frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 11 September 2014
doi: 10.3389/fpls.2014.00461

Meta-analysis of metabolome QTLs in *Arabidopsis*: trying to estimate the network size controlling genetic variation of the metabolome

Bindu Joseph[1], Susanna Atwell[1], Jason A. Corwin[1], Baohua Li[1] and Daniel J. Kliebenstein[1,2] *

[1] Department of Plant Sciences, University of California Davis, Davis, CA, USA
[2] DynaMo Center of Excellence, University of Copenhagen, Frederiksberg, Denmark

Edited by:
Zoran Nikoloski, Max Planck Institute of Molecular Plant Physiology, Germany

Reviewed by:
Jose M. Jimenez-Gomez, Max Planck Institute for Plant Breeding, Germany
Sabrina Kleessen, Max Planck Institute of Molecular Plant Physiology, Germany

***Correspondence:**
Daniel J. Kliebenstein, Department of Plant Sciences, University of California Davis, One Shields Avenue, Davis, CA 95616, USA
e-mail: kliebenstein@ucdavis.edu

A central goal of systems biology is to develop models that are both predictive and accurately describe the biological system. One complexity to this endeavor is that it is possible to develop models that appear predictive even if they use far fewer components than the biological system itself uses for the same process. This problem also occurs in quantitative genetics where it is often possible to describe the variation in a system using fewer genes than are actually variable due to the complications of linkage between causal polymorphisms and population structure. Thus, there is a crucial need to begin an empirical investigation into the true number of components that are used by biological systems to determine a phenotypic outcome. In this study, we use a meta-analysis of directly comparable metabolomics quantitative studies using quantitative trait locus mapping and genome wide association mapping to show that it is currently not possible to estimate how many genetic loci are truly polymorphic within *Arabidopsis thaliana*. Our analysis shows that it would require the analysis of at least a 1000 line bi-parental population to begin to estimate how many polymorphic loci control metabolic variation within *Arabidopsis*. Understanding the base number of loci that are actually involved in determining variation in metabolic systems is fundamental to developing systems models that are truly reflective of how metabolism is modulated within a living organism.

Keywords: QTL, quantitative genetics, metabolite, glucosinolate, RIL population

INTRODUCTION

A central goal of systems biology is to develop models that are both predictive and accurately describe the biological system. Complicating this endeavor is the observation that it is possible to develop highly accurate models that use far fewer components than the biological system itself uses for the same process. The absence of components means that the models typically are not predictive when moving into new areas (phenotypes, environments, species) other than the explicit conditions in which the model was developed. This is because in these different areas there are new and previously unrecognized components that need to be included to make the model accurate. Thus, while it is possible to create highly accurate models, these models ability to be predictive into new untested conditions is frequently hindered. Solving this requires developing a base understanding of how large a true molecular network is within a biological system to ensure that the models are of similar scale.

This reduction conundrum also occurs in quantitative genetics where it is often possible to accurately describe the variation in a system using far fewer genetic loci than may actually be causing the phenotypic variation. This potential arises from the fact that most genetic populations are smaller than required to allow all independent components varying within the population to behave independently (Brem et al., 2005; West et al., 2007; Buckler et al., 2009; Chan et al., 2010b, 2011; Bloom et al., 2013; Albert et al., 2014). This lack of independence arises from the fact that genes are genetically linked upon chromosomes and there has not been sufficient recombination to separate them (Falconer and Mackay, 1996; Mackay, 2001; Manolio et al., 2009). Additionally, most populations do not have sufficient numbers of individuals to fully sample the genetic matrix. For example it would require ~33000 yeast lines to sample all possible combinations of 15 loci once (Albert et al., 2014). Another factor affecting independence in quantitative genetics is that there may be natural or artificial selection structuring the genome and further decreasing randomness and hence independence (Platt et al., 2010a,b). Thus, there is a crucial need to begin an empirical investigation into the true number of variable causal loci that may be present in any mapping population.

A key approach to study the number of loci causing variation in a phenotype is the use of structured mapping populations. Modeling studies often test how the size of a mapping population and its recombination design impact the ability to find quantitative trait loci (QTLs). However, these modeling studies are typically built on the assumption that existing populations have largely discovered what is available to be discovered within a specific population. This arises because studies often bootstrap the analysis by taking a subset of the population and then testing how many of the final QTLs were found (de Koning et al., 1998; Charmet, 2000; Perretant et al., 2000; Doerge, 2002; West et al., 2007). These analyses always show that only a smaller subset of the lines were necessary to identify what was found in the full population. This is then frequently misinterpreted to mean that all potential QTLs present within this population were found. However, this simply means that the researcher did not need the full population to find what they found. In contrast, this backward bootstrapping has no predictive capacity to provide information on what additional information might have been found if the population had been even larger than tested. Thus there is a need to conduct an empirical analysis of how population size may influence the number of QTLs being detected as a first step to figuring out how many loci may exist within a population or species.

In the report, we conduct a meta-analysis of metabolite QTL mapping studies within simple bi-parental recombinant inbred line (RIL) populations of *Arabidopsis thaliana* to begin studying what is necessary to estimate the number of loci that affect a trait in a species. We show that the different RIL populations have similar genetic architecture suggesting that they can be treated as a randomized sample of the species. Using multiple populations of differing sizes measured for the same phenotypes with the same experimental and statistical approaches, we show that the number of QTL identified increases with population size but that it is currently impossible to tell if this relationship is linear or log-linear. Separating these two models is essential to developing future populations but will require a bi-parental RIL population of at least 1000 lines. Interestingly, the new QTLs identified with the increasing population sizes were not of small effect but instead they were of similar effect to the QTLs found in smaller populations. In contrast, most models assume that as more QTL are found they are of smaller and smaller effect. This means that we may be vastly underestimating the genetic potential present within any single RIL population. If we are underestimating a simple bi-parental population than there is an even larger issue of underestimation with more complex multi-parent or genome wide association (GWA) mapping population. New empirical and modeling studies taking into account these meta-analysis observations will be needed to design optimal mapping populations for future analysis.

MATERIALS AND METHODS

METABOLOMICS META-ANALYSIS OF RIL POPULATIONS

We obtained all the QTL mapping data from two previous experiments looking at metabolomics QTLs in the Bay × Sha and Kas × Tsu RIL populations (Loudet et al., 2002; McKay et al., 2008; Rowe et al., 2008; Juenger et al., 2010; Joseph et al., 2013a). These experiments were done in the same growth chamber using the same experimental protocols optimizing the ability to directly compare the results. Additional metabolomics QTL studies in *Arabidopsis* RIL populations were not included either because they used different experimental designs that prevented the ability to compare the results or the appropriate data were not available (Keurentjes et al., 2006; Lisec et al., 2008, 2009; Sulpice et al., 2009; Brotman et al., 2011). All metabolomics were conducted at the University of California Davis metabolome facility following the same published protocols as described in the direct citations for each dataset (Weckwerth et al., 2004; Fiehn et al., 2005, 2008; Rowe et al., 2008; Chan et al., 2010a; Joseph et al., 2013a).

METABOLOMICS META-ANALYSIS OF GWA IN *Arabidopsis*

To compare the genetic architecture of metabolome variation in RILs with GWA populations, we obtained all the metabolite variation data from a previous GWA analysis of 96 accessions that were done in the same growth chamber using the same experimental protocols (Chan et al., 2010a). These 96 accessions were the same as described in other GWA analysis (Atwell et al., 2010). This allowed us to optimize the comparability of the results.

GLUCOSINOLATE META-ANALYSIS OF RIL POPULATIONS

For our meta-analysis of glucosinolate QTL analysis, we obtained all QTL mapping data from previous experiments looking at glucosinolate QTLs in the L*er* × Col-0, L*er* × Cvi, Bay × Sha, Da(1)-12 × Ei-2, and Kas × Tsu RIL populations (Kliebenstein et al., 2001b, 2002a,b; Wentzell et al., 2007; Joseph et al., 2013b). The number of QTLs found for each trait were available for all populations but the estimated additive effect per locus was only available for the Bay × Sha, Da(1)-12 × Ei-2, and Kas × Tsu RIL populations (Kliebenstein et al., 2001b, 2002a,b; Wentzell et al., 2007; Joseph et al., 2013b). These QTL studies were all conducted with the same technical platform and similar replication allowing for an optimal comparison of the results (Kliebenstein et al., 2001b, 2002a,b; Wentzell et al., 2007; Joseph et al., 2013b). For all experiments, they were conducted using the same established high-throughput glucosinolate extraction protocol with the same quantification approaches and level of replication (Kliebenstein et al., 2001a,b,c; Reichelt et al., 2002).

STATISTICAL ANALYSIS

All statistical analysis and visualizations were conducted within the R software (R Development Core Team, 2014).

RESULTS

COMPARATIVE METABOLOME HERITABILITY ACROSS POPULATIONS

To begin investigating how population size and diversity may influence metabolome QTL identification, we compiled data from two metabolomics studies in which the same experimental design, metabolome analysis protocol and growth chambers were used (Rowe et al., 2008; Joseph et al., 2013a). This minimizes any technical or environmental difference between the experiments that could influence the comparison. The two metabolomics QTL studies used RIL populations of different sizes; the Kas × Tsu *A. thaliana* RIL population had 316 lines while the Bay × Sha population had only 210 lines measured (Loudet et al., 2002; McKay et al., 2008; Rowe et al., 2008; Joseph et al., 2013a,b). Further, the two populations are highly diverse with minimal shared regions of high or low polymorphism indicating that we can treat them as a random sampling of potential RILs that may be generated from *Arabidopsis* (**Figure 1**; Atwell et al., 2010). 258 predominantly primary metabolites were detected in the QTL mapping experiments for both populations allowing these metabolites to be used for a direct comparison of the genetics controlling the plant metabolome between these two populations (Rowe et al., 2008; Joseph et al., 2013a). The two populations showed a highly similar heritability distribution for the metabolome variation, 21% for Kas × Tsu and 25% for Bay × Sha (**Figure 2A**; Rowe et al., 2008; Joseph et al., 2013a). In contrast, a direct comparison of

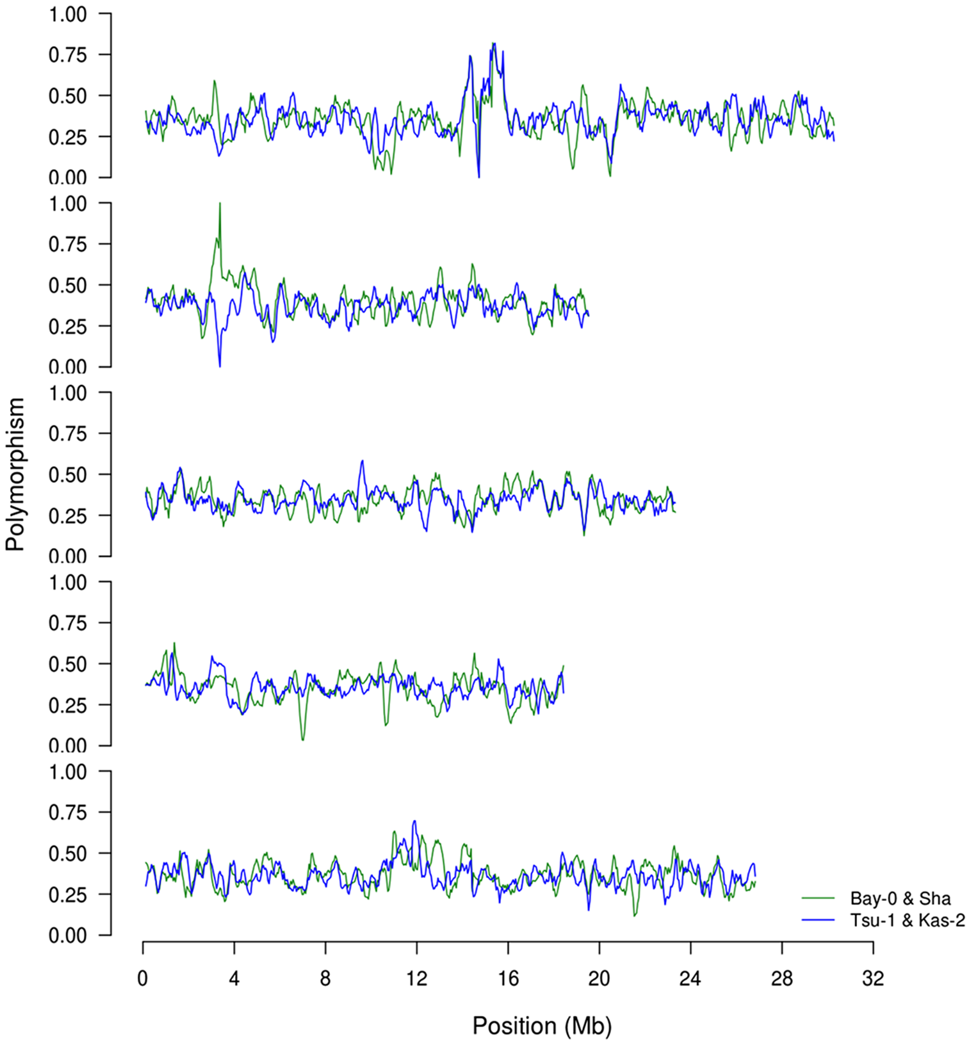

FIGURE 1 | Comparison of SNP distributions in the recombinant inbred line (RIL) populations. We obtained previously published single nucleotide polymorphism calls between the four genotypes, Bay, Sha, Kas, and Tsu. Using this data, we estimated the distribution of SNPs on chromosomes 1–5 (top to bottom) based on a sliding window of 250 Kb along the chromosome. The frequency is shown as the fraction of polymorphisms per 100 bp between the two parents for BayxSha (green) and KasxTsu (blue) RIL populations.

the heritability for each specific metabolite in the two populations showed that there was no significant correlation across the metabolites (**Figure 2B**, Pearson correlation, $r^2 = 0.04$, P = NS; **Figure 2B**; Rowe et al., 2008; Joseph et al., 2013a). Thus, while the genetics affecting the metabolome has similar overall heritability in the two populations, this genetic diversity affects different metabolites in the two populations. As such the size of the RIL population did not influence the distribution of heritability's.

SHARED STRUCTURE BUT DIFFERENT SPECIFICS OF METABOLOMIC DIVERSITY IN RIL POPULATIONS

Heritability showed that the overall genetic architecture affecting the metabolome is similar between the two populations but that the affected metabolites are different. We further tested this using the genetic variance present in each metabolite across the RILs. To do this, we compared the genetic CV for all 258 metabolites across the two populations. As with heritability, the overall distribution of metabolic variance was very similar between the two populations (**Figure 3**). The larger Kas × Tsu population was slightly skewed toward metabolites with larger population variance while the smaller Bay × Sha population had a slight enrichment in metabolites with lower genetic variance (**Figure 3**; Rowe et al., 2008; Joseph et al., 2013a). As with heritability, plotting the genetic CV of the 258 metabolites detected in both populations showed no correlation indicating that different specific metabolites are affected by this similar genetic architecture (**Figure 4**, Pearson correlation).

Within the Bay × Sha population the highly variable metabolites were enriched in sugars and amine metabolic processes (**Figure 4**). In contrast, the metabolites specifically variable in the Kas × Tsu population are more associated with stress responses

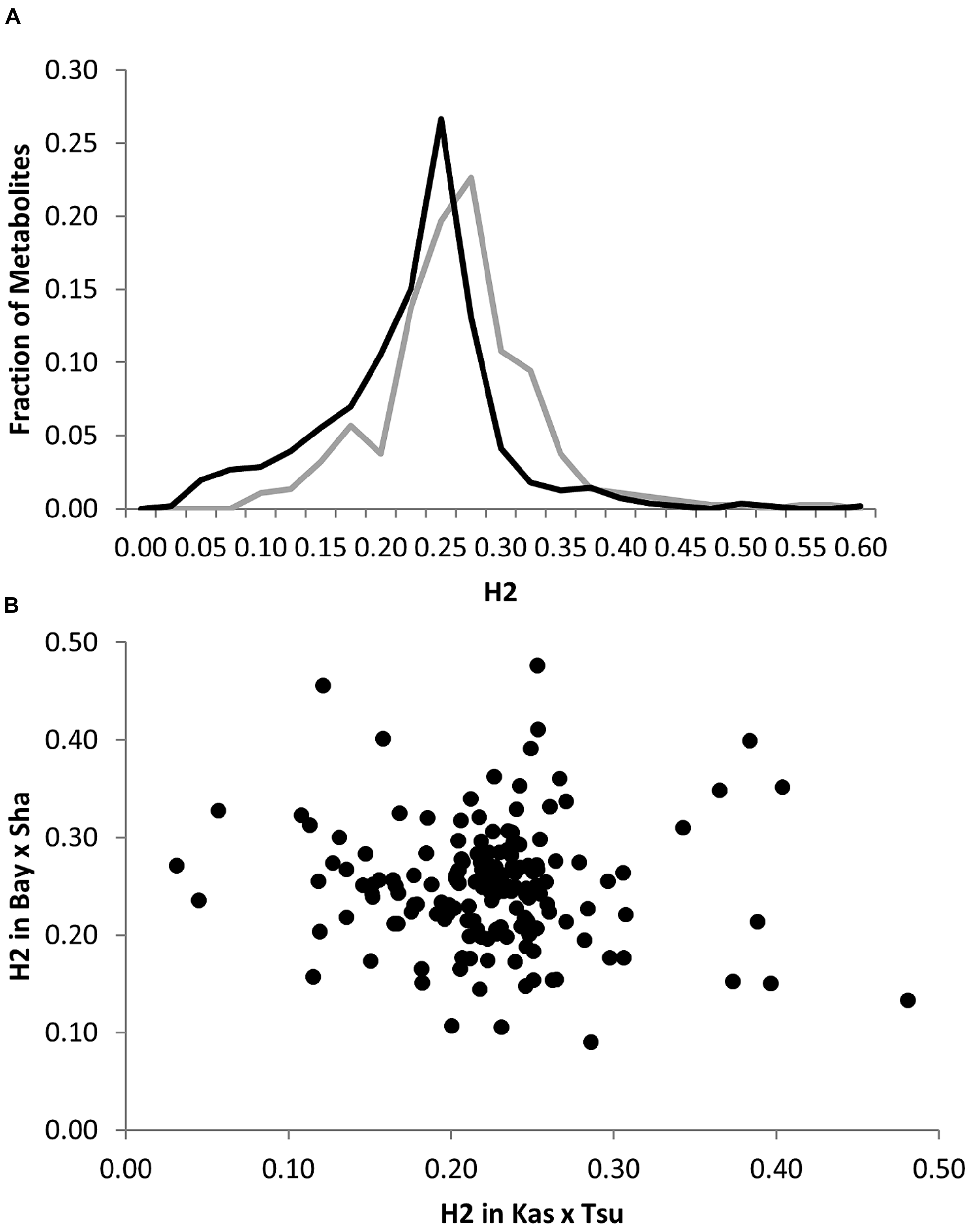

FIGURE 2 | Comparative metabolite genetics. (A) Comparison of estimated metabolite heritability's in the Bay × Sha (gray) and Kas × Tsu (black) RIL populations. **(B)** Scatter plot of heritability for 258 metabolites where both heritability and replicate effect could be estimated in the Bay × Sha and Kas × Tsu RIL populations. No significant correlation was found (Pearson correlation).

like Shikimate, putrescine, SA, and isonicotinic acid or lipid metabolism (**Figure 4**). Interestingly, the known compounds that displayed elevated genetic variance in both populations are key energy balance compounds like asparagine, pyruvate, glucose-6-phosphate, fructose-6-phosphate and galactose-6-phosphate (**Figure 4**). Typically, these central energy flux components are considered constrained in their function which should limit their diversity but this does not appear to be the case in *Arabidopsis*. The lack of correlation amongst specific metabolites for heritability or genetic CV between the two populations argues that each population has specific genetic polymorphisms that alter distinct metabolites in each population. These genetic polymorphisms are largely not shared between the two populations (**Figure 1**). Thus, while the specific genetic variation in each RIL population affects different metabolites, the overall genetic architecture (the distribution of heritabilities and CV) of each population is similar. The fact that the overall genetic architecture of the RIL populations is comparable suggests that we can treat diverse RIL populations in *Arabidopsis* as random sample of the potential genetic diversity within the species.

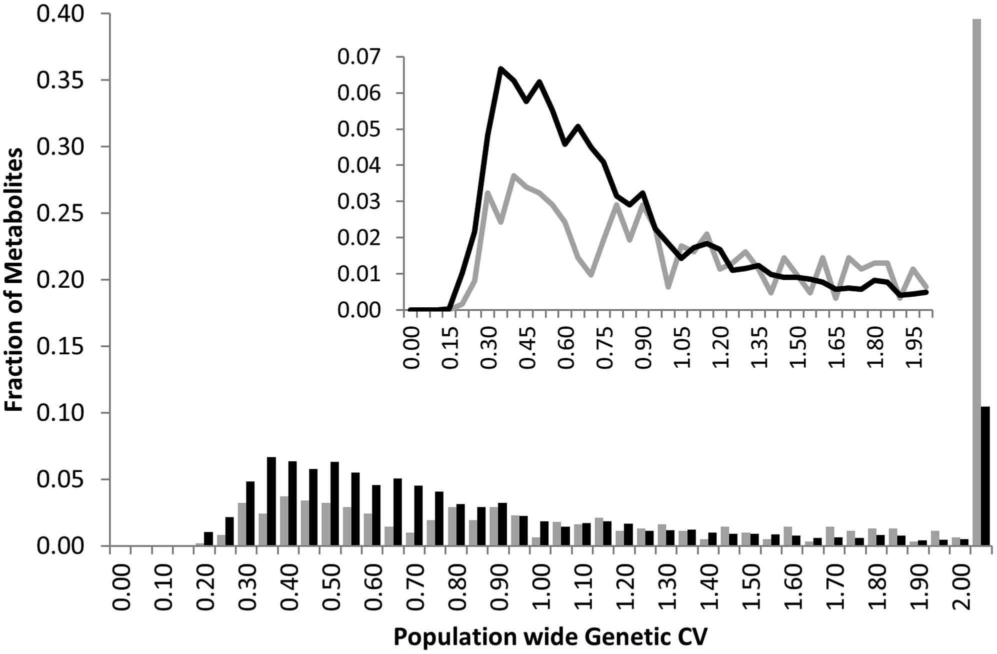

FIGURE 3 | Comparative metabolite genetics. Comparison of population CV across the metabolites within the Bay × Sha (gray) and Kas × Tsu (black) RIL populations. Inset shows a zoomed in view of the majority of the distribution.

COMPARATIVE METABOLOME GENETICS ACROSS POPULATIONS

To begin testing how an increase in the number of lines between the two populations increases the ability to identify QTLs, we compared the number of QTLs identified for the 258 common metabolites. For both populations we had previously used the same composite interval mapping (CIM) approach to identify and call significant QTLs. In combination with the similar range of heritability's and genetic variances, this identical statistical approach allows us to conduct a direct comparison where the only major difference in the two populations is line number (Jansen, 1994; Zeng et al., 1999a; Broman et al., 2003; Rowe et al., 2008; Joseph et al., 2013a). There was 54% more QTLs per metabolite detected in the 316 line Kas × Tsu population than for the 211 line Bay × Sha population (**Figure 5**). The Kas × Tsu population had an average of 1.22 ± 0.02 QTLs per metabolite in comparison to 0.79 ± 0.1 QTLs per metabolite found in the Bay × Sha RIL (**Figure 6**; Avg ± S.E.). The main difference between the two populations was the number of metabolites with at least one detected QTL; Kas × Tsu had a QTL detected for 75% of metabolites while for Bay × Sha this was only 44%. There was also an increase in the number of metabolites with two or more QTLs (**Figure 5**). Interestingly, 316 lines represents about a 50% increase in the number of lines which is similar to the 50% increase in QTL detected suggesting the potential for a linear relationship between the number of lines present in a population to the number of metabolite QTL detected.

RIL POPULATION SIZE AND QTL DETECTION USING TARGETED METABOLITE ANALYSIS

To better test how RIL population size influences the ability to identify QTLs we obtained data on QTL mapping for aliphatic and indolic glucosinolate accumulation within five different *Arabidopsis* populations (Lister and Dean, 1993; Alonso-Blanco et al., 1998; Kliebenstein et al., 2001b, 2002a,b; Loudet et al., 2002; Pfalz et al., 2007; Wentzell et al., 2007; McKay et al., 2008; Joseph et al., 2013b). These RIL populations differ in size from 100 to 411 lines and the glucosinolates were measured in the same tissue with similar replication using the same technical platform. Additionally, the glucosinolate QTL mapping was done with the same algorithm for all experiments (Kliebenstein et al., 2001b, 2002a,b; Wentzell et al., 2007; Joseph et al., 2013b). This allows us to conduct a direct comparison of QTL detection where the major difference is solely due to differences in the populations. Comparing the number of QTL identified to the number of lines in each population showed that QTL identification significantly increased with population size (**Figure 6**; Pearson Correlation, $P < 0.001$). This increase in QTL identification with population size was found for both aliphatic and indolic glucosinolates (**Figure 6**; Pearson Correlation, $P < 0.001$ for both). Within these populations, the aliphatic and indolic glucosinolates are controlled by different causal loci suggesting that they are behaving as independent measures of the relationship between power to identify QTL and population size within these populations (Kliebenstein et al., 2001b, 2002a,b; Pfalz et al., 2007; Wentzell et al., 2007; Joseph et al., 2013b).

When analyzing the relationship between QTL number found and population size for the glucosinolate phenotypes we found a log-linear relationship between the two parameters as has often been found in modeling studies (**Figure 6**; Pearson correlation using log adjusted values, $P < 0.001$ for both; Falconer and Mackay, 1996; Mackay, 2001; Doerge, 2002; Stich et al., 2007;

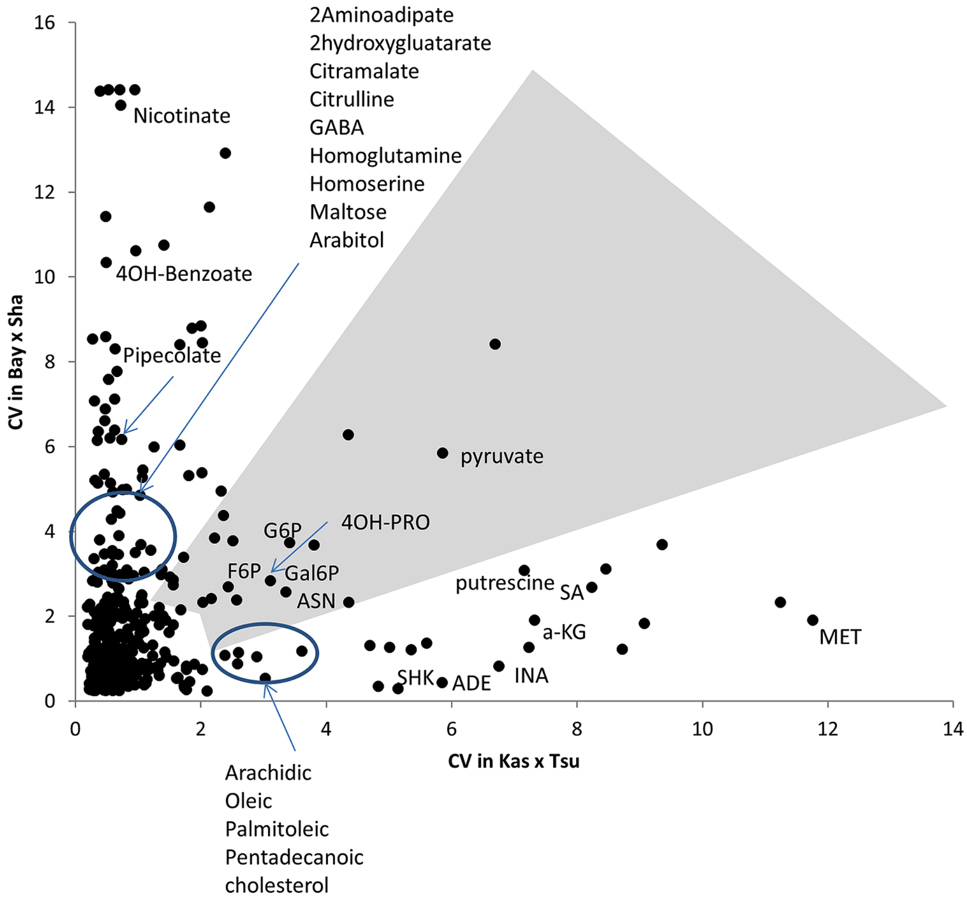

FIGURE 4 | Different modules of metabolic variation in different populations. Comparison of metabolite CV across all the RILs in the Bay × Sha and Kas × Tsu populations for 258 metabolites where CV could be estimated in both populations and that showed significant heritability's in both populations. The gray polygon shows metabolites with elevated phenotypic variation in both populations. Below the polygon are metabolites with high variation only in Kas × Tsu while to the left are metabolites that show elevated variation in only Bay × Sha. Known metabolites are labeled while unknown metabolites are unlabeled. Elevated variance is defined as being in the top 10th percentile of metabolites within a population.

West et al., 2007; McMullen et al., 2009; Klasen et al., 2012). However, a linear regression was also an equal statistical fit to the data (**Figure 6**). Thus, the existing data cannot differentiate between a linear and log-linear relationship of QTL number detected to population size. This is even though the large population is considered to be sufficient to fully sample the potential QTL in a population (**Figure 6**; Falconer and Mackay, 1996; Mackay, 2001; Doerge, 2002; Stich et al., 2007; West et al., 2007; McMullen et al., 2009; Klasen et al., 2012). Using the regression estimates we projected the linear and log-linear regression models with their 95% confidence intervals to larger population sizes to test what population size would be required to differentiate between these two different regressions (**Figure 7**). Even though both the aliphatic and indolic glucosinolates had different specific regression estimates the two traits generated the same estimate that it would require minimally between 950 and 1000 individuals in a single bi-parental RIL population to test which regression model more accurately approximates the relationship between population size and the power to identify new QTLs (**Figure 7**).

RIL POPULATION SIZE AND QTL EFFECT SIZE

A common assumption in QTL mapping is that small RIL populations will identify larger effect loci and the larger populations simply add smaller effect loci (Beavis, 1994, 1998; Zu, 2003). Using our meta-analysis data, we tested if there was a relationship between population size and the additive effect of the QTLs found. We chose to focus on additive effect rather than percent of total variance (r^2) explained because the additive effect of an individual locus is not dependent upon the total population variance. In contrast, the r^2 per locus is determined both by the effect of the locus and the total population variance (Beavis, 1994, 1998; Falconer and Mackay, 1996; Mackay, 2001; Zu, 2003). Thus, additive effect allows for more independence when comparing across populations. We thus compared line number against the additive effect size of each QTL found across three of the populations. Two of the five populations did not have the estimated additive effect size for the QTLs and were not used. Unexpectedly, this comparison showed no statistically significant relationship between population size and the additive effect of the identified QTLs for either the aliphatic or indolic glucosinolates (**Figure 8**;

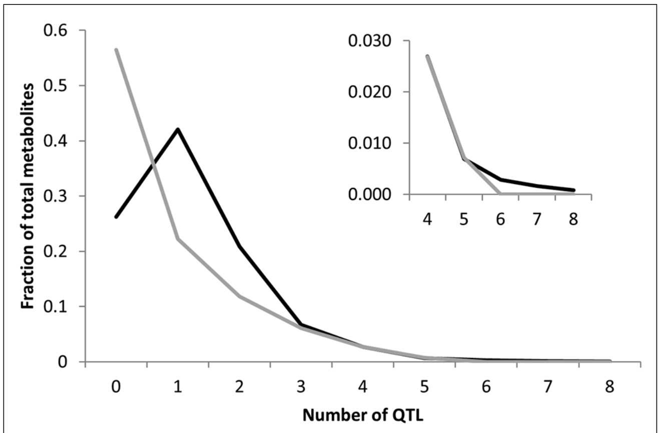

FIGURE 5 | Comparison of QTL detection across populations. Shown is the frequency of metabolites that detected a given number of nuclear genome Quantitative trait loci (QTLs) in the Bay × Sha RIL population (gray) and the Kas × Tsu RIL population. Inset shows a magnification of the *X* axis in the four to eight QTL region.

Pearson and Spearman Rank correlation tests). Thus, the new QTL found as population size increased were not of smaller effect but instead of similar effect size to those found with smaller population sizes.

COMPARATIVE METABOLIC DIVERSITY BETWEEN RIL AND GWA POPULATIONS

A common concern affecting RIL populations is that there are only two alleles per locus and thus might be limited in their genetic variation relative to species-wide diversity (Kim et al., 2007; Nordborg and Weigel, 2008). A proposed solution to this limitation is the use of unstructured GWA mapping populations that sample greater genetic diversity (Atwell et al., 2010; Chan et al., 2010a, 2011). This bi-allelic aspect of RIL populations is suggested to constrain RIL populations to simply sampling a subset of phenotypic variation in the more genetically diverse unstructured GWA populations. To assess how this bi-allelic structure may or may not constrain a RIL population, we compared the metabolomic analysis of two RIL populations with a GWA population using 135 metabolites detected in all three experiments (Rowe et al., 2008; Atwell et al., 2010; Chan et al., 2010a). These 135 metabolites included most of the known primary compounds (Rowe et al., 2008; Atwell et al., 2010; Chan et al., 2010a). Using the available data, we determined genetic CV for all 135 metabolites measured in each of the populations (Rowe et al., 2008; Atwell et al., 2010; Chan et al., 2010a; **Figure 9**). Comparing the genetic CV showed that the RIL populations could capture a majority of the genetic variance controlling metabolite variation within *Arabidopsis* (**Figure 9**). Additionally, both RIL populations identified variance not present in the accessions as they had at least 14 metabolites showing twice the genetic variance found in the accessions (**Figure 9**). In contrast, there were only five metabolites showing elevated variation in the accessions that was not captured in the two RIL populations (oxoproline, glycerol and three unknowns; **Figure 9**). Thus, while RILs have lower genetic diversity at individual loci than the accessions, this does not limit the associated phenotypic diversity.

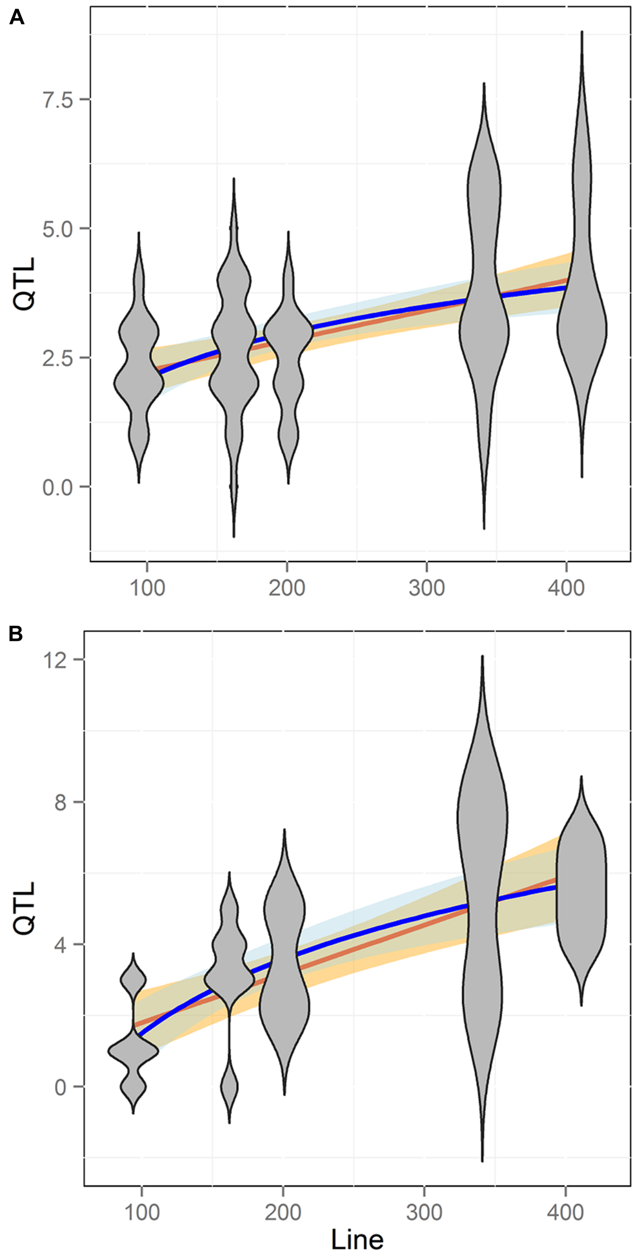

FIGURE 6 | Comparison of QTL detection across populations. Shown is the number of QTL identified across five different RIL populations of *Arabidopsis* with different line numbers. Blue shows the estimated log-linear relationship of QTL # identified per Line # while Orange shows the estimated linear relationship. SE of the estimated relationships are shown in filled color. Slopes and confidence intervals were obtained using Pearson correlation of either the linear or log-adjusted data. Plots are shown in linear scales for comparison purposes. **(A)** Aliphatic glucosinolate traits **(B)** Indolic glucosinolate traits.

DISCUSSION

ESTIMATING THE NUMBER OF QTLS PER BI-PARENTAL POPULATION

Obtaining an accurate estimate of how many QTLs exist within a single population is a key parameter for any quantitative modeling study. However, there is no empirical understanding for how this

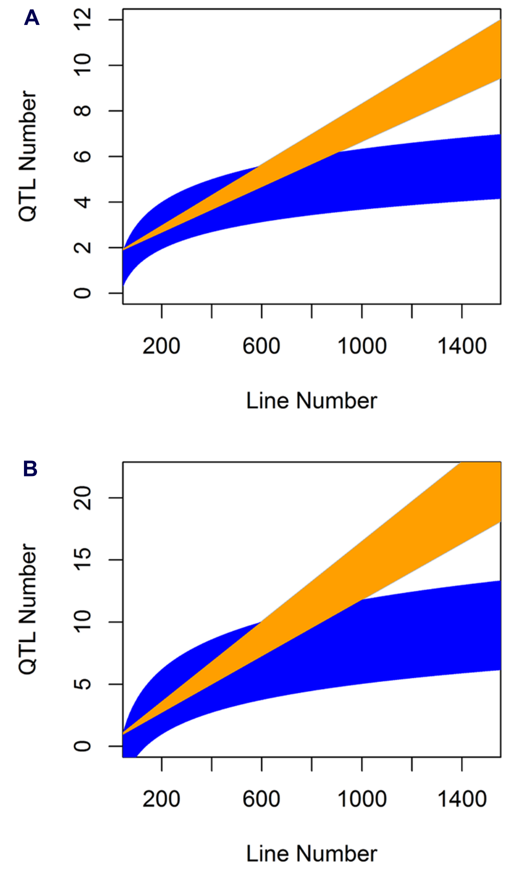

FIGURE 7 | Comparative prediction of QTL detection across population size. The slope and 95% confidence interval in the slope for both the linear and log-linear regression of QTL number by line number was obtained for the aliphatic **(A)** and indolic **(B)** glucosinolates across five populations where they were measured with the same experimental design and technical protocol. These estimated slopes were then use to plot the predicted QTL number detected per populations of a given line under the linear and log-linear predictions. Yellow shows the linear prediction and blue the log-linear prediction.

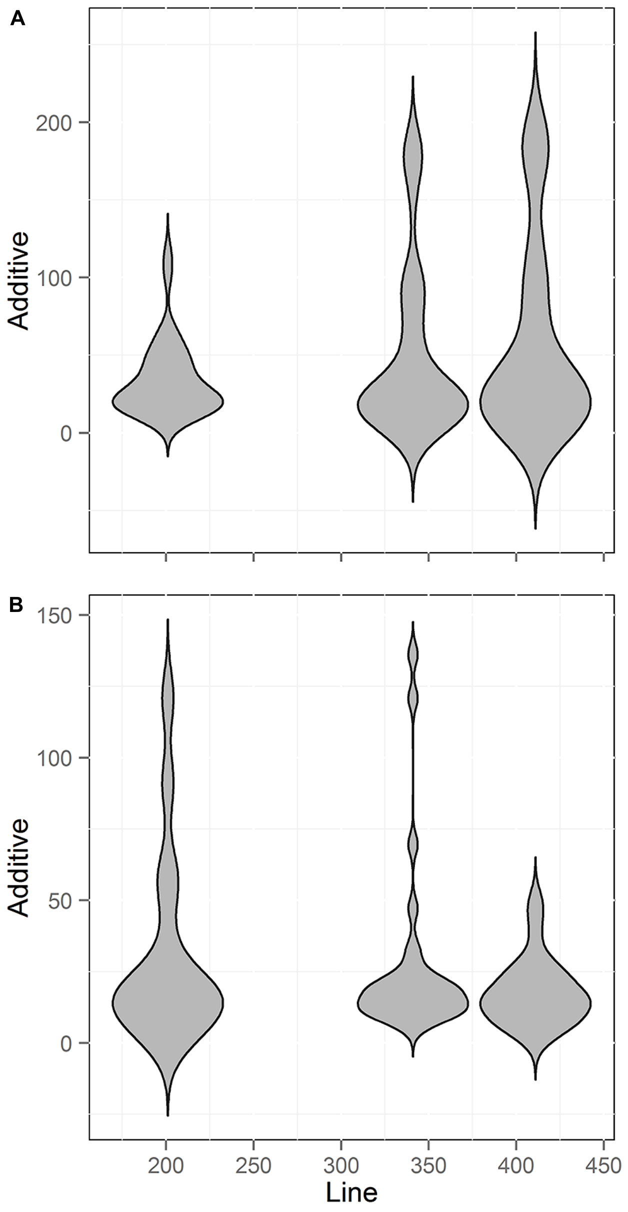

FIGURE 8 | Comparison of QTL Effect sizes by population size. Shown is the number of QTL identified across three different RIL populations of *Arabidopsis* with different line numbers. Two smaller populations could not be included because estimates of additive effects were not published. **(A)** Aliphatic glucosinolate traits **(B)** Indolic glucosinolate traits.

parameter, QTLs per population, should be set within modeling studies. This leads to a massive range of values in these modeling studies complicating any true comparison between these modeling studies (Zeng et al., 1999b; Otto and Jones, 2000; Xu, 2003; Slate, 2013; Guo et al., 2014). Even the analysis of individual large populations reporting large numbers of loci controlling nearly all of the heritability are likely under-estimates due to insufficient recombination between loci (McMullen et al., 2009; Huang et al., 2010; Kump et al., 2011; Tian et al., 2011; Bloom et al., 2013; Albert et al., 2014). In addition to modeling studies, understanding how many QTLs exist within a single population is fundamental to knowing how to design future mapping populations especially with the interest in multi-parent populations (Kover et al., 2009; McMullen et al., 2009). To date, these populations have been modeled and structured based on the assumption that existing bi-parental populations have identified the majority of identifiable QTLs in those populations (Kover et al., 2009; McMullen et al., 2009).

Thus, we conducted a meta-analysis of existing metabolite QTL mapping within simple bi-parental *Arabidopsis* populations to begin an empirical assessment of how many QTLs are present in a given mapping population. As expected, including more lines in a bi-parental did lead to more identified QTLs. However, the relationship between line number and QTL number equally fit a linear and log-linear relationship even with 411 lines (**Figure 6**). Extending these linear and log-linear relationships showed that it would require 950 or more lines for even a simple bi-parental population before we could tell which relationship accurately describes

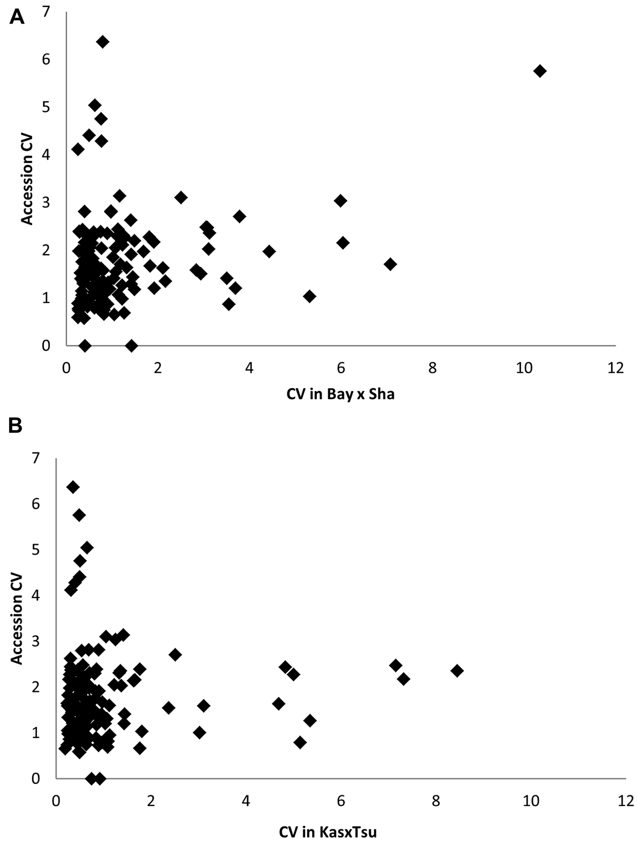

FIGURE 9 | Comparative phenotype diversity in 96 *Arabidopsis* accessions to RIL populations. (A) Comparison of Genetic CV for 135 metabolites measured in the Bay × Sha RILs and96 *Arabidopsis* accessions. **(B)** Comparison of Genetic CV for 135 metabolites measured in the Kas × Tsu RILs and96 *Arabidopsis* accessions.

the true number of metabolite QTL per population (**Figure 7**). Accurately testing which model best describes QTL number per a bi-parental population is essential given that most multi-parent populations are stopping at 1000 lines under the assumption that this is sufficient to identify all the QTLs (Flint et al., 2005; Kover et al., 2009). However, our meta-analysis suggests that 1000 lines are only just sufficient in a bi-parental population much less a population with multiple alleles per locus. This indicates that at least one bi-parental population of more than 1000 lines should be developed in *Arabidopsis* to tests if the QTL identification power of a population is a linear or log-linear relationship to its size.

The above analysis is focused on metabolite related QTLs and as such may not be broadly reflective of all traits. For example, expression linked traits show higher average heritability than the metabolomic traits but lower heritability than the glucosinolate related traits used in this study (Keurentjes et al., 2007; West et al., 2007; Rowe et al., 2008; Kliebenstein, 2009; Jimenez-Gomez et al., 2011; Joseph et al., 2013a). Most physiological and defense related phenotypes have heritability's similar to these metabolomic results within these populations (Joseph et al., 2013b). Given the potential for different genetic architectures in each trait, it argues that there needs to be a broader effort to test if the if the link between QTL identification power and population size is affected by the phenotype being studied.

QTL EFFECT SIZE DISTRIBUTION

In contrast to most quantitative theory, our analysis found that larger population sizes identified QTLs with a similar distribution of effect sizes as smaller populations (Beavis, 1994, 1998). A possible explanation for how the new QTLs found with increasing line numbers is that there might be a significant confounding issue of linked QTLs in most existing populations (Yamamoto et al., 2014). If two QTLs were linked and had opposing effects there is a high likelihood that in small populations the region would be missed. This is because the two QTLs effects would cancel each other out and the lack of recombination would not allow either locus to be detected. Thus, when these loci are identified in larger populations they could have effect sizes similar to QTLs that are not linked. This linkage of opposing effects has previously been found using larger *Arabidopsis* populations (Wentzell et al., 2007; Rowe and Kliebenstein, 2008; Rowe et al., 2008). Another possibility is that if two QTLs are linked with a similar direction of effects, they might be identified as a single locus in smaller populations but with an inordinately large effect size (Yamamoto et al., 2014). Then upon increasing the number of lines, the two loci would separate into two QTLs of smaller effect than the original locus (Hansen et al., 2007). Combinations of loci with both similar and opposing effects have been found when conducting fine-scale dissection of *Arabidopsis* QTLs (Kroymann and Mitchell-Olds, 2005; Wentzell et al., 2008). Thus, the base assumption that increasing the population sizewill solely identify QTLs of smaller effect size is not supported by this empirical meta-analysis.

CONCLUSION

Our meta-analysis of QTL analysis using metabolite phenotypes across multiple *Arabidopsis* RIL populations shows that it is not possible to accurately estimate how many QTLs are present in a single population with two alleles per locus. This is in contrast to the myriad of modeling studies that make explicit assumptions about this variable. Future work will be required to extend these populations to larger sizes to provide a direct and empirical estimate of how many QTLs may exist in a population. Conducting these experiments will then provide a more firm foundation for extensions of similar estimates into multi-parent populations and further into the entire species. It is only then that we will have a true view of how many naturally variable genes causally affect variation within the plant metabolome.

ACKNOWLEDGMENTS

This work was supported by the US National Science Foundation (DBI Award 0820580 and MCB Award 1330337) and by the Danish National Research Foundation (DNRF99).

REFERENCES

Albert, F. W., Treusch, S., Shockley, A. H., Bloom, J. S., and Kruglyak, L. (2014). Genetics of single-cell protein abundance variation in large yeast populations. *Nature* 506, 494–497. doi: 10.1038/nature12904

Alonso-Blanco, C., Peeters, A. J. M., Koornneef, M., Lister, C., Dean, C., Van Den Bosch, N., et al. (1998). Development of an AFLP based linkage map of

Ler, Col and Cvi *Arabidopsis thaliana* ecotypes and construction of a Ler/Cvi recombinant inbred line population. *Plant J.* 14, 259–271. doi: 10.1046/j.1365-313X.1998.00115.x

Atwell, S., Huang, Y., Vilhjalmsson, B. J., Willems, G., Horton, M., Li, Y., et al. (2010). Genome-wide association study of 107 phenotypes in a common set of *Arabidopsis thaliana* in-bred lines. *Nature* 465, 627–631. doi: 10.1038/nature08800

Beavis, W. D. (1994). "The power and deceit of QTL experiments: lessons from comparitive QTL studies," in *Proceedings of the Forty-Ninth Annual Corn and Sorghum Industry Research Conference*, (Washington, DC: American Seed Trade Association), 250–266.

Beavis, W. D. (1998). "QTL analyses: power, precision, and accuracy," in *Molecular Dissection of Complex Traits*, ed. A. H. Paterson (New York, NY: CRC Press), 145–162.

Bloom, J. S., Ehrenreich, I. M., Loo, W., Vo Lite, T. L., and Kruglyak, L. (2013). Finding the sources of missing heritability in a yeast cross. *Nature* 494, 234–237. doi: 10.1038/nature11867

Brem, R. B., Storey, J. D., Whittle, J., and Kruglyak, L. (2005). Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436, 701–703. doi: 10.1038/nature03865

Broman, K. W., Wu, H., Sen, Ś., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19, 889–890. doi: 10.1093/bioinformatics/btg112

Brotman, Y., Riewe, D., Lisec, J., Meyer, R. C., Willmitzer, L., and Altmann, T. (2011). Identification of enzymatic and regulatory genes of plant metabolism through QTL analysis in *Arabidopsis. J. Plant Physiol.* 168, 1387–1394. doi: 10.1016/j.jplph.2011.03.008

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., et al. (2009). The genetic architecture of maize flowering time. *Science* 325, 714–718. doi: 10.1126/science.1174276

Chan, E. K., Rowe, H. C., Corwin, J. A., Joseph, B., and Kliebenstein, D. J. (2011). Combining genome-wide association mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana. PLoS Biol.* 9:e1001125. doi: 10.1371/journal.pbio.1001125

Chan, E. K., Rowe, H. C., Hansen, B. G., and Kliebenstein, D. J. (2010a). The complex genetic architecture of the metabolome. *PLoS Genet.* 6:e1001198. doi: 10.1371/journal.pgen.1001198

Chan, E. K. F., Rowe, H. C., and Kliebenstein, D. J. (2010b). Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. *Genetics* 185, 991–1007. doi: 10.1534/genetics.109.108522

Charmet, G. (2000). Power and accuracy of QTL detection: simulation studies of one-QTL models. *Agronomie* 20, 309–323. doi: 10.1051/agro:2000129

de Koning, D. J., Visscher, P. M., Knott, S. A., and Haley, C. S. (1998). A strategy for QTL detection in half-sib populations. *Anim. Sci.* 67, 257–268. doi: 10.1017/S1357729800010018

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nat. Rev. Genet.* 3, 43–52. doi: 10.1038/nrg703

Falconer, D. S., and Mackay, T. F. C. (1996). *Introduction to Quantitative Genetics.* (Harlow: Longman).

Fiehn, O., Wohlgemuth, G., and Scholz, M. (2005). "Setup and annotation of metabolomic experiments by integrating biological and mass spectrometric metadata," in *Data Integration in the Life Sciences*, eds B. Ludäscher and L. Raschid (Berlin-Heidelberg: Springer), 224–239. doi: 10.1007/11530084_18

Fiehn, O., Wohlgemuth, G., Scholz, M., Kind, T., Lee, D. Y., Lu, Y., et al. (2008). Quality control for plant metabolomics: reporting MSI-compliant studies. *Plant J.* 53, 691–704. doi: 10.1111/j.1365-313X.2007.03387.x

Flint, J., Valdar, W., Shifman, S., and Mott, R. (2005). Strategies for mapping and cloning quantitative trait genes in rodents. *Nat. Rev. Genet.* 6, 271–286. doi: 10.1038/nrg1576

Guo, G., Zhao, F., Wang, Y., Zhang, Y., Du, L., and Su, G. (2014). Comparison of single-trait and multiple-trait genomic prediction models. *BMC Genet.* 15:30. doi: 10.1186/1471-2156-15-30

Hansen, B. G., Kliebenstein, D. J., and Halkier, B. A. (2007). Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in *Arabidopsis. Plant J.* 50, 902–910. doi: 10.1111/j.1365-313X.2007.03101.x

Huang, X. H., Wei, X. H., Sang, T., Zhao, Q. A., Feng, Q., Zhao, Y., et al. (2010). Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42, U961–U976. doi: 10.1038/ng.695

Jansen, R. C. (1994). Controlling the type-I and type-II errors in mapping quantitative trait loci. *Genetics* 138, 871–881.

Jimenez-Gomez, J. M., Corwin, J. A., Joseph, B., Maloof, J. N., and Kliebenstein, D. J. (2011). Genomic analysis of QTLs and genes altering natural variation in stochastic noise. *PLoS Genet.* 7:e1002295. doi: 10.1371/journal.pgen.1002295

Joseph, B., Corwin, J. A., Li, B., Atwell, S., and Kliebenstein, D. J. (2013a). Cytoplasmic genetic variation and extensive cytonuclear interactions influence natural variation in the metabolome. *eLife* 2, e00776.

Joseph, B., Corwin, J. A., Zuest, T., Li, B., Iravani, M., Schaepman-Strub, G., et al. (2013b). Hierarchical nuclear and cytoplasmic genetic architectures for plant growth and defense within *Arabidopsis. Plant Cell* 25, 1929–1945. doi: 10.1105/tpc.113.112615

Juenger, T. E., Sen, S., Bray, E., Stahl, E., Wayne, T., Mckay, J., et al. (2010). Exploring genetic and expression differences between physiologically extreme ecotypes: comparative genomic hybridization and gene expression studies of Kas-1 and Tsu-1 accessions of *Arabidopsis thaliana. Plant Cell Environ.* 33, 1268–1284. doi: 10.1111/j.1365-3040.2010.02146.x

Keurentjes, J. J. B., Fu, J. Y., De Vos, C. H. R., Lommen, A., Hall, R. D., Bino, R. J., et al. (2006). The genetics of plant metabolism. *Nat. Genet.* 38, 842–849. doi: 10.1038/ng1815

Keurentjes, J. J. B., Fu, J. Y., Terpstra, I. R., Garcia, J. M., Van Den Ackerveken, G., Snoek, L. B., et al. (2007). Regulatory network construction in *Arabidopsis* by using genome-wide gene expression quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.* 104, 1708–1713. doi: 10.1073/pnas.0610429104

Kim, S., Plagnol, V., Hu, T. T., Toomajian, C., Clark, R. M., Ossowski, S., et al. (2007). Recombination and linkage disequilibrium in *Arabidopsis thaliana. Nat. Genet.* 39, 1151–1155. doi: 10.1038/ng2115

Klasen, J. R., Piepho, H. P., and Stich, B. (2012). QTL detection power of multi-parental RIL populations in *Arabidopsis thaliana. Heredity* 108, 626–632. doi: 10.1038/hdy.2011.133

Kliebenstein, D. J. (2009). A quantitative genetics and ecological model system: understanding the aliphatic glucosinolate biosynthetic network via QTLs. *Phytochem. Rev.* 8, 243–254. doi: 10.1007/s11101-008-9102-8

Kliebenstein, D., Lambrix, V., Reichelt, M., Gershenzon, J., and Mitchell-Olds, T. (2001a). Gene duplication and the diversification of secondary metabolism: side chain modification of glucosinolates in *Arabidopsis thaliana. Plant Cell* 13, 681–693. doi: 10.1105/tpc.13.3.681

Kliebenstein, D. J., Gershenzon, J., and Mitchell-Olds, T. (2001b). Comparative quantitative trait loci mapping of aliphatic, indolic and benzylic glucosinolate production in *Arabidopsis thaliana* leaves and seeds. *Genetics* 159, 359–370.

Kliebenstein, D. J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D., Gershenzon, J., et al. (2001c). Genetic control of natural variation in *Arabidopsis thaliana* glucosinolate accumulation. *Plant Physiol.* 126, 811–825. doi: 10.1104/pp.126.2.811

Kliebenstein, D., Pedersen, D., Barker, B., and Mitchell-Olds, T. (2002a). Comparative analysis of quantitative trait loci controlling glucosinolates, myrosinase and insect resistance in *Arabidopsis thaliana. Genetics* 161, 325–332.

Kliebenstein, D. J., Figuth, A., and Mitchell-Olds, T. (2002b). Genetic architecture of plastic methyl jasmonate responses in *Arabidopsis thaliana. Genetics* 161, 1685–1696.

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., et al. (2009). A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana. PLoS Genet.* 5:e1000551. doi: 10.1371/journal.pgen.1000551

Kroymann, J., and Mitchell-Olds, T. (2005). Epistasis and balanced polymorphism influencing complex trait variation. *Nature* 435, 95–98. doi: 10.1038/nature03480

Kump, K. L., Bradbury, P. J., Wisser, R. J., Buckler, E. S., Belcher, A. R., Oropeza-Rosas, M. A., et al. (2011). Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. *Nat. Genet.* 43, U163–U120. doi: 10.1038/ng.747

Lisec, J., Meyer, R. C., Steinfath, M., Redestig, H., Becher, M., Witucka-Wall, H., et al. (2008). Identification of metabolic and biomass QTL in *Arabidopsis thaliana* in a parallel analysis of RIL and IL populations. *Plant J.* 53, 960–972. doi: 10.1111/j.1365-313X.2007.03383.x

Lisec, J., Steinfath, M., Meyer, R. C., Selbig, J., Melchinger, A. E., Willmitzer, L., et al. (2009). Identification of heterotic metabolite QTL in *Arabidopsis thaliana* RIL and IL populations. *Plant J.* 59, 777–788. doi: 10.1111/j.1365-313X.2009.03910.x

Lister, C., and Dean, D. (1993). Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana. Plant J.* 4, 745–750. doi: 10.1046/j.1365-313X.1993.04040745.x

Loudet, O., Chaillou, S., Camilleri, C., Bouchez, D., and Daniel-Vedele, F. (2002). Bay-0 × Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in *Arabidopsis. Theor. Appl. Genet.* 104, 1173–1184. doi: 10.1007/s00122-001-0825-9

Mackay, T. F. C. (2001). The genetic architecture of quantitative traits. *Annu. Rev. Genet.* 35, 303–339. doi: 10.1146/annurev.genet.35.102401.090633

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

McKay, J. K., Richards, J. H., Nemali, K. S., Sen, S., Mitchell-Olds, T., Boles, S., et al. (2008). Genetics of drought adaptation in *Arabidopsis thaliana* II. QTL analysis of a new mapping population Kas-1 × Tsu-1. *Evolution* 62, 3014–3026. doi: 10.1111/j.1558-5646.2008.00474.x

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H. H., Sun, Q., et al. (2009). Genetic properties of the maize nested association mapping population. *Science* 325, 737–740. doi: 10.1126/science.1174320

Nordborg, M., and Weigel, D. (2008). Next-generation genetics in plants. *Nature* 456, 720–723. doi: 10.1038/nature07629

Otto, S. P., and Jones, C. D. (2000). Detecting the undetected: estimating the total number of loci underlying a quantitative trait. *Genetics* 156, 2093–2107.

Perretant, M. R., Cadalen, T., Charmet, G., Sourdille, P., Nicolas, P., Boeuf, C., et al. (2000). QTL analysis of bread-making quality in wheat using a doubled haploid population. *Theor. Appl. Genet.* 100, 1167–1175. doi: 10.1007/s0012200 51420

Pfalz, M., Vogel, H., Mitchell-Olds, T., and Kroymann, J. (2007). Mapping of QTL for resistance against the crucifer specialist herbivore *Pieris brassicae* in a new *Arabidopsis* inbred line population, Da(1)-12 × Ei-2. *PLoS ONE* 2:e578. doi: 10.1371/journal.pone.0000578

Platt, A., Horton, M., Huang, Y. S., Li, Y., Anastasio, A. E., Mulyati, N. W., et al. (2010a). The scale of population structure in *Arabidopsis thaliana. PLoS Genet.* 6:e1000843. doi: 10.1371/journal.pgen.1000843

Platt, A., Vilhjalmsson, B. J., and Nordborg, M. (2010b). Conditions under which genome-wide association studies will be positively maisleading. *Genetics* 186, 1045–1052. doi: 10.1534/genetics.110.121665

R Development Core Team. (ed.) (2014). "R: a language and environment for statistical computing," in *R.F.F.S. Computing* (Vienna: R Foundation for Statistical Computing).

Reichelt, M., Brown, P. D., Schneider, B., Oldham, N. J., Stauber, E., Tokuhisa, J., et al. (2002). Benzoic acid glucosinolate esters and other glucosinolates from *Arabidopsis thaliana. Phytochemistry* 59, 663–671. doi: 10.1016/S0031-9422(02)00014-6

Rowe, H. C., Hansen, B. G., Halkier, B. A., and Kliebenstein, D. J. (2008). Biochemical networks and epistasis shape the *Arabidopsis thaliana* metabolome. *Plant Cell* 20, 1199–1216. doi: 10.1105/tpc.108.058131

Rowe, H. C., and Kliebenstein, D. J. (2008). Complex genetics control natural variation in *Arabidopsis thaliana* resistance to Botrytis cinerea. *Genetics* 180, 2237–2250. doi: 10.1534/genetics.108.091439

Slate, J. (2013). From beavis to beak color: a simulation study to examine how much QTL mapping can reveal about the genetic architecture of quantitative traits. *Evolution* 67, 1251–1262. doi: 10.1111/evo.12060

Stich, B., Yu, J. M., Melchinger, A. E., Piepho, H. P., Utz, H. F., Maurer, H. P., et al. (2007). Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy. *Genetics* 176, 563–570. doi: 10.1534/genetics.106.067033

Sulpice, R., Pyl, E. T., Ishihara, H., Trenkamp, S., Steinfath, M., Witucka-Wall, H., et al. (2009). Starch as a major integrator in the regulation of plant growth. *Proc. Natl. Acad. Sci. U.S.A.* 106, 10348–10353. doi: 10.1073/pnas.0903478106

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., et al. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* 43, U159–U113. doi: 10.1038/ng.746

Weckwerth, W., Loureiro, M. E., Wenzel, K., and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U.S.A.* 101, 7809–7814. doi: 10.1073/pnas.0303415101

Wentzell, A. M., Boeye, I., Zhang, Z. Y., and Kliebenstein, D. J. (2008). Genetic networks controlling structural outcome of glucosinolate activation across development. *PLoS Genet.* 4:e1000234. doi: 10.1371/journal.pgen.1000234

Wentzell, A. M., Rowe, H. C., Hansen, B. G., Ticconi, C., Halkier, B. A., and Kliebenstein, D. J. (2007). Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. *PLoS Genet.* 3:e162. doi: 10.1371/journal.pgen.0030162

West, M. A. L., Kim, K., Kliebenstein, D. J., Van Leeuwen, H., Michelmore, R. W., Doerge, R. W., et al. (2007). Global eQTL mapping reveals the complex genetic architecture of transcript level variation in *Arabidopsis. Genetics* 175, 1441–1450. doi: 10.1534/genetics.106.064972

Xu, S. Z. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* 163, 789–801.

Yamamoto, E., Iwata, H., Tanabata, T., Mizobuchi, R., Yonemaru, J., Yamamoto, T., et al. (2014). Effect of advanced intercrossing on genome structure and on the power to detect linked quantitative trait loci in a multi-parent population: a simulation study in rice. *BMC Genet.* 15:50. doi: 10.1186/1471-2156-15-50

Zeng, Z.-B., Kao, C.-H., and Basten, C. J. (1999a). Estimating the genetic architecture of quantitative traits. *Genet. Res.* 75, 345–355.

Zeng, Z. B., Kao, C. H., and Basten, C. J. (1999b). Estimating the genetic architecture of quantitative traits. *Genet. Res.* 74, 279–289. doi: 10.1017/S0016672399004255

Zu, S. (2003). Theoretical basis of the Beavis effect. *Genetics* 165, 2259–2268.

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: 23 June 2014; accepted: 26 August 2014; published online: 11 September 2014.

Citation: Joseph B, Atwell S, Corwin JA, Li B and Kliebenstein DJ (2014) Meta-analysis of metabolome QTLs in Arabidopsis: trying to estimate the network size controlling genetic variation of the metabolome. Front. Plant Sci. ***5****:461. doi: 10.3389/fpls.2014.00461*

This article was submitted to Plant Systems Biology, a section of the journal Frontiers in Plant Science.