Mb	35568	1719	927	512	10747
MJBJ	23415788	2.809 Mb	34528	1646	890	490	13792
TDBG	27731985	3.327 Mb	36281	1705	929	503	8860
THBJ	33225893	3.987 Mb	39509	1852	1034	555	10136
TNRJ	28404510	3.408 Mb	38673	1454	824	485	10696

example, PTTG_05760 had significant matches in all six races. The hits in MLDS, TDBG, and TNRJ matched the Race1 (BBBD) protein sequence exactly, while hits from MHDS, MJBJ, and THBJ had a non-synonymous amino acid substitution (serine to alanine). Race1 (BBBD), MLDS, TDBG, and TNRJ all have a low (avirulent) infection type on TcLr16. MHDS, MJBJ, and THBJ all have a high (virulent) infection type on TcLr16. Thus, the analysis considers this protein to be a candidate avirulence protein. This analysis identified 15 unique secreted peptides (**Table 3**). Functions for these proteins were predicted using PHYRE2 (Kelley and Sternberg, 2009). Twelve proteins had no significant similarities in the database. One showed significant similarity to cerato-platanin, a fungal elicitor in the Barwin-like endoglucanase superfamily (Pazzagli et al., 1999). A second protein contained a tyrosinase domain and the last showed significant similarity to a gibberellin receptor.

qRT-PCR primers were designed for the 15 genes identified as putative avirulence effectors. These were tested on cDNA generated from RNA from Thatcher wheat inoculated separately with the six leaf rust races and collected daily for 6 days. Following testing for specificity and appropriate efficiency, the primers were used to measure expression of seven genes. ΔCq was calculated against the rust reference gene histone H4 (PtHisH4). PtHisH4 was chosen for its stable expression among the races. Temporal expression profiles for these genes are presented in **Figures 3A–G**. There were various expression patterns with the seven genes, but were similar within all six races. PTTG_05706 was transcribed at a higher rate than PtHisH4 and expression was maintained throughout the 6 days (**Figure 3A**). PTTG_12153 was expressed at a low level for the first 3 days, but transcription spiked at day 4 and maintained this level through day 6. PTTG_11899 and PTTG_09426 both showed a general trend of reduced transcription from day 2 to day 6 (**Figures 3C,D**). The final three genes had differences in expression between the races. MLDS had a higher expression of PTTG_03539 at day 3, while TNRJ spiked at day 4 while the other five races had very low expression. Expression of PTTG_03539 varied between all races at day 5 and 6 (**Figure 3E**). TDBG and MHDS expressed PTTG_25269 at a high rate on day 2, while TNRJ was higher at day 3 (**Figure 3F**). The peak expression of PTTG_12522 was steady for TNRJ till day four (**Figure 3G**).

Table 3 | Putatively secreted peptides corresponding to virulence shifts.

Gene name[a]	Putative function[b]	Corresponding *Lr* genes[c]
PTTG_05870	Unknown	*Lr2a, Lr2c, Lr9, Lr3ka, Lr11, Lr30*
PTTG_00023	Unknown	*Lr9, Lr3ka, Lr11, Lr30*
PTTG_25426	Unknown	*Lr9, Lr3ka, Lr11, Lr30*
PTTG_28391	Unknown	*Lr9, Lr3ka, Lr11, Lr30*
PTTG_11899	Unknown	*Lr9, Lr3ka, Lr11, Lr30*
PTTG_03539	Tyrosinase	*Lr9, Lr3ka, Lr11, Lr30*
PTTG_05706	Unknown	*Lr16*
PTTG_12153	Unknown	*Lr16, Lr26*
PTTG_12522	Unknown	*Lr16, Lr26*
PTTG_25271	Barwin-like endoglucanase[d]	*Lr16*
PTTG_26277	Unknown	*Lr16, Lr26, Lr17, LrB*
PTTG_09426	Unknown	*Lr16, Lr26*
PTTG_25509	Gibberellin receptor[e]	*Lr24*
PTTG_25269	Unknown	*Lr17, LrB*
PTTG_02284	Unknown	*Lr16*

[a] *Gene names are from the version 2 annotation of the leaf rust genome (http://www.broadinstitute.org/scientific-community/data).*

[b] *Putative functions as predicted by PHYRE2 (Kelley and Sternberg, 2009). Functions with a partial coverage confidence of 100% are reported.*

[c] *Leaf rust resistance genes matching the virulence/avirulence pattern for the protein presented as determined by the presence of non-synonymous amino acid substitutions among the six races.*

[d] *Barwin-like endoglucanase protein superfamily includes the fungal elicitor cerato-platanin family (Pazzagli et al., 1999).*

[e] *Partial coverage of the amino acid sequence (22%) was predicted to have structure similarity to a probable gibberellin receptor. Other high confidence hits included hormone-sensitive esterases and lipases.*

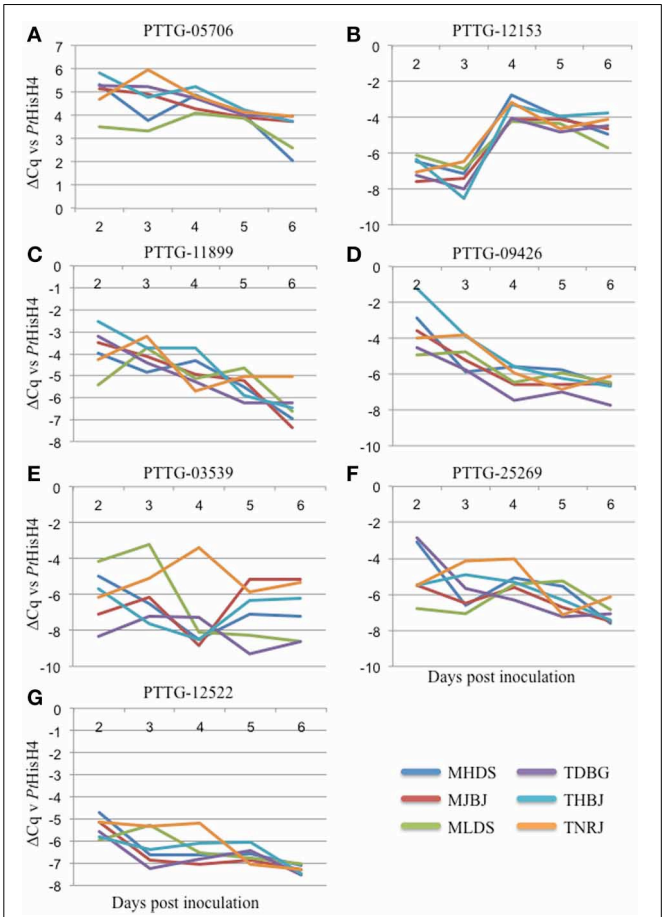


FIGURE 3 | Gene expression profiles for seven avirulence candidates from *P triticina*. Plant samples were taken at 24 h intervals post inoculation (X axis). Real-time PCR was used to quantify expression of each gene. *P. triticina* Histone H4 (PtHisH4) was used as the internal control for normalization. ΔCq vs. PtHisH4 was used to plot expression (Y axis).

DISCUSSION AND FUTURE DIRECTIONS

Secreted peptides of pathogenic fungi are of particular interest as some have been demonstrated to act as avirulence effectors in other studied systems (Dodds et al., 2004; De Wit et al., 2009; Nirmala et al., 2011). In order to advance knowledge of the mechanisms underlying virulence and avirulence in the wheat-*Pt* pathosystem, we have employed an RNA-Seq approach to identify genes and proteins that are expressed by *Pt* during compatible interactions and may in turn be perceived by the host during incompatible interactions. To date, there are 88 wheat leaf rust resistance genes described and used in various breeding programs (http://www.shigen.nig.ac.jp/wheat/komugi/, accessed 30 September 2013). Of these, at least 27 have been backcrossed into the Thatcher background, providing a uniform genetic background against which their contribution to disease resistance can be assessed (Kolmer, 1996). Because of the biotrophic nature of the fungus, these isolines are also the only means of studying the differing virulence reactions to resistance genes. Sixteen of these isolines are routinely used as differentials for race determination during annual surveys. The races' differential reaction to the presence of these resistance genes will allow for identification of potential avirulence effectors represented in the transcriptomes.

RNA sequencing has been employed to identify *Pt* proteins involved during infection. At 6 days post inoculation, many of the genes needed for infection are presumed to be active. Urediniospores are being formed, feeding structures are established, but hyphae are growing and haustoria are still being formed. Isolating RNA at this stage would presumably represent many of the genes of the asexual cycle. This study was focused on identifying the secretome of *Pt* and variants associated with virulence shifts. We identified 543 expressed genes that may code for secreted peptides. The six races in this study shared 445 of those genes. The remaining 98 genes, while not represented in the assembly among all six races, may still be expressed in individual races. For example, while contigs matching PTTG_25269 were only found in three of the six races, the qRT-PCR data indicated that the gene representing this contig was expressed in all of the races. This is an important point to consider when analyzing transcriptome assemblies; the representation of the transcriptome may not be complete, depending on the sequencing depth.

Among 543 expressed genes, SNPs resulting in non-synonymous amino acid changes relative to the reference race 1 predicted proteome were observed in the assembly data that correlated with virulence shifts for 11 different leaf rust resistance genes. Since the reference Race1 (BBBD) is avirulent on 15 of the 16 leaf rust genes tested, the virulence shift predictions were based on the avirulent races having the reference race-type allele, and the virulent races having an alternate allele that may evade recognition by the resistance gene, or have an unrecognized altered activity.

Of the 15 identified, three had high confidence functional predictions from PHYRE2 (Kelley and Sternberg, 2009), including a tyrosinase, a putative gibberellin receptor, and a Barwin-like endoglucanase. Appropriately performing qRT-PCR primers were developed for seven of the genes. The expression profiles for these genes showed differences that may correlate with their activity. For example, PTTG_12153 is upregulated between three and four DAI. This corresponds to a shift from mainly haustorial feeding to urediniospore production. While the function of this gene is currently unknown, future studies may indicate a role in this capacity. One of the proteins has structural similarity to a Barwin-like endoglucanase, a family containing cerato-platanin, a known fungal elicitor (Pazzagli et al., 1999). In *Arabidopsis*, ectopic expression of this elicitor has protective effects against other pathogens (Yang et al., 2009). However, we were unable to develop qRT-PCR primers to detect expression for this gene. In the future, it will be cloned and characterized from genomic DNA. PTTG_05706 showed the highest level of expression, was rapidly induced within the first 48 h, and maintained a high expression level throughout the course of infection. PTTG_12522 was induced at two DAI, and showed stable expression for the remainder of the measured period. Since wheat leaf rust completes its entire infection cycle in the 6 day timeframe (penetration, haustoria formation, uredinia production, pre-pustule eruption), the data presented here should be a good indication of when and how these genes are functioning during infection. PTTG_03539 shows a spike in expression for race TNRJ at day 4 relative to the other races tested. Additionally, PTTG_25269 shows a spike and sustained high expression in TNRJ in days 3 and 4. The altered expression of these genes in this race may be responsible for its observed aggressive nature in the field. Five random leaves were selected for RNA extraction at each time point. This sampling scheme normalizes plant-to-plant variation in the expression data. Genes showing interesting changes in expression patterns, such as PTTG_03539, will be examined more closely in the future. Temporally variable expression of secreted proteins that may be effectors or avirulence genes of rust pathogens in different plant species has been observed in other studies, suggesting that these genes may affect host colonization or recognition of the pathogen by the host in a time-dependent manner (Duplessis et al., 2011b; Cantu et al., 2013; Tremblay et al., 2013).

The work reported here generated six secretomes of *Pt* including 543 predicted secreted proteins. RNAseq allows a much deeper sampling of the RNA species and resulting in a higher probability of an RNA to be identified. However, our work shows that assemblies may not reliably detect low level transcripts. Association of SNPs with virulence shifts have shown that 11 of the 15 avirulence candidates could correlate to reactions by multiple resistance genes. More races from known genetic lineages need to be sequenced so that stronger associations can be made. The putative effector genes in this study will be cloned from cDNA and used in experiments to determine if they have a role in the resistance response in wheat. A better understanding of the molecular determinants in disease resistance and susceptibility can generate additional tools for practical application in breeding programs and race identification.

MATERIALS AND METHODS

PLANT GROWTH CONDITIONS, INOCULATION, AND TISSUE HARVEST

The hard red spring wheat variety, Thatcher (*Triticum aestivum* L.) was used in all of the experiments. Seeds were planted in Metro Mix 360 (SunGro, Vancouver, Canada) and grown in a growth chamber at 18°C with 16 h day/8 h night cycles. At the 2–3 leaf stage, plants were inoculated by suspending 5 mg urediniospores per mL Soltrol 170 (Philips 66, Bartlesville, OK) and spraying onto the plants using an atomizer at 40 psi. Following inoculation,

plants were incubated in a dew chamber at 100% humidity for 24 h at 18°C. Plants were then moved back into growth chambers. Leaves from 30 inoculated plants per race were collected and pooled 6 DAI and immediately frozen in liquid nitrogen. For time course expression studies, 30 plants were inoculated per race as described. Five random leaves per race were collected and stored as described at 2, 3, 4, 5, and 6 DAI. Total RNA was extracted from tissue with the *mir*Vana miRNA isolation kit (AM1560, Life Technologies, Carlsbad, CA) according to the manufacturer's recommendations with the omission of the miRNA enrichment step. RNA was quantified with a Nanodrop ND1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA).

RNA SEQUENCING

Total RNA was submitted to Cofactor Genomics (St. Louis, MO). The RNA was sequenced according to in house protocols steps summarized as follows. Whole transcriptome RNA was extracted from total RNA by removing large and small ribosomal subunit RNA (rRNA) using the RiboMinus Eukaryote Kit (Invitrogen, Carlsbad, CA). 5 μg of total RNA was hybridized to rRNA-specific biotin labeled probes at 70°C for 5 min. The rRNA-probe complexes were removed by strepavidin-coated magnetic beads. The rRNA-free transcriptome RNA was concentrated by ethanol precipitation. Double-stranded cDNA was treated with a mix of T4 DNA polymerase, Klenow large fragment and T4 polynucleotide kinase to create blunt-ended DNA, to which was subsequently added a single A base at the 3′ end using Klenow fragment (3′ to 5′ exo-) and dATP. A-tailed DNA was ligated with paired end adaptors using T4-DNA ligase provided with the Illumina RNA-seq kit (Illumina, San Diego, CA). Size selection of adaptor-ligated cDNA was performed by cutting the target fragment out of a 4–12% acrylamide gel. The amplified cDNA library with ideal fragment size was obtained by in-gel PCR using the Phusion High-Fidelity system (New England Biolabs, Ipswitch, MA). Size-fractionated, amplified DNA was sequenced according to the Illumina RNAseq protocol (Illumina).

cDNA SYNTHESIS—ILLUMINA SEQUENCING

1 μg of rRNA depleted RNA was fragmented by incubation in fragmentation buffer included in the Illumina RNA-seq kit (Illumina) for 5 min at 94°C. Fragmented RNA was purified by ethanol precipitation. First strand cDNA was prepared by priming the fragmented RNA with random hexamers, followed by reverse transcription with Superscript II (Invitrogen, Carlsbad, CA). Second strand cDNA was synthesized by incubating first strand cDNA with second strand buffer, RNase Out and dNTP from the Illumina RNA-seq kit on ice for 5 min. The reaction mix was then treated with DNA Pol I and RNaseH at 16°C for 2.5 h (Invitrogen).

DATA PROCESSING AND ASSEMBLY

Bases with quality scores less than 20 were trimmed from both ends of raw sequencing reads (fastq_quality_trim -q 20 -t 30). Trimmed reads with a length greater than 30 and 80% of bases with quality scores greater than 20 were retained for assembly (fastq_quality_filter -p 20 -q 80). These tools are part of the FASTX-Toolkit (http://hannonlab.cshl.edu/fastxtoolkit/). Quality-filtered reads were assembled into transcripts using Trinity v2011059 (Grabherr et al., 2011), a de Brujin graph-based assembler. The -jaccard_clip option was used to minimize fusion transcripts resulting from overlapping UTR regions from the fungal transcriptome. Assembled transcripts were aligned against the *Pt* Race1 (BBBD) reference genome and retained in a separate FASTA file.

Six frame translations of the six race transcriptome was conducted with a custom Python script [http://www.python.org, (Cock et al., 2009)]. Peptides greater than 20 amino acids in any frame were written to a text file and converted to a BLASTP (Altschul et al., 1990) database. Predicted proteins from the race 1 reference genome were used to generate a list of putatively secreted proteins using SignalP (Petersen et al., 2011). Proteins meeting criteria defined by Joly et al. (2010) as putatively secreted were used in a BLASTP query (Altschul et al., 1990) against the six race translated transcriptome using an expect value of 1e-30. A Python script (http://www.python.org) was used to parse the resulting BLASTP XML output. To identify proteins with SNPs relative to the race 1 reference, a ratio of identical matches over the aligned length was taken from the parsed BLASTP output. Proteins in any race with a ratio less than one were chosen for further analysis.

EXPRESSION PROFILING, cDNA SYNTHESIS—qRT-PCR

First strand cDNA was prepared by priming 1 μg total RNA with random hexamers, followed by reverse transcription with Superscript II (Invitrogen) according to the manufacturer's recommendations. Primers for qRT-PCR were designed from the assembled contigs and used to assess differences in gene expression between the races. Bullseye EvaGreen qPCR mastermix for iCycler (BioRad, La Jolla, CA) was used in all reactions. Three technical replicates were performed for each reaction. All primers were assayed for efficiency prior to use in experiments and had efficiencies within the range 90–110%. The resulting Cq value for the target gene was subtracted from the Cq value of the internal rust reference gene histone H4. Primer sequences used in this study are provided in **Supplementary Table 2**.

Mention of a trademark of a proprietary product does not constitute a guarantee of warranty of the product by the United States Department of Agriculture, and does not imply its approval to the exclusion of other products that may also be suitable. USDA is an equal opportunity provider and employer.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Robert Bowden for his comments during the preparation of this manuscript. This work was funded through USDA-ARS CRIS 5430-21000-010-00D. This is a joint contribution of the United States Department of Agriculture, Agriculture Research Service and the Kansas Agricultural Experiment Station. Contribution no. 14-114J.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2013.00520/abstract

Supplementary Table 1 | Putatively secreted proteins shared by six races.

Supplementary Table 2 | Primers used in this study.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Bolton, M. D., Kolmer, J. A., and Garvin, D. F. (2008). Wheat leaf rust caused by *Puccinia triticina*. *Mol. Plant Pathol.* 9, 563–575. doi: 10.1111/j.1364-3703.2008.00487.x

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis f. sp. tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis f. sp. tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., et al. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* 434, 980–986. doi: 10.1038/nature03449

De Wit, P. J., Mehrabi, R., Van den Burg, H. A., and Stergiopoulos, I. (2009). Fungal effector proteins: past, present and future. *Mol. Plant Pathol.* 10, 735–747. doi: 10.1111/j.1364-3703.2009.00591.x

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16, 755–768. doi: 10.1105/tpc.020040

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011a). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011b). *Melampsora larici-populina* transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A. S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant-rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Flor, H. H. (1942). Inheritance of pathogenicity in a cross between physiologic races 22 and 24 of *Melampsora lini*. *Phytopathology* 35, 5.

Flor, H. H. (1971). Current status of gene-for-gene concept. *Annu. Rev. Phytopathol.* 9, 275. doi: 10.1146/annurev.py.09.090171.001423

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Hacquard, S., Joly, D. L., Lin, Y. C., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Joly, D. L., Feau, N., Tanguay, P., and Hamelin, R. C. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11:422. doi: 10.1186/1471-2164-11-422

Kamper, J., Kahmann, R., Bolker, M., Ma, L. J., Brefort, T., Saville, B. J., et al. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* 444, 97–101. doi: 10.1038/nature05248

Kelley, L. A., and Sternberg, M. J. (2009). Protein structure prediction on the Web: a case study using the PHYRE server. *Nat. Protoc.* 4, 363–371. doi: 10.1038/nprot.2009.2

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Mol. Plant Microbe Interact.* 18, 1130–1139. doi: 10.1094/MPMI-18-1130

Kemen, E., Kemen, A., Ehlers, A., Voegele, R., and Mendgen, K. (2013). A novel structural effector from rust fungi is capable of fibril formation. *Plant J.* 75, 767–780. doi: 10.1111/tpj.12237

Kolmer, J. A. (1996). Genetics of resistance to wheat leaf rust. *Annu. Rev. Phytopathol.* 34, 435–455. doi: 10.1146/annurev.phyto.34.1.435

McIntosh, R. A., Wellings, C. R., and Park, R. F. (1995). *Wheat Rusts: An Atlas of Resistance Genes*. Melbourne: CSIRO; Dordrecht: Kluwer Academic Publ. doi: 10.1007/978-94-011-0083-0

Mur, L. A. J., Kenton, P., Lloyd, A. J., Ougham, H., and Prats, E. (2008). The hypersensitive response; the centenary is upon us but how much do we know? *J. Exp. Bot.* 59, 501–520. doi: 10.1093/jxb/erm239

Nirmala, J., Drader, T., Lawrence, P. K., Yin, C., Hulbert, S., Steber, C. M., et al. (2011). Concerted action of two avirulent spore effectors activates Reaction to *Puccinia graminis 1* (Rpg1)-mediated cereal stem rust resistance. *Proc. Natl. Acad. Sci. U.S.A.* 108, 14676–14681. doi: 10.1073/pnas.1111771108

Pazzagli, L., Cappugi, G., Manao, G., Camici, G., Santini, A., and Scala, A. (1999). Purification, characterization, and amino acid sequence of cerato-platanin, a new phytotoxic protein from Ceratocystis fimbriata f. sp. platani. *J. Biol. Chem.* 274, 24959–24964. doi: 10.1074/jbc.274.35.24959

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins-a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Samborski, D. J., and Dyck, P. L. (1968). Inheritance of virulence in wheat leaf rust on standard differential wheat varieties. *Can. J. Genet. Cytol.* 10, 24–32.

Statler, G. D. (2000). Inheritance of virulence of *Puccinia triticina* culture X47, the F1 of the cross 71-112 x 70-1. *Can. J. Plant Pathol.* 22, 276–279. doi: 10.1080/07060660009500475

Stergiopoulos, I., and de Wit, P. J. (2009). Fungal effector proteins. *Annu. Rev. Phytopathol.* 47, 233–263. doi: 10.1146/annurev.phyto.112408.132637

Tremblay, A., Hosseini, P., Li, S., Alkharouf, N. W., and Matthews, B. F. (2013). Analysis of *Phakopsora pachyrhizi* transcript abundance in critical pathways at four time-points during infection of a susceptible soybean cultivar using deep sequencing. *BMC Genomics* 14:614. doi: 10.1186/1471-2164-14-614

Yang, Y., Zhang, H., Li, G., Li, W., Wang, X., and Song, F. (2009). Ectopic expression of MgSM1, a Cerato-platanin family protein from *Magnaporthe grisea*, confers broad-spectrum disease resistance in Arabidopsis. *Plant Biotechnol. J.* 7, 763–777. doi: 10.1111/j.1467-7652.2009.00442.x

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 October 2013; accepted: 02 December 2013; published online: 13 January 2014.

*Citation: Bruce M, Neugebauer KA, Joly DL, Migeon P, Cuomo CA, Wang S, Akhunov E, Bakkeren G, Kolmer JA and Fellers JP (2014) Using transcription of six Puccinia triticina races to identify the effective secretome during infection of wheat. Front. Plant Sci. **4**:520. doi: 10.3389/fpls.2013.00520*

This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.

Copyright © 2014 Bruce, Neugebauer, Joly, Migeon, Cuomo, Wang, Akhunov, Bakkeren, Kolmer and Fellers. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 26 June 2014
doi: 10.3389/fpls.2014.00299

Duplications and losses in gene families of rust pathogens highlight putative effectors

Amanda L. Pendleton[1], Katherine E. Smith[2], Nicolas Feau[3], Francis M. Martin[4], Igor V. Grigoriev[5], Richard Hamelin[3], C. Dana Nelson[2], J. Gordon Burleigh[1,6,7] and John M. Davis[1,7,8]*

[1] *Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL, USA*
[2] *Southern Research Station, Southern Institute of Forest Genetics, USDA Forest Service, Saucier, MS, USA*
[3] *Department of Forest Sciences, University of British Columbia, Vancouver, BC, Canada*
[4] *Interactions Arbres/Microorganismes, Laboratoire d'Excellence ARBRE, INRA-Nancy, UMR Institut National de la Recherche Agronomique - Université de Lorraine, Champenoux, France*
[5] *US Department of Energy, Joint Genome Institute, Walnut Creek, CA, USA*
[6] *Biology Department, University of Florida, Gainesville, FL, USA*
[7] *Genetics Institute, University of Florida, Gainesville, FL, USA*
[8] *School of Forest Resources and Conservation, University of Florida, Gainesville, FL, USA*

Edited by:
Sébastien Duplessis, INRA, France

Reviewed by:
Dag Ahren, Lund University, Sweden
Yao-Cheng Lin, VIB/Ghent University, Belgium

****Correspondence:***
John M. Davis, School of Forest Resources and Conservation, University of Florida, 365 Newins-Ziegler Hall, Gainesville, FL 32611, USA
e-mail: jmdavis@ufl.edu

Rust fungi are a group of fungal pathogens that cause some of the world's most destructive diseases of trees and crops. A shared characteristic among rust fungi is obligate biotrophy, the inability to complete a lifecycle without a host. This dependence on a host species likely affects patterns of gene expansion, contraction, and innovation within rust pathogen genomes. The establishment of disease by biotrophic pathogens is reliant upon effector proteins that are encoded in the fungal genome and secreted from the pathogen into the host's cell apoplast or within the cells. This study uses a comparative genomic approach to elucidate putative effectors and determine their evolutionary histories. We used OrthoMCL to identify nearly 20,000 gene families in proteomes of 16 diverse fungal species, which include 15 basidiomycetes and one ascomycete. We inferred patterns of duplication and loss for each gene family and identified families with distinctive patterns of expansion/contraction associated with the evolution of rust fungal genomes. To recognize potential contributors for the unique features of rust pathogens, we identified families harboring secreted proteins that: (i) arose or expanded in rust pathogens relative to other fungi, or (ii) contracted or were lost in rust fungal genomes. While the origin of rust fungi appears to be associated with considerable gene loss, there are many gene duplications associated with each sampled rust fungal genome. We also highlight two putative effector gene families that have expanded in *Cqf* that we hypothesize have roles in pathogenicity.

Keywords: effectors, rust pathogens, secretome, genome evolution, comparative genomics

INTRODUCTION

Rust fungi are plant infecting filamentous fungi in the order Pucciniales (Basidiomycota) that are unified by obligate biotrophy (Voegele and Mendgen, 2011). This form of pathogenicity requires a live host to establish a parasitic relationship. This is accomplished through the establishment of a molecularly intimate interaction at the host-pathogen interface characterized by the secretion of an arsenal of proteins from the pathogen that suppress host defense mechanisms and promote the acquisition of essential nutrients by the pathogen (Dodds et al., 2009; Stergiopoulos and de Wit, 2009). Such proteins, termed effectors, are thought to establish and maintain a compatible interaction between the pathogen and host. The processes that drive evolution of effector diversity are of great interest because pathogen's effector genes and host resistance genes are the interacting "gene-for-gene" pairs that drive coevolution in these pathosystems (Jones and Dangl, 2006; Stergiopoulos and de Wit, 2009).

Secreted proteins can be identified from whole genome sequences through the utilization of bioinformatic tools to isolate proteins with N-terminal secretion signals. Bioinformatic pipelines can then be used to narrow predicted secreted protein sets to putative effectors. These proteins contain features of known effectors such as elevated cysteine content (greater than 2%), that would enable the formation of stabilizing disulfide bridges (Stergiopoulos and de Wit, 2009), and protein domains associated with pathogenicity. Length is a criteria used to identify small secreted proteins (SSPs) from within putative effector protein sets, as SSPs are effector-like proteins with lengths less than 300 amino acids. Sequence comparisons alone do not provide a reliable means to identify putative effectors since some known effectors are lineage-specific while others are conserved across taxa (Rep, 2005; Saunders et al., 2012; Giraldo and Valent, 2013). Candidate effectors, a further distinction, are putative effectors that have additional support for roles in pathogenicity

(i.e., induced transcription or elevated expression *in planta*). Genetic evidence for functional redundancy of effectors, presumably due to multigene families of effector proteins, whose members share similar functions, has been reported in several pathogens (Kamper et al., 2006; Rafiqi et al., 2012; Saitoh et al., 2012; Giraldo and Valent, 2013). This suggests it would be useful to characterize families of proteins with effector-like characteristics so as to identify families that have expanded during evolution in association with the acquisition of pathogenic life history characteristics. Examining the evolutionary history of protein families across a set of diverse fungal taxa should help identify lineage-specific, putative effector protein families, families that may have evolved similar functions in more distantly related taxa, and families that may exhibit functional redundancy.

Cronartium quercuum f. sp. *fusiforme* (*Cqf*) is a rust pathogen that has a complex life cycle with five spore types and exhibits alternation between two hosts, oak (*Quercus* spp.) and southern pines (*Pinus* spp.). The fungus incites fusiform rust disease on southern pines, leading to significant economic losses to the forest products industry. The impact of the disease on pine production has motivated extensive research on the genetic interaction between *Cqf* and pine. The objective of this study is to identify putative effector gene families in the *Cqf* genome through comparative genomic analyses between *Cqf* and 15 other fungal taxa, including two other rust pathogens. We have identified families that have expanded in *Cqf* that we hypothesize are involved in conditioning stem gall phenotypes observed on the pine host. Our analyses provide a more thorough perspective on *Cqf* and rust pathogen evolution and also highlight the evolutionary patterns of putative effector families that *Cqf* employs to establish disease on two taxonomically diverse host species.

MATERIALS AND METHODS

GENE FAMILY CONSTRUCTION

Complete proteomes were downloaded from the public databases of the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/genome), U.S. Department of Energy's Joint Genome Institute (jgi.doe.gov/fungi), and the Broad Institute (www.broadinstitute.org). Sixteen proteomes were obtained: (Basidiomycota) *Cronartium quercuum* f.sp. *fusiforme* G11 version 1.0 (*Cqf*; unpublished, jgi.doe.gov/Cronartium), *Melamspora larici-populina* version 1.0 (*Mlp*; Duplessis et al., 2011a,b), *Puccinia graminis* f.sp. *tritici* CRL 75-36-700-3 race SCCL (*Pgt*; Duplessis et al., 2011a,b), *Mixia osmundae* IAM 14324 version 1.0 (*Mos*; Toome et al., 2014), *Sporobolomyces roseus* version 1.0 (*Sro*; with permission; jgi.doe.gov/fungi), *Rhodotorula graminis* strain WP1 version 1.1 (*Rgr*; with permission; jgi.doe.gov/fungi), *Ustilago maydis* strain 521 (*Uma*; Kamper et al., 2006), *Malasezzia globosa* CBS 7966 (*Mgl*; Xu et al., 2007), *Pisolithus tinctorius* Marx 270 version 1.0 (Pti; with permission; jgi.doe.gov/fungi), *Phanerochaete chrysosporium* version 2.0 (*Pch*; Martinez et al., 2004), *Heterobasidion irregulare* version 2.0 (*Hir*; Olson et al., 2012), *Serpula lacrymans* S7.3 version 2.0 (*Sla*; Eastwood et al., 2011), *Agaricus bisporus* var. *bisporus* H97 version 2.0 (*Abi*; Morin et al., 2012), *Laccaria bicolor* version 2.0 (*Lbi*; Martin et al., 2008), *Amanita muscaria* Koide version 1.0 (*Amu*; with permission; jgi.doe.gov/fungi), and (Ascomycota) *Saccharomyces cerevisiae* S288C (*Sce*; Goffeau et al., 1996), for a total of 200,313 proteins. Gene families were delineated by OrthoMCL v.5.0 software (Li et al., 2003) using default parameters (minimum *e*-value of 1e-05, minimum similarity of 50%).

SECRETOME PREDICTION

The collective set of secreted proteins, or the secretome, of *Cqf* was identified bioinformatically. Annotation of a secreted protein is determined by signal peptide (SignalP 3.0 and 4.0; Bendtsen et al., 2004; Petersen et al., 2011), protein localization (TargetP 1.1; Emanuelsson et al., 2000), and transmembrane domain (TMHMM 2.0; Krogh et al., 2001) bioinformatics prediction software (Feau et al. *in prep.*). Proteins predicted by TargetP 1.1 to be targeted for the mitochondrion (with RC values between 1 and 3) were discarded and residual proteins are submitted to TMHMM 2.0. If no TM-domain is identified in the protein, or a TM-domain is predicted in the N-terminal region of the protein (i.e., in the first 70 amino acids), the protein is re-oriented toward SignalP 4.0; in any other case, the protein is discarded. SignalP 4.0 either implements the SignalP-TM network to discriminate between a true signal peptide and an N-terminal trans-membrane region or the SignalP-noTM network if the program does not identify a TM-like domain in the N-terminal region of the protein. In this last case (i.e., if the the SignalP-noTM network is implemented by SignalP 4.0), the protein is re-oriented toward SignalP 3.0 and a signal peptide prediction is positive if either both NN and HMM converged in a positive result or if NN D-score returns a positive result with a D-score ≥ 0.5.

ESTIMATION OF GENE TREES

The protein sequences from each gene family were aligned using MUSCLE (Edgar, 2004). We assembled a collection of amino acid alignments from gene families with at least four sequences. For each of the gene family alignments, we performed a maximum likelihood (ML) search to find the optimal topology using RAxML v.7.2.8 with the PROTCATJTT model (Stamatakis et al., 2005). Gene tree estimates often contain much error and can be improved with knowledge of the underlying species tree (e.g., Rasmussen and Kellis, 2011). We constructed a species tree from a phylogenetic matrix of 2404 single copy genes with sequences from at least eight fungal taxa. We performed a ML search using RAxML v.7.2.8 with the PROTCATJTT model on the concatenated single gene matrix to estimate the species tree. For each of the gene trees, we used TreeFix version 1.1.8 (Wu et al., 2013) to improve on the ML topology given the species tree. TreeFix searches for a statistically equivalent rooted gene tree topology that minimizes the number of duplications and losses implied by the species tree. For 10 of the gene families, the TreeFix runs did not complete in 1 week. For these gene trees, we rooted the ML tree with a root that minimizes the number of implied duplications and losses using the program OptRoot (www.wehe.us). For all of the gene trees output from TreeFix or OptRoot, the locations of the implied duplications and losses were mapped on the species tree using URec version 1.02 (Gorecki and Tiuryn, 2007).

FUNCTIONAL ANNOTATION OF PROTEINS

Functional annotations were obtained from the Joint Genome Institute's (JGI) Mycocosm (jgi.doe.gov/fungi; Grigoriev et al.,

2014) for the 16 organisms included in the phylogenetic and gene family analyses. Protein domains were identified using the online InterPro interface (http://www.ebi.ac.uk/interpro/; Hunter et al., 2012). Transmembrane domain regions were identified in amino acid sequences of proteins using TMpred (www.ch.embnet.org/software/TMPRED_form.html; Hofmann, 1993). Glycosylphosphatidylinositol (GPI) anchor sites were predicted using big-PI Predictor (http://mendel.imp.ac.at/gpi/gpi_server.html; Eisenhaber et al., 1999).

RESULTS

GENE FAMILY ANALYSIS

The OrthoMCL analysis of the proteomes identified 19,489 gene families that contained 152,964 proteins. This protein count was ∼76% (152,964/200,313) of the total proteins input into OrthoMCL analysis. Protein counts per gene family ranged from 2 (minimum size for a gene family) to 343 proteins, and the average family size was 7.8 proteins. Approximately 42% of the gene families had proteins encoded from only a single taxon, and families with proteins encoded in two or three taxa were the next most abundant families (**Figure 1**). Relatively few families contained proteins detected in 4–14 taxa, but more families contained proteins detected in 15 or 16 taxa (∼12% of all families; 2,277/19,489). The families broadly conserved across all 16 sampled taxa are likely to contain core essential fungal proteins. The remaining ∼24% of input proteins that did not group into families are considered true singletons, as they lack homologs within their own proteome or in the other taxa.

To highlight gene families specific to the rust pathogen lineage, we compared gene family conservation between four pathogen genomes belonging to the subphylum Pucciniomycotina, which include three rust pathogens (*Cqf*, *Mlp*, and *Pgt*; Pucciniales) and a non-rust fern pathogen, *Mixia osmundae* (*Mos*; Mixiales). We selected the 4673 gene families containing proteins from at least one of these four pathogens (and no proteins from other sampled taxa) from the complete OrthoMCL family dataset. These families contained 22,784 proteins and exhibited varying patterns of conservation across the four taxa (**Figure 2A**). Most prominently, 14,978 of the 22,784 proteins (65.7%) were encoded in only one of the four pathogen genomes, illustrating high levels of species specificity (**Figure 2A**). Fewer proteins were shared between two or more rust fungi in this subset of families (7512/22,784 or 33.0%) (**Figure 2A**). Of the 19,485 families determined by OrthoMCL, 656 families (or 3.4%; **Figure 2B**) consisted of gene models found only in the three rust pathogen genomes, where each of the three rust pathogens had a representative gene model in the family. A total of 3466 proteins (**Figure 2A**) were ascribed to these rust pathogen-specific families. The largest family contained 249 proteins, and the smallest had 3 proteins. These 656 families represent the "core" rust pathogen protein set. Of the sampled genomes, the two pathogens with the most uniquely shared families are *Cqf* and *Mlp*, which have 878 conserved families.

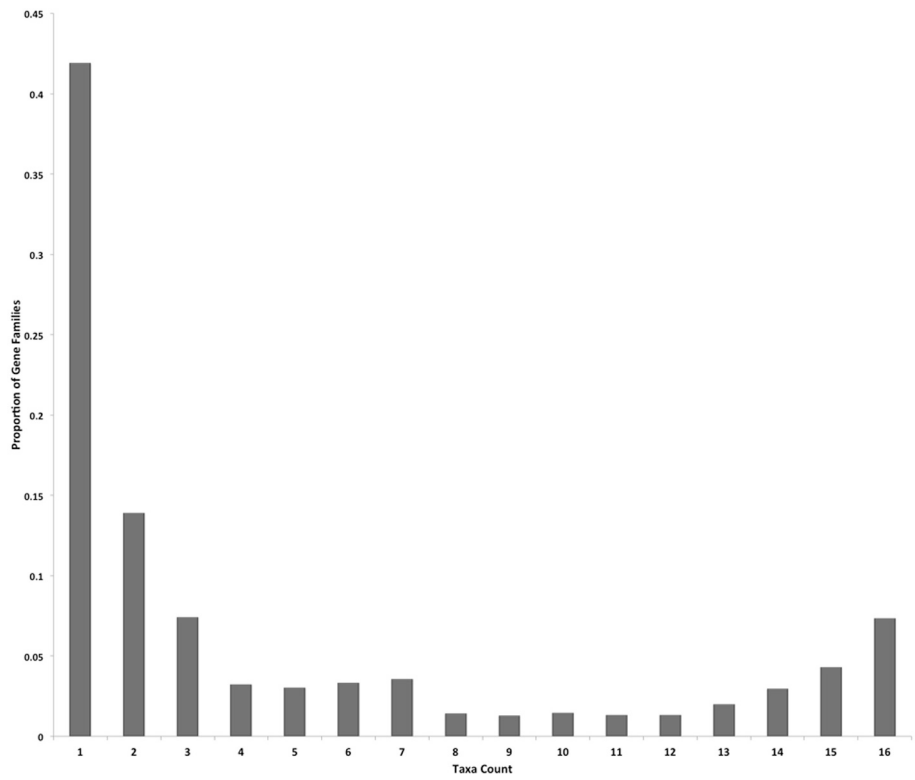


FIGURE 1 | Gene families are predominantly species-specific in the sampled taxa. The proportions of gene families with proteins encoded in one through 16 fungal taxa genomes (taxa count) are displayed for the 19,489 OrthoMCL gene families.

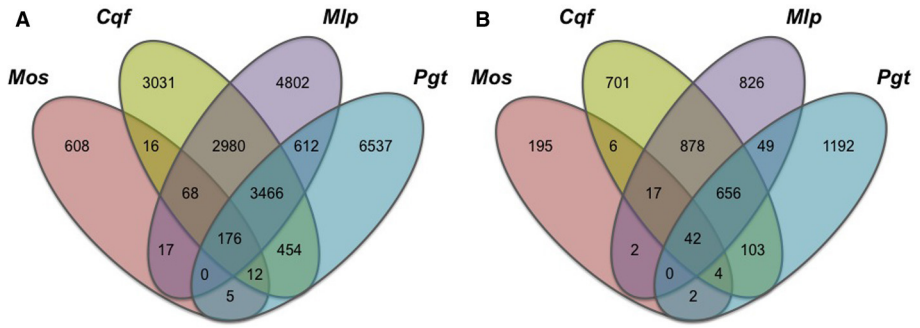


FIGURE 2 | Conserved proteins and families within only four Pucciniomycete pathogen genomes are mostly species-specific. Gene family (OrthoMCL) conservation within Pucciniomycete pathogens; *Mixia osmundae* (Mos), *Cronartium quercuum* f.sp. *fusiforme* (Cqf), *Melampsora larici-populina* (Mlp), *and Puccinia graminis* f.sp. *tritici* (Pgt). The values indicate the total number of **(A)** gene models or **(B)** gene families conserved in only these four species and absent in the remaining 12 fungal taxa included in the OrthoMCL analysis.

IDENTIFYING PUTATIVE EFFECTORS

We identified gene families encoding putative effectors in the *Cqf* genome. To highlight putative effectors, the predicted secretome (predicted secreted proteins; see Methods) was analyzed for cysteine content and family-level conservation. The *Cqf* secretome harbors 666 SSPs, which are secreted proteins with fewer than 300 amino acids (aa). The range in protein lengths within the secretome was 51–1716 aa with a median length of 249 aa.

Analysis of *Cqf* gene families elucidated the evolutionary histories of secreted putative effectors. To identify putative effector families within the *Cqf* genome, we selected gene families with at least two secreted proteins, as these families would then contain at least two paralogous putative effectors and the family would have therefore expanded in the *Cqf* genome. In total, 132 putative effector families were identified. Sixty-five of these families were conserved effector families, with proteins from two or more fungal taxa. These families had sequences from 6.94 taxa on average (**Table 1**) and represent potential effectors with functions that can occur in a wide range of hosts. Alternatively, 67 novel effector families were considered to be evolutionary innovations since the family members consisted of only *Cqf* proteins (**Table 2**). The average family size for conserved effector families (18.23 proteins) was significantly larger than *Cqf*-specific families (3.54 proteins; *t*-test, p-value < 0.001). However, there was no difference in the number of *Cqf* proteins per family in conserved (mean = 5.02 proteins) and *Cqf*-specific families (mean = 2.4 proteins). Families where all *Cqf* protein members are predicted to be secreted were found in both candidate effector family types and at proportions that were not significantly different from one another (conserved families = 40/65, *Cqf*-specific = 44/67; **Tables 1, 2**). Evidence for potential sub- and/or neofunctionalization was observed in 23 of the 67 (34.3%) *Cqf*-specific putative effector families, as only a subset of proteins within these families received secretion predictions, suggesting distinct biological roles among family members.

GENE GAINS AND LOSSES

Gene gain and loss was quantified across all 16 sampled fungal taxa. We mapped the gene trees from gene families with at least four proteins onto a species tree to determine the patterns of duplication and loss across the 16 fungal taxa. In total, we examined 10,371 gene trees containing 131,863 protein sequences. These gene trees implied a minimum of 49,539 duplications (i.e., gene family gains) and 21,789 losses (i.e., gene family contractions and/or entire family loss). Over 93.9% of the duplications and 67.9% of the losses are species-specific, occurring in a single lineage at the tips of the species tree (**Figure 3**). The number of species-specific duplications was positively correlated with the size of a taxon's proteome ($R^2 = 0.93$), suggesting that gene duplication is a mechanism for proteomic expansion and diversification for the selected fungal taxa (**Figure 4**). There was no obvious relationship between proteome size and species-specific duplication with life history forms (i.e., symbiotic, pathogenic, or free-living) (**Figure 4**). Species-specific losses were not correlated with the proteome size, but the rust pathogen lineage exhibited fewer losses than other sampled taxa (**Figure 5**).

We identified genes that were gained and lost specifically in the rust pathogen clade. There were many gene losses (1217 events within 1148 families) associated origin of the rust pathogen clade within Pucciniomycotina (*Cqf*, *Mlp*, and *Pgt*) compared to the number of gains (248 events, 142 families) (site R in **Figure 3**). The number of taxa represented in these 1148 families range from 2 to 16 species, with the largest proportion of families (10.6%) having representatives from all 16 taxa in the analysis (**Figure 6**). Families lost genes at the origin of the rust fungi appear to occur in few of the sampled fungal lineages than those that had duplications in the rust fungi. Fifty percent of duplicated families contain proteins from 14 or more sampled taxa (**Figure 6**). Though the disproportionate level of gene losses prior to the common ancestor of rust pathogens is striking, each of the three rust fungal species shows evidence of high species-specific rates of duplication (**Figure 3**). In fact, 32.1% of all the duplications across the tree are specific to only one of the rust species (**Figure 3**).

The proteome of the *Cqf* rust pathogen is enriched for novel proteins whose expansion has presumably contributed to specialization in its pathosystem. Numerous species-specific duplications have occurred following within the *Cqf* lineage (2730 duplication events in 549 families; **Figure 3**). Of the 549 families

Table 1 | Conserved putative effector families have broad and narrow taxonomic distributions.

OrthoMCL group ID	Total proteins in family	*Cqf* proteins in family	Proteins in *Cqf* secretome	Number of taxa represented in family
5485	12	7*	7	3
2168	17	6	5	3
2725	16	13	5	3
1053	94	6	4	13
2731	16	5	4	5
5853	11	5	4	5
6604	9	5	4	2
1101	65	62	3	2
1281	33	3*	3	3
1293	32	5	3	10
1397	27	3*	3	12
1831	19	3*	3	6
2730	16	3*	3	10
5067	13	4	3	7
6199	10	3*	3	3
6608	9	3*	3	3
6599	9	3*	3	5
7036	8	4	3	2
9412	5	3*	3	2
10,437	4	3*	3	2
1014	170	69	2	10
1219	39	3	2	14
1382	28	2*	2	13
1410	27	3	2	15
1444	26	2*	2	14
1428	26	2*	2	10
1507	24	3	2	15
1546	23	2*	2	13
1541	23	2*	2	16
1593	22	2*	2	13
1703	21	2*	2	8
1898	19	2*	2	14
1829	19	2*	2	15
1957	18	2*	2	14
2180	17	3	2	9
2191	17	2*	2	10
2172	17	2*	2	16
4359	15	2*	2	13
3833	15	2*	2	12
3825	15	2*	2	11
4560	14	2*	2	10
5075	13	2*	2	7
5492	12	3	2	6
5799	12	2*	2	10
6367	10	3	2	6
6213	10	3	2	3
6206	10	2*	2	3
6601	9	7	2	2
6606	9	3	2	7
6629	9	2*	2	6
7040	8	2*	2	2
7590	7	2*	2	3
9416	5	4	2	2
9414	5	4	2	2
9843	5	2*	2	3
9462	5	2*	2	2
9446	5	2*	2	2
9431	5	2*	2	4
9424	5	2*	2	3
10,479	4	3	2	2
10,474	4	3	2	2
10,432	4	3	2	2
11,968	3	2*	2	2
11,958	3	2*	2	2
11,908	3	2*	2	2
Total	1185 proteins	326 proteins	162 proteins	-
Average	18.23 proteins per family	5.02 *Cqf* proteins per family	2.5 secreted proteins per family	6.94 taxa per family

Gene families with greater than two Cqf predicted secreted proteins are listed. Data is ranked by the number of Cqf secreted proteins. The total number of proteins in each family is provided as well as the number of proteins belonging to the Cqf secretome (i.e., predicted secreted proteins). Asterisks adjacent to total protein counts indicate families where all members are Cqf secretome members. If no asterisk is present, only a portion of the family received secretion predictions. Family 5485 (bold) will be detailed later in article.

that have undergone *Cqf*-specific duplications, 248 (or 45.17%) contain proteins not observed in any other analyzed fungal taxa. These 248 novel families comprise 14.5% of the annotated *Cqf* proteome (2017/13,903 proteins), highlighting the rapid expansion of novel, likely pathogenicity-related gene families. The vast majority (98.8%) of these novel families do not have BLASTp hits in the NCBI non-redundant database or have hits to unknown proteins (minimum *e*-value of 1e-10; **Table 3**) and 94.5% do not contain InterPro domains (unpublished, jgi.doe.gov/Cronartium; Hunter et al., 2012). Since the families that are unique to the *Cqf* lineage are largely uncharacterized, they likely follow the assumptions for putative pathogenicity factors or effectors. Nearly 12% (234/2,017) of proteins encoded in the 248 novel *Cqf* families are members of the predicted *Cqf* secretome. This is significantly greater than the ∼8% of entire *Cqf* proteome that also belongs to the secretome (Chi-square = 25.418, p-value = 0.0001). The protein characteristics of these secreted proteins are effector-like, as the average cysteine content is 2.2% and the median protein length is 272 amino acids.

The families that were duplicated in the *Cqf* lineage and contain sequences from other taxa exhibit patterns of conservation that differ from the families duplicated or depleted in rust

Table 2 | Potential sub- and neo-functionalization within *Cqf*-specific putative effector families.

Gene family ID	*Cqf* proteins in family	Proteins in *Cqf* secretome within family
8514	6*	6
9417	5*	5
9892	5*	5
2663	17	4
6030	11	4
9436	5	4
9897	5	4
7037	8	3
9890	5	3
9891	5	3
10,467	4	3
11,111	4	3
11,876	3*	3
11,879	3*	3
11,934	3*	3
12,788	3*	3
12,794	3*	3
12,804	3*	3
12,812	3*	3
5202	13	2
7921	7	2
7928	7	2
8502	6	2
9880	5	2
10,433	4	2
10,463	4	2
11,112	4	2
11,928	3	2
11,944	3	2
12,777	3	2
12,803	3	2
12,807	3	2
12,814	3	2
14,526	2*	2
14,527	2*	2
14,537	2*	2
14,552	2*	2
14,554	2*	2
14,563	2*	2
14,570	2*	2
14,577	2*	2
14,589	2*	2
14,623	2*	2
15,977	2*	2
15,979	2*	2
15,980	2*	2
15,992	2*	2
16,012	2*	2
16,034	2*	2
16,036	2*	2
16,051	2*	2
16,052	2*	2
16,078	2*	2
16,079	2*	2
16,080	2*	2
16,081	2*	2
16,091	2*	2
16,101	2*	2
16,102	2*	2
16,106	2*	2
16,119	2*	2
16,123	2*	2
16,129	2*	2
16,146	2*	2
16,160	2*	2
16,163	2*	2
16,191	2*	2
Total	237 proteins	164 proteins
Average	3.54 *Cqf* proteins per family	2.4 secreted proteins per family

Cqf-specific gene families with greater than two predicted secreted proteins are listed. The total number of proteins in each Cqf-specific family is provided as well as the number of proteins belonging to the Cqf secretome (i.e., predicted secreted proteins) are indicated. Asterisks adjacent to total protein counts indicate families where all members are Cqf secretome members. If no asterisk is present, only a portion of the family received secretion predictions. Family 9417 (bold) will be detailed later in article.

pathogens (site R, **Figure 3**). Instead, these *Cqf*-specific duplicated families ($n = 549$ families) are predominantly conserved in not only *Cqf*, but also 2-3 taxa (**Figure 6**).

Cqf PUTATIVE EFFECTOR GENE FAMILIES—DISTRIBUTION AND EXPANSION

Family 5485 is the largest family represented in the predicted *Cqf* secretome. The family contains 12 orthologous proteins (7 *Cqf*, 4 *Mlp*, and 1 *Pgt* proteins). Eleven of the 12 proteins have predicted N-terminal signal peptides (SignalP 4.0; Petersen et al., 2011), and all seven members from *Cqf* are annotated as belonging to the *Cqf* secretome. Domain architecture and conservation data for Family 5485 proteins helps to predict their biological functions and putative roles in establishing infection. Additionally, 11 of the 12 proteins in this family contained three multicopper oxidase (MCO) domains and the remaining protein (Pgt_20719) contained two of the three domains (**Figure 7A**). The Interpro domains identified include: Cupredoxin domain (IPR008972), Multicopper Oxidase, Type 1 (IPR001117), and Multicopper Oxidase, Type 2 (IPR011706), and Multicopper Oxidase, Type 3 domain (IPR011707) (**Figure 7A**). A Copper-Binding Site domain (IPR002355) was identified in only three *Cqf* family members. These three proteins have a distinct phylogenetic history from other family members (**Figure 7B**). Generally, the

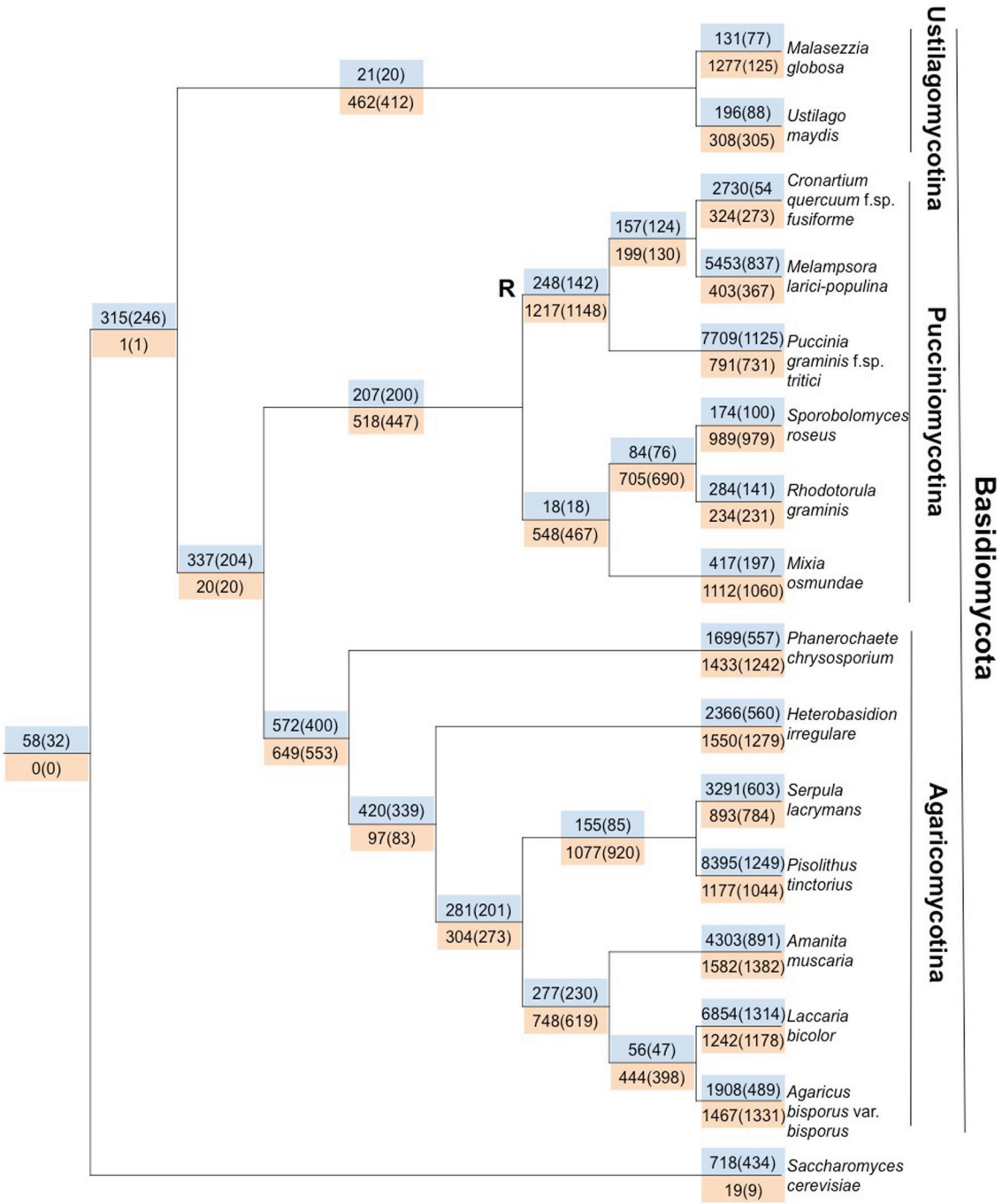


FIGURE 3 | Gene loss and gain in the Basidiomycota fungal lineage highlights shared loss in the rust pathogen lineage and high levels of species-specific gain. Mapping putative gene duplications and losses across 16 fungal taxa. Values in blue are associated with gains/duplications, whereas orange indicates loss. Outside of parentheses are the number of gain or loss events that have occurred on the branch preceding a node, and within parentheses are the number of gene families associated with duplications or losses. The node denoted with R indicates the last common ancestor of the rust pathogens.

phylogenetic relationships, as well as the genomic colocalization of the proteins in this family mirrors the domain architecture, providing insight into how these proteins evolved (**Figure 7A**). Several additional families of MCOs are present in the *Cqf* genome (i.e., Families 5853, 1542, and 1053), however, by definition, Family 5485 has a distinct evolutionary history from other families as evidenced by distinct family placement by OrthoMCL.

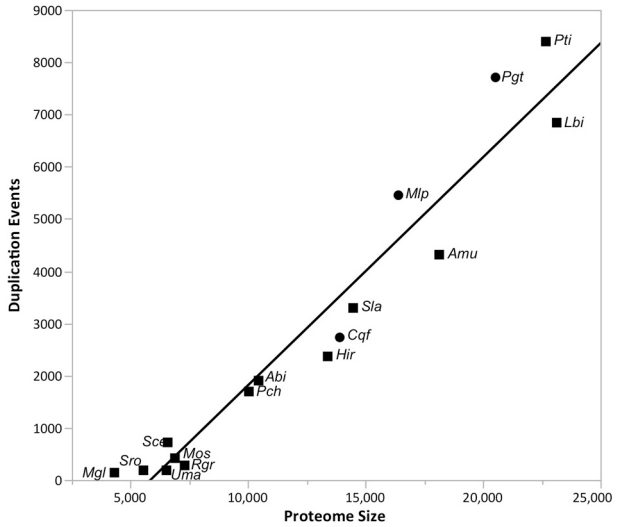


FIGURE 4 | The proteome size is positively correlated with the number of gene family duplications across taxa. The linear relationship ($R^2 = 0.93$) between the proteome size (i.e., protein count) and the number of species-specific duplications detected in the analyses from **Figure 1** is depicted in this figure. Rust pathogens (*Cqf, Mlp,* and *Pgt*) are indicated with circles and non-rust pathogens with squares. The line of best fit (black) is indicated. Please reference Methods for species abbreviations.

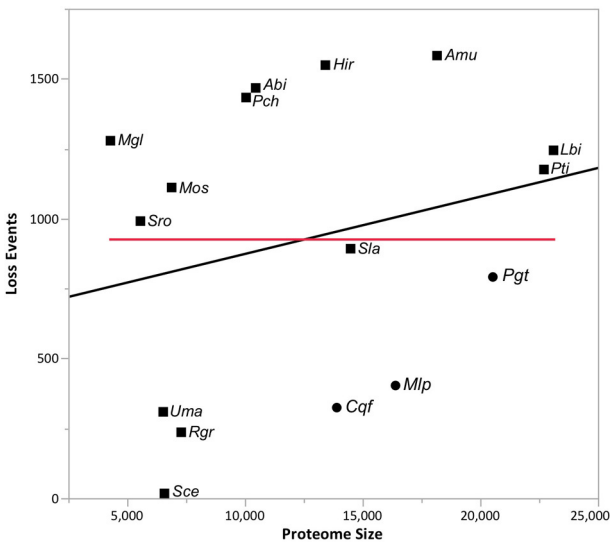


FIGURE 5 | Lack of relationship between proteome size and gene family losses across taxa. The relationship ($R^2 = 0.06$) between the proteome size (i.e., protein count) and the number of species-specific losses detected in the analyses from **Figure 1** is depicted in this figure. The line of best fit (black) and fit mean (red) are also shown. Rust pathogens (*Cqf, Mlp,* and *Pgt*) are indicated with circles and non-rust pathogens with squares. Please reference Methods for species abbreviations.

Family 9417 is the third largest family in the *Cqf* secretome, with all five of its proteins predicted as secreted. This family contains putative effectors likely involved in the establishment of disease, as all members have signal peptides, short lengths (average 207 aa), and high cysteine content (average 6.5%). All five family members contain at least one fungal extracellular membrane (CFEM) domain (Interpro IPR008427). Five additional proteins encoded in the *Cqf* proteome contain CFEM domains. Two of the five do not belong to a gene family, and the remaining three proteins each were ascribed to different families containing orthologs from multiple fungal taxa, unlike *Cqf*-specific family 9417. Similar to Family 5485, proteins of Family 9417 also colocalize in the genome, as three members are located on scaffold 43 of the *Cqf* assembly and the remaining two proteins are adjacent to one another on scaffold 5 (**Figure 8**). Protein members of Family 9417 adhere to consensus domain structure and subcellular targeting of previously identified CFEM proteins. Online prediction algorithms detected transmembrane domain regions (Tmpred; Hofmann, 1993) and glycosylphosphatidylinositol (GPI) anchor sites (big-Pi Predictor; Nielsen et al., 1997) in a subset of the family proteins (**Figure 8**). All proteins, excluding Cqf91696, were predicted to have N-terminal transmembrane helices spanning amino acids 3-23 for both proteins. Two members, Cqf712797 and Cqf651034 had C-terminal GPI anchor sites at amino acids 223 and 302, respectively (*p*-values 1.25E-04 and 2.10E-04). Only Cqf91696 had no bioinformatic evidence of association with the fungal membrane.

DISCUSSION

This study provides the first detailed analysis of the secretome of the fusiform rust pathogen, *Cqf*, since the recent assembly and annotation of a draft reference genome (unpublished, jgi.doe.gov/Cronartium). Additional criteria used in isolating putative effectors from within the *Cqf* genome and its corresponding secretome, included proteins exhibiting rust pathogen-specific and *Cqf*-specific gene family membership. Following gene family constructions, we highlighted putative effectors with paralogs (within *Cqf*) or orthologs/paralogs (between *Cqf* and other taxa) within the *Cqf* secretome. Over half (51%) of proteins considered to be effector-like (small, cysteine-rich, secreted proteins) belong to gene families. This is comparable to results found in the hemibiotrophic pathogens *Phytophthora ramorum* and *P. sojae* where 77% of their secretomes are found in multigene families (Tyler et al., 2006). These findings demonstrate the value of an evolutionary perspective for highlighting families harboring putative *Cqf* effectors. Altogether, the large-scale comparative genomics analyses in this study help elucidate the unique patterns of evolution in a rust proteome and its associated secretome.

PUTATIVE EFFECTOR FAMILIES

With the completion of the *Cqf* draft genome, it is important to identify proteins that may be involved in establishing disease, such as effectors, on oak and pine hosts. Based on the evolutionary forces presumed to act on effectors, in combination with a trio of rust pathogen genomes facilitating comparative analyses, we can now do experiments not previously feasible. We suggest this is a reasonable approach to identifying putative effectors that complements more conventional methods. Previous studies searched for effectors within other systems based on the presence of a signal peptide, cysteine richness, and short protein lengths (<300 aa) (Joly et al., 2010; Cantu et al., 2011; Duplessis et al., 2011b;

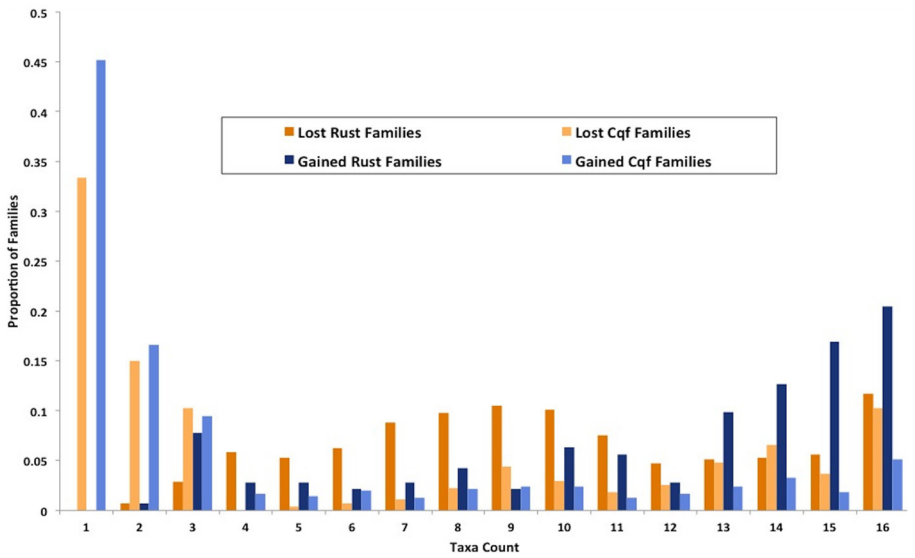


FIGURE 6 | Families gained and lost in rust pathogens and *Cqf* have varying levels of taxonomic conservation. The proportion of gene families (y-axis) that contain protein members from 1 to 16 fungal taxa (x-axis) among those that have either expanded or contracted in the rust fungi or *Cqf*.

Table 3 | *Cqf*-specific duplicated gene families contain predominantly uncharacterized proteins.

Gene family annotation	Number of gene families (proteins)
No hits	183 (1510)
Unknown protein	57 (430)
20S Proteasome subunit alpha 6	2 (9)
Zinc finger CCHC-type protein	1 (33)
HIV-1 retropepsin, polyprotein	1 (13)
Polysaccharide lyase family 4	1 (7)
Reverse transcriptase	1 (5)
MFS transporter, inorganic phosphate transporter	1 (5)
CFEM domain containing protein	1 (5)

Functional annotation of the 248 gene families duplicated only in Cqf by BLASTp against the non-redundant NCBI database (minimum e-values of 1e-10). A family was ascribed a function if more than two proteins in the family received the same top annotated BLASTp hit. The number of proteins within families is indicated in parentheses.

Hacquard et al., 2012; Saunders et al., 2012). This study highlights the usefulness of comparative genomic analyses to examine the evolutionary history of each secretome member, and that this approach can also be complemented with structural characteristics of predicted secreted proteins. The rationale behind these comparisons is that effector families conserved in rust fungi and unique to *Cqf* are candidates for conditioning rust pathogen and *Cqf* infection strategies, respectively.

We observed species-specific proteomic gene family gains/duplications in the *Cqf* lineage, a subset of which represents putative effectors. The paralogous nature (i.e., multi-copy) of their protein family members indicates functional redundancy, which is consistent with other pathogenic fungi (Kamper et al., 2006; Saitoh et al., 2012). We have identified two lines of evidence that point toward neo- and sub-functionalization in *Cqf* putative effector families. First, differential subcellular localization predictions have been observed within putative effector families. In about 34% of *Cqf*-specific families, only a subset of proteins are secreted from the fungal cell, while remaining family members are not predicted for secretion, thus remaining within the fungal cell. This pattern suggests that secreted proteins with effector function may have evolved from non-secreted proteins without an effector function or vice versa. Second, changes in domain architecture of proteins within putative effector families also points to neo- or subfunctionalization. For example family 5485 contains MCO laccase-like enzymes and a single clade of three *Cqf* proteins that have acquired a MCO copper binding site in the evolution of this family. It is possible that these proteins have novel or distinct functions within *Cqf* than their paralogs within the genome. This family is a strong putative effector family because all *Cqf* members belong to the predicted secretome and it has undergone *Cqf*-specific family duplications. Protein members within this family co-localize in the genome, possibly resulting from tandem duplication from non-equal crossing over. Various functions have been ascribed to previously identified fungal MCOs including lignin degradation (Leonowicz et al., 2001; Lundell et al., 2010), melanin synthesis (Langfelder et al., 2003), fruiting body formation (Kues and Liu, 2000), and pathogenicity on hosts (Zhu and Williamson, 2004). This family has expanded in *Cqf,* the first sequenced rust pathogen that forms stem galls in woody tissues, and we hypothesize that these enzymes play a role in gall formation. The most common function for laccases/MCOs in basidiomycete fungi is lignin metabolism (Thurston, 1994; Kües and Rühl, 2011). However, this gene family exhibits a lack of conservation with known

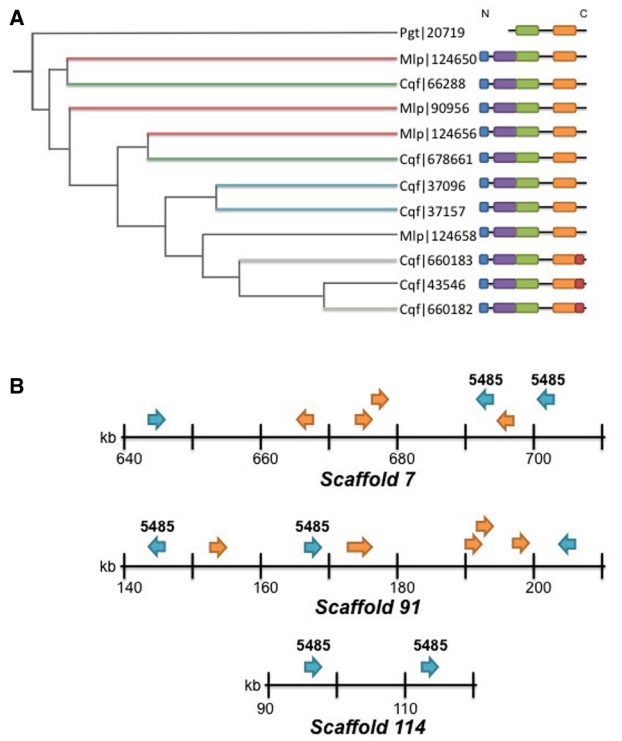


FIGURE 7 | Domain structure, phylogenetic relationships, and colocalization of copper oxidases. (A) Phylogenetic relationship estimated with TreeFix (Wu et al., 2013) for the 12 members of Family 5485. Tree reconstructed with Tree of Life viewer (Letunic and Bork, 2011). See methods for additional details for gene tree estimation. Branch lengths are not informative for this tree. Domain architectures are indicated for each protein: N-terminal signal peptides per SignalP 4.0 (blue; Petersen et al., 2011); (MCu1—IPR01117) Multicopper oxidase Type 1 (green); (MCu2—IPR011706) Multicopper oxidase Type 2 (orange); (MCu3—IPR011707) Multicopper oxidase Type 3 (purple); and (CuBS—IPR002355) Multicopper oxidase, copper-binding site as determined by InterPro (red; Hunter et al., 2012). The colocalization of proteins on scaffolds is indicated by branches sharing the identical colors. Thin black branches do not colocalize. **(B)** Co-localization of proteins within MCO Family 5485 on scaffolds 7, 91, and 114. Family 5485 members are denoted with the ID above each gene model. Secretome members are blue arrows and non-secreted proteins are orange arrows. Gene orientation on the scaffold is indicated with arrows. Note: gene lengths are not to scale.

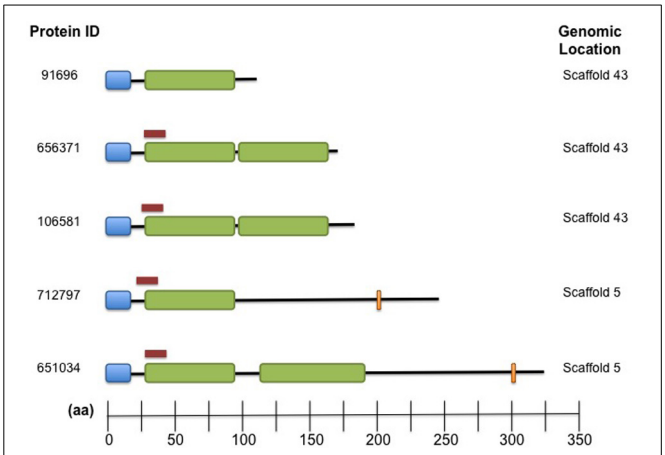


FIGURE 8 | Family 9417 proteins harbor both signal peptides and CFEM domains domains and colocalize in the *Cqf* genome. Domain architecture of Family 9417 protein members where signal peptides (blue), CFEM Interpro domains (green), GPI anchor sites (orange), and predicted transmembrane domains (red) are indicated for each protein. Three proteins colocalize together on scaffold 43, the remaining on scaffold 5 (indicated on right).

MCOs of lignin-degrading wood rots (*P. chrysosporium* and *S. lacrymans*), which points to the possibility that these enzymes may be involved in pathogenicity or may metabolize a plant substrate other than lignin. On both hosts, *Cqf* infects primary tissue that lacks high levels of lignification such as spongy mesophyll cells of oak leaves (Mims et al., 1996) and vascular cambium of pine (Gray et al., 1982). If the Family 5485 enzymes are involved in lignin degradation, the enzymatic activity may occur late in gall development on the pine host, where the tissues are more heavily lignified due to secondary wall formation. Though their biochemical targets are unknown *in planta*, we hypothesize that Family 5485 enzymes are secreted during infection and condition the gall phenotype on the pine host. Further studies are required to elucidate their true role in disease.

A second gene family that has expanded in the *Cqf* lineage is Family 9417, which includes five *Cqf*-specific paralogs that co-localize in the genome. Similar to Family 5485, differential domain architecture within this family implies that neo- or subfunctionalization may have occurred. Family 9417 contains putative effectors that harbor conserved, fungal-specific CFEM-domains. These domains exhibit a characteristic cysteine distribution and have a broad taxonomic conservation in fungi (Kulkarni et al., 2003; Martin et al., 2008; Perez et al., 2011). Predicted functions of proteins harboring CFEM domains include critical roles in appressorial development (Choi and Dean, 1997; DeZwaan et al., 1999), signal transducers, adhesion and cell-surface receptors (Kulkarni et al., 2003). In contrast to Family 5485 proteins, which may interact with the host during infection, the molecular target for Family 9417 proteins could be fungal. We hypothesize these proteins are secreted and may play roles during infection of the host.

EVOLUTION OF GENE GAIN AND LOSS

Patterns of gene family loss and gain for rust fungi highlight major shifts in their proteomes, possibly associated with the rust pathogen's obligate biotrophic lifestyle. The origin of the rust pathogen clade is associated with nearly five times more losses, or family contractions, than duplications. There are many possible mechanisms for gene loss in rust fungi. For this reason, further investigations are required to both identify specific mechanisms and quantify their levels of effects on gene family evolution in rust fungi. However, we hypothesize that the since obligate biotrophy has evolved multiple times in fungi (Spanu, 2012), the skew toward gene loss in the rust pathogen lineage might be associated with the shift from the life history of its ancestral state to that of the obligate biotrophic pathogens we observe today. These lost and/or contracted families exhibit broad taxonomic

conservation and may have been constituents of the ancestral "core" fungal gene set, suggesting that they are unnecessary for obligate biotrophic but may be necessary for free-living and symbiotic species. For example, enzymes integral to the sulfur and nitrogen assimilation pathways are missing in *Cqf* (unpublished, jgi.doe.gov/Cronartium), *Mlp*, and *Pgt* (Duplessis et al., 2011b). This also suggests that evolution for obligate biotrophy drives toward an irreversible life history shift (Spanu, 2012).

Although the rusts have undergone considerable gene family losses and contractions, they exhibit some of the largest proteomes in fungi. Much of their proteome size appears to be due to species-specific duplications. Nearly one-third (32.1%) of all observed duplications across all of the sampled basidiomycete fungi are rust taxon-specific duplications. The high levels of species-specific duplication yield disproportionately greater numbers of newly-evolved genes in the rust pathogen genomes compared to ancient or conserved genes (genes shared with older lineages) in each proteome. The presence of so many species-specific duplications suggests that the rusts have highly labile genomes. This is consistent with the large (>10%) genomic size variation detected in progeny from a single *Cqf* cross relative to parental isolates (Anderson et al., 2010). Such rapid changes, occurring in the span of a single generation, could facilitate the gene gains and losses observed in our analyses. The close association with hosts may foster a labile and diverse genome, enabling the parasites to rapidly adapt to the continually evolving host resistance pathways.

COMPARATIVE ANALYSIS AND GENETIC MAPPING TO VALIDATE PUTATIVE EFFECTORS

Further characterization of putative effectors in *Cqf* could be accomplished with analysis of selection potentially arising from host resistance mechanisms (Allen et al., 2004; Aguileta et al., 2009; Barrett et al., 2009; Thrall et al., 2012). In addition, expression analysis can be informative, since secreted proteins with specific expression profiles during infection are stronger effector candidates (Ellis et al., 2009). Time-course experiments have been successful in other rust pathogen systems to elucidate the effector-like proteins involved in multiple or highly specific stages during infection (Joly et al., 2010; Duplessis et al., 2011a; Bruce et al., 2014). Also, resequencing of closely related rust pathogens such as *Cronartium ribicola*, *C. flaccidum*, and *Peridermium harknessii* (Vogler and Bruns, 1998) would improve precision of gene family delineations and identification of true singleton *Cqf* effectors, which are likely to be more newly evolved than effectors in families, and may therefore be products of highly-specific host-*Cqf* coevolution. Finally, a subset of the predicted effectors are avirulence proteins and are, by definition, involved in genotype-specific "gene-for-gene" interactions with hosts. These putative avirulence effectors can be validated through genetic mapping to their corresponding host resistance genes, an approach that has previously been successful in identifying the first avirulence protein locus in *Cqf* (Kubisiak et al., 2011). Altogether, these validation approaches will yield true members of the *Cqf* secretome and provide additional insight into the biological functions for effectors infecting oak and pine.

ACKNOWLEDGMENTS

We acknowledge funding from the USDA Forest Service, Southern Research Station (Agreement 11-CA-11330126-120) (to John M. Davis). This work was partly supported by the French National Research Agency through the Laboratory of Excellence ARBRE (ANR-11-LABX-0002-01) (to Francis M. Martin). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank the following fungal genome project PIs for pre-publication access to genome sequence information: Drs. Sharon Doty (*Rhodotorula graminis*) and Ken Wolfe (*Sporobolomyces roseus*). The *Cqf* genome assembly and annotations can be interactively accessed through the JGI fungal genome portal MycoCosm at jgi.doe.gov/Cronartium (unpublished). We acknowledge Alicia Clym and Andrea Aerts at JGI for their roles in assembling and annotating the *Cqf* reference genome, respectively. Dr. Annegret Kohler at INRA, as well as Dr. Alan Kuo at JGI, are acknowledged for their leading roles in the acquisition of mycorrhizal genome sequence data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00299/abstract

REFERENCES

Aguileta, G., Hood, M., Refregier, G., Giraud, T., Kader, J., and Delseny, M. (2009). Genome evolution in plant pathogenic and symbiotic fungi. *Adv. Bot. Res.* 49, 151–193. doi: 10.1016/S0065-2296(08)00603-4

Allen, R., Bittner-Eddy, P., Grenvitte-Briggs, L., Meitz, J., Rehmany, A., Rose, L., et al. (2004). Host-parasite coevolutionary conflict between Arabidopsis and downy mildew. *Science* 306, 457, 1957–1960. doi: 10.1126/science.1104022

Anderson, C., Kubisiak, T., Nelson, C., Smith, J., and Davis, J. (2010). Genome size variation in the pine fusiform rust pathogen *Cronartium quercuum* f.sp. fusiforme as determined by flow cytometry. *Mycologia* 102, 1295–1302. doi: 10.3852/10-040

Barrett, L., Thrall, P., Dodds, P., van der Merwe, M., Linde, C., Lawrence, G., et al. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol. Biol. Evol.* 26, 2499–2513. doi: 10.1093/molbev/msp166

Bendtsen, J., Nielsen, H., Von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340, 783–795. doi: 10.1016/j.jmb.2004.05.028

Bruce, M., Neugebauer, K. A., Joly, D. L., Migeon, P., Cuomo, C. A., Wang, S., et al. (2014). Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat. *Front. Plant Sci.* 4:520. doi: 10.3389/fpls.2013.00520

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K., et al. (2011). Next Generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Choi, W., and Dean, R. (1997). The adenylate cyclase gene MAC1 of *Magnaporthe grisea* controls appressorium formation and other aspects of growth and development. *Plant Cell* 9, 1973–1983. doi: 10.1105/tpc.9.11.1973

DeZwaan, T., Carroll, A., Valent, B., and Sweigard, J. (1999). *Magnaporthe grisea Pth11p* is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues. *Plant Cell* 11, 2013–2030. doi: 10.1105/tpc.11.10.2013

Dodds, P., Rafiqi, M., Gan, P., Hardham, A., Jones, D., and Ellis, J. (2009). Effectors of biotrophic fungi and oomycetes: pathogenicity factors and triggers of host resistance. *New Phytol.* 183,993–999. doi: 10.1111/j.1469-8137.2009.02922.x

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011b). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., et al. (2011a). *Melampsora larici-populina* transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. *Mol. Plant Microbe Interact.* 24, 808–818. doi: 10.1094/MPMI-01-11-0006

Eastwood, D. C., Floudas, D., Binder, M., Majcherczyk, A., Schneider, P., Aerts, A., et al. (2011). The plant cell wall–decomposing machinery underlies the functional diversity of forest fungi. *Science* 333, 762–765. doi: 10.1126/science.1205411

Edgar, R. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Eisenhaber, B., Bork, P., and Eisenhaber, F. (1999). Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.* 292, 741–758.

Ellis, J., Rafiqi, M., Gan, P., Chakrabarti, A., and Dodds, P. (2009). Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Curr. Opin. Plant Biol.* 12, 399–405. doi: 10.1016/j.pbi.2009.05.004

Emanuelsson, O., Nielsen, H., Brunak, S., and Von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* 300, 1005–1016. doi: 10.1006/jmbi.2000.3903

Giraldo, M., and Valent, B. (2013). Filamentous plant pathogen effectors in action. *Nat. Rev. Microbiol.* 11, 800–814. doi: 10.1038/nrmicro3119

Goffeau, A., Barrell, B., Bussey, H., Davis, R., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 546–567.

Gorecki, P., and Tiuryn, J. (2007). URec: a system for unrooted reconciliation. *Bioinformatics* 23, 511–512. doi: 10.1093/bioinformatics/btl634

Gray, D., Amerson, H., and Van Dyke, C. (1982). An ultrastructural comparison of monokaryotic and dikaryotic haustoria formed by the fusiform rust fungus *Cronartium-quercuum* f-sp *fusiforme*. *Can. J. Bot.* 60, 2914–2922. doi: 10.1139/b82-352

Grigoriev, I., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., et al. (2014). MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42, D699–D704. doi: 10.1093/nar/gkt1183

Hacquard, S., Joly, D., Lin, Y., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (Poplar Leaf Rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hofmann, K. (1993). TMbase-A database of membrane spanning protein segments. *Biol. Chem.* 374, 166.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T., Bateman, A., et al. (2012). InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40, D306–D312. doi: 10.1093/nar/gkr948

Joly, D., Feau, N., Tanguay, P., and Hamelin, R. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11: 422. doi: 10.1186/1471-2164-11-422

Jones, J., and Dangl, J. (2006). The plant immune system. *Nature* 444, 323–329. doi: 10.1038/nature05286

Kamper, J., Kahmann, R., Bolker, M., Ma, L., Brefort, T., Saville, B., et al. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* 444, 97–101. doi: 10.1038/nature05248

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Kubisiak, T., Anderson, C., Amerson, H., Smith, J., Davis, J., and Nelson, C. (2011). A genomic map enriched for markers linked to Avr1 in *Cronartium quercuum* f.sp. *fusiforme*. *Fungal Genet. Biol.* 48, 266–274. doi: 10.1016/j.fgb.2010.09.008

Kues, U., and Liu, Y. (2000). Fruiting body production in basidiomycetes. *Appl. Microbiol. Biotechnol.* 54, 141–152. doi: 10.1007/s002530000396

Kües, U., and Rühl, M. (2011). Multiple multi-copper oxidase gene families in basidiomycetes-what for? *Curr. Genomics* 12, 72–94. doi: 10.2174/138920211795564377

Kulkarni, R., Kelkar, H., and Dean, R. (2003). An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. *Trends Biochem. Sci.* 28, 118–121. doi: 10.1038/nature03449

Langfelder, K., Streibel, M., Jahn, B., Haase, G., and Brakhage, A. (2003). Biosynthesis of fungal melanins and their importance for human pathogenic fungi. *Fungal Genet. Biol.* 38, 143–158. doi: 10.1016/S1087-1845(02)00526-1

Leonowicz, A., Cho, N., Luterek, J., Wilkolazka, A., Wojtas-Wasilewska, M., Matuszewska, A., et al. (2001). Fungal laccase: properties and activity on lignin. *J. Basic Microb.* 41, 185–227. doi: 10.1002/1521-4028(200107)41:3/4<185::AID-JOBM185>3.0.CO;2-T

Letunic, I., and Bork, P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res.* 39, W475–W478. doi: 10.1093/nar/gkr201

Li, L., Stoeckert, C., and Roos, D. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.122450

Lundell, T., Makela, M., and Hilden, K. (2010). Lignin-modifying enzymes in filamentous basidiomycetes - ecological, functional and phylogenetic review. *J. Basic Microb.* 50, 5–20. doi: 10.1002/jobm.200900338

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E., Duchaussoy, F., et al. (2008). The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452, 88–92. doi: 10.1038/nature06556

Martinez, D., Larrondo, L. F., Putnam, N., Gelpke, M. D. S., Huang, K., Chapman, J., et al. (2004). Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. *Nat. Biotechnol.* 22, 695–700. doi: 10.1038/nbt967

Mims, C. W., Liljebjelke, K. A., and Covert, S. F. (1996). Ultrastructure of telia and teliospores of the rust fungus *Cronartium quercuum f.* sp. *fusiforme*. *Mycologia* 88, 47–56. doi: 10.2307/376078

Morin, E., Kohler, A., Baker, A. R., Foulongne-Oriol, M., Lombard, V., Nagye, L. G., et al. (2012). Genome sequence of the button mushroom Agaricus bisporus reveals mechanisms governing adaptation to a humic-rich ecological niche. *Proc. Natl. Acad. Sci. U.S.A.* 109, 17501–17506. doi: 10.1073/pnas.1206847109

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6. doi: 10.1093/protein/10.1.1

Olson, Å., Aerts, A., Asiegbu, F., Belbahri, L., Bouzid, O., Broberg, A., et al. (2012). Insight into trade off between wood decay and parasitism from the genome of a fungal forest pathogen. *New Phytol.* 194, 1001–1013. doi: 10.1111/j.1469-8137.2012.04128.x

Perez, A., Ramage, G., Blanes, R., Murgui, A., Casanova, M., and Martinez, J. P. (2011). Some biological features of *Candida albicans* mutants for genes coding fungal proteins containing the CFEM domain. *FEMS Yeast Res.* 11, 273–284. doi: 10.1111/j.1567-1364.2010.00714.x

Petersen, T., Brunak, S., Von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Rafiqi, M., Ellis, J., Ludowici, V., Hardham, A., and Dodds, P. (2012). Challenges and progress towards understanding the role of effectors in plant-fungal interactions. *Curr. Opin. Plant Biol.* 15, 477–482. doi: 10.1016/j.pbi.2012.05.003

Rasmussen, M., and Kellis, M. (2011). A Bayesian approach for fast and accurate gene tree reconstruction. *Mol. Biol. Evol.* 28, 273–290. doi: 10.1093/molbev/msq189

Rep, M. (2005). Small proteins of plant-pathogenic fungi secreted during host colonization. *FEMS Microbiol. Lett.* 253, 19–27. doi: 10.1016/j.femsle.2005.09.014

Saitoh, H., Fujisawa, S., Mitsuoka, C., Ito, A., Hirabuchi, A., Ikeda, K., et al. (2012). Large-scale gene disruption in *Magnaporthe oryzae* identifies MC69, a secreted protein required for infection by monocot and dicot fungal pathogens. *PLoS Pathog.* 8:e1002711. doi: 10.1371/journal.ppat.1002711

Saunders, D. G., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:e29847. doi: 10.1371/journal.pone.0029847

Spanu, P. D. (2012). The genomics of obligate (and nonobligate) biotrophs. *Annu. Rev. Phytopathol.* 50, 91–109. doi: 10.1146/annurev-phyto-081211-173024

Stamatakis, A., Ludwig, T., and Meier, H. (2005). RAxML-III: a program for sequential, parallel and distributed inference of large phylogenetic trees. *Concurr. Comput. Pract. Exp.* 17, 1705–1723. doi: 10.1093/bioinformatics/bti191

Stergiopoulos, I., and de Wit, P. (2009). Fungal effector proteins. *Annu. Rev. Phytopathol.* 47, 233–263. doi: 10.1146/annurev.phyto.112408.132637

Thrall, P., Laine, A., Ravensdal, M., Nemri, A., Dodds, P., Barrett, L., et al. (2012). Rapid genetic change underpins antagonistic coevolution in a natural host-pathogen metapopulation. *Ecol. Lett.* 15, 425–435. doi: 10.1111/j.1461-0248.2012.01749.x

Thurston, C. (1994). The structure and function of fungal laccases. *Microbiol* 140, 19–26. doi: 10.1099/13500872-140-1-19

Toome, M., Ohm, R. A., Riley, R. W., James, T. Y., Lazarus, K. L., Henrissat, B., et al. (2014). Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. *New Phytol.* 202, 554–564. doi: 10.1111/nph.12653

Tyler, B., Tripathy, S., Zhang, X., Dehal, P., Jiang, R., Aerts, A., et al. (2006). *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313, 1261–1266. doi: 10.1126/science.1128796

Voegele, R., and Mendgen, K. (2011). Nutrient uptake in rust fungi: how sweet is parasitic life? *Euphytica* 179, 41–55. doi: 10.1007/s10681-011-0358-5

Vogler, D., and Bruns, T. (1998). Phylogenetic relationships among the pine stem rust fungi (*Cronartium* and *Peridermium* spp.). *Mycologia* 90, 244–257. doi: 10.2307/3761300

Wu, Y., Rasmussen, M., Bansal, M., and Kellis, M. (2013). TreeFix: statistically informed gene tree error correction using species trees. *Syst. Biol.* 62, 110–120. doi: 10.1093/sysbio/sys076

Xu, J., Saunders, C. W., Hu, P., Grant, R. A., Boekhout, T., Kuramae, E. E., et al. (2007). Dandruff-associated Malassezia genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. *Proc. Natl. Acad. Sci. U.S.A.* 104, 18730–18735. doi: 10.1073/pnas.0706756104

Zhu, X., and Williamson, P. (2004). Role of laccase in the biology and virulence of *Cryptococcus neoformans*. *FEMS Yeast Res.* 5, 1–10. doi: 10.1016/j.femsyr.2004.04.004

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 01 April 2014; accepted: 06 June 2014; published online: 26 June 2014.
Citation: Pendleton AL, Smith KE, Feau N, Martin FM, Grigoriev IV, Hamelin R, Nelson CD, Burleigh JG and Davis JM (2014) Duplications and losses in gene families of rust pathogens highlight putative effectors. Front. Plant Sci. ***5****:299. doi: 10.3389/fpls.2014.00299*
This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.
Copyright © 2014 Pendleton, Smith, Feau, Martin, Grigoriev, Hamelin, Nelson, Burleigh and Davis. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 01 September 2014
doi: 10.3389/fpls.2014.00372

Diversifying selection in the wheat stem rust fungus acts predominantly on pathogen-associated gene families and reveals candidate effectors

Jana Sperschneider[1]*, Hua Ying[2], Peter N. Dodds[2], Donald M. Gardiner[3], Narayana M. Upadhyaya[2], Karam B. Singh[1,4], John M. Manners[2] and Jennifer M. Taylor[2]

[1] *Plant Industry, Centre for Environment and Life Sciences, Commonwealth Scientific and Industrial Research Organisation, Perth, WA, Australia*
[2] *Plant Industry, Black Mountain Laboratories, Commonwealth Scientific and Industrial Research Organisation,Canberra, ACT, Australia*
[3] *Plant Industry, Queensland Bioscience Precinct, Commonwealth Scientific and Industrial Research Organisation, Brisbane, QLD, Australia*
[4] *University of Western Australia Institute of Agriculture, University of Western Australia, Crawley, WA, Australia*

Edited by:
Sébastien Duplessis, Institut national de la recherche agronomique, France

Reviewed by:
Sylvain Raffaele, Institut national de la recherche agronomique, France
Gabriela Aguileta, Universite Paris Sud, France

****Correspondence:***
Jana Sperschneider, CSIRO Plant Industry, Centre for Environment and Life Sciences, Underwood Avenue, Floreat, WA 6014, Australia
e-mail: jana.sperschneider@csiro.au

Plant pathogens cause severe losses to crop plants and threaten global food production. One striking example is the wheat stem rust fungus, *Puccinia graminis* f. sp. *tritici*, which can rapidly evolve new virulent pathotypes in response to resistant host lines. Like several other filamentous fungal and oomycete plant pathogens, its genome features expanded gene families that have been implicated in host-pathogen interactions, possibly encoding effector proteins that interact directly with target host defense proteins. Previous efforts to understand virulence largely relied on the prediction of secreted, small and cysteine-rich proteins as candidate effectors and thus delivered an overwhelming number of candidates. Here, we implement an alternative analysis strategy that uses the signal of adaptive evolution as a line of evidence for effector function, combined with comparative information and expression data. We demonstrate that *in planta* up-regulated genes that are rapidly evolving are found almost exclusively in pathogen-associated gene families, affirming the impact of host-pathogen co-evolution on genome structure and the adaptive diversification of specialized gene families. In particular, we predict 42 effector candidates that are conserved only across pathogens, induced during infection and rapidly evolving. One of our top candidates has recently been shown to induce genotype-specific hypersensitive cell death in wheat. This shows that comparative genomics incorporating the evolutionary signal of adaptation is powerful for predicting effector candidates for laboratory verification. Our system can be applied to a wide range of pathogens and will give insight into host-pathogen dynamics, ultimately leading to progress in strategies for disease control.

Keywords: rust, effector, adaptation, avirulence, selection, *Puccinia graminis*, fungal pathogens

INTRODUCTION

The basidiomycete *Puccinia graminis* f. sp. *tritici* is the causal agent of wheat and barley stem rust. Wheat stem rust is a major threat to wheat production worldwide and hence to global food security, and has a long history of evolution to new virulent pathotypes in response to deployment of host genetic resistance (Leonard and Szabo, 2005; Pardey et al., 2013). The Ug99 strain, which first emerged in east Africa around 1999, has potential to infect most of the world's wheat (Singh et al., 2011). *Puccinia graminis* f. sp. *tritici* features a remarkably large genome of 89 Mb, which is more than four times larger than the related genome of the basidiomycete smut fungus *Ustilago maydis* (Kämper et al., 2006). Innovation of gene content has been attributed to large sets of lineage-specific expanded gene families, which are thought to drive adaptation and pathogen-associated processes (Duplessis et al., 2011; Raffaele and Kamoun, 2012).

Filamentous plant pathogens use molecules called effectors that modify host defense-related signaling, cell structure, metabolism and function to effect successful infection (Koeck et al., 2011; Giraldo and Valent, 2013). These pathogens deliver effector molecules either to the host apoplast or translocate them directly into the host cytoplasm, often through specialized infection structures such as haustoria. Plants defend themselves against pathogen attacks by using surface and intracellular recognition mechanisms, part of which directly recognize effectors and trigger resistance reactions (Dodds and Rathjen, 2010). Over time these interactions lead to an evolutionary arms race between host and pathogen (Anderson et al., 2010). Therefore, effectors are expected to be amongst the most rapidly evolving genes in pathogen genomes in a recurring strategy to circumvent plant resistance.

Advances in next-generation sequencing technologies are yielding a growing number of sequenced pathogen genomes,

which allow a comprehensive and objective study of the evolutionary arms race between host and pathogen through the analysis of large-scale divergence genomic data. Genes undergoing purifying selection evolve slowly to maintain their conserved function. Weakened purifying selection occurs when a gene can mutate freely and randomly with little penalty because it has little or restricted functional significance. On the other hand, positive natural selection occurs when changes with functional consequences are favored as it produces high variability and adaptability. This is often observed in pathogenicity-related genes of microbes, which must adapt to the changing host environment and avoid immune recognition. The relative rates of synonymous (d_S) and non-synonymous substitutions (d_N) in genes is commonly used to assess selection. A d_N/d_S ratio <1 indicates purifying selection and functional conservation, $d_N/d_S = 1$ is consistent with neutral evolution, and $d_N/d_S > 1$ is indicative of diversifying selection or potential functional divergence.

A diverse range of methods for estimating diversifying selection (counting methods, pairwise d_N/d_S ratios, maximum likelihood codon specific estimates) are available (Aguileta et al., 2009). Calculation of a global d_N/d_S ratio is relatively simple, but has low sensitivity since in many cases diversifying selection acts only on certain sites of the protein domains. Computational diversifying selection analysis estimated by maximum likelihood is a more sophisticated procedure to test for selection pressure on individual codons that can also take into account the lineage of genes within a family (Yang and Nielsen, 2000, 2002). These methods are powerful but they strongly depend on the quality of the multiple sequence alignment, the phylogenetic tree, the level of sequence divergence and the sample size. The software package PAML (Yang, 1997, 2007) implements sophisticated methods for estimating the d_N/d_S ratio and aims to detect positively selected sites in a number of genes by using varying d_N/d_S ratios among sites and Bayes Empirical Bayes (BEB) analysis (Yang et al., 2005). Sites with posterior probability of greater than 95% (or 99%) are inferred as positively selected codons.

Signatures of diversifying selection have been predicted computationally in several filamentous plant pathogen effectors (Ma and Guttman, 2008; Aguileta et al., 2009). In the highly polymorphic phytotoxin-like *scr74* gene family of the oomycete *Phytophthora infestans*, evidence of diversifying selection in the mature protein region was found (Liu et al., 2005). Diversifying selection was also detected in RXLR effector paralogs, acting on the C-terminal regions of the proteins (Win et al., 2007). In the fungal host-specific necrotrophic effector *ToxA*, produced by the wheat pathogens *Pyrenophora tritici-repentis* and *Parastagonospora nodorum*, two codons were predicted to be under diversifying selection using polymorphism data (Stukenbrock and McDonald, 2007). Likewise, four codons were predicted to be under diversifying selection in the necrotrophic effector *Tox1*, produced by *Parastagonospora nodorum* (Liu et al., 2012). Six genes encoding cell wall degrading enzymes in *Zymoseptoria tritici* were also found to undergo diversifying selection in either host adaptation or host evasion processes (Brunner et al., 2013). Plant resistance (*R*) genes control recognition of pathogens carrying specific avirulence (*Avr*) effectors in a gene-for-gene model. For several avirulence proteins, signatures of diversifying selection have been detected. For instance, the *AvrL567*, *AvrP123*, and *AvrP4* genes are highly polymorphic in *Melampsora lini* and show evidence of diversifying selection consistent with an evolutionary arms race (Dodds et al., 2006; Barrett et al., 2009). Signatures of diversifying selection in *Phytophthora* effectors have also been linked to the ability to jump to another host (Dong et al., 2014).

The prediction of pathogenicity-associated proteins such as fungal effectors is an ongoing challenge. The most common technique is to return a set of proteins that have a predicted secretion signal and additionally are small and cysteine-rich, despite increasing awareness in the literature that not all effectors share these features (Ellis et al., 2009; Sperschneider et al., 2013). For example, the flax rust avirulence effectors AvrL567 and AvrM are devoid of cysteines (Dodds et al., 2004; Catanzariti et al., 2006). Furthermore, the number of effector candidates will be overwhelmingly large, making target selection for functional testing in the laboratory difficult. For example, Duplessis et al. (2011) report 1106 proteins in *P. graminis* f. sp. *tritici* that fit the following criteria as candidate effectors: a predicted secretion signal, no transmembrane region, no GPI-anchor site and size smaller than 300 amino acids. These candidate effectors were predicted to form 164 clusters, varying in membership from 2 to 44, and form a striking 10% of all of the expanded families in the stem rust genome. Another bioinformatic pipeline uses Markov and hierarchical clustering to identify protein families of wheat stem rust and poplar leaf rust, and it prioritizes effector candidates using a ranking system based on eight criteria associated with effector properties (Saunders et al., 2012). In both approaches, selection of gene families as candidate effectors strongly depends on the accuracy of the secretion prediction tools and the validity of the thresholds used for small size and cysteine-rich.

Instead of making *a priori* assumptions on effector candidate properties solely on the sequence level, in this work we combine three layers of evidence (taxonomic information, *in planta* upregulation, diversifying selection) to predict effector candidates in a pathogen genome that is highly capable of host adaptation. First, we use comparative information from publicly available fungal genomes to separate the predicted gene families into those that are specific to pathogenic fungi and into those that are found across pathogenic and non-pathogenic fungi. Both pathogen-associated and fungal gene families are then grouped using unsupervised clustering based on a broad set of 35 sequence-derived protein features to find putative effector protein clusters (Sperschneider et al., 2013). If small, cysteine-rich proteins with a secretion signal are a class of proteins that are distinct from the remaining proteins, they can be expected to show up as a cluster with enrichment in those features. To add further evidence for a protein's involvement in pathogenicity, expression data and signatures of diversifying selection are incorporated to prioritize effector candidates. A combination of these three layers of evidence leads to a small set of effector candidates that have a likely role of interacting with the host plant and are well supported candidates for future laboratory testing.

MATERIALS AND METHODS

PREDICTION OF PATHOGEN-ASSOCIATED AND FUNGAL GENE FAMILIES

For the prediction of gene families, we used Tribe-MCL (Enright et al., 2002) with all-vs.-all phmmer bit scores (Finn et al., 2011) on the *P. graminis* f. sp. *tritici* protein set. Both Tribe-MCL and phmmer were run with default parameters. For each member in a gene family, we recorded the phmmer hit distribution to 72 publicly available fungal genomes from the JGI MycoCosm (Grigoriev et al., 2012), supplemented by five genomes of *Fusarium pseudograminearum* and the two genomes *Fusarium acuminatum* and *Fusarium incarnatum—F. equiseti* (Gardiner et al., 2012; Moolhuijzen et al., 2013). This resulted in a set of 78 fungal genomes which includes 24 non-pathogens (saprophytes, non-pathogenic yeasts) and 54 pathogens (pathogens, parasites, ectomycorrhizal, symbionts, mycoparasitic) (see **Table S1** for genome list and classification). For a given protein x_i with its corresponding list of phmmer hits that cover at least 60% of query and target sequence, pathogen precision is calculated as follows:

$$P(i) = 100 \times \frac{\#\text{hits to fungal pathogens}}{\#\text{ hits}}$$

For a gene family, $P(i)$ is calculated by taking the average pathogen precision over its protein members. Gene families with $P(i) > 90\%$ were classified as pathogen-associated.

PREDICTION OF PUTATIVE EFFECTOR CLUSTERS AND EXPRESSION DATA

For each pathogen-associated or fungal gene family, a representative member was chosen and a previously developed protein *k*-means clustering method was applied to the pathogen-associated and fungal gene family representatives to find putative effector protein clusters across the gene families (Sperschneider et al., 2013). Each representative gene family member was assigned a 35-dimensional feature vector based on the average values for the corresponding gene family members. The 35-dimensional feature vector contains the protein length, the *D*-score returned by SIGNALP 4.1 (Petersen et al., 2011), the score for extracellular localization site prediction calculated by WoLF PSORT (Horton et al., 2007), the molecular weight, protein charge, isoelectric point, amino acid composition as well as the classification of amino acid composition (tiny, small, aliphatic, aromatic, non-polar, polar, charged, basic, and acidic). The features were calculated using pepstats from the EMBOSS package (Rice et al., 2000). All of these 35 features were calculated for each gene family member and the average values across the whole gene family were assigned as the feature vector for the gene family representative. The elbow plot method was used to estimate the number of clusters for the *k*-means clustering method, both using SciPy. After clustering, Mann–Whitney *U*-tests using R were conducted for each feature in the 35-dimensional vector to test whether the distribution within a cluster is identical to the full background distribution, i.e. all clusters. Highly significant *p*-values for both directions (lesser and greater, $p < 2.2 \times 10^{-16}$) were recorded. Putative effector clusters were chosen based on elevated levels of expression during infection using stem rust expression data. Stem rust expression data from Duplessis et al. (2011) was accessed at the GEO database (http://www.ncbi.nlm.nih.gov/geo/) as series GSE25020 and the tool GEO2R was used to identify differentially expressed stem rust genes by comparing *in planta* wheat infection to germinated urediniospores. Genes with fold change of at least two were reported as up-regulated during infection. *P*-values were adjusted using the Benjamini & Hochberg method, genes with $p < 0.05$ were reported as significant and genes with $p < 0.00001$ were labeled as highly significant. RNAseq expression data of isolated haustoria was compared to that of germinated urediniospores of wheat stem rust strain 21–0 and transcripts with normalized fold changes >2 were reported as haustorial up-regulated genes (Upadhyaya et al., in press).

DIVERSIFYING SELECTION ANALYSIS

The analysis of diversifying selection was performed for the Pucciniomycotina branch of the fungal kingdom, which includes the following genomes:

- *Puccinia graminis* f. sp. *tritici* (Duplessis et al., 2011).
- *P. triticina* 1-1 BBBD Race 1 [*Puccinia* Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)].
- *P. striiformis* PST-78 [*Puccinia* Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)].
- *Melampsora lini* (Nemri et al., 2014).
- Melampsora laricis-populina (Duplessis et al., 2011).
- *Sporobolomyces roseus* [unpublished data, with permission from the JGI MycoCosm (Grigoriev et al., 2012)].

For each protein in *P. graminis* f. sp. *tritici*, phmmer (Finn et al., 2011) was run against the Pucciniomycotina genomes, and all significant protein hits (*E*-value $< 10^{-5}$) and per-domain hits per protein were recorded. Significant protein hits were kept if the combined domain hits for query and target cover more than 60% of the sequences, respectively. This ensured that for the subsequent diversifying selection analysis only reliable and well-conserved multiple alignments were used. Protein multiple sequence alignments were inferred using PRANK with the +F option (Loytynoja and Goldman, 2005), which has been shown to outperform other alignment methods in diversifying selection analyses (Fletcher and Yang, 2010). Alignment columns with more than 70% gap characters were masked using custom Python scripts for preparation of phylogenetic tree prediction. Phylogenetic trees were calculated using the Phyml package version 20120412 (Guindon et al., 2010). Trees were midpoint rooted and orthologs were derived with the species overlap method using ETE (Huerta-Cepas et al., 2010). Each ortholog set was aligned, gaps were masked and phylogenetic trees were predicted as described above. The gaps in the ortholog protein alignments were used to produce a coding sequence alignment using Pycogent (Knight et al., 2007) for input to PAML. Site-specific diversifying selection on ortholog sets was calculated using PAML, with alignment gaps removed. Two likelihood ratio tests of site-specific diversifying selection were used:

model M1 (neutral) to model M2 (selection) and model M7 (beta) to M8 (beta&ω) and *P*-values < 0.05 were reported as significant.

RESULTS

A PIPELINE FOR PREDICTING EFFECTOR CANDIDATES IN EXPANDED PATHOGEN GENOMES USING MULTIPLE LINES OF EVIDENCE

The wheat stem rust *Puccinia graminis* f. sp. *tritici* has shown the ability to rapidly overcome previously resistant wheat cultivars and, like many other filamentous plant pathogens, features a large genome with expanded gene families that have been linked to host-pathogen interactions (Duplessis et al., 2011; Raffaele and Kamoun, 2012). Adaptive selection and genetic variation are the driving forces behind the evolutionary arms race between host and pathogen, and effector proteins can be expected to be amongst the most rapidly evolving genes in pathogen genomes. Existing effector prediction pipelines that focus on the set of secreted, small, cysteine-rich proteins are prone to return an overwhelming number of candidate effectors and furthermore, not all true effectors will fit these criteria. In the following, we show that a small set of strong effector candidates can be predicted by combining multiple lines of evidence (pathogen-specific, highly expressed during infection, rapidly evolving), without making *a priori* assumptions about their protein properties.

We prioritize effector candidates in the stem rust *P. graminis* f. sp. *tritici* by linking comparative data, patterns of diversifying selection, and expression data during infection to a gene family's likelihood of participation in host-pathogen interactions (**Figure 1**). Using comparative information from 78 fungal genomes (pathogens and non-pathogens, **Table S1**), gene families were divided into those that can be associated with pathogenicity and those that are not related to pathogenicity (with the latter referred to as fungal genes). To dissect the nature of the gene families on a protein similarity level, a gene family representative was randomly chosen and the averages of 35 sequence-derived features for the gene family (e.g., average signal peptide prediction score, average molecular weight, average amino acid composition, see Materials and Methods) were assigned to each representative. This analysis then allowed us to use a previously developed unsupervised *k*-means clustering technique of sequence-derived features (Sperschneider et al., 2013). Enrichment and depletion analysis of the predicted protein clusters could then show whether putative effector clusters with certain characteristic features (e.g. secretion signals, small amino acids and cysteines) were present in an unbiased way, and whether commonalities or differences across pathogen-associated and fungal gene families existed. To include further evidence for pathogenicity function, ortholog sets were analyzed using prediction of their likelihood of undergoing diversifying selection. Pairwise d_N/d_S ratios from entire genes can only indicate diversifying selection if it has acted on all or the majority of the codons, which is rarely the case. Diversifying selection prediction is more powerful using site-specific or branch-specific models, which allow varying d_N/d_S ratios across codons or lineages. Proteins that were up-regulated *in planta* and were observed as part of pathogen-associated gene families predicted to undergo diversifying selection were selected as strong effector candidates for *P. graminis* f. sp. *tritici*.

PATHOGEN-ASSOCIATED AND FUNGAL GENE FAMILIES FORM DISTINCT PROTEIN CLUSTERS

Lineage-specific expanded gene families in *P. graminis* f. sp. *tritici* are thought to drive genome innovation, adaptation and pathogen-associated processes. To investigate gene families that can be associated with pathogenicity, the set of 15,979 *P. graminis*

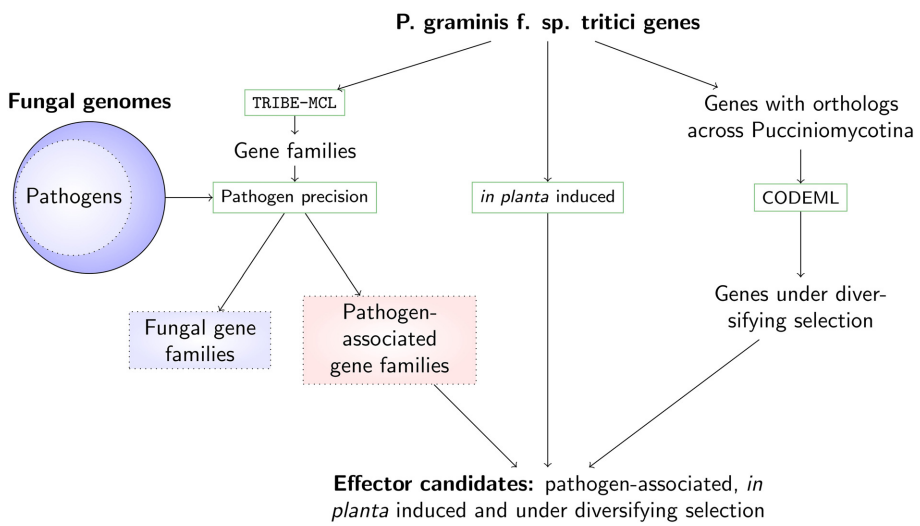


FIGURE 1 | A pipeline for predicting effector candidates in expanded pathogen genomes incorporating comparative analysis, expression data and evolutionary support. The stem rust genes are grouped into gene families using TRIBE-MCL (Enright et al., 2002) and then split into families associated with pathogenicity (including those that are specific to *P. graminis* f. sp. *tritici*) and into fungal families, using comparative information from 78 fungal genomes. To limit the number of putative effector candidates, we use two additional lines of evidence: up-regulation during *in planta* infection and signatures of diversifying selection. Genes that are rapidly evolving, induced during infection and conserved predominantly across pathogenic fungi are returned as prime effector candidates.

f. sp. *tritici* genes was first collapsed into gene families using Tribe-MCL with all-vs.-all phmmer bit scores (Enright et al., 2002; Finn et al., 2011). This resulted in 2351 predicted gene families covering a total of 15,387 genes with an additional 592 single genes. The gene families were then examined regarding their conservation across the fungal kingdom using 78 publicly available fungal genomes (Materials and Methods). The 78 fungal genomes were classified as either non-pathogenic (saprophytes, non-pathogenic yeasts) or pathogenic (pathogens and plant-associated fungi such as parasites, ectomycorrhizal, symbionts, mycoparasitic). This method suggests 1210 pathogen-associated gene families (including those that are specific to *P. graminis* f. sp. *tritici*) and 1141 fungal gene families not associated with pathogenicity.

An ongoing question is whether fungal effector proteins share characteristics on the sequence or structure level. To begin to address this question, an existing *k*-means clustering technique based on 35 sequence-derived protein features was applied (Sperschneider et al., 2013). Briefly, each protein has 35 features ranging from signal peptide prediction score to amino acid composition. Unsupervised *k*-means clustering is used to predict protein groups that share features distinctive from other groups. From this, enrichment or depletion in characteristic features was calculated for each cluster. However, expanded gene families in the stem rust genome are expected to share a high degree of sequence similarity and a clustering step based on sequence-derived features is likely to group existing gene families together instead of looking for functional commonalities between gene families potentially involved in pathogenicity. Therefore, a representative member was chosen for each gene family and was assigned the average 35 sequence-derived feature vector for the corresponding gene family. This methodology avoids clustering according to sequence similarity within the gene families.

For the 1210 pathogen-associated gene families, the *k*-means clustering returns 12 protein clusters, whereas the 1141 fungal gene families are predicted to form 10 clusters (**Table 1**). Enrichment and depletion analysis shows the characteristic features of each cluster and reveals two clusters of secreted proteins, one across the pathogen-associated gene families and one across the fungal gene families (**Table 1**, C6 and C7). The pathogen-associated secreted cluster C6 is enriched in tiny and non-polar amino acids as well as cysteines and glycines. The enrichment in cysteines and small size are features that are commonly associated with fungal effector proteins. Interestingly, only 66.5% of proteins in cluster C6 are predicted to contain a signal peptide by SignalP 4.1 (Petersen et al., 2011), suggesting that they may be

Table 1 | Properties for the clusters of pathogen-associated and fungal gene families are shown.

Cluster	# of gene families	# of proteins	Enrichment ↑	Depletion ↓
PATHOGEN-ASSOCIATED GENE FAMILIES				
C1	131	1012	–	Protein charge, basic
C2	88	751	Polar, charged, acidic, D, E	Protein charge
C3	53	324	Protein charge, basic, K	–
C4	72	342	Molecular weight	–
C5	110	533	R	Molecular weight, acidic, N, D
C6	117	871	Secretion, extracellular, tiny, non-polar, C, G	Molecular weight, charged, acidic, R, E
C7	155	1033	Polar, S	Aliphatic, V
C8	70	389	Tiny, small	Aromatic, charged, basic
C9	73	307	Aliphatic, non-polar, I	Charged, acidic, D
C10	18	112	–	–
C11	153	1709	Aromatic, charged, acidic, E, I, K, F	Tiny, small, A, P, S, T
C12	170	1468	–	–
FUNGAL GENE FAMILIES				
C1	131	1045	–	Tiny, small, P, S
C2	70	416	Molecular weight	–
C3	183	1113	Aromatic, H, W	–
C4	151	1031	Small, polar, P, S	Aliphatic, I, V
C5	103	493	Polar, charged, basic, D, E, K	Aliphatic, aromatic
C6	135	683	–	Protein charge
C7	73	451	Secretion, extracellular, tiny, small, G	Charged, basic, R, E
C8	135	573	Aliphatic, non-polar, A, G, V	–
C9	100	513	Aliphatic, aromatic, non-polar, I, L, F	Charged, basic, acidic, D, E, Q, K
C10	60	218	Protein charge, charged, basic, R, K	–

For each characteristic in the 35-dimensional feature vector, Mann–Whitney U-tests were used to test whether the distribution within a cluster is identical to the full background distribution for all clusters and highly significant p-values for both directions (lesser ↓ and greater ↑) are shown. Secretion refers to the predicted Signal P score and extracellular score to the WoLF PSORT score. The following amino acid memberships are used: tiny (A, C, G, S, T), small (A, C, D, G, N, P, S, T, V), aliphatic (A, I, L, V), aromatic (F, H, W, Y), polar (D, E, H, K, N, Q, R, S, T), charged (D, E, H, K, R), basic (H, K, R), and acidic (D, E).

clustered according to unifying features that go beyond the signal peptide prediction score. The fungal gene families also contain a cluster with enrichment in secretion signals and tiny and small amino acids as well as glycines (**Table 1**, C7), however without enrichment in cysteines.

We searched the Pfam database for significant hits (E-value $< 10^{-5}$) to all proteins in the secreted clusters. The pathogen-associated secreted cluster C6 has significant Pfam domain hits only for 53 of its 871 proteins (6.1%), which is expected as the vast majority of fungal effector proteins are known to lack functional annotation. On the other hand, the fungal secreted cluster C7 could be confidently annotated with Pfam domain hits for 327 of its 451 proteins (72.5%). **Table 2** shows the most frequent Pfam domain hits for the two clusters. The fungal secreted cluster C7 contains a large number of proteins involved in phospholipase activity (PF01735), glycosyl hydrolase activity (PF00704 and PF00150), proteolysis (PF00026) and superoxide dismutase activity (PF00080). The pathogen-associated secreted protein cluster C6 only has a few Pfam domain hits, which does not allow for functional annotation.

Taken together, this analysis suggests that there are distinct groups of secreted proteins that can be separated by means of comparative genomics. We predicted a cluster of fungal gene families with an enriched secretion signal that contains proteins with putative enzymatic roles related to plant biopolymers. In contrast, across the pathogen-associated gene families, we predicted a cluster with enrichment in secretion signals, tiny and non-polar amino acids as well as cysteines and glycines. The vast majority of these proteins lack functional annotation. This further supports the hypothesis that secreted, small, cysteine-rich proteins play a role specifically in the pathogenicity of *P. graminis* f. sp. *tritici*.

SMALL, CYSTEINE-RICH PROTEINS IN PATHOGEN-ASSOCIATED GENE FAMILIES SHOW ELEVATED LEVELS OF UP-REGULATION DURING INFECTION

To investigate whether differences in expression levels can be found between fungal and pathogen-associated gene families, the number of up-regulated genes was calculated for each cluster given in **Table 1**. We used expression data from Duplessis et al. (2011) to identify differentially expressed stem rust genes by comparing *in planta* wheat infection to germinated urediniospores. Genes which have differential expression with $p < 0.05$ are reported as significant and those with $p < 0.00001$ as highly significant, both requiring a fold change >2.

At significance threshold $p < 0.05$, up-regulation was found for 1476 of 8851 (16.6%) genes in the pathogen-associated gene family clusters, whereas up-regulation was detected for 1295 of 6536 (19.8%) genes in the fungal gene family clusters (**Figure 2**). A striking difference in the distribution of up-regulated stem

Table 2 | The most frequent Pfam domain hits are shown for the two clusters with an enriched secretion signal across the pathogen-associated and fungal gene families.

Cluster	Pfam domains	Pfam ID	# of proteins	InterPro description
PATHOGEN-ASSOCIATED C6				
	Triglyceride lipase	PF01764	4	Triglyceride lipases are lipolytic enzymes that hydrolyse ester linkages of triglycerides
	Copper/zinc superoxide dismutase	PF00080	3	Metalloproteins that prevent damage by oxygen-mediated free radicals
	Dioxygenase	PF00775	2	Dioxygenases catalyze the incorporation of both atoms of molecular oxygen into substrates using a variety of reaction mechanisms
	Glycosyl hydrolases family 43	PF04616	2	Widespread group of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety
FUNGAL C7				
	Lysophospholipase	PF01735	18	Phospholipase activity
	Glycosyl hydrolases family 18	PF00704	17	Widespread group of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates, or between a carbohydrate and a non-carbohydrate moiety
	Cellulase (glycosyl hydrolase family 5)	PF00150	16	Degradation of cellulose and xylans
	Eukaryotic aspartyl protease	PF00026	16	Aspartic-type endopeptidase activity, proteolysis
	Copper/zinc superoxide dismutase	PF00080	16	Metalloproteins that prevent damage by oxygen-mediated free radicals

For the two clusters, a Pfam search was performed for all members. The proteins in the pathogen-associated cluster C6 predominantly lack functional annotation, whereas the fungal secreted cluster C7 has Pfam domain hits for the majority of its proteins. For each cluster, the top five Pfam domain hits are shown if they have at least two members.

rust genes with high significance ($p < 0.00001$) can be observed across the clusters (**Figure 3**). The pathogen-associated, secreted, cysteine-rich cluster C6 contains the highest number of genes that are up-regulated. However, we do observe highly significant up-regulation during infection also across the fungal gene families. This can be expected as the stem rust activates a diverse number of proteins during infection and not all of these will act as an effector or have a function directly related to pathogenesis. For example, haustoria play a major role in nutrient uptake from the host during infection and show high expression of sugar and amino acid transporters (Garnica et al., 2013). Furthermore, not all fungal effector proteins can be expected to be part of the small, cysteine-rich cluster C6. Additional pathogen-associated clusters show potential of containing effector candidates, with elevated levels of highly significant up-regulation during infection in clusters C1, C7, and C11 (**Table 1**, **Figure 3**). 9.7, 4.7, and 14.4% of proteins in clusters C1, C7, and C11, respectively, are predicted to be secreted by SignalP 4.1 and could thus contain putative effector candidates. Therefore, signatures of diversifying selection are used in the following as an additional line of evidence for pathogenicity function.

IN PLANTA INDUCED GENES UNDERGOING DIVERSIFYING SELECTION ARE PREDOMINANTLY FOUND AMONGST PATHOGEN-ASSOCIATED GENE FAMILIES

Because of the potential and pressure for co-evolution with the host, fungal pathogen effectors and avirulence genes may undergo accelerated rates of diversification and are thus likely to show signs of site-specific diversifying selection. In order to have sufficient evolutionary divergence, diversifying selection analysis was performed for the publicly available genomes of the Pucciniomycotina branch of the fungal kingdom (see Materials and Methods). In total, 10,213 stem rust genes with at least two orthologs across the Pucciniomycotina branch of the fungal kingdom were analyzed for signatures of site-specific diversifying selection. 4752 of these genes belong to the pathogen-associated gene families, 5177 belong to the fungal gene families and 284 genes are not part of gene families. In total, 387 of the 10,213 genes (3.8%) were found to show site-specific diversifying selection using two likelihood ratio tests of CODEML in the PAML software. Despite the higher number of genes in the fungal gene families, site-specific diversifying selection was predominantly detected across pathogen-associated gene families (**Table 3**). In total, 341 of the 387 (88.1%) rapidly evolving genes are part of

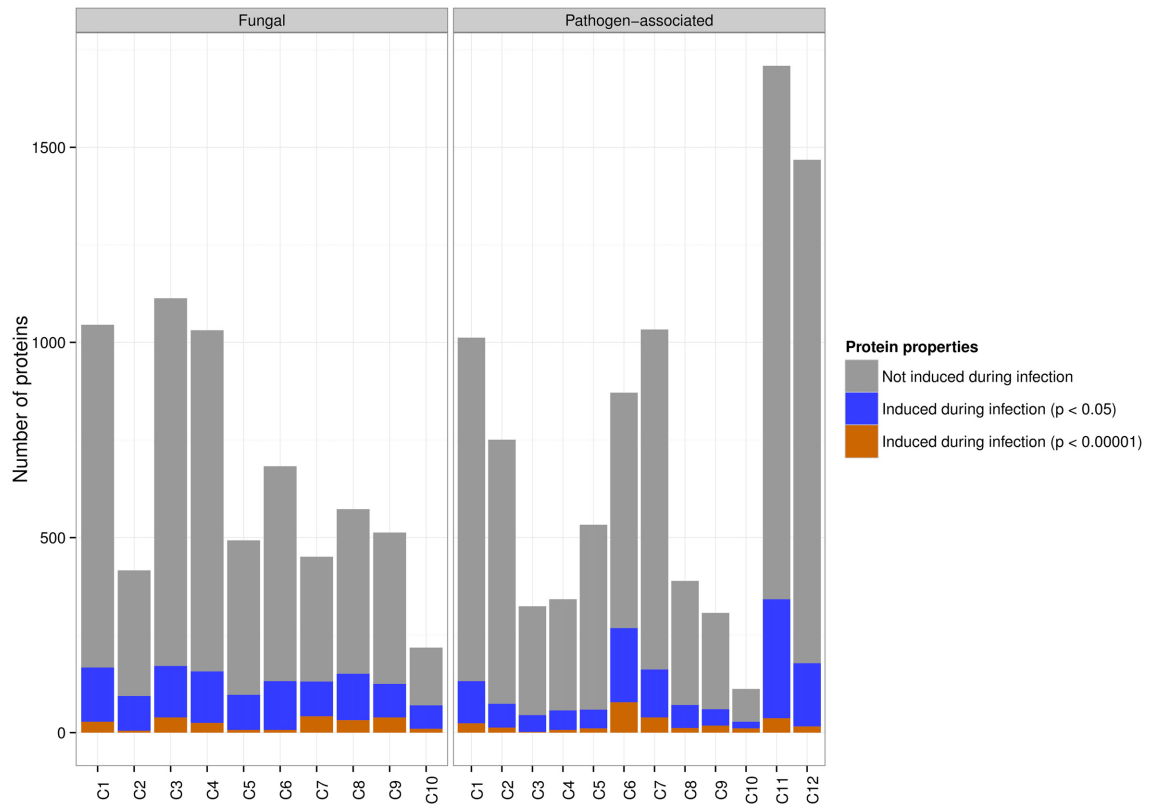


FIGURE 2 | Both fungal and pathogen-associated wheat stem rust gene families are *in planta* induced. The majority of pathogen-associated and fungal genes are not up-regulated during infection. At a significance threshold of $p < 0.05$, 19.8% of genes from fungal gene families are up-regulated during infection and are distributed across all clusters. 16.6% of genes from pathogen-associated gene families are up-regulated during infection. At a significance threshold of $p < 0.00001$, the highest number of genes up-regulated during infection is found in the pathogen-associated cluster C6 (secreted, cysteine-rich).

pathogen-associated gene families, whereas only 43 are part of fungal gene families.

Despite the fairly similar distribution of *in planta* up-regulated genes across the pathogen-associated and fungal gene families, site-specific diversifying selection using PAML was almost exclusively detected across pathogen-associated gene families.

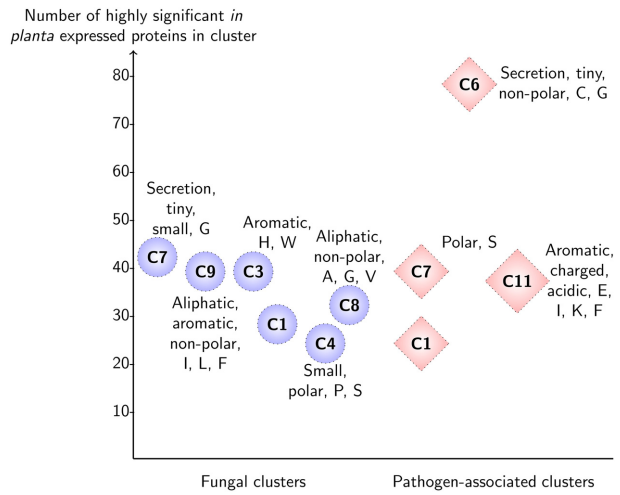


FIGURE 3 | Highly significant *in planta* expression for predicted clusters of the fungal gene families and pathogen-associated gene families in wheat stem rust. For each cluster, the enriched sequence-derived features and number of highly significant up-regulated genes are shown ($p < 0.00001$). Note that for clarity, only clusters which have more than 20 up-regulated proteins are shown. The pathogen-associated cluster C6 sits at the top with 78 proteins that are up-regulated with high significance and has enrichment in features that are associated with effector proteins (secreted, small size, cysteine-rich).

Table 3 | Diversifying selection in the wheat stem rust is predominantly detected across pathogen-associated gene families.

	Additional criteria	# of genes	# of genes under diversifying selection
P. graminis f. sp. *tritici* genes with at least two orthologs	–	10,213	387 (3.8%)
	Member of pathogen-associated gene family	4752	**341 (7.2%)**
	Member of fungal gene family	5177	43 (0.8%)
	Not a member of gene family	284	3 (1.1%)

10,213 P. graminis f. sp. tritici genes with at least two orthologs were analyzed for site-specific diversifying selection using two likelihood ratio tests of CODEML. The majority of rapidly evolving genes are part of pathogen-associated gene families.

In particular, 81 up-regulated genes ($p < 0.05$) from pathogen-associated gene families were predicted to undergo diversifying selection whereas only 11 up-regulated genes from fungal gene families are rapidly evolving. A more stringent significance threshold of $p < 0.00001$ returned 14 genes from pathogen-associated gene families and two genes from fungal gene families that are undergoing diversifying selection (**Figure 4**).

Only one of the up-regulated genes ($p < 0.05$) in the fungal secreted cluster C7 is predicted to undergo site-specific diversifying selection, which indicates that these secreted proteins might have roles that do not require adaptation to the host or the environment. In contrast, the pathogen-associated secreted, cysteine-rich cluster C6 contains 15 up-regulated genes that are predicted to be undergoing diversifying selection. However, other pathogen-associated clusters such as C11 also contain a high number of rapidly evolving genes that are up-regulated during infection (**Figure 4**), which underlines the need to look beyond secreted, small and cysteine-rich as criteria for effector candidates.

Taken together, these results suggest that diversifying selection in *P. graminis* f. sp. *tritici* occurs predominantly across gene families that can be associated with pathogenicity via comparative genomics. However, signatures of diversifying selection cannot exclusively be linked to host adaptation or a role in pathogenicity. Therefore, a combination of pathogen-association, signatures of diversifying selection and expression data will be used to identify a small set of highly likely effector candidates.

PATHOGEN-ASSOCIATION, RAPID EVOLUTION AND UP-REGULATION DURING *IN PLANTA* INFECTION AND IN HAUSTORIA DEFINES PRIME EFFECTOR CANDIDATES

Prime fungal effector candidates can be expected to be highly expressed during infection and to be experiencing diversifying selection, reflecting the evolutionary arms race between host and pathogen. When searching for signatures of adaptive evolution in proteins that are highly induced during infection, we found site-specific diversifying selection almost exclusively in those that are pathogen-associated. To further prioritize the effector candidates, we used haustorial expression data of germinated urediniospores versus isolated haustoria of wheat stem rust strain 21-0 (RNAseq, Upadhyaya et al., in press). Forty six genes were found to be under diversifying selection, up-regulated during *in planta* infection and to have a fold change >2 when comparing expression in haustoria to germinated urediniospores (**Table 4**). Forty two of these 46 candidates are part of pathogen-associated gene families and are thus strong candidates for having a role in pathogenicity.

A common feature of fungal effectors is that they currently lack functional annotation. Using the motif search tool MEME (Bailey et al., 2009), we could not detect a common sequence motif for the 46 proteins under diversifying selection. Only four of the 46 genes could be functionally annotated using Pfam searches. PGTG_17076 has a Pfam hit to seed maturation proteins (PF04927), PGTG_05197 to sugar transporter proteins (PF00083), PGTG_07078 to ribosomal proteins (PF00338) and PGTG_17927 to the alpha-kinase family (PF02816). Only three of the 46 proteins under diversifying selection have orthologies outside the concentrated group of *P. graminis* f. sp. *tritici*,

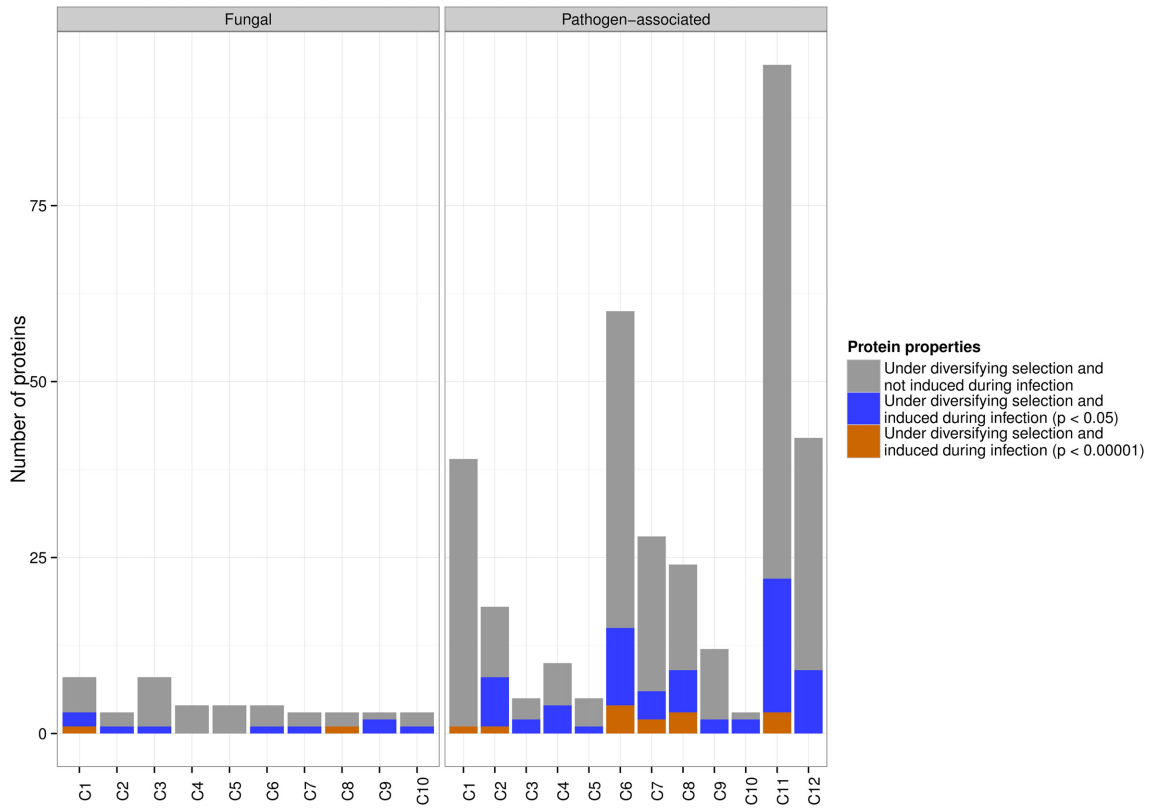


FIGURE 4 | The distribution of pathogen-associated and fungal genes with *in planta* up-regulation and diversifying selection across the clusters. Diversifying selection is predominantly predicted for genes that are part of pathogen-associated gene families. At a significance threshold of $p < 0.05$, 81 pathogen-associated genes are rapidly evolving, compared to only 11 fungal genes. At a significance threshold of $p < 0.00001$, 14 rapidly evolving genes are predicted across the pathogen-associated gene families, with the majority of these being part of clusters C6, C7, C8 and C11.

P. striiformis, and *P. triticina*. The remaining proteins under diversifying selection share orthology with the wheat pathogens stripe rust and leaf rust, but not with other rust fungi. Similarly, all of the known rust Avr proteins in *M. lini* occur in families that are restricted to this species or its relative *M. laricipopulina* and are not shared across other rust genera (Nemri et al., 2014). Thus, the observed genus level specificity of these candidates is consistent with their suspected roles in host-pathogen interactions.

In particular, PGTG_08638, PGTG_16225, PGTG_04972, PGTG_14091, PGTG_09276, and PGTG_05592 are highly induced in haustoria (fold change > 100, **Table 4**). Indeed, one variant of protein PGTG_08638 was identified in a functional screen of candidate effectors from the stem rust fungus *P. graminis* f. sp. *tritici* as an inducer of host genotype-specific hypersensitive cell death in wheat (Upadhyaya et al., 2013). This is in agreement with our PAML analysis, which detects diversifying selection only for the three orthologs (stem rust PGTG_08638, stripe rust PSTG_14557, leaf rust PTTG 06270), but not for the other variants (**Figure 5**). From this data, one could speculate that a gene duplication event with successive diversification has led to a gain of function in the stem rust variant PGTG_08638. This protein is recognized inside wheat cells (Upadhyaya et al., 2013) and would therefore belong to a class of effectors that are delivered into host cells from haustoria. The other haustorially expressed gene candidates described here are likely also to be in this class.

By combining comparative information, expression data and signatures of diversifying selection, we have dramatically reduced the number of effector candidates reported in earlier studies (Duplessis et al., 2011; Saunders et al., 2012) to a small set of strong effector candidates. The computational prediction of an effector candidate in this set (PGTG_08638) that has been shown to induce host genotype-specific hypersensitive cell death in wheat (Upadhyaya et al., 2013) emphasizes that this set of 46 effector candidates could be the basis of a priority list for laboratory verification. Furthermore, protein PGTG_08638 is likely to have been missed by effector prediction pipelines that select candidates based on a secretion signal, small size and high number of cysteines, as it only contains one cysteine residue in its sequence of 292 amino acids. Indeed, in the work by Saunders et al. (2012), PGTG_08638 was predicted to be in a tribe that was ranked below the cut-off score and was thus dismissed as an effector candidate. Our results also confirm the presence of adaptive diversification of wheat stem rust gene families involved in pathogenicity and will lead to further insight into how this successful pathogen overcomes host resistance.

Table 4 | *P. graminis* f. sp. *tritici* genes that are predicted to undergo site-specific diversifying selection with significant *in planta* up-regulation (wheat infection to germinated urediniospores, $p < 0.05$, fold change > 2) and a fold change > 2 when comparing isolated haustoria to germinated urediniospores.

Pathogen-associated cluster	Protein	Secretion signal	aas	Cys	Rust orthologs	PFAM domain	Wheat/ Spores (FC)	Haustoria/ Spores (FC)
C11	PGTG_08638	Yes	292	1	Stripe, leaf	–	187.1	1085.9
Fungal	PGTG_16225	Yes	310	7	Stripe, leaf	–	43.8	239
C11	PGTG_04972	–	340	–	Stripe, leaf	–	21.8	147.7
No gene family	PGTG_14091	Yes	502	1	Stripe, leaf	–	25.8	145.5
C12	PGTG_09276	–	134	–	Stripe	–	4.7	142.4
C6	PGTG_05592	Yes	257	10	Stripe	–	48.2	118.7
C1	PGTG_05174	–	239	1	Stripe, leaf, flax	–	32.6	81.2
C6	PGTG_11727	Yes	336	11	Stripe	–	9	56
C2	PGTG_06244	Yes	277	1	Leaf	–	4.2	53.7
Fungal	PGTG_17076	–	67	–	Stripe, leaf	Seed maturation protein (PF04927, 2.9e-07)	118.8	50.8
C6	PGTG_03859	Yes	165	11	Stripe, leaf	–	19.8	44.8
C11	PGTG_16303	Yes	441	4	Stripe	–	3.1	31.6
C8	PGTG_10538	Yes	411	–	Stripe, leaf	–	31.9	31.6
C8	PGTG_10539	Yes	392	2	Leaf	–	13.9	30.1
C10	PGTG_09318	Yes	87	8	Stripe, leaf	–	77	19.2
C11	PGTG_00341	–	361	2	Stripe, leaf	–	4.9	16.6
C2	PGTG_10398	–	1011	3	Stripe, leaf	–	22.8	13.5
C11	PGTG_14389	–	298	1	Stripe	–	3.9	11.4
C6	PGTG_17308	Yes	125	7	Stripe	–	5.9	8.3
C4	PGTG_01642	–	1329	2	Stripe, leaf	–	2.7	7.3
C11	PGTG_10056	Yes	766	12	Stripe, leaf	–	37.1	7.3
C7	PGTG_03213	–	565	–	Stripe, leaf	–	25.6	7.2
C11	PGTG_15791	–	376	4	Stripe	–	2.7	7.2
C8	PGTG_13414	Yes	197	2	Stripe, leaf	–	45.8	6.6
C4	PGTG_01631	Yes	807	5	Stripe, leaf	–	7.5	5.6
C11	PGTG_15481	Yes	502	3	Stripe, leaf	–	2.3	5.6
C11	PGTG_17733	Yes	482	–	Leaf	–	9.5	5.3
C8	PGTG_10625	Yes	474	3	Stripe	–	5	5.3
C11	PGTG_18622	–	300	2	Stripe, leaf	–	2.5	4.9
C2	PGTG_07786	–	854	3	Stripe, leaf	–	2.6	4.4
C11	PGTG_12173	–	629	4	Stripe, leaf	–	6.1	3.9
C2	PGTG_07911	Yes	324	4	Stripe, leaf	–	7	3.9
C6	PGTG_16750	Yes	124	8	Leaf	–	4.1	3.8
C6	PGTG_04109	Yes	100	9	Leaf	–	5.1	3.8
C6	PGTG_05119	Yes	141	6	Stripe, leaf	–	61.3	3.7
Fungal	PGTG_05197	–	537	9	Leaf	Sugar (and other) transporter (PF00083, 3e-76)	5.7	3.7
C7	PGTG_07078	–	324	–	Stripe, leaf, flax	Ribosomal protein S10p/S20e (PF00338, 1.2e-16)	2.6	3.2
C12	PGTG_17001	–	878	13	Stripe, leaf	–	2.6	3
C11	PGTG_15702	–	334	15	Stripe, leaf, flax, poplar	–	2.3	2.8
C3	PGTG_06359	–	734	13	Stripe, leaf	–	6.2	2.7
C6	PGTG_17534	Yes	114	4	Stripe	–	18.1	2.4
C11	PGTG_14388	Yes	353	1	Leaf	–	2.5	2.2
C9	PGTG_19205	–	299	5	Leaf	–	9.7	2.1
C9	PGTG_14673	Yes	189	10	Stripe	–	6.4	2.1
C11	PGTG_15389	–	667	2	Leaf	–	2.2	2.1
C8	PGTG_10832	–	149	3	Leaf	–	2	2
C12	PGTG_17927	–	711	9	Stripe	Alpha-kinase family (PF02816,1.4e-32)	5.5	2

For each protein, its signal peptide prediction by Signal P 4.1, protein sequence length (aas), number of cysteines (Cys) and distribution of wheat stem rust orthologs across the Pucciniomycotina genomes are given. Stripe stands for P. striiformis PST-78, leaf for P. triticina 1-1 BBBD Race 1, flax for M. lini and poplar for M. laricis-populina. Note that no proteins undergoing diversifying selection were found in pathogen-associated cluster C5. Protein expression levels are given for the two experiments (FC: fold change). The genes are ordered by decreasing haustorial differential expression fold change.

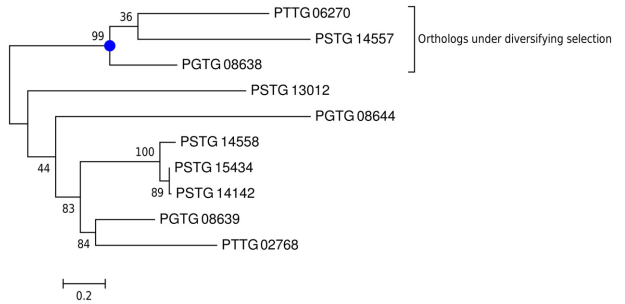


FIGURE 5 | For the gene family of protein PGTG_08638, diversifying selection is only detected for the orthologs. A phylogenetic tree for the gene family of PGTG_08638 was predicted using PhyML and branch support values are shown. PAML detects site-specific diversifying selection only for the branch with the three orthologs (stem rust PGTG 08638, stripe rust PSTG_14557, leaf rust PTTG 06270), but not for the other variants PGTG_08639 and PGTG_08644 and their orthologs.

DISCUSSION

Understanding how filamentous fungal plant pathogens cause disease is of utmost importance due to the devastating losses they cause in important crop plants. In particular, the wheat stem rust *Puccinia graminis* f. sp. *tritici* has demonstrated the ability to rapidly overcome resistant host lines and is currently threatening the majority of the global wheat cultivars (Singh et al., 2011). The prediction of genes involved in host-pathogen interactions, in particular effector proteins that interact directly with target host defense proteins, has thus far been difficult for fungal pathogens due to a lack of signature sequence motifs. Previous efforts to understand virulence largely relied on the prediction of secreted, small and cysteine-rich proteins as candidate effectors (Ellis et al., 2009), which is likely to return an overwhelming number of candidates for experimental verification. Furthermore, there is growing evidence that effectors can have unconventional characteristics, such as no predicted signal peptide, a low number of cysteine residues or a large size (Sperschneider et al., 2013).

In this work, we introduced an alternative bioinformatics pipeline that prioritizes effector candidates by combining multiple lines of evidence such as taxonomic information, expression data and evolutionary signatures of diversifying selection. This approach was applied to the wheat stem rust *P. graminis* f. sp. *tritici*, which features lineage-specific expanded gene families that are thought to drive genome innovation, adaptation and pathogenicity-associated processes (Duplessis et al., 2011; Raffaele and Kamoun, 2012). Unsupervised clustering of the predicted gene families based on sequence-derived protein features revealed that pathogen-associated and fungal gene families form distinct clusters and that elevated levels of up-regulation during infection can be found across both classes. Our observation is that expression data is likely to be the line of evidence with the greatest detection power but probably the least discriminatory as a large number of genes can be expected to be up-regulated during infection that do not necessarily correspond to effectors. The ability of expression data to capture effectors highly depends on choosing the right time points and on the power of the experimental setup to mimic infection under field conditions. In pathogens where haustorial tissue can be extracted such as the wheat stem rust fungus, expression data is likely to be a powerful line of evidence. In other pathogens there can be issues with capturing early time points for effector expression and thus, the benefits of using expression data may be of lesser value. In contrast, *in planta* induced genes undergoing diversifying selection are predominantly found amongst pathogen-associated gene families. This indicates that pathogen-associated gene families in *P. graminis* f. sp. *tritici* might be involved in host adaptation or pathogen specialization. The diversifying selection analysis is likely to be highly discriminatory, but may miss effector candidates where the selection signal is not strong. This approach clearly becomes more powerful the more species are available for a certain lineage.

In particular, this analysis revealed a set of 42 effector candidates in *P. graminis* f. sp. *tritici* that are part of pathogen-associated gene families, up-regulated during infection and at the same time rapidly evolving in a suspected evolutionary arms race with the host (**Table 4**). Furthermore, we find that the majority of these effector candidates are part of pathogen-associated clusters that lack enrichment in features commonly associated with fungal effectors such as small size and high number of cysteines. Recently, one of our predicted effector candidates was shown to have a variant that induces host genotype-specific hypersensitive cell death in wheat in a functional screen of candidate effectors from the stem rust fungus *P. graminis* f. sp. *tritici* (Upadhyaya et al., 2013). Therefore, we are confident that this small set of effector candidates should be the future target for laboratory experiments.

We propose that there is a need to expand the criteria for predicting effectors beyond secreted, small and cysteine-rich. Instead, a combination of the three signals (taxonomic information, expression data and diversifying selection) can be a powerful and unbiased predictor for genes involved in host-pathogen interactions. Our diversifying selection pipeline can be applied to a wide range of pathogens for which divergence data is available and can give insight into the evolutionary processes between host and pathogen. In particular, our work indicates that elements of the genome linked to pathogenicity are evolutionarily dynamic, possibly in mechanisms relating to host adaptation or pathogen specialization. Therefore, effector candidates in plant pathogens that show signs of rapid evolution are promising targets for disease control in crop plants.

ACKNOWLEDGMENTS

We thank Dr. James Hane, Dr. Adnane Nemri and Dr. Lars Jermiin for providing helpful comments on earlier drafts of this manuscript. Jana Sperschneider was supported by the CSIRO Transformational Biology Capability platform. Donald M. Gardiner was partially supported by the Australian Grains Research and Development Corporation. Narayana M. Upadhyaya and Peter N. Dodds thank the Two Blades Foundation for financial support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/Journal/10.3389/fpls.2014.00372/abstract

Table S1 | The list of fungal genomes used for comparative classification of gene families into pathogen-associated and fungal.

REFERENCES

Aguileta, G., Refregier, G., Yockteng, R., Fournier, E., and Giraud, T. (2009). Rapidly evolving genes in pathogens: methods for detecting positive selection and examples among fungi, bacteria, viruses and protists. *Infect. Genet. Evol.* 9, 656–670. doi: 10.1016/j.meegid.2009.03.010

Anderson, J. P., Gleason, C. A., Foley, R. C., Thrall, P. H., Burdon, J. B., and Singh, K. B. (2010). Plants versus pathogens: an evolutionary arms race. *Funct. Plant Biol.* 37, 499–512. doi: 10.1071/FP09304

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., et al. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 37, W202–W208. doi: 10.1093/nar/gkp335

Barrett, L. G., Thrall, P. H., Dodds, P. N., Van Der Merwe, M., Linde, C. C., Lawrence, G. J., et al. (2009). Diversity and evolution of effector loci in natural populations of the plant pathogen *Melampsora lini*. *Mol. Biol. Evol.* 26, 2499–2513. doi: 10.1093/molbev/msp166

Brunner, P. C., Torriani, S. F., Croll, D., Stukenbrock, E. H., and McDonald, B. A. (2013). Coevolution and life cycle specialization of plant cell wall degrading enzymes in a hemibiotrophic pathogen. *Mol. Biol. Evol.* 30, 1337–1347. doi: 10.1093/molbev/mst041

Catanzariti, A. M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini* AvrL567 avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell* 16, 755–768. doi: 10.1105/tpc.020040

Dodds, P. N., Lawrence, G. J., Catanzariti, A. M., Teh, T., Wang, C. I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8888–8893. doi: 10.1073/pnas.0602577103

Dodds, P. N., and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nat. Rev. Genet.* 11, 539–548. doi: 10.1038/nrg2812

Dong, S., Stam, R., Cano, L. M., Song, J., Sklenar, J., Yoshida, K., et al. (2014). Effector specialization in a lineage of the Irish potato famine pathogen. *Science* 343, 552–555. doi: 10.1126/science.1246300

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Ellis, J. G., Rafiqi, M., Gan, P., Chakrabarti, A., and Dodds, P. N. (2009). Recent progress in discovery and functional analysis of effector proteins of fungal and oomycete plant pathogens. *Curr. Opin. Plant Biol.* 12, 399–405. doi: 10.1016/j.pbi.2009.05.004

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584. doi: 10.1093/nar/30.7.1575

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* 39, W29–W37. doi: 10.1093/nar/gkr367

Fletcher, W., and Yang, Z. (2010). The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* 27, 2257–2267. doi: 10.1093/molbev/msq115

Gardiner, D. M., McDonald, M. C., Covarelli, L., Solomon, P. S., Rusu, A. G., Marshall, M., et al. (2012). Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathog.* 8:e1002952. doi: 10.1371/journal.ppat.1002952

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PLoS ONE* 8:e67150. doi: 10.1371/journal.pone.0067150

Giraldo, M. C., and Valent, B. (2013). Filamentous plant pathogen effectors in action. *Nat. Rev. Microbiol.* 11, 800–814. doi: 10.1038/nrmicro3119

Grigoriev, I. V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., et al. (2012). The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* 40, D26–D32. doi: 10.1093/nar/gkr947

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321. doi: 10.1093/sysbio/syq010

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Huerta-Cepas, J., Dopazo, J., and Gabaldon, T. (2010). ETE: a python environment for tree exploration. *BMC Bioinformatics* 11:24. doi: 10.1186/1471-2105-11-24

Kämper, J., Kahmann, R., Bolker, M., Ma, L. J., Brefort, T., Saville, B. J., et al. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* 444, 97–101. doi: 10.1038/nature05248

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., et al. (2007). PyCogent: a toolkit for making sense from sequence. *Genome Biol.* 8, R171. doi: 10.1186/gb-2007-8-8-r171

Koeck, M., Hardham, A. R., and Dodds, P. N. (2011). The role of effectors of biotrophic and hemibiotrophic fungi in infection. *Cell Microbiol.* 13, 1849–1857. doi: 10.1111/j.1462-5822.2011.01665.x

Leonard, K. J., and Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Mol. Plant Pathol.* 6, 99–111. doi: 10.1111/j.1364-3703.2005.00273.x

Liu, Z., Bos, J. I., Armstrong, M., Whisson, S. C., Da Cunha, L., Torto-Alalibo, T., et al. (2005). Patterns of diversifying selection in the phytotoxin-like scr74 gene family of *Phytophthora infestans*. *Mol. Biol. Evol.* 22, 659–672. doi: 10.1093/molbev/msi049

Liu, Z., Zhang, Z., Faris, J. D., Oliver, R. P., Syme, R., McDonald, M. C., et al. (2012). The cysteine rich necrotrophic effector SnTox1 produced by *Stagonospora nodorum* triggers susceptibility of wheat lines harboring Snn1. *PLoS Pathog.* 8:e1002467. doi: 10.1371/journal.ppat.1002467

Loytynoja, A., and Goldman, N. (2005). An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U.S.A.* 102, 10557–10562. doi: 10.1073/pnas.0409137102

Ma, W., and Guttman, D. S. (2008). Evolution of prokaryotic and eukaryotic virulence effectors. *Curr. Opin. Plant Biol.* 11, 412–419. doi: 10.1016/j.pbi.2008.05.001

Moolhuijzen, P. M., Manners, J. M., Wilcox, S. A., Bellgard, M. I., and Gardiner, D. M. (2013). Genome sequences of six wheat-infecting fusarium species isolates. *Genome Announc.* 1, e00670–e00713. doi: 10.1128/genomeA.00670-13

Nemri, A., Saunders, D. G., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Pardey, P. G., Beddow, J. M., Kriticos, D. J., Hurley, T. M., Park, R. F., Duveiller, E., et al. (2013). Agriculture. Right-sizing stem-rust research. *Science* 340, 147–148. doi: 10.1126/science.122970

Petersen, T. N., Brunak, S., Von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10, 417–430. doi: 10.1038/nrmicro2790

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the european molecular biology open software suite. *Trends Genet.* 16, 276–277. doi: 10.1016/S0168-9525(00)02024-2

Saunders, D. G., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Raffaele, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:e29847. doi: 10.1371/journal.pone.0029847

Singh, R. P., Hodson, D. P., Huerta-Espino, J., Jin, Y., Bhavani, S., Njau, P., et al. (2011). The emergence of Ug99 races of the stem rust fungus is a threat to world wheat production. *Annu. Rev. Phytopathol.* 49, 465–481. doi: 10.1146/annurev-phyto-072910-095423

Sperschneider, J., Gardiner, D. M., Taylor, J. M., Hane, J. K., Singh, K. B., and Manners, J. M. (2013). A comparative hidden Markov model analysis pipeline identifies proteins characteristic of cereal-infecting fungi. *BMC Genomics* 14:807. doi: 10.1186/1471-2164-14-807

Stukenbrock, E. H., and McDonald, B. A. (2007). Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. *Mol. Plant Pathol.* 8, 321–332. doi: 10.1111/j.1364-3703.2007.00396.x

Upadhyaya, N. M., Garnica, D. P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., et al. (in press). Comparative genomics of Australian isolates of the wheat stem rust pathogen Puccinia graminis f. sp tritici reveals extensive polymorphism in candidate effector genes. *Front. Plant Sci.*

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M., Ellis, J., and Dodds, P. (2013). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Mol. Plant Microbe Interact* 27, 255–264. doi: 10.1094/MPMI-07-13-0187-FI

Win, J., Morgan, W., Bos, J., Krasileva, K. V., Cano, L. M., Chaparro-Garcia, A., et al. (2007). Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *Plant Cell* 19, 2349–2369. doi: 10.1105/tpc.107.051037

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Yang, Z., and Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43. doi: 10.1093/oxfordjournals.molbev.a026236

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917. doi: 10.1093/oxfordjournals.molbev.a004148

Yang, Z., Wong, W. S., and Nielsen, R. (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* 22, 1107–1118. doi: 10.1093/molbev/msi097

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 29 May 2014; accepted: 11 July 2014; published online: 01 September 2014.
*Citation: Sperschneider J, Ying H, Dodds PN, Gardiner DM, Upadhyaya NM, Singh KB, Manners JM and Taylor JM (2014) Diversifying selection in the wheat stem rust fungus acts predominantly on pathogen-associated gene families and reveals candidate effectors. Front. Plant Sci. **5**:372. doi: 10.3389/fpls.2014.00372*
This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.
Copyright © 2014 Sperschneider, Ying, Dodds, Gardiner, Upadhyaya, Singh, Manners and Taylor. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read, for greatest visibility

COLLABORATIVE PEER-REVIEW
Designed to be rigorous – yet also collaborative, fair and constructive

FAST PUBLICATION
Average 85 days from submission to publication (across all journals)

COPYRIGHT TO AUTHORS
No limit to article distribution and re-use

TRANSPARENT
Editors and reviewers acknowledged by name on published articles

SUPPORT
By our Swiss-based editorial team

IMPACT METRICS
Advanced metrics track your article's impact

GLOBAL SPREAD
5'100'000+ monthly article views and downloads

LOOP RESEARCH NETWORK
Our network increases readership for your article

Frontiers
EPFL Innovation Park, Building I • 1015 Lausanne • Switzerland
Tel +41 21 510 17 00 • Fax +41 21 510 17 01 • info@frontiersin.org
www.frontiersin.org

Find us on