frontiers in Plant Science

OPINION
published: 17 May 2017
doi: 10.3389/fpls.2017.00826

How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes

Andrea Minio[1], *Jerry Lin*[1], *Brandon S. Gaut*[2] *and Dario Cantu*[1]*

[1] *Department of Viticulture and Enology, University of California, Davis, Davis, CA, United States,* [2] *Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, United States*

OPEN ACCESS

Edited by:
José Tomás Matus, Centre for Research in Agricultural Genomics, Spain

Reviewed by:
Jordi Garcia-Mas, Institute for Research and Technology in Food and Agriculture, Spain
Michela Troggio, Fondazione Edmund Mach, Italy

***Correspondence:**
Dario Cantu
dacantu@ucdavis.edu

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

Received: *04 April 2017*
Accepted: *02 May 2017*
Published: *17 May 2017*

Citation:
Minio A, Lin J, Gaut BS and Cantu D (2017) How Single Molecule Real-Time Sequencing and Haplotype Phasing Have Enabled Reference-Grade Diploid Genome Assembly of Wine Grapes. Front. Plant Sci. 8:826. doi: 10.3389/fpls.2017.00826

Keywords: heterozygosity, inbreeding depression, cabernet sauvignon, comparative genomics, grape pan-genome

HIGH HETEROZYGOSITY IS A CHALLENGE FOR GRAPE GENOME ASSEMBLY

Domesticated grapevines (*Vitis vinifera*) have relatively small genomes of about 500 Mb (Lodhi and Reisch, 1995; Jaillon et al., 2007; Velasco et al., 2007), which is similar to other small-genomes species like rice (430 Mb; Goff et al., 2002), medicago (500 Mb; Tang et al., 2014), and poplar (465 Mb; Tuskan et al., 2006). Despite their small genome size, the sequencing and assembling of grapevine genomes is difficult because of high levels of heterozygosity. The high heterozygosity in domesticated grapes may be due, in part, to their domestication from an obligately outcrossing, dioecious wild progenitor. Domesticated grapes can be selfed, in theory, because their mating system transitioned to hermaphroditic, self-fertile flowers during domestication. In practice, however, selfed progeny tend to be non-viable, presumably due to a high deleterious recessive load and resulting inbreeding depression. As a consequence of these fitness effects, most grape cultivars are crosses between distantly related parents (Strefeler et al., 1992; Ohmi et al., 1993; Bowers and Meredith, 1997; Sefc et al., 1998; Lopes et al., 1999; Di Gaspero et al., 2005; Tapia et al., 2007; Ibáñez et al., 2009; Cipriani et al., 2010; Myles et al., 2011; Lacombe et al., 2013).

One such cultivar is Cabernet Sauvignon, one of the most widely cultivated wine grape cultivars. Cabernet Sauvignon was produced from a cross between Sauvignon Blanc and Cabernet Franc sometime before the seventeenth century in the Aquitaine region of France (Bowers and Meredith, 1997). Whether a spontaneous hybrid or a product of human breeding, all of the Cabernet Sauvignon grown around the world is thought to have resulted from this single hybridization event. Just as the parents of Cabernet Sauvignon have been identified, the genetic origin of many other important wine grape cultivars is known, and they often originate from the direct crossing of common, distantly-related cultivars (Strefeler et al., 1992; Ohmi et al., 1993; Qu et al., 1996; Bowers and Meredith, 1997; Sefc et al., 1998; Lopes et al., 1999; Crespan and Milani, 2001; Vouillamoz et al., 2003, 2004; Di Gaspero et al., 2005; Vouillamoz and Grando, 2006; Lacombe et al., 2007, 2013; Tapia et al., 2007; Boursiquot et al., 2009; Ibáñez et al., 2009; Cipriani et al., 2010; Myles et al., 2011;

García-Muñoz et al., 2012). Due to this intraspecific hybridization process, levels of heterozygosity in grape cultivars can easily exceed 11% (Jaillon et al., 2007).

High heterozygosity is challenging for genome assembly, because heterozygous genomes typically produce more fragmented sequences than haploid or homozygous genomes of similar size and complexity (Yu et al., 2005; Argout et al., 2011; The Tomato Genome Consortium, 2012). The goal of standard assembly approaches is to collapse homologous regions with sufficient similarity into haploid consensus sequences, but divergent haplotypes in heterozygous regions typically result in multiple, difficult to resolve assembly paths which must then be assembled separately. Additionally, the boundaries between haploid consensus contigs and heterozygous regions cannot be resolved with a unique path; as a result they are left unlinked, which breaks assembly contiguity (**Figure 1A**). Altogether, elevated heterozygosity increases fragmentation and inflates the size of the total assembly, potentially doubling the genome size if the majority of the two homologous genomes are assembled separately (Huang et al., 2012; Li et al., 2012; Safonova et al., 2015). Fragmentation and retention of redundant regions can also lead to inaccurate gene models, apparent paralogous genes and duplicated blocks, incorrect gene copy number, and synteny breaks.

INITIAL ATTEMPTS TO SEQUENCE THE GRAPE GENOME

Despite the challenges in assembling heterozygous genomes, the commercial and cultural importance of the grapevine has led to several sequencing attempts. Two genome reference drafts for the common grapevine were released in 2007 (Jaillon et al., 2007; Velasco et al., 2007). Remarkably, these were the first genomes of any fruiting crop to be sequenced and only the fourth for flowering plants. These reference genomes, both of which utilized the Pinot Noir cultivar, were assembled using different approaches to address heterozygosity. The first genome by Jaillon et al. reduced heterozygosity by inbreeding a line of Pinot Noir (var. PN40024) to ∼7% heterozygosity (Jaillon et al., 2007). To produce the second genome, Velasco et al. sequenced a Pinot Noir clone (ENTAV115) directly then assembled contigs that represented separate homologous chromosomes (Velasco et al., 2007). Unsurprisingly, these early efforts are poor by current standards. The PN40024 genome had ∼8.4-fold coverage and was assembled into 19,577 contigs with a contig N_{50} of only 65.9 kbp. Later sequencing increased coverage to up to 12x and the contig N_{50} of the PN40024 genome to 102.7 kb (**Figure 1B**). The ENTAV115 genome used both Sanger paired-reads and 454 sequencing to achieve a total coverage of ∼4.2x. Although riddled with gaps and potentially omitting large regions of repetitive sequences where genes could be located, the two genomes provided valuable insights into grape genomes. Together they revealed that the Pinot Noir genome features: (i) ∼30,000 protein-coding genes, comparable to Arabidopsis but about 75% of rice and poplar; (ii) a high proportion of repetitive elements comprising an estimated ∼40% of the genome; (iii) complex patterns of gene duplications consistent with one or more paleopolypoidy events; (iv) expansion of gene families that influence the organoleptic properties of the berry; (v) a typical number (∼200) of NBS-LRR genes, which often function in disease resistance, and (vi) a standard complement of genes involved in disease signaling pathways. Despite its limitations, the PN40024 genome assembly has proven to be invaluable to the grape research community. Cited in over 2,000 articles, it has served as a reference in more than 3,000 genome-wide transcriptional analyses.

Following the publication of the PN40024 genome in 2007, no genome reference of equivalent or greater quality has been released for *V. vinifera*. Only a handful of studies have attempted to use *bona fide* genome-wide approaches to measure diversity within the species (Giannuzzi et al., 2011; Da Silva et al., 2013; Di Genova et al., 2014; Cardone et al., 2016). With the advent of second generation short read sequencing, attempts were made to perform *de novo* assembly and reference based resequencing of grape cultivars. These attempts failed to provide a high quality representation of the sequenced grape genotypes. A *de novo* approach was adopted to assemble the genome sequence of Thompson Seedless, a ubiquitous multipurpose cultivar. Despite an enormous sequencing depth (327x), the short fragment size did not permit resolution of repetitive regions, resulting in an extremely fragmented assembly (Di Genova et al., 2014; **Figure 1B**). For the wine grape cultivar Tannat (Da Silva et al., 2013), the authors applied a reference based assembly approach, which had proved to be effective in assembling multiple Arabidopsis genotypes (Gan et al., 2011). However, reference-based assembly failed to reconstruct genotype specific sequences with Tannat data, demonstrating that large scale resequencing initiatives like the 1,000 Human Genome project (Auton et al., 2015) and the 1,001 Arabidopsis Genomes project (Alonso-Blanco et al., 2016) would not succeed for *Vitis*. In fact, while the approach supported variant calling with *de novo* assembly to resolve regions highly divergent in sequence between Tannat and PN40024, it was unable to recover regions absent in the reference but present in Tannat. Consequently, over 10% of the gene space was not represented in the assembly, illustrating that the genomic sequence of one cultivar is insufficient for representing the total variability of the species. To improve representation of the *V. vinifera* pan-genome and encompass the variability of the species, we need the complete *de novo* assembled genomes of additional genotypes. Moreover, as grape cultivars are intraspecific hybrids of different genotypes, assembly of each genome should include a diploid representation of the genome to preserve information about the characteristics of each haplotype.

RECENT DEVELOPMENTS IN GRAPE GENOME SEQUENCING

Single Molecule Real Time (SMRT) DNA sequencing (Pacific Biosciences) has emerged as a leading technology for characterizing complex structural variations, supporting and refining the assembly of complex genomes in hybrid fashion or alone for reconstructing highly continuous assemblies of both

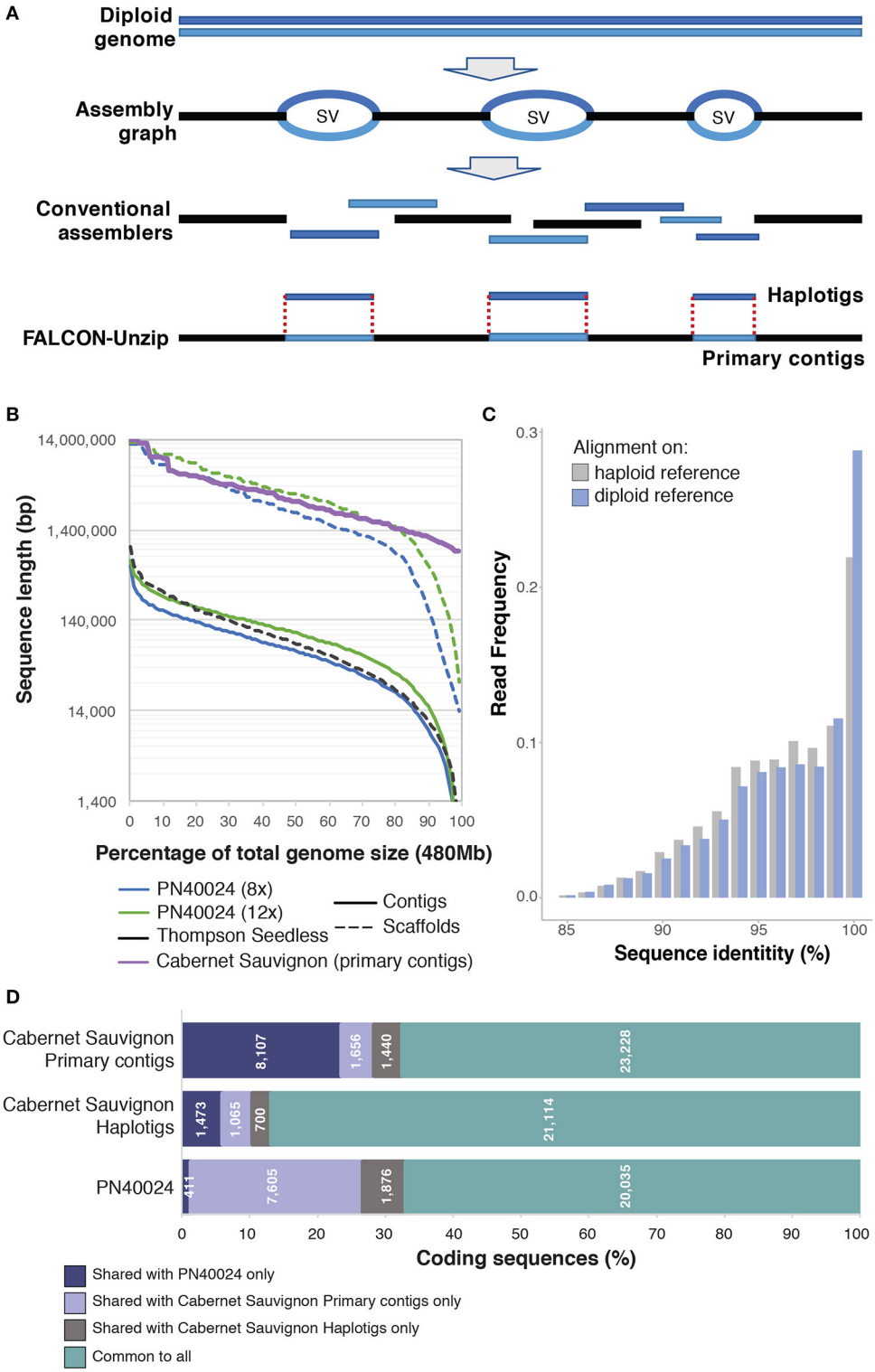

FIGURE 1 | Comparison of conventional assemblers and their application to the grapevine genome using FALCON-Unzip and its results on Cabernet Sauvignon. (A) A diagram comparing how conventional assemblers and FALCON-Unzip resolve homozygous and heterozygous regions of diploid genomes. **(B)** Comparison of sequence length distribution between the primary contigs of Cabernet Sauvignon assembled with FALCON-Unzip and other *Vitis vinifera* genome contig and scaffold assemblies. **(C)** Distribution of sequence identity between RNAseq reads and reference when mapping is done only on primary contigs or on a combination of primary contigs and haplotigs. **(D)** Shared coding genes sequences between Cabernet Sauvignon haplotypes and PN40024. Predicted coding

(Continued)

FIGURE 1 | Continued
sequences from the Cabernet Sauvignon primary contigs were aligned using GMAP (Wu and Watanabe, 2005) to the Cabernet Sauvignon haplotigs and the PN40024 chromosomes to identify the shared part of the represented gene space. Only alignments with identity ≥80% and coverage ≥66% were considered. In similar fashion, coding sequences from the Cabernet Sauvignon haplotigs were aligned against the primary contigs and the PN40024 chromosomes, and coding sequences from PN40024 were aligned against both primary contigs and haplotigs of Cabernet Sauvignon.

small and highly repetitive genomes (Chin et al., 2013; Doi et al., 2014; Huddleston et al., 2014, 2016; Gordon et al., 2016; Ricker et al., 2016; Seo et al., 2016; Vij et al., 2016). The advantage of SMRT technology arises from the delivery of long reads, currently averaging over 30 kbp and potentially approaching 100 kbp. In addition to facilitating assembly of more contiguous genomes, long reads carry the necessary information to phase haplotypes over multiple kilobase distances. The open-source software, FALCON-unzip (Chin et al., 2016), was developed specifically to utilize the long reads generated using SMRT sequencing technology and assemble diploid genomes into highly contiguous and correctly phased diploid genomes. The algorithm first constructs a string graph composed of "haploid consensus" contigs together with bubbles representing structural variant sites between homologous loci. Sequenced reads are then phased and separated for each haplotype on the basis of heterozygous positions. Phased reads are finally used to assemble the backbone sequence (primary contigs) and the alternative haplotype sequences (haplotigs) (**Figure 1A**). The combination of primary contigs and haplotigs constitute the final diploid assembly with phased single-nucleotide polymorphisms and structural variants between the two haplotypes.

We have recently reported the assembly using SMRT technology and FALCON-unzip of the highly heterozygous diploid genome of Cabernet Sauvignon (Chin et al., 2016), one of the most widely cultivated wine grape cultivars. As it is the progeny of Cabernet Franc and Sauvignon Blanc, two cultivars with extremely divergent phenotypical traits, reconstructing the diploid structure of Cabernet Sauvignon is necessary for identifying the alleles inherited from the parent cultivars. We sequenced the Cabernet Sauvignon genome with a coverage depth of ∼140x using SMRT sequencing technology. Sequencing reads were then assembled using FALCON-unzip into a highly contiguous genome that integrated phased haplotype information. FALCON-unzip generated a set of primary contigs (591.4 Mbp in 718 contigs with $N_{50} = 2.17$ Mbp, **Figure 1B**) that covers one of the two haplotypes, and a set of correlated haplotigs (367.8 Mbp in 2,037 contigs with $N_{50} = 0.80$ Mbp). The assembled sequences exceed PN40024 contigs and Thompson Seedless scaffolds by nearly two orders of magnitude in size (**Figure 1B**), ranking this assembly not only as the best *V. vinifera* genome assembly but also among the highest quality plant genomes published to date, including other genomes sequenced with SMRT technology (Sakai et al., 2015; VanBuren et al., 2015; Jiao et al., 2016; The UC Davis Coffee Genome Project, 2017). Symptomatic of the extreme divergence in allele sequences in *Vitis*, the length of the primary assembly was inflated with respect to the expected genome size, illustrating one of the challenges of sequencing highly heterozygous genomes (Chin et al., 2016). After manual removal of un-phased haplotigs, the primary assembly is an ideal candidate for scaffolding or hybrid assembly with optical maps to produce a genome assembly of even higher quality.

Preliminary gene model prediction identified over 34,000 protein coding sequences on the primary assembly of the Cabernet Sauvignon genome and nearly 24,000 on the haplotigs (Chin et al., 2016). Just a few hundred of PN40024 annotated coding genes did not find any suitable alignment on the Cabernet Sauvignon assembly (411 genes; identity ≥80% and coverage ≥66%), but nearly 4,900 Cabernet Sauvignon loci could not be found on the PN40024 genome (**Figure 1D**). These results are in accordance with other studies that reported presence/absence polymorphisms of gene models between wine grape cultivars (Da Silva et al., 2013; Venturini et al., 2013; Jiao et al., 2015), but the high number of genes not found in PN40024 likely reflects its incompleteness. Moreover, nearly 2,100 coding sequences identified in the Cabernet Sauvignon haplotigs were not found on the primary assembly (**Figure 1D**). While limited by the preliminary status of the annotation, these observations point to a high degree of structural variation between homologous chromosomes. Moreover, these structural variations are likely to have functional consequences since they encompass coding sequences. The variability between haplotypes may also impact and potentially confound the analysis of RNAseq data. In the worst case, the expression of haplotype-specific loci that are not represented on the reference genome would be assigned to the most similar genomic region of the reference, which is likely to generate expression mismeasurement artifacts. As shown in **Figure 1C**, in the presence of a diploid reference (primary contigs plus haplotigs), about 10% more RNAseq reads map at ≥99% identity. This observation suggests that when both alleles are represented in the reference reads align to their respective haplotype; RNAseq can therefore be used to determine allelic specific gene expression.

CONCLUSIONS

Genome resequencing projects of both prokaryotic and eukaryotic organisms have clearly shown that one genome sequence is insufficient to properly describe the genetic characteristics of a species (Tettelin et al., 2005; Donati et al., 2010). In order to grasp comprehensive genetic variability and complete gene pools in outcrossing species, such as grape, we also need to go beyond the generation of haploid consensus sequences and focus our efforts to begin assembling diploid genome sequences with phased haplotypes. As discussed in this article, long read sequences and bioinformatic tools that take advantage of them have solved a critical bottleneck in

grape genomics. As long-range scaffolding technologies, such as those based on proximity ligation–based methods like Hi-C (Putnam et al., 2016) or optical maps (Hastie et al., 2013; Yoon et al., 2016) are optimized for highly heterozygous plant genomes, we expect that reference-grade genome references will quickly become available for many grape species and cultivars of interest. This genomic information will allow us to identify core sequences that are common to all cultivars, as well as dispensable sequences comprising partially shared and non-shared genes that contribute to inter-cultivar phenotypic variation. This genomic information will also enable the identification of the genetic bases of economically important traits to accelerate the breeding of new cultivars and rootstocks.

AUTHOR CONTRIBUTIONS

DC and AM conceived the article. Figure was prepared by AM and DC. AM, JL, BG, and DC wrote the first draft of the manuscript. DC revised and finalized.

ACKNOWLEDGMENTS

The genome sequencing of Cabernet Sauvignon in the Cantu lab is supported by J. Lohr Vineyards and Wines and by E. & J. Gallo Winery. Part of this work is carried out in collaboration with UC Davis Chile and funded by the Chilean Economic Development Agency (CORFO).

REFERENCES

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M. M., et al. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. doi: 10.1016/j.cell.2016.05.063

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., et al. (2011). The genome of *Theobroma cacao*. *Nat. Genet.* 43, 101–108. doi: 10.1038/ng.736

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., et al. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. doi: 10.1038/nature15393

Boursiquot, J. M., Lacombe, T., Laucou, V., Julliard, S., Perrin, F. X., Lanier, N., et al. (2009). Parentage of merlot and related winegrape cultivars of southwestern france: discovery of the missing link. *Aust. J. Grape Wine Res.* 15, 144–155. doi: 10.1111/j.1755-0238.2008.00041.x

Bowers, J. E., and Meredith, C. P. (1997). The parentage of a classic wine grape, Cabernet Sauvignon. *Nat. Genet.* 16, 84–87. doi: 10.1038/ng0597-84

Cardone, M. F., D'Addabbo, P., Alkan, C., Bergamini, C., Catacchio, C. R., Anaclerio, F., et al. (2016). Inter-varietal structural variation in grapevine genomes. *Plant J.* 88, 648–661. doi: 10.1111/tpj.13274

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi: 10.1038/nmeth.4035

Cipriani, G., Spadotto, A., Jurman, I., Di Gaspero, G., Crespan, M., Meneghetti, S., et al. (2010). The SSR-based molecular profile of 1005 grapevine (*Vitis vinifera* L.) accessions uncovers new synonymy and parentages, and reveals a large admixture amongst varieties of different geographic origin. *Theor. Appl. Genet.* 121, 1569–1585. doi: 10.1007/s00122-010-1411-9

Crespan, M., and Milani, N. (2001). The Muscats: a molecular analysis of synonyms, homonyms and genetic relationships within a large family of grapevine cultivars. *Vitis* 40, 23–30.

Da Silva, C., Zamperin, G., Ferrarini, A., Minio, A., Dal Molin, A., Venturini, L., et al. (2013). The high polyphenol content of grapevine cultivar tannat berries is conferred primarily by genes that are not shared with the reference genome. *Plant Cell* 25, 4777–4788. doi: 10.1105/tpc.113.118810

Di Gaspero, G., Cipriani, G., Marrazzo, M. T., Andreetta, D., Prado Castro, M. J., Peterlunger, E., et al. (2005). Isolation of (AC)n-microsatellites in *Vitis vinifera* L. and analysis of genetic background in grapevines under marker assisted selection. *Mol. Breed.* 15, 11–20. doi: 10.1007/s11032-004-1362-4

Di Genova, A., Almeida, A. M., Muñoz-Espinoza, C., Vizoso, P., Travisany, D., Moraga, C., et al. (2014). Whole genome comparison between table and wine grapes reveals a comprehensive catalog of structural variants. *BMC Plant Biol.* 14:7. doi: 10.1186/1471-2229-14-7

Doi, K., Monjo, T., Hoang, P. H., Yoshimura, J., Yurino, H., Mitsui, J., et al. (2014). Rapid detection of expanded short tandem repeats in personal genomics using hybrid sequencing. *Bioinformatics* 30, 815–822. doi: 10.1093/bioinformatics/btt647

Donati, C., Hiller, N. L., Tettelin, H., Muzzi, A., Croucher, N. J., Angiuoli, S. V., et al. (2010). Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol.* 11:R107. doi: 10.1186/gb-2010-11-10-r107

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* 477, 419–423. doi: 10.1038/nature10414

García-Muñoz, S., Lacombe, T., de Andrés, M. T., Gaforio, L., Muñoz-Organero, G., Laucou, V., et al. (2012). Grape varieties (*Vitis vinifera* L.) from the Balearic Islands: genetic characterization and relationship with Iberian Peninsula and Mediterranean Basin. *Genet. Resour. Crop Evol.* 59, 589–605. doi: 10.1007/s10722-011-9706-5

Giannuzzi, G., Addabbo, P. D., Gasparro, M., Martinelli, M., Carelli, F. N., Antonacci, D., et al. (2011). Analysis of high-identity segmental duplications in the grapevine genome. *BMC Genomics.* 12:436. doi: 10.1186/1471-2164-12-436

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

Gordon, D., Huddleston, J., Chaisson, M. J., Hill, C. M., Kronenberg, Z. N., Munson, K. M., et al. (2016). Long-read sequence assembly of the gorilla genome. *Science* 352:aae0344. doi: 10.1126/science.aae0344

Hastie, A. R., Dong, L., Smith, A., Finklestein, J., Lam, E. T., Huo, N., et al. (2013). Rapid genome mapping in nanochannel arrays for highly complete and accurate de novo sequence assembly of the complex Aegilops tauschii genome. *PLoS ONE* 8:e55864. doi: 10.1371/journal.pone.0055864

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., et al. (2012). HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* 22, 1581–1588. doi: 10.1101/gr.133652.111

Huddleston, J., Chaisson, M. J., Meltz Steinberg, K., Warren, W., Hoekzema, K., Gordon, D. S., et al. (2016). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* 27, 677–685. doi: 10.1101/gr.214007.116.

Huddleston, J., Ranade, S., Malig, M., Antonacci, F., Chaisson, M., Hon, L., et al. (2014). Reconstructing complex regions of genomes using long-read sequencing technology. *Genome Res.* 24, 688–696. doi: 10.1101/gr.168450.113

Ibáñez, J., Vargas, A. M., Palancar, M., Borrego, J., and De Andrés, M. T. (2009). Genetic relationships among table-grape varieties. *Am. J. Enol. Vitic.* 60, 35–42.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi: 10.1038/nature06148

Jiao, C., Gao, M., Wang, X., and Fei, Z. (2015). Transcriptome characterization of three wild Chinese Vitis uncovers a large number of distinct disease related genes. *BMC Genomics* 16:223. doi: 10.1186/s12864-015-1442-3

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Campbell, M. S., et al. (2016). The complex sequence landscape of maize revealed by single molecule technologies. *bioRxiv* 73, 1–19. doi : 10.1101/079004

Lacombe, T., Boursiquot, J. M., Laucou, V., Dechesne, F., Varès, D., and This, P. (2007). Relationships and genetic diversity within the accessions related to malvasia held in the Domaine de Vassal grape germplasm repository. *Am. J. Enol. Vitic.* 58, 124–131.

Lacombe, T., Boursiquot, J. M., Laucou, V., Di Vecchi-Staraz, M., Péros, J. P., and This, P. (2013). Large-scale parentage analysis in an extended set of grapevine cultivars (*Vitis vinifera* L.). *Theor. Appl. Genet.* 126, 401–414. doi: 10.1007/s00122-012-1988-2

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., et al. (2012). Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief. Funct. Genomics* 11, 25–37. doi: 10.1093/bfgp/elr035

Lodhi, M. A., and Reisch, B. I. (1995). Nuclear DNA content of Vitis species, cultivars, and other genera of the Vitaceae. *Theor. Appl. Genet.* 90, 11–16. doi: 10.1007/BF00220990

Lopes, M. S., Sefc, K. M., Eiras Dias, E., Steinkellner, H., Laimer Câmara Machado, M., and Câmra Machado, A. (1999). The use of microsatellites for germplasm management in a Portuguese grapevine collection. *Theor. Appl. Genet.* 99, 733–739. doi: 10.1007/s001220051291

Myles, S., Boyko, A. R., Owens, C. L., Brown, P. J., Grassi, F., Aradhya, M. K., et al. (2011). Genetic structure and domestication history of the grape. *Proc. Natl. Acad. Sci. U.S.A.* 108, 3530–3535. doi: 10.1073/pnas.1009363108

Ohmi, C., Wakana, A., Shiraishi, S., and Alexandria, M. (1993). Study of the parentage of grape cultivars by genetic interpretation of GPI-2 and PGM-2 isozymes. *Euphytica* 65, 195–202. doi: 10.1007/BF00023083

Putnam, N. H., Connell, B. O., Stites, J. C., Rice, B. J., Hartley, P. D., Sugnet, C. W., et al. (2016). Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* 26, 342–350. doi: 10.1101/gr.193474.115

Qu, X., Lu, J., Lamikanra, O., Science, V., and Florida, A. (1996). Genetic diversity in muscadine and american bunch grapes based on randomly amplified polymorphic DNA (RAPD) analysis. *J. Am. Soc. Hort. Sci.* 121, 1020–1023.

Ricker, N., Shen, S. Y., Goordial, J., Jin, S., and Fulthorpe, R. R. (2016). PacBio SMRT assembly of a complex multi-replicon genome reveals chlorocatechol degradative operon in a region of genome plasticity. *Gene* 586, 239–247. doi: 10.1016/j.gene.2016.04.018

Safonova, Y., Bankevich, A., and Pevzner, P. A. (2015). dipSPAdes: assembler for highly polymorphic diploid genomes. *J. Comput. Biol. A J. Comput. Mol. Cell Biol.* 22, 528–545. doi: 10.1089/cmb.2014.0153

Sakai, H., Naito, K., Ogiso-Tanaka, E., Takahashi, Y., Iseki, K., Muto, C., et al. (2015). The power of single molecule real-time sequencing technology in the de novo assembly of a eukaryotic genome. *Sci. Rep.* 5:16780. doi: 10.1038/srep16780

Sefc, K. M., Steinkellner, H., Glössl, J., Kampfer, S., and Regner, F. (1998). Reconstruction of a grapevine pedigree by microsatellite analysis. *Theor. Appl. Genet.* 97, 227–231. doi: 10.1007/s001220050889

Seo, J., Rhie, A., Kim, J., Lee, S., Sohn, M., Kim, C.-U., et al. (2016). De novo assembly and phasing of a Korean human genome. *Nature* 538, 243–247. doi: 10.1038/nature20098

Strefeler, M. S., Weeden, N. F., and Reisch, B. I. (1992). Inheritance of chloroplast DNA in two full-sib Vitis populations. *Vitis* 31, 183–187.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014). An improved genome release (version Mt4.0) for the model legume Medicago truncatula. *BMC Genomics* 15:312. doi: 10.1186/1471-2164-15-312

Tapia, A. M., Cabezas, J. A., Cabello, F., Lacombe, T., Martínez-Zapater, J. M., Hinrichsen, P., et al. (2007). Determining the Spanish origin of representative ancient American grapevine varieties. *Am. J. Enol. Vitic.* 58, 242–251.

The Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–41. doi: 10.1038/nature11119

The UC Davis Coffee Genome Project (2017). *Coffea arabica UCDv0.5 (Coffee bean)–Phytozome.jgi.doe.gov*. Available online at: https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Carabica_er

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.* 102, 13950–13955. doi: 10.1073/pnas.0506758102

Tuskan, G., Difazio, A., Jansson, S., Bohlmann, S., Grigoriev, J., Hellsten, I., et al. (2006). The genome of black cottonwood, populus trichocarpa (torr. and Gray). *Science* 313, 1596–1604. doi: 10.1126/science.1128691

VanBuren, R., Bryant, D., Edger, P. P., Tang, H., Burgess, D., Challabathula, D., et al. (2015). Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature* 527, 508–511. doi: 10.1038/nature15714

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., et al. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* 2:e1326. doi: 10.1371/journal.pone.0001326

Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G. B., Fasoli, M., Santo, S. D., et al. (2013). De novo transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. *BMC Genomics* 14:41. doi: 10.1186/1471-2164-14-41

Vij, S., Kuhl, H., Kuznetsova, I. S., Komissarov, A., Yurchenko, A. A., Van Heusden, P., et al. (2016). Chromosomal-level assembly of the asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS Genet.* 12:e1005954. doi: 10.1371/journal.pgen.1005954

Vouillamoz, J. F., and Grando, M. S. (2006). Genealogy of wine grape cultivars: "Pinot" is related to "Syrah." *Heredity (Edinb).* 97, 102–110. doi: 10.1038/sj.hdy.6800842

Vouillamoz, J. F., Maigre, D., and Meredith, C. P. (2004). Identity and parentage of two alpine grape cultivars from Switzerland (*Vitis vinifera* L. Lafnetscha and Himbertscha). *Vitis - J. Grapevine Res.* 43, 81–87.

Vouillamoz, J., Maigre, D., and Meredith, C. P. (2003). Microsatellite analysis of ancient alpine grape cultivars: pedigree reconstruction of *Vitis vinifera* L. "Cornalin du Valais." *Theor. Appl. Genet.* 107, 448–454. doi: 10.1007/s00122-003-1265-5

Wu, T. D., and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859–1875. doi: 10.1093/bioinformatics/bti310

Yoon, S., Kim, S. Y., and Nam, D. (2016). Improving gene-set enrichment analysis of RNA-Seq data with small replicates. *PLoS ONE* 11, 1–16. doi: 10.1371/journal.pone.0165919

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., et al. (2005). The genomes of Oryza sativa: a history of duplications. *PLoS Biol.* 3:e38. doi: 10.1371/journal.pbio.0030038

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.