Local adaptation shapes metabolic diversity in the global population of *Arabidopsis thaliana*

Rik Kooke[1,2,], Willem Kruijer[2], Henriette D.L.M. van Eekelen[3], Frank F.M. Becker[1], Ron Wehrens[2], Robert D. Hall[3,4], Roland Mumm[3], Ric C.H. de Vos[3], Fred A. van Eeuwijk[2] and Joost J.B. Keurentjes[1]

[1] Laboratory of Genetics, Wageningen University & Research, Wageningen, the Netherlands, NL-6708 PB

[2] Biometris, Wageningen University & Research, Wageningen, the Netherlands, NL-6708 PB

[3] Business Unit Bioscience, Wageningen Research, Wageningen, the Netherlands, NL-6708 PB

[4] Laboratory of Plant Physiology, Wageningen University & Research, Wageningen, the Netherlands, NL-6708 PB

ORCID Identifier

R. Kooke

W. Kruijer

H.D.L.M. van Eekelen

F.F.M. Becker

R. Wehrens

R.D. Hall 0000-0002-5786-768X

R. Mumm

R.C.H. de Vos 0000-0002-2181-5624

F.A. van Eeuwijk

J.J.B. Keurentjes 0000-0001-8918-0711

Corresponding Author

Joost J.B. Keurentjes

Droevendaalsesteeg 1

NL-6708 PB Wageningen

The Netherlands

T: 0317 483149

E: Joost.Keurentjes@wur.nl

Classification

Biological Sciences - Genetics

Keywords

Arabidopsis thaliana, Metabolomics, Secondary metabolism, Genetics, Adaptation

Abstract

The biosynthesis, structure and accumulation of secondary metabolites in plants are largely controlled by genetic factors, which can vary substantially among genotypes within a species. Here we studied a global population of *Arabidopsis thaliana* accessions for qualitative and quantitative variation in volatile and non-volatile secondary metabolites using essentially untargeted metabolomics. Genome-wide association (GWA) mapping revealed that metabolic variation mainly traces back to genetic variation in dedicated biosynthesis genes. Effect sizes of genetic variants, estimated by a Bayesian procedure, indicate that most of the genetic variation in the accumulation of secondary metabolites is explained by large-effect genes and defined by multiple polymorphisms. The various genetic variants resulted from independent mutation events and combined into distinctive haplotypes, which are representative for specific geographical regions. A strong relationship between the effect-size of regulatory loci, their allele frequencies and fixation index indicates that selection forces discriminate between haplotypes, resulting in different phytochemical profiles. Finally, we demonstrate that haplotype frequencies deviate from neutral theory predictions, suggesting that metabolic profiles are shaped by local adaptation and co-evolution of independent loci.

Introduction

The success and evolution of life on earth relies almost completely on the ability of plants, as primary components of the food chain and net-producers of oxygen, to grow and flourish in a wide diversity of environments and conditions. For this, plants have adapted their morphology, developmental timing and metabolism to some of the most extreme settings (Cannell et al., 2020). Adaptation also includes the sometimes multi-trophic interactions with the biotic and abiotic environment. Most of the properties of plants, therefore, display an enormous variation in expression among and even within species. Although monogenic qualitative traits do occur (*e.g.,* disease resistance or specialized organogenesis), the majority of traits are polygenic, or even omnigenic (Boyle et al., 2017, Chateigner et al., 2020). Such so-called quantitative traits have been shaped for millions of years through evolutionary processes, like mutation and recombination, drift, dispersal and natural selection, and their genetic architecture approaches an infinitesimal model, in which an infinite number of genes with infinitely small effects determines the resulting phenotype (Rockman, 2012, Olson-Manning et al., 2012).

Arabidopsis thaliana is no exception and is deservedly considered the reference plant species for studies on natural variation (Alonso-Blanco et al., 2009). *A. thaliana* diverged five to six million years ago from other species and despite relatively recent bottleneck events, due to glacial-interglacial climate changes, has adapted to a wide range of environmental settings spanning nearly all terrestrial habitats across the globe (Lyu, 2017, Shimizu and Purugganan, 2005). The species displays a wide diversity in the manifestation of traits and the abundant genetic resources available enabled the analysis of linkage between sequence diversity and natural variation in adaptive properties (Bergelson and Roux, 2010, Kover and Mott, 2012, Trontin et al., 2011). Being sessile organisms, unable to escape

environmental threats, plants have evolved an enormous arsenal of phytochemical compounds, predominantly secondary metabolites, to combat risks and hazards (Fang et al., 2019). Natural variation in chemical profiles offers resilience to species and even increases evolvability in changing environments (Payne and Wagner, 2019, Segrè et al., 2002). Because plants need to be able to respond quickly to environmental fluctuations, a flexible and diverse chemical composition is advantageous and it is reasonable to assume that secondary metabolites are key targets upon which selection acts (Kooke and Keurentjes, 2012, Wu et al., 2018). Gaining insight into the genetic regulation of plant metabolites, being the last layer of the underlying molecular network, provides an alternative strategy to decipher the evolutionary forces that have shaped natural variation in quantitative traits (Chae et al., 2012, Kooke and Keurentjes, 2012).

Indeed, much of the metabolic disparity can be attributed to variation in genetic control and several mapping studies have associated local sequence diversity with metabolic traits and related higher order phenotypes (Chan et al., 2010, Fu et al., 2009, Joseph et al., 2015, Keurentjes et al., 2006, Keurentjes et al., 2008, Kliebenstein, 2009, Yu et al., 2020, Zhang et al., 2020). For instance, metabolic variation has been associated successfully with genomic diversity in genome wide association studies (GWAS) using linear mixed models (LMM), such as EMMAX (Fusari et al., 2017, Kerwin et al., 2015, Wu et al., 2016, Strauch et al., 2015, Li et al., 2020). Such approximate methods have demonstrated their effectiveness as a powerful tool to identify genetic associations, while considering relatedness among samples and accounting for population stratification and other confounding factors (Kang et al., 2010, Korte and Farlow, 2013, Lippert et al., 2011, Listgarten et al., 2010). However, GWA methods rely on frequentist statistics, which suffers from overfitting, bias in the estimation of effect sizes and low power of detecting rare alleles (Josephs et al., 2017, Zhou and Stephens, 2012). These drawbacks hinder the accurate estimation of the effect of genetic variation in an evolutionary context, which can be largely overcome by whole genome regression (WGR) methods, which include all genetic polymorphisms simultaneously into a single statistical model (de Los Campos et al., 2013). WGR provides accurate estimations of effect sizes of genetic polymorphisms, which is instrumental in the elucidation of the genetic architecture of traits and the selective forces acting on this.

Here, we aimed to determine the relative abundance of volatile and non-volatile secondary metabolites in a large global population of Arabidopsis accessions and provide accurate estimates of the effect of genetic polymorphisms on variation in these phytochemical profiles by applying a Bayesian WGR model (Moser et al., 2015). We further provide insight into the genetic architecture of metabolic traits by demonstrating that genes involved in metabolic biosynthesis pathways are the prime targets of evolutionary forces. Finally, we reveal that local adaptive processes and climatic variables shape global variation in plant secondary metabolism.

Results

Natural variation in plant secondary metabolism is driven by pathway diversity

To analyze the extent of natural variation in secondary plant metabolism, a global collection of 359 Arabidopsis accessions (Baxter et al., 2010, Horton et al., 2012, Li et al., 2010) was evaluated in duplicate for metabolic content

in four-week-old rosettes of plants grown under standard conditions in short days. This population has previously been analyzed extensively for numerous morphological, metabolic and stress-related traits (Bac-Molenaar et al., 2015b, Davila Olivas et al., 2017, Fusari et al., 2017, Thoen et al., 2017, Wu et al., 2018, Kooke et al., 2016), although the systematic effect-size of genomic variation and genetic architecture was hardly established in these studies. Here, we investigated plant metabolic content in two independent replicate samples of each accession through both untargeted UPLC-Orbitrap-FTMS and SPME-GCMS based profiling, enabling the detection and relative quantification of non-volatile and volatile secondary metabolites, respectively (Salem et al., 2020, Kooke et al., 2019, Wehrens et al., 2016). A total number of, respectively, 567 and 603 non-volatile and volatile compounds were detected of which many could be putatively annotated using public and in-house databases (Perez de Souza et al., 2017, Tikunov et al., 2012, Kooke et al., 2019, Witjes et al., 2019). Widespread quantitative and qualitative variation in chemical profiles was observed among the genotypes of the population (Figure 1A; Supplemental Table 1). A large number of volatile and non-volatile metabolites were commonly detected in all accessions analyzed (32% and 17%, respectively), while 9% and 16% of the compounds were detected in less than 50 accessions, respectively (Figure 1C). For instance, each accession contained on average eight of the 94 rare non-volatile metabolites, but the accessions Zdr-6 and Mt-0 contained no less than 38 and 39 of these compounds, respectively. The lowest number (228) of non-volatile metabolites was detected in Cvi-0, although this accession contained the highest level of 3-butenyl glucosinolate. This pattern of variation in phytochemical profiles is consistent with earlier findings in a much smaller set of accessions (Keurentjes et al., 2006). Moreover, the variation in abundance of detected metabolites was substantial, with similar median coefficients of variation of 101% and 82% for volatile and non-volatile compounds, respectively. However, the broad-sense heritability (H^2) of non-volatile metabolites was on average much higher than that of the volatile metabolites, possibly due to stronger stochasticity in the release of volatile compounds. Nonetheless, median heritabilities of 40% (volatiles) and 64% (non-volatiles) demonstrate that a large part of the variation in secondary metabolites is heritable (Figure 1B).

With a view to identifying the causal genetic factors, we associated the observed variation in secondary metabolite profiles with ~215K single nucleotide polymorphisms (SNPs) using a state-of-the-art Bayesian WGR method implemented in the Bayes-R software (hereafter referred to as WGR), which explicitly partitions the genetic variance into contributions of loci with large, intermediate or small effects (Moser et al., 2015). Since Bayesian WGR methods do not provide significance measures, a conventional GWAS using the EMMA-X approach (Kang et al., 2010) (hereafter referred to as GWA) was performed on the same data for reference purposes. Because WGR provides only SNP effect estimates, we set a conservative arbitrary effect size threshold of $|\beta| > 0.01$ for SNPs substantially contributing to explained variance, although the exact explained variance in metabolite accumulation depends also on the minor allele frequency (MAF) and total variance.

For roughly half of the 567 non-volatile metabolites (53%) at least a single SNP was detected that explained a substantial part of the variability (WGR: 300 metabolites (SNP effect size $|\beta| > 0.01$); GWA: 303 metabolites (SNP significance $P_{BONF} < 0.05$)). The proportion of the 603 volatile metabolites for which variation could be associated to

one or more SNPs was substantially lower (WGR: 157 metabolites (26%, $|\beta| > 0.01$); GWA: 134 metabolites (22%, $P_{BONF} < 0.05$)) (Supplemental Table 2). The difference in mapping power between the two types of metabolites reflects their contrast in heritability, while the discrepancy between the WGR and GWA approach might be an effect of low minor allele frequencies. Rare large-effect loci are better detected using WGR while the power to detect true positives with LMM GWAS declines with decreasing MAFs. The latter would suggest that variation in volatile metabolites, in contrast to variation in non-volatile metabolites, is more frequently driven by rare alleles. The overlap in metabolites for which variation could be associated to genetic variation by both WGR and GWA was 83% for the non-volatiles and 57% for the volatiles, emphasizing the differences between metabolite types and mapping approach.

To assign genomic functions causal for the observed variation in metabolic profiles, candidate genes in linkage disequilibrium (LD) with the large-effect SNPs detected by WGR ($|\beta| > 0.01$) were identified. For many metabolites multiple QTLs were detected, often represented by series of large-effect SNPs in LD. In total, 1,097 unique candidate genes explaining variation in the 301 non-volatile metabolites were assigned, whereas 651 candidate genes were assigned explaining variation in the 154 volatile metabolites (Supplemental Table 2). Only 120 of the assigned genes explain variation in both non-volatiles and volatiles, illustrating that there is little overlap in the genetic regulation of volatile and non-volatile secondary metabolites.

At a number of loci, variation in several metabolites mapped to the same candidate gene. A QTL hotspot was defined as a gene locus substantially associated with variation in at least five metabolites by a minimum of two large-effect SNPs ($|\beta| > 0.01$). This criterion was met by nine loci after WGR analysis of non-volatile metabolites and three loci for volatile metabolites, with a small overlap of two hotspots detected with both platforms (Supplemental Table 2, Figure 2A). With the exception of *ACD6* and *GLABRA1*, which are involved in trade-offs between growth and defence (Fusari et al., 2017, Todesco et al., 2010) and trichome development (Wang et al., 2019, Herman and Marks, 1989, Marks and Feldmann, 1989), respectively, all assigned hotspot candidate genes are directly related to the biosynthesis of secondary metabolites. Most prominent is the glucosinolate metabolism (Figure 2C), which is represented by three QTL hotspots, explaining variation in 82 metabolites. While variation in various non-volatile intact glucosinolates maps to all three QTL hotspots, variation in the well-known volatile glucosinolate breakdown products, such as nitriles and isothiocyanates (Hansen et al., 2008, Textor et al., 2004, Kliebenstein et al., 2001) maps predominantly to the *AOP* and *MAM* loci. Other QTL hotspots mainly represent phenolic compounds, such as phenylpropanoids and flavonoids (Goujon et al., 2003, Ishihara et al., 2016, Kim et al., 2004, Ross et al., 1999). The analysis of QTL hotspots illustrates that many secondary metabolites are regulated simultaneously by natural variation in a small number of biosynthesis genes upstream of branching points in metabolic pathways. However, enrichment analysis demonstrates that many more genes involved in secondary metabolic processes are assigned as candidate genes explaining variation in non-volatile metabolites than can be expected by chance (43 genes, $P_{BENJAMINI} < 0.01$) (Supplemental Table 3, Figure 2B) (Huang et al., 2008). A similar observation was made for candidate genes

177 explaining variation in volatile compounds (Supplemental Table 3, Figure 2B). In addition, many genes assigned to
178 explain variation in specific annotated compounds, after WGR analysis, are directly involved in their biosynthesis
179 pathway (Supplemental Table 4). These results strongly suggest that, although variation in a few genes has a large
180 broad-spectrum effect on metabolic profiles, the widespread natural variation in plant secondary metabolite content
181 is largely driven by variation in a high number of specific biosynthesis genes, consistent with a scale-free model of
182 biological networks. Moreover, these analyses provide statistical evidence that many of the associations detected by
183 WGR mapping are true positives, indicating that variation in a large number of secondary metabolite biosynthesis
184 genes is maintained in nature by selective processes (Bac-Molenaar et al., 2015a).
185
186 **The role of genetic heterogeneity and inconsistent impact in controlling metabolic diversity**
187 A large advantage of GWR over conventional GWA approaches is the ability to assess more precisely the effect size
188 (β) of each SNP as a proportion of the total genetic variance (σ^2_g) explaining quantitative variation in traits. To gain
189 insight into the genetic architecture regulating the variation in metabolic content, SNPs were partitioned into
190 categories of uninformative ($\beta = 0\sigma^2_g$), small ($|\beta| > 0.0001\sigma^2_g$), intermediate ($|\beta| > 0.001\sigma^2_g$) and large ($|\beta| > 0.01\sigma^2_g$)
191 effect sizes for each metabolite using GWR (Moser et al., 2015). Although the correct estimation of SNP effect sizes
192 is compromised in GWA mapping by dependency on MAF, a significant correlation was observed between the effect
193 sizes of the most informative SNPs obtained by WGR and GWA for both volatiles ($R^2 = 0.33$) and non-volatiles ($R^2 =$
194 0.65) (Supplemental figure 1). Such a relationship was absent when all informative SNPs were included. This is partly
195 due to the strong LD between many SNPs at particular loci and emphasizes the importance of bias reduction, as
196 applied in WGR. Importantly, regardless of the mapping approach followed, the same candidate genes, assumingly
197 causal for the observed variation in metabolite abundance, were generally assigned.
198 The observed discrepancy in heritability and number of detected QTLs between volatiles and non-volatiles is also
199 reflected in the division of SNP effect sizes. On average, the number and proportion of detected large-effect SNPs
200 per metabolite is higher for non-volatiles (*nV*) than for volatiles (*V*): (n_{nV} = 1292.8 ($|\beta| > 0.0001$), 52.8 ($|\beta| > 0.001$),
201 3.5 ($|\beta| > 0.01$) vs. n_V = 1225.4 ($|\beta| > 0.0001$), 32.9 ($|\beta| > 0.001$), 1.8 ($|\beta| > 0.01$) (Supplemental table 2). In total,
202 52% of the genetically explained variation in non-volatile metabolites is determined by large-effect SNPs, whereas
203 this is only 38% for volatiles. In comparison to other metabolites, the genetically explained variation of metabolites
204 for which a QTL ($|\beta| > 0.01$) was detected was on average more strongly determined by SNPs of large effect (*nV*: 65%
205 *vs.* 38%; *V*: 56% *vs.* 32%) with higher maximum absolute SNP effect sizes (*nV*: $|\beta|$ = 0.061 *vs.* 0.004; *V*: $|\beta|$ = 0.050
206 *vs.* 0.006).
207 However, analogous to the differences in regulation between volatiles and non-volatiles, large inconsistencies do
208 also occur between and within metabolic classes. For instance, 68% of the genetically explained variation in aliphatic
209 glucosinolates is accounted for by large-effect SNPs, whereas this is only 57% and 53% for phenolics and volatile
210 glucosinolate derivatives, respectively. In addition, within the metabolic class of aliphatic glucosinolates, 85% of the

genetic variation in short-chain C3-OH glucosinolates is explained by large-effect SNPs, in contrast to 31% for the long-chain C5-C8 MT/MS glucosinolates (Supplemental Figure 2).

The analysis of the regulatory landscape of plant secondary metabolism revealed that many metabolites are regulated by a majority of unique biosynthesis genes and a few hub genes, controlling variation in the abundance of numerous different metabolites. We investigated this observation further by accounting for the effect sizes of SNPs and their associated candidate genes in regulating plant metabolic content. We defined major genes as genes associated with QTLs represented by SNPs with extraordinary high effect sizes ($|\beta| > 0.05$) and a high likelihood of contributing to the explained variance (Bayes posterior inclusion probability (PIP) > 0.5). As such, 55 major QTLs controlling variation in non-volatiles were detected including eleven of the twelve QTL hotspots (Supplemental Table 5). Similarly, 39 major QTL controlling variation in volatile metabolites were detected, among which the QTL hotspots *AOP* and *MAM* (Supplemental Table 5). Strikingly, non-volatile metabolites (117, *i.e.,* 20.6%) are much more often regulated by major-effect QTLs than volatiles (35, *i.e.,* 5.8%). The variation in non-volatile aliphatic glucosinolates and volatile glucosinolate derivatives, for instance, is often determined by one to three QTLs with a major effect and a number of other loci with moderate effects. In contrast, phenolics are mostly regulated by one major QTL and a variety of smaller-effect loci (Supplemental Tables 2, 4 and 6). These analyses highlight the differences in genetic architecture controlling the variation in metabolic content and argues for a power-law distribution of effect sizes in metabolic networks, suggesting that the genetic regulation of metabolic variation leans more towards an additive than an infinite model.

An important topic in the evolution of species is whether adaptive traits have spread through natural populations by drift, recombination and selection of a single alteration or by convergent evolution, in which similar effects are repeatedly obtained by multiple independent mutation events. We addressed this issue by investigating genetic and allelic heterogeneity in the regulation of metabolic content. For this, we performed a detailed analysis of the major-effect WGR QTLs, which are represented by multiple large-effect SNPs ($|\beta| > 0.01$). Notably, collocating SNPs often differ markedly in effect-size and MAF, and sometimes are not even in LD with each other. This strongly suggests that different haplotypes of particular loci exist within the investigated population, with unequal selection forces acting on the distinctive SNPs. Exemplary, at a number of hotspot QTLs the genetic variation explaining the abundance of a specific metabolite is often determined predominantly by a single major-effect SNP and several smaller-effect SNPs. However, this major-effect SNP is not necessarily the same for other metabolites controlled by the identical hotspot QTL (Supplemental Table 2). So, where independent mutation events at a single locus might lead to allelic heterogeneity in the regulation of one metabolite it might cause haplotype diversity in the regulation of another and, depending on the adaptive value of one or the other metabolite, results in different selective forces on sequence variation.

A different pattern, suggesting alternative modes of regulation, emerges from the analysis of the QTL hotspot at the bottom of chromosome four, explaining variation in 22 non-volatile metabolites. This QTL is represented by two

large-effect SNPs in the candidate gene *ACD6*, which are not in LD with each other (R^2=0.11) (Supplemental Figure 3A). However, the first SNP is in strong LD with various polymorphisms in the *ACD6* promoter, while the second SNP is in strong LD with non-synonymous SNPs in the last exon of the coding region, resulting in the existence of four different haplotypes. Strikingly, 68% of the accessions belonged to haplotype IV, whereas 25%, 6%, and 1% of the investigated accessions belonged to haplotypes III, I, and II, respectively. Haplotype analysis of variation in a representative metabolite controlled by this locus revealed that this metabolite was detected in 90% (18/20) of the accessions of haplotype I, whereas this was only the case for 40% (35/88) and 14% (33/237) of the accessions of the haplotype classes III and IV, respectively. This metabolite could not be detected in any of the accessions (0/2) belonging to haplotype class II (Supplemental Figure 3B). Finally, the abundance of this metabolite was substantially higher in haplotype class I than in any of the other haplotype classes (Supplemental Figure 3C). This indicates an intragenic epistatic interaction effect between the two gene domains. The non-synonymous gene body SNPs might determine the functionality of the gene product and are epistatic over the promoter SNPs, which, in turn, control the level of transcription and, hence, the quantity of the gene product. The variation between haplotypes in the efficiency to produce and accumulate specific metabolites strongly suggests that selective forces have established the observed differences in the global population haplotype frequencies.

Likewise, variation in nine non-volatile metabolites is associated with three SNPs collocating at the *UGT89A2* locus, which together explain up to 37% of the genetic variance in a mixed linear model although none of the SNPs explains more than 24% for any metabolite on their own (Supplemental Table 7). In this case two of the three SNPs exert similar effects on the phenotype, again suggesting allelic heterogeneity, while a specific combination of genotypes is only observed in one accession, suggesting strong negative selection on that haplotype. These analyses indicate that multiple evolutionary events can shape the allelic diversity of genes and might act in concert to obtain distinctive phenotypic levels of metabolites. In addition, it demonstrates that, in these cases, selection acts on haplotypes, determined by multiple SNPs, rather than on single unique SNPs.

In other instances, multiple large-effect SNPs at a single locus are associated to more than one likely candidate gene, suggesting genetic heterogeneity or even channelling of biosynthesis genes (Witjes et al., 2019). The QTL hotspot at the top of chromosome five, explaining variation in the highest number of metabolites, predominantly glucosinolates, links to three genes (*MAM*, *TT16* and *DMR6*), which all might be involved in metabolite biosynthesis (Falcone Ferreyra et al., 2015, Kroymann et al., 2001, Xu et al., 2017). Variation in the majority of metabolites that map to this locus is best explained by large-effect SNPs that are most strongly associated with *MAM*, although quantitative variation in a minor number of metabolites is stronger linked to large-effect SNPs at the position of *TT16* or *DMR6* (Supplemental Figure 4). Moreover, both *TT16* and *DMR6* are not in strong LD with *MAM* (R^2 < 0.23). A similar observation can be made for metabolite variation mapping to the QTL hotspot at the top of chromosome four (Supplemental Figure 5). Here, variation in metabolite abundance is most strongly associated to large-effect SNPs at the position of either the *AOP* gene or *DAAR1*, both of which have been implicated in metabolite biosynthesis

(Kliebenstein et al., 2001, Strauch et al., 2015). Finally, variation in at least 15 unannotated non-volatile metabolites is associated with no less than seven independent SNPs co-locating at the SAMT/BAMT locus at the lower arm of chromosome five, together explaining up to 57% of the genetic variation in a mixed linear model (Supplemental Table 8). Two of these SNPs are in close proximity and share similar effects on a number of metabolites but display very different MAFs and are not in LD with each other (R^2 = 0.07), again suggesting allelic heterogeneity. The other five SNPs are more separated and span a region of multiple genes with different metabolic conversion activities, including S-adenosylmethione-dependent methyltransferases (*AT5G37970, AT5G37990* and *AT5G38020*), an UDP-glycosyl transferase (*AT5G38010*), oxidoreductases (*AT5G37940*, *AT5G37960*, *AT5G37980* and *AT5G38000*), an arabinosidase (*AT5G37920*) and a general transferase (*AT5G37950*). The metabolites mapping to this locus are not consistent in the SNPs with the strongest associations and thus might be regulated by different genes in this region.

These results illustrate that WGR mapping can be instrumental in disentangling the genetic architecture of metabolic trait regulation by providing a higher resolution of assigning candidate genes to detected genomic associations, through inclusion of accurate SNP effect-size estimates.

Polygenic adaptation of plant secondary metabolism to local settings

Evolutionary theory predicts that selective forces shift allele frequencies of non-neutral genetic variants in admixed populations (Walsh and Lynch, 2018) and there is no reason to expect other effects on natural variation in secondary metabolism. Indeed, a positive and significant correlation was observed between the BayesR effect-size $|\beta|$ and MAF of the SNPs most strongly associated ($|\beta|> 0.01$ and $PIP_{large} > 0.5$) with variation in the accumulation of both volatile (R^2=0.32) and non-volatile (R^2=0.20) secondary metabolites (Supplemental Figure 6). Such a correlation was absent when all, mostly neutral, SNPs were taken into account, suggesting that balancing selection is a strong driver of maintaining natural variation in secondary metabolism (Benderoth et al., 2006). Following this reasoning it is plausible to assume local adaptation, which should be reflected in the fixation index (F_{ST}) of selective loci (Holsinger and Weir, 2009). As expected, SNP F_{ST} values calculated for five European regions correlate positively with the effect-size of the SNPs most informative ($|\beta|> 0.01$ and $PIP_{large} > 0.5$) for variation in volatile and non-volatile metabolites (R^2=0.42 and 0.22, respectively) (Supplemental Figure 7). Illustratively, approximately 4% of all investigated SNPs pass the F_{ST} threshold value of 0.2, generally considered as indicative for selection, while almost 13% of the SNPs with an effect-size $|\beta|> 0.01$ and PIP > 0.5, explaining variation in non-volatile abundance, meet this criterion. This proportion was even higher for SNPs explaining variation in volatiles (21.9%), although the number of qualifying SNPs was much lower. Moreover, a strong enrichment ($P < 0.005$) of 160 genes involved in secondary metabolism was observed among the list of candidate genes assigned to SNPs with F_{ST} values above 0.2, compared to the 132 genes expected by chance. In addition, strong selection ($0.21< F_{ST} <0.44$) was observed for half the number of QTL-hotspots, including all three loci explaining variation in glucosinolate metabolism, with large-effect SNPs in the 0.05‰ extreme end of the F_{ST} distribution ($F_{ST} > 0.3$). Other loci explaining variation in glucosinolate, phenylpropanoid, and flavonoid

accumulation also displayed high F_{ST} values (Supplemental table 9). Together, these analyses illustrate that local adaptation is an important driver for the evolution of natural variation in secondary metabolism.

The high F_{ST} values observed for some SNPs suggests the existence of population structure and a relationship between metabolic profiles and the geographic origin of accessions. To investigate this further a two-dimensional hierarchical clustering was performed on the geographic distribution of analyzed accessions and the accumulation of detected compounds therein. Not surprisingly, given their strong enrichment in the phytochemical profile, aliphatic glucosinolates and their volatile derivatives were among the strongest determinants of genotype clustering (Supplemental Figure 8A). The aliphatic glucosinolates were further categorized on chain length and side-chain modification and based on this classification four main groups of genotypes, each with a different organization of glucosinolate biosynthesis alleles and enriched in accessions from specific geographic regions, could be distinguished (Supplemental Figure 8B). For instance, alkenyl and long-chain glucosinolates group closely together and are, on average, detected at high levels in the majority of genotypes from the UK and France (group 1). From the 105 accessions within this group, 45% is of French origin (*i.e.,* 77% of all French accessions) while 27% originates from the UK (*i.e.,* 60% of all accessions from the UK) (Supplemental Table 1). The qualitative and quantitative variation in glucosinolate content is largely explained by allelic variation in five biosynthesis genes (*viz. MAM*, *AOP*, *TT16*, *DMR6*, and *GS-OH*). Both the *TT16* locus and the *DMR6* locus have not been reported earlier in glucosinolate studies, probably due to the close vicinity to *MAM*. Both loci are, however, not in very strong LD with *MAM* ($R^2 < 0.23$). A candidate gene for *TT16* might be coding for a succinyl CoA ligase involved in the TCA cycle (AT5G23250), or a nicotinamidase involved in coniferin metabolism (AT5G23220/30). The *DMR6* gene is co-expressed with *GS-OH*, shares protein domains with *AOP* and *GS-OH* and exhibits, like *AOP* and *GS-OH*, oxidoreductase activity (Warde-Farley et al., 2010). Given the variation in C4-8 alkenyl glucosinolates explained by *DMR6*, the loss of an active *GS-OH* allele might be compensated by a similar function of *DMR6*. To test this hypothesis, a population of doubled haploid (DH) lines was generated from a cross between the accessions UKNW06-460 and Tamm-2 and subjected to metabolic analyses. Unlike the reference accession Col-0, which lacks expression of functional AOP (Kliebenstein et al., 2001), these accessions share functional *AOP* and *MAM* genes, necessary for the production of 3-butenyl, which is the precursor for the synthesis of 2-hydroxy-3-butenyl. However, UKNW06-460 and Tamm-2 differ in functionality of the *GSOH* and *DMR6* loci. Tamm-2 lacks functional alleles of both *GSOH* and *DMR6* and is unable to synthesise 2-hydroxy-3-butenyl despite the fair amount of available 3-butenyl, whereas UKNW06-460 contains both a functional *GSOH* and *DMR6* allele, which allows it to produce 2-hydroxy-3-butenyl in addition to 3-butenyl (Supplemental figure 9A). Recombinant DH lines were divided in four haplotype classes based on functionality of the *GSOH* and *DMR6* alleles. With a few exceptions, haplotypes lacking a functional *DMR6* allele (haplotype I and II) did not produce detectable levels of 3-butenyl and 2-hydroxy-3-butenyl. Haplotypes containing a functional allele of *DMR6* (haplotype III and IV), however, were able to synthesise high levels of both compounds irrespective of the presence of a functional *GSOH* allele (haplotype IV) or not (Haplotype III), although the production of 2-hydroxy-3-butenyl is much higher when both genes are functional (Supplemental figure 9B). While the model of redundant functions of the *GSOH* and *DMR6* gene

does not unambiguously explain the observed metabolic profiles in natural accessions and DH lines, possibly due to additional segregating genetic modifiers, metabolic feedback, genotype mis-calling or metabolite mis-annotation, the results clearly indicate similarities in function of the *GSOH* and *DMR6* genes.

Group 1 is largely represented by haplotypes containing the Columbia (Col-0) reference allele for *MAM* and *AOP* and the non-Col-0 *TT16* allele, combined with either the non-Col-0 *GS-OH* or *DMR6* allele (Figure 3). A second group consists mainly of accessions from Sweden and the USA, of which the majority contain high levels of C3 and C4 alkenyl glucosinolates. The accumulation of these glucosinolates is determined by a cooperation of the *AOP* and *MAM* loci and haplotypes carrying the Col-0 *AOP* allele and the non-Col-0 *MAM* allele, synthesizing the highest levels of alkenyl glucosinolates, dominate this group (Figure 3). Two other, North-East European, groups are characterized by an overrepresentation of accessions from the Czech Republic (group 3) and Germany (group 4), respectively. Haplotypes in the Czech group share the non-Col-0 *AOP* allele, which is associated with high levels of C3-hydroxyl glucosinolates, while haplotypes in the German group share the Col *TT16* allele and produce large amounts of C4-methylthio/methylsulfinyl glucosinolates (Figure 3). These analyses illustrate that glucosinolate profiles are strongly determined by local adaptive processes that have shaped the allelic diversity at various biosynthesis loci. These loci explain on average a large proportion (41%) of the observed variation in aliphatic glucosinolates (Supplemental Table 10) and, depending on the specific haplotype, result in a semi-qualitative distribution in almost discrete classes of accumulation (Figure 3). For most aliphatic glucosinolates this variation is almost completely explained by additive effects of contributing loci, while epistatic interactions play a very minor role (Supplemental Table 10).

Additive effects increase the selection coefficient of specific loci, which drives the co-evolution of certain allele combinations in particular environments, as was previously observed for glucosinolate metabolism (Brachi et al., 2015). In line with this, the occurrence of some haplotypes deviates significantly from estimates based on population allele frequencies. For instance, the non-Col-0 *GS-OH* allele almost exclusively occurs together with the Col-0 *MAM* and *AOP* allele, while the combination of the nonCol-0 *AOP* allele and the Col-0 *MAM* allele is extremely rare (Figure 3). Interestingly, a sixth locus, encoding an epithiospecifier protein (*ESP*) involved in glucosinolate breakdown and explaining a large fraction of the variation in C4-8 alkenyl glucosinolates acts hypostatic to the *GS-OH* and *MAM* loci (Supplemental Table 11). A co-evolution of *ESP* with *GS-OH* and *MAM*, and to a lesser extent *DMR6*, is further suggested by moderate to high levels of LD between SNPs representing the various loci, even though located on different chromosomes (Supplemental Figure 10). The active non*Col-0* ESP allele is found in all, but one accessions from France and the UK (Supplemental Table 12), suggesting a favourable production of nitriles over isothiocyanates in these regions. In contrast, the vast majority of accessions from the Czech Republic carry the weak *Col-0* allele, further confirming strong population structure for the accumulation of these compounds.

Analogous to the glucosinolates and their volatile derivatives many other metabolites, including phenylpropanoids and flavonoids, display strong correlations with geographical clines and climate parameters (Supplemental Table 13). In addition, the loci explaining most of this variation, like *BGLU6* and *UGT78D2*, exhibit high F_{ST} values and a strong correlation of allelic variation with latitude and longitude (Supplemental Tables 9 and 13).

385 This study demonstrates that natural variation in plant secondary metabolism is governed by allelic variation in many
386 biosynthesis genes, of which the effect size could be accurately determined by Bayesian statistics. A substantial part
387 of the allelic diversity is likely shaped by local adaptation to resident climates with strong selection for specific
388 haplotypes and metabolic profiles.
389
390

391 ## Discussion

392 In this study we provide evidence of selective forces acting on adaptive allelic variation in biosynthesis genes of
393 secondary metabolites, shaping regional variation in phytochemical profiles. Moreover, we were able to estimate
394 accurately the effect size of genetic variants and establish the genetic architecture of the regulatory landscape of
395 plant secondary metabolism. Using state-of-the-art untargeted metabolomics more than a thousand volatile and
396 non-volatile secondary metabolites could be comparatively quantified in a global collection of more than 350
397 Arabidopsis accessions. For most of these compounds, heritable qualitative and quantitative variation was observed
398 and for approximately half of the non-volatile, and a quarter of the volatile metabolites, this variation could be
399 explained by genetic variation in specific genomic loci, according to a Bayesian WGR approach. A majority of these
400 loci could be tied to candidate genes involved in the biosynthesis pathway of the metabolite under study, suggesting
401 that most of the identified associations are true positives. Furthermore, many metabolites mapped to the same
402 genomic loci, resulting in a total of ten hotspots explaining variation in no less than 44 and 120 volatile and non-
403 volatile metabolites, respectively. Surprisingly, very little overlap was observed in QTL hotspots and candidate genes
404 assumed to be causal for the observed variation in volatile and non-volatile metabolites. This suggests an
405 independent regulation of these two types of compounds, which can be explained by regulation *in cis* of key
406 biosynthesis genes of specific metabolites or modules rather than a systematic overarching regulation *in trans* of
407 whole plant secondary metabolism.
408 The WGR enabled estimation of the effect-size of genetic variation allowed the classification of contributing loci and
409 a more detailed analysis of specific loci. Apparently, one-third to half of the secondary metabolites is regulated by
410 large-effect polymorphisms, although large differences occur between and within metabolic classes. This supports
411 the generally accepted view that the genetic basis of plant secondary metabolism is founded on a few large-effect
412 loci, numerous small effect-loci, and genetic and environmental interactions (Olson-Manning et al., 2012).
413 Interestingly, on several loci, multiple independent mutation events could be detected, resulting in a variety of
414 different haplotypes in the global Arabidopsis population. Although allelic or genetic heterogeneity might occur,
415 signatures of selection indicate that, depending on the geographic origin of the accessions, unequal selective forces
416 act on the uncovered genetic variation, resulting in different effects on the accumulation of metabolites. These
417 findings correspond well with the observed dependency of fitness effects of metabolic profiles on field conditions
418 (Kerwin et al., 2015). Indeed, a positive correlation was detected between the effect-size, F_{ST} value and allele
419 frequency of genetic variation and a strong enrichment of QTL hotspots and candidate genes involved in secondary

metabolism with high F_{ST} values. The data presented here support the hypothesis that natural variation in secondary metabolism is maintained by fluctuating or balancing selection throughout the long evolutionary history of Arabidopsis (Kliebenstein et al., 2016).

The relationship between the geographic origin of accessions and their metabolic profile can, at least for aliphatic glucosinolates, largely be explained by additive effects of genetic variation. This is consistent with reports of allelic variation for morphological and physiological traits being correlated with climate and geographical variables (Baxter et al., 2010, Dubin et al., 2015, Hancock et al., 2011, Kooke et al., 2016, Li et al., 2010). In addition, the observed frequency of haplotypes differs from expectations of neutral theory, indicating local adaptation and selection of favourable combinations of alleles of different genes. Moreover, significant LD could be detected between loci involved in the joint regulation of specific metabolites, even when these are located on different chromosomes, speculating on the co-evolution of genes establishing biosynthesis pathways. Co-evolution and epistatic selection have previously been suggested for important loci controlling glucosinolate metabolism, such as *GS-OH*, *AOP* and *MAM* (Brachi et al., 2015), and our data suggest that other genes, such as *DMR6, ESP* and *CYP79F2*, have undergone similar selective processes and might have an impact on fitness as well. Interestingly, recent analyses indicate a role for epigenetic regulation of metabolic content as well, which, in contrast to the irreversible adjustments described here, might provide a more flexible adaptation to alternating conditions (Kooke et al., 2019).

Our results provide compelling additional evidence that natural variation in the accumulation of secondary metabolites in *Arabidopsis thaliana* is governed by local adaptation through evolutionary selection of polygenic variation.

Material and Methods

Plant growth conditions

Seeds of 359 Arabidopsis natural accessions belonging to the HapMap panel (Horton et al., 2012, Li et al., 2010) were sown on filter paper with demi water, stratified at 4˚C in dark conditions for five days and transferred to a culture room (16h LD, 24˚C) for 42h to induce seed germination. Eight replicates per accession were transplanted to wet Rockwool blocks of 4x4 cm in a completely randomized block design in a climate chamber. Chamber climate conditions were as follows: 12 h short days, light intensity 125 $\mu mol.m^{-2}.s^{-1}$, temperature 20°C day/18°C night, relative humidity 70%. All plants were watered daily for five minutes with 1/1000 Hyponex solution (Hyponex, Osaka, Japan). Two replicates of each accession were harvested in bulk to serve as reference material in metabolite analyses. The other plants were harvested in two replicate pools of three plants 28 days after sowing. The harvest time was set to the end of the light period.

Metabolomics

Samples were ground in liquid nitrogen and an aliquot of all samples was mixed to generate a large pool needed for preparing the quality controls (QCs). These were independently and simultaneously weighed and extracted with the study samples (5–6 times per batch) and injected at regular intervals within the analysis series.

For the detection of non-volatiles, aqueous-methanol extracts were prepared from 50 mg frozen ground material to which 200 µl of 94 % MeOH containing 0.125 % formic acid was added (De Vos et al., 2007). Batch sizes ranged from 78 to 80 samples, with the exception of the last batch, batch 10, containing 48 samples (Wehrens et al., 2016). After sonication and filtering, the crude extracts were analysed as described previously (van Duynhoven et al., 2014) using UPLC (Waters Aquity) coupled to a high-resolution Orbitrap FTMS (Thermo). A 20 min gradient of 5–35 % acetonitril, acidified with 0.1 % formic acid, at a flow rate of 400 µl/min was used to separate compounds on a 2.1 x 150 mm^2 C18-BEH column (1.7 µm particle size) at 40 °C. Metabolites were detected using a LTQ-Orbitrap hybrid MS system operating in negative electrospray ionization mode heated at 300 °C with a source voltage of 4.5 kV. The transfer tube in the ion source was replaced and the FTMS recalibrated after each sample batch, without stopping the UPLC system. After pre-processing raw data files in an untargeted manner using a Metalign-MSClust based workflow (Kooke et al., 2019, Lommen and Kools, 2012, Tikunov et al., 2012), metabolites occurring in fewer than 20 different genotypes were removed, leading to a data matrix containing relative intensities of 567 reconstructed metabolites in 761 samples (including 51 QCs). The percentage of non-detects (*i.e.*, an intensity value below the arbitrarily set detection threshold) in this matrix is 48 %. For individual metabolites, the fraction of non-detects can be much larger, and in this data set is up to 97 %.

The detection of volatiles is based on aliquots of the same Arabidopsis material as described for the non-volatiles. The aim here was to analyse volatile organic compounds (VOCs) present in the leaf material using solid phase microextraction (SPME) of the headspace. Extracts of 50 mg from frozen ground material were analysed on a GC–MS system (Agilent GC7890A with a quadrupole MSD Agilent 5978C) in fifteen batches of 34-99 samples, with, on average, 15 study samples per QC, as described earlier (Cordovez et al., 2015, Mumm et al., 2016, Verhoeven et al., 2012). In contrast to these studies, the temperature program of the GC oven started at 45 °C (2 min hold) and rose first with 8–190 °C.min^{-1}, followed by 25–280 °C (2 min hold). This data set contains information on 753 samples (including 50 QCs) with, in total, 40 % non-detects, similar to what was found for the non-volatiles. For individual metabolites, the percentage of non-detects goes up to 97 %. As for the non-volatiles, only those volatile metabolites were retained that were present in at least 20 different genotypes, in this case 603 metabolites.

In all subsequent analysis, log-transformed metabolite intensity values were used. Accessions in which the relative abundance of the assayed metabolite did not pass the detection threshold were assigned the threshold value of 3.0 and 0.45 for non-volatiles and volatiles, respectively.

Descriptive statistics

The variance components for all the individual traits were used to calculate the broad-sense heritability, H^2, in analysis of variance (ANOVA) according to the formula $H^2 = \sigma^2_G / (\sigma^2_G + \sigma^2_E)$, with $\sigma^2_G = (MS(G) - MS(E))/r$, $\sigma^2_E = MS(E)$,

where r is the number of replicates and $MS(G)$ and $MS(E)$ are the mean sums of squares for genotype and residual error, respectively. Narrow-sense heritability, h^2, was defined as $h^2 = \sigma^2_A/(\sigma^2_G + \sigma^2_E)$, which takes only the additive genetic effects (σ^2_A) in account. Marker-based estimates of narrow-sense heritability were obtained from the mixed model, $y_{i,j} = \mu + G_i + E_{i,j}$, $(i = 1,\ldots,n, j = 1,\ldots,r)$, $G \sim N(0, \sigma^2_A K)$, $E\_i,j \sim N(0, \sigma^2_E)$, where $y_{i,j}$ is the phenotypic response of replicate j of genotype i, μ is the intercept, $G = (G_1,\ldots,G_n)$ is the vector of random genetic effects, and the errors $E_{i,j}$ have independent normal distributions with variance σ^2_E which is the residual variance for a single individual (Kruijer et al., 2014). The vector G has a multivariate $N(0, \sigma^2_A K)$ distribution, and the genetic relatedness matrix K is estimated from standardized SNP-scores. The coefficient of variation (CV_G) was calculated as $CV_G = (\sigma_G/\bar{X})*100\%$.

Genome-wide association mapping

All accessions were genotyped with 214,051 SNPs (Li et al., 2010) of which, after removal of SNPs with a minor allele frequency (MAF) < 0.05, 199,589 were used for genome wide association mapping.

All traits were analysed with a Bayesian statistical model, BayesR, which uses a Gibbs sampling approach to estimate variant effects that are modelled as a mixture distribution of four normal distributions. SNPs were assigned a prior variance of $0\sigma^2_g$, $0.0001\sigma^2_g$, $0.001\sigma^2_g$, and $0.01\sigma^2_g$, across the four distributions, respectively, where σ^2_g is the total genetic variance. This allows many uninformative variants to be dropped from the model and permits remaining variants to have moderate to large effects. Gibbs sampling was performed for 50,000 iterations, after 20,000 burn-in iterations. Conventional GWAS was performed using the R-package statgenGWAS (https://github.com/Biometris/statgenGWAS/), following the approach of previous studies (Kang et al., 2010, Kruijer et al., 2014).

Enrichment analysis

Enrichment analysis was performed using the functional annotation tool in DAVID 6.8 (https://david.ncifcrf.gov/home.jsp) (Huang et al., 2008) for the candidate gene lists of the volatile and non-volatile compounds separately. The gene lists were run against a background set of Arabidopsis TAIR IDs. The gene ontology category GOTERM_BP_FAT was explored and the significance threshold for enrichment was set at $P < 0.01$.

Fixation index

A large subset of the global Arabidopsis population was divided into five sub-populations based on the countries of origin that were best represented in the collection, *viz.* Germany (57 accessions), France (56 accessions), Sweden (48 accessions), UK (47 accessions), and the Czech Republic (29 accessions). These accessions together comprised 68% (237 out of 350 genotypes) of the entire population. F_{ST} values were calculated using the following formula: $F_{ST} = (\pi_{Between} - \pi_{Within})/\pi_{between}$, where $\pi_{Between}$ and π_{Within} represent the expected heterozygosity (*i.e.,* genetic diversity) between populations and the expected heterozygosity within populations, respectively.

Validation of *GSOH*/*DMR6* redundancy

To test for redundancy of the *GSOH* and *DMR6* gene functions three accessions were selected that differed in functional alleles for these two genes and *AOP*. Col-0 (CS76113) does not express functional alleles of *AOP* and is unable to synthesize 3-butenyl, the precursor for the synthesis of 2-hydroxy-3-butenyl (Kliebenstein et al., 2001). UKNW06-460 (CS76279) carries both a functional *GSOH* and *DMR6* gene, whereas Tamm-2 (CS76244) lacks functional alleles for both genes. From a cross between Tamm-2 (♀) and UKNW06-460 (♂) 88 double haploid lines were generated (Filiault et al., 2017), which were grown in a climate-controlled growth chamber for 28 days in short-day conditions (12 h, 125 µM, 70% RH, 20/18°C D/N). After harvesting, plants were snap-freezed and stored in -80°C upon further analysis. Subsequently, all DH lines were genotyped with KASPar™ genotyping technology (KBiosciences), distinguishing the parental haplotype for *GSOH* and *DMR6* at two SNP-positions for each gene. Primers were developed on the following sequences: GS-OH_1 (Chr2: 10799538 bp) CS76244/CS76279 > gagagacttctcaattcaaaaacagacatggaggatctta/gtagcgagactaaatcaagagacagcggtgaaggaatatc; GS-OH_2 (Chr2: 10832548 bp) CS76244/CS76279 > tgtcaggatcaaattcaatttcaataaccagtcacaatatctttcaatta/tctggaatttctatctacatttagcataacctttcatattatttttttacac; DMR6_1 (Chr5: 8379646) CS76244/CS76279 > atctaagactccattgttatcctatccacaagtatgtcc/aatgagtggccgtcaaaccctccttctttcaagtaagca; DMR6_2 (Chr5: 8366765) CS76244/CS76279 > gttggtccagacttacctaataggctaatgcctta/ttcattggagtttgctgatcttatgatgacgtttt, according to standard protocol (Smith and Maughan, 2015). DH lines were classified in one of four different genotypic classes (haplotype I: *GSOH* inactive/*DMR6* inactive; haplotype II: *GSOH* active/*DMR6* inactive; haplotype III: *GSOH* inactive/*DMR6* active; haplotype IV: *GSOH* active/*DMR6* active. For each haplotype three independent pools of two plants each (*i.e.*, 3 x 2 different DH lines) were collected from which a single methanol extract was obtained as described above. Extracts were then subjected to UPLC FTMS as described above and accumulation of 3-butenyl and 2-hydroxy-3-butenyl was quantified as the percentage of detection saturation.

Climate Data

Climate data for the collection origin of each accession was obtained from the Climate Research Unit at the University of East Anglia. Data were extracted for nine climate variables giving the average per month over a 30-year (1961-1990) period (New et al., 2002). From these 9 variables, most other variables were extracted. Day length (spring) and relative humidity (spring) from the site of 306 accessions were obtained from the NCEP-NCAR climate reanalysis project (Hancock et al., 2011, Kistler et al., 2001) and the FAO GeoNetwork (http://fao.org.geonetworks/srv/en/main.home).

Funding

This research was part of the "Learning from Nature" program, supported by the Dutch Technology Foundation (STWGrant 1099), which is part of the Netherlands Organization for Scientific Research (NWO). Further support was obtained from a Centre of BioSystems Genomics (CBSG) Metabolomics Hotel Project (TD16-5), A CBSG Arabidopsis project (AA3-WU-PL), a Netherlands Metabolic Centre (NMC) project (Population Metabolomics – Arabidopsis 3370046800) and a booster grant (050-040-213) of the Netherlands Genomics Initiative (NGI) to the Consortium for Improving Plant Yield (CIPY).

Acknowledgements

RK was funded by grants from CBSG (AA3-WU-PL) and STW (STWGrant 1099).

References

ALONSO-BLANCO, C., AARTS, M. G., BENTSINK, L., KEURENTJES, J. J., REYMOND, M., VREUGDENHIL, D. & KOORNNEEF, M. 2009. What has natural variation taught us about plant development, physiology, and adaptation? *Plant Cell,* 21**,** 1877-96.

BAC-MOLENAAR, J. A., FRADIN, E. F., RIENSTRA, J. A., VREUGDENHIL, D. & KEURENTJES, J. J. B. 2015a. GWA Mapping of Anthocyanin Accumulation Reveals Balancing Selection of MYB90 in Arabidopsis thaliana. *PLoS One,* 10**,** e0143212.

BAC-MOLENAAR, J. A., VREUGDENHIL, D., GRANIER, C. & KEURENTJES, J. J. B. 2015b. Genome-wide association mapping of growth dynamics detects time-specific and general quantitative trait loci. *Journal of Experimental Botany,* 66**,** 5567-5580.

BAXTER, I., BRAZELTON, J. N., YU, D., HUANG, Y. S., LAHNER, B., YAKUBOVA, E., LI, Y., BERGELSON, J., BOREVITZ, J. O., NORDBORG, M., VITEK, O. & SALT, D. E. 2010. A coastal cline in sodium accumulation in Arabidopsis thaliana is driven by natural variation of the sodium transporter AtHKT1;1. *PLoS Genet,* 6**,** e1001193.

BENDEROTH, M., TEXTOR, S., WINDSOR, A. J., MITCHELL-OLDS, T., GERSHENZON, J. & KROYMANN, J. 2006. Positive selection driving diversification in plant secondary metabolism. *Proc Natl Acad Sci U S A,* 103**,** 9118-23.

BERGELSON, J. & ROUX, F. 2010. Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. *Nat Rev Genet,* 11**,** 867-79.

BOYLE, E. A., LI, Y. I. & PRITCHARD, J. K. 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell,* 169**,** 1177-1186.

BRACHI, B., MEYER, C. G., VILLOUTREIX, R., PLATT, A., MORTON, T. C., ROUX, F. & BERGELSON, J. 2015. Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. *Proceedings of the National Academy of Sciences,* 112**,** 4032-4037.

CANNELL, N., EMMS, D. M., HETHERINGTON, A. J., MACKAY, J., KELLY, S., DOLAN, L. & SWEETLOVE, L. J. 2020. Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant Evolution. *Curr Biol,* 30**,** 1783-1800.e11.

CHAE, L., LEE, I., SHIN, J. & RHEE, S. Y. 2012. Towards understanding how molecular networks evolve in plants. *Curr Opin Plant Biol,* 15**,** 177-84.

CHAN, E. K., ROWE, H. C., HANSEN, B. G. & KLIEBENSTEIN, D. J. 2010. The complex genetic architecture of the metabolome. *PLoS Genet,* 6**,** e1001198.

CHATEIGNER, A., LESAGE-DESCAUSES, M. C., ROGIER, O., JORGE, V., LEPLÉ, J. C., BRUNAUD, V., ROUX, C. P., SOUBIGOU-TACONNAT, L., MARTIN-MAGNIETTE, M. L., SANCHEZ, L. & SEGURA, V. 2020. Gene expression predictions and networks in natural populations supports the omnigenic theory. *BMC Genomics,* 21**,** 416.

CORDOVEZ, V., CARRION, V. J., ETALO, D. W., MUMM, R., ZHU, H., VAN WEZEL, G. P. & RAAIJMAKERS, J. M. 2015. Diversity and functions of volatile organic compounds produced by Streptomyces from a disease-suppressive soil. *Front Microbiol,* 6**,** 1081.

DAVILA OLIVAS, N. H., KRUIJER, W., GORT, G., WIJNEN, C. L., VAN LOON, J. J. A. & DICKE, M. C. 2017. Genome-wide association analysis reveals distinct genetic architectures for single and combined stress responses in Arabidopsis thaliana. *New Phytologist,* 213**,** 838-851.

DE LOS CAMPOS, G., PÉREZ, P., VAZQUEZ, A. I. & CROSSA, J. 2013. Genome-enabled prediction using the BLR (Bayesian Linear Regression) R-package. *Methods Mol Biol,* 1019**,** 299-320.

DE VOS, R. C., MOCO, S., LOMMEN, A., KEURENTJES, J. J. B., BINO, R. J. & HALL, R. D. 2007. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat Protoc,* 2**,** 778-91.

DUBIN, M. J., ZHANG, P., MENG, D., REMIGEREAU, M.-S., OSBORNE, E. J., PAOLO CASALE, F., DREWE, P., KAHLES, A., JEAN, G., VILHJÁLMSSON, B., JAGODA, J., IREZ, S., VORONIN, V., SONG, Q., LONG, Q., RÄTSCH, G., STEGLE, O., CLARK, R. M. & NORDBORG, M. 2015. DNA methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. *eLife,* 4**,** e05255.

FALCONE FERREYRA, M. L., EMILIANI, J., RODRIGUEZ, E. J., CAMPOS-BERMUDEZ, V. A., GROTEWOLD, E. & CASATI, P. 2015. The Identification of Maize and Arabidopsis Type I FLAVONE SYNTHASEs Links Flavones with Hormones and Biotic Interactions. *Plant Physiol,* 169**,** 1090-107.

FANG, C., FERNIE, A. R. & LUO, J. 2019. Exploring the Diversity of Plant Metabolism. *Trends Plant Sci,* 24**,** 83-98.

FILIAULT, D. L., SEYMOUR, D. K., MARUTHACHALAM, R. & MALOOF, J. N. 2017. The Generation of Doubled Haploid Lines for QTL Mapping. *Methods Mol Biol,* 1610**,** 39-57.

FU, J., KEURENTJES, J. J. B., BOUWMEESTER, H., AMERICA, T., VERSTAPPEN, F. W., WARD, J. L., BEALE, M. H., DE VOS, R. C., DIJKSTRA, M., SCHELTEMA, R. A., JOHANNES, F., KOORNNEEF, M., VREUGDENHIL, D., BREITLING, R. & JANSEN, R. C. 2009. System-wide molecular evidence for phenotypic buffering in Arabidopsis. *Nat Genet,* 41**,** 166-7.

FUSARI, C. M., KOOKE, R., LAUXMANN, M. A., ANNUNZIATA, M. G., ENKE, B., HOEHNE, M., KROHN, N., BECKER, F. F. M., SCHLERETH, A., SULPICE, R., STITT, M. & KEURENTJES, J. J. B. 2017. Genome-Wide Association Mapping Reveals That Specific and Pleiotropic Regulatory Mechanisms Fine-Tune Central Metabolism and Growth in Arabidopsis. *The Plant Cell,* 29**,** 2349-2373.

GOUJON, T., SIBOUT, R., POLLET, B., MABA, B., NUSSAUME, L., BECHTOLD, N., LU, F., RALPH, J., MILA, I., BARRIÈRE, Y., LAPIERRE, C. & JOUANIN, L. 2003. A new Arabidopsis thaliana mutant deficient in the expression of O-methyltransferase impacts lignins and sinapoyl esters. *Plant Molecular Biology,* 51**,** 973-989.

HANCOCK, A. M., BRACHI, B., FAURE, N., HORTON, M. W., JARYMOWYCZ, L. B., SPERONE, F. G., TOOMAJIAN, C., ROUX, F. & BERGELSON, J. 2011. Adaptation to climate across the Arabidopsis thaliana genome. *Science,* 334**,** 83-6.

HANSEN, B. G., KERWIN, R. E., OBER, J. A., LAMBRIX, V. M., MITCHELL-OLDS, T., GERSHENZON, J., HALKIER, B. A. & KLIEBENSTEIN, D. J. 2008. A Novel 2-Oxoacid-Dependent Dioxygenase Involved in the Formation of the Goiterogenic 2-Hydroxybut-3-enyl Glucosinolate and Generalist Insect Resistance in Arabidopsis. *Plant Physiology,* 148**,** 2096-2108.

HERMAN, P. L. & MARKS, M. D. 1989. Trichome Development in Arabidopsis thaliana. II. Isolation and Complementation of the GLABROUS1 Gene. *Plant Cell,* 1**,** 1051-1055.

HOLSINGER, K. E. & WEIR, B. S. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet,* 10**,** 639-50.

HORTON, M. W., HANCOCK, A. M., HUANG, Y. S., TOOMAJIAN, C., ATWELL, S., AUTON, A., MULIYATI, N. W., PLATT, A., SPERONE, F. G., VILHJÁLMSSON, B. J., NORDBORG, M., BOREVITZ, J. O. & BERGELSON, J. 2012. Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nat Genet,* 44**,** 212-6.

HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. 4**,** 44.

ISHIHARA, H., TOHGE, T., VIEHÖVER, P., FERNIE, A. R., WEISSHAAR, B. & STRACKE, R. 2016. Natural variation in flavonol accumulation in Arabidopsis is determined by the flavonol glucosyltransferase BGLU6. *Journal of Experimental Botany,* 67**,** 1505-1517.

JOSEPH, B., CORWIN, J. A. & KLIEBENSTEIN, D. J. 2015. Genetic variation in the nuclear and organellar genomes modulates stochastic variation in the metabolome, growth, and defense. *PLoS Genet,* 11**,** e1004779.

JOSEPHS, E. B., STINCHCOMBE, J. R. & WRIGHT, S. I. C. 2017. What can genome-wide association studies tell us about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytologist,* 214**,** 21-33.

KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S. Y., FREIMER, N. B., SABATTI, C. & ESKIN, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet,* 42**,** 348-54.

KERWIN, R., FEUSIER, J., CORWIN, J., RUBIN, M., LIN, C., MUOK, A., LARSON, B., LI, B., JOSEPH, B., FRANCISCO, M., COPELAND, D., WEINIG, C. & KLIEBENSTEIN, D. J. 2015. Natural genetic variation in Arabidopsis thaliana defense metabolism genes modulates field fitness. *eLife,* 4**,** e05604.

KEURENTJES, J. J. B., FU, J., DE VOS, C. H., LOMMEN, A., HALL, R. D., BINO, R. J., VAN DER PLAS, L. H., JANSEN, R. C., VREUGDENHIL, D. & KOORNNEEF, M. 2006. The genetics of plant metabolism. *Nat Genet,* 38**,** 842-9.

KEURENTJES, J. J. B., SULPICE, R., GIBON, Y., STEINHAUSER, M. C., FU, J., KOORNNEEF, M., STITT, M. & VREUGDENHIL, D. 2008. Integrative analyses of genetic variation in enzyme activities of primary carbohydrate metabolism reveal distinct modes of regulation in Arabidopsis thaliana. *Genome Biol,* 9**,** R129.

KIM, S.-J., KIM, M.-R., BEDGAR, D. L., MOINUDDIN, S. G. A., CARDENAS, C. L., DAVIN, L. B., KANG, C. & LEWIS, N. G. 2004. Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 1455-1460.

KISTLER, R., KALNAY, E., COLLINS, W., SAHA, S., WHITE, G., WOOLLEN, J., CHELLIAH, M., EBISUZAKI, W., KANAMITSU, M., KOUSKY, V., VAN DEN DOOL, H., JENNE, R. & FIORINO, M. 2001. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society,* 82**,** 247-267.

KLIEBENSTEIN, D. 2009. Advancing genetic theory and application by metabolic quantitative trait loci analysis. *Plant Cell,* 21**,** 1637-46.

KLIEBENSTEIN, D. J., CACHO, N. I. & STANISLAV, K. 2016. Chapter Three - Nonlinear Selection and a Blend of Convergent, Divergent and Parallel Evolution Shapes Natural Variation in Glucosinolates. *Advances in Botanical Research.* Academic Press.

KLIEBENSTEIN, D. J., LAMBRIX, V. M., REICHELT, M., GERSHENZON, J. & MITCHELL-OLDS, T. 2001. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in Arabidopsis. *Plant Cell,* 13**,** 681-93.

KOOKE, R. & KEURENTJES, J. J. B. 2012. Multi-dimensional regulation of metabolic networks shaping plant development and performance. *Journal of Experimental Botany,* 63**,** 3353-3365.

KOOKE, R., KRUIJER, W., BOURS, R., BECKER, F. F. M., KUHN, A., GEEST, H. V. D., BUNTJER, J., DOESWIJK, T., GUERRA, J., BOUWMEESTER, H. J., VREUGDENHIL, D. & KEURENTJES, J. J. B. 2016. Genome-wide association mapping and genomic prediction elucidate the genetic architecture of morphological traits in Arabidopsis thaliana. *Plant Physiology*.

KOOKE, R., MORGADO, L., BECKER, F., VAN EEKELEN, H., HAZARIKA, R., ZHENG, Q., DE VOS, R. C. H., JOHANNES, F. & KEURENTJES, J. J. B. 2019. Epigenetic mapping of the Arabidopsis metabolome reveals mediators of the epigenotype-phenotype map. *Genome Res,* 29**,** 96-106.

KORTE, A. & FARLOW, A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods,* 9**,** 29.

KOVER, P. X. & MOTT, R. 2012. Mapping the genetic basis of ecologically and evolutionarily relevant traits in Arabidopsis thaliana. *Curr Opin Plant Biol,* 15**,** 212-7.

KROYMANN, J., TEXTOR, S., TOKUHISA, J. G., FALK, K. L., BARTRAM, S., GERSHENZON, J. & MITCHELL-OLDS, T. 2001. A gene controlling variation in Arabidopsis glucosinolate composition is part of the methionine chain elongation pathway. *Plant Physiol,* 127**,** 1077-88.

KRUIJER, W., BOER, M., MALOSETTI, M., FLOOD, P. J., ENGEL, B., KOOKE, R., KEURENTJES, J. J. B. & EEUWIJK, F. V. 2014. Marker-Based Estimation of Heritability in Immortal Populations. *Genetics*.

LI, X., TIEMAN, D., LIU, Z., CHEN, K. & KLEE, H. J. 2020. Identification of a lipase gene with a role in tomato fruit short-chain fatty acid-derived flavor volatiles by genome-wide association. *The Plant Journal,* 104**,** 631-644.

LI, Y., HUANG, Y., BERGELSON, J., NORDBORG, M. & BOREVITZ, J. O. 2010. Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. *Proc Natl Acad Sci U S A,* 107**,** 21199-204.

LIPPERT, C., LISTGARTEN, J., LIU, Y., KADIE, C. M., DAVIDSON, R. I. & HECKERMAN, D. 2011. FaST linear mixed models for genome-wide association studies. *Nat Methods,* 8**,** 833-5.

LISTGARTEN, J., KADIE, C., SCHADT, E. E. & HECKERMAN, D. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc Natl Acad Sci U S A,* 107**,** 16465-70.

LOMMEN, A. & KOOLS, H. J. 2012. MetAlign 3.0: performance enhancement by efficient use of advances in computer hardware. *Metabolomics,* 8**,** 719-726.

LYU, J. 2017. Arabidopsis evolution: Roots in Africa. *Nat Plants,* 3**,** 17091.

MARKS, M. D. & FELDMANN, K. A. 1989. Trichome Development in Arabidopsis thaliana. I. T-DNA Tagging of the GLABROUS1 Gene. *Plant Cell,* 1**,** 1043-1050.

MOSER, G., LEE, S. H., HAYES, B. J., GODDARD, M. E., WRAY, N. R. & VISSCHER, P. M. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS genetics,* 11**,** e1004969.

MUMM, R., HAGEMAN, J. A., CALINGACION, M. N., DE VOS, R. C. H., JONKER, H. H., ERBAN, A., KOPKA, J., HANSEN, T. H., LAURSEN, K. H., SCHJOERRING, J. K., WARD, J. L., BEALE, M. H., JONGEE, S., RAUF, A., HABIBI, F., INDRASARI, S. D., SAKHAN, S., RAMLI, A., ROMERO, M., REINKE, R. F., OHTSUBO, K., BOUALAPHANH, C., FITZGERALD, M. A. & HALL, R. D. 2016. Multi-platform metabolomics analyses of a broad collection of fragrant and non-fragrant rice varieties reveals the high complexity of grain quality characteristics. *Metabolomics,* 12**,** 38.

NEW, M., LISTER, D., HULME, M. & MAKIN, I. 2002. A high-resolution data set of surface climate over global land areas. *Climate Research,* 21**,** 1-25.

OLSON-MANNING, C. F., WAGNER, M. R. & MITCHELL-OLDS, T. 2012. Adaptive evolution: evaluating empirical support for theoretical predictions. 13**,** 867.

PAYNE, J. L. & WAGNER, A. 2019. The causes of evolvability and their evolution. *Nat Rev Genet,* 20**,** 24-38.

PEREZ DE SOUZA, L., NAAKE, T., TOHGE, T. & FERNIE, A. R. 2017. From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web resources for mass spectral plant metabolomics. *Gigascience,* 6**,** 1-20.

ROCKMAN, M. V. 2012. THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT'S GOLD DOES NOT GLITTER. *Evolution,* 66**,** 1-17.

ROSS, J. R., NAM, K. H., D'AURIA, J. C. & PICHERSKY, E. 1999. S-Adenosyl-l-Methionine:Salicylic Acid Carboxyl Methyltransferase, an Enzyme Involved in Floral Scent Production and Plant Defense, Represents a New Class of Plant Methyltransferases. *Archives of Biochemistry and Biophysics,* 367**,** 9-16.

SALEM, M. A., PEREZ DE SOUZA, L., SERAG, A., FERNIE, A. R., FARAG, M. A., EZZAT, S. M. & ALSEEKH, S. 2020. Metabolomics in the Context of Plant Natural Products Research: From Sample Preparation to Metabolite Analysis. *Metabolites,* 10.

SEGRÈ, D., VITKUP, D. & CHURCH, G. M. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A,* 99**,** 15112-7.

SHIMIZU, K. K. & PURUGGANAN, M. D. 2005. Evolutionary and ecological genomics of Arabidopsis. *Plant Physiol,* 138**,** 578-84.

SMITH, S. M. & MAUGHAN, P. J. 2015. SNP genotyping using KASPar assays. *Methods Mol Biol,* 1245**,** 243-56.

STRAUCH, R. C., SVEDIN, E., DILKES, B., CHAPPLE, C. & LI, X. 2015. Discovery of a novel amino acid racemase through exploration of natural variation in Arabidopsis thaliana. *Proc Natl Acad Sci U S A,* 112**,** 11726-31.

TEXTOR, S., BARTRAM, S., KROYMANN, J., FALK, K. L., HICK, A., PICKETT, J. A. & GERSHENZON, J. 2004. Biosynthesis of methionine-derived glucosinolates in Arabidopsis thaliana: recombinant expression and characterization of methylthioalkylmalate synthase, the condensing enzyme of the chain-elongation cycle. *Planta,* 218**,** 1026-1035.

THOEN, M. P. M., DAVILA OLIVAS, N. H., KLOTH, K. J., COOLEN, S., HUANG, P.-P., AARTS, M. G. M., BAC-MOLENAAR, J. A., BAKKER, J., BOUWMEESTER, H. J., BROEKGAARDEN, C., BUCHER, J., BUSSCHER-LANGE, J., CHENG, X., FRADIN, E. F., JONGSMA, M. A., JULKOWSKA, M. M., KEURENTJES, J. J. B.,

LIGTERINK, W., PIETERSE, C. M. J., RUYTER-SPIRA, C., SMANT, G., TESTERINK, C., USADEL, B., VAN LOON, J. J. A., VAN PELT, J. A., VAN SCHAIK, C. C., VAN WEES, S. C. M., VISSER, R. G. F., VOORRIPS, R., VOSMAN, B., VREUGDENHIL, D., WARMERDAM, S., WIEGERS, G. L., VAN HEERWAARDEN, J., KRUIJER, W., VAN EEUWIJK, F. A. & DICKE, M. C. 2017. Genetic architecture of plant stress resistance: multi-trait genome-wide association mapping. *New Phytologist,* 213**,** 1346-1362.

TIKUNOV, Y. M., LAPTENOK, S., HALL, R. D., BOVY, A. & DE VOS, R. C. H. 2012. MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics,* 8**,** 714-8.

TODESCO, M., BALASUBRAMANIAN, S., HU, T. T., TRAW, M. B., HORTON, M., EPPLE, P., KUHNS, C., SURESHKUMAR, S., SCHWARTZ, C., LANZ, C., LAITINEN, R. A., HUANG, Y., CHORY, J., LIPKA, V., BOREVITZ, J. O., DANGL, J. L., BERGELSON, J., NORDBORG, M. & WEIGEL, D. 2010. Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature,* 465**,** 632-6.

TRONTIN, C., TISNÉ, S., BACH, L. & LOUDET, O. 2011. What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants? *Curr Opin Plant Biol,* 14**,** 225-31.

VAN DUYNHOVEN, J., VAN DER HOOFT, J. J., VAN DORSTEN, F. A., PETERS, S., FOLTZ, M., GOMEZ-ROLDAN, V., VERVOORT, J., DE VOS, R. C. & JACOBS, D. M. 2014. Rapid and sustained systemic circulation of conjugated gut microbial catabolites after single-dose black tea extract consumption. *J Proteome Res,* 13**,** 2668-78.

VERHOEVEN, H. A., JONKER, H., DE VOS, R. C. & HALL, R. D. 2012. Solid phase micro-extraction GC-MS analysis of natural volatile components in melon and rice. *Methods Mol Biol,* 860**,** 85-99.

WALSH, B. & LYNCH, M. 2018. *Evolution and Selection of Quantitative Traits,* New York, U.S.A., Oxford University Press.

WANG, Z., YANG, Z. & LI, F. 2019. Updates on molecular mechanisms in the development of branched trichome in Arabidopsis and nonbranched in cotton. *Plant Biotechnol J,* 17**,** 1706-1722.

WARDE-FARLEY, D., DONALDSON, S. L., COMES, O., ZUBERI, K., BADRAWI, R., CHAO, P., FRANZ, M., GROUIOS, C., KAZI, F., LOPES, C. T., MAITLAND, A., MOSTAFAVI, S., MONTOJO, J., SHAO, Q., WRIGHT, G., BADER, G. D. & MORRIS, Q. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res,* 38**,** W214-20.

WEHRENS, R., HAGEMAN, J. A., VAN EEUWIJK, F., KOOKE, R., FLOOD, P. J., WIJNKER, E., KEURENTJES, J. J. B., LOMMEN, A., VAN EEKELEN, H. D., HALL, R. D., MUMM, R. & DE VOS, R. C. 2016. Improved batch correction in untargeted MS-based metabolomics. *Metabolomics,* 12**,** 88.

WITJES, L., KOOKE, R., VAN DER HOOFT, J. J. J., DE VOS, R. C. H., KEURENTJES, J. J. B., MEDEMA, M. H. & NIJVEEN, H. 2019. A genetical metabolomics approach for bioprospecting plant biosynthetic gene clusters. *BMC Res Notes,* 12**,** 194.

WU, S., ALSEEKH, S., CUADROS-INOSTROZA, Á., FUSARI, C. M., MUTWIL, M., KOOKE, R., KEURENTJES, J. J. B., FERNIE, A. R., WILLMITZER, L. & BROTMAN, Y. 2016. Combined Use of Genome-Wide Association Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in Arabidopsis thaliana. *PLoS Genet,* 12**,** e1006363.

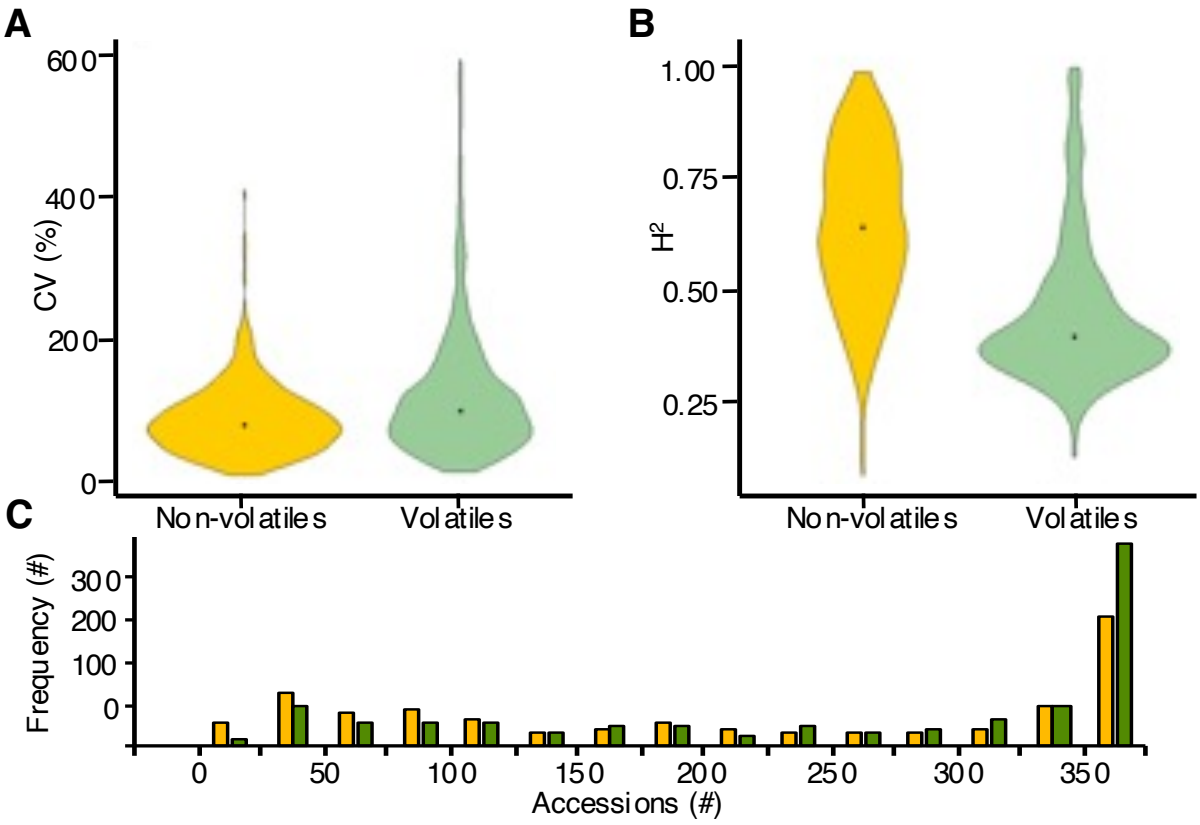
WU, S., TOHGE, T., CUADROS-INOSTROZA, Á., TONG, H., TENENBOIM, H., KOOKE, R., MÉRET, M., KEURENTJES, J. J. B., NIKOLOSKI, Z., FERNIE, A. R., WILLMITZER, L. & BROTMAN, Y. 2018. Mapping the Arabidopsis Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions. *Mol Plant,* 11**,** 118-134.

XU, W., BOBET, S., LE GOURRIEREC, J., GRAIN, D., DE VOS, D., BERGER, A., SALSAC, F., KELEMEN, Z., BOUCHEREZ, J., ROLLAND, A., MOUILLE, G., ROUTABOUL, J. M., LEPINIEC, L. & DUBOS, C. 2017. TRANSPARENT TESTA 16 and 15 act through different mechanisms to control proanthocyanidin accumulation in Arabidopsis testa. *J Exp Bot,* 68**,** 2859-2870.

YU, X., XIAO, J., CHEN, S., YU, Y., MA, J., LIN, Y., LI, R., LIN, J., FU, Z., ZHOU, Q., CHAO, Q., CHEN, L., YANG, Z. & LIU, R. 2020. Metabolite signatures of diverse Camellia sinensis tea populations. *Nat Commun,* 11**,** 5586.

ZHANG, W., ALSEEKH, S., ZHU, X., ZHANG, Q., FERNIE, A. R., KUANG, H. & WEN, W. 2020. Dissection of the domestication-shaped genetic architecture of lettuce primary metabolism. *The Plant Journal,* 104**,** 613-630.

ZHOU, X. & STEPHENS, M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet,* 44**,** 821-4.



821

822 **Figure 1:** Genetic variation for metabolic accumulation in a global population of *Arabidopsis thaliana*.
823 Distribution of the coefficient of variation (**A**) and the broad-sense heritability (**B**) of the relative abundance of volatile
824 and non-volatile metabolites. (**C**) Frequency distribution of the number of accessions in which a specific metabolite
825 was detected. Green and yellow items represent volatile and non-volatile metabolites, respectively.

826

827

828

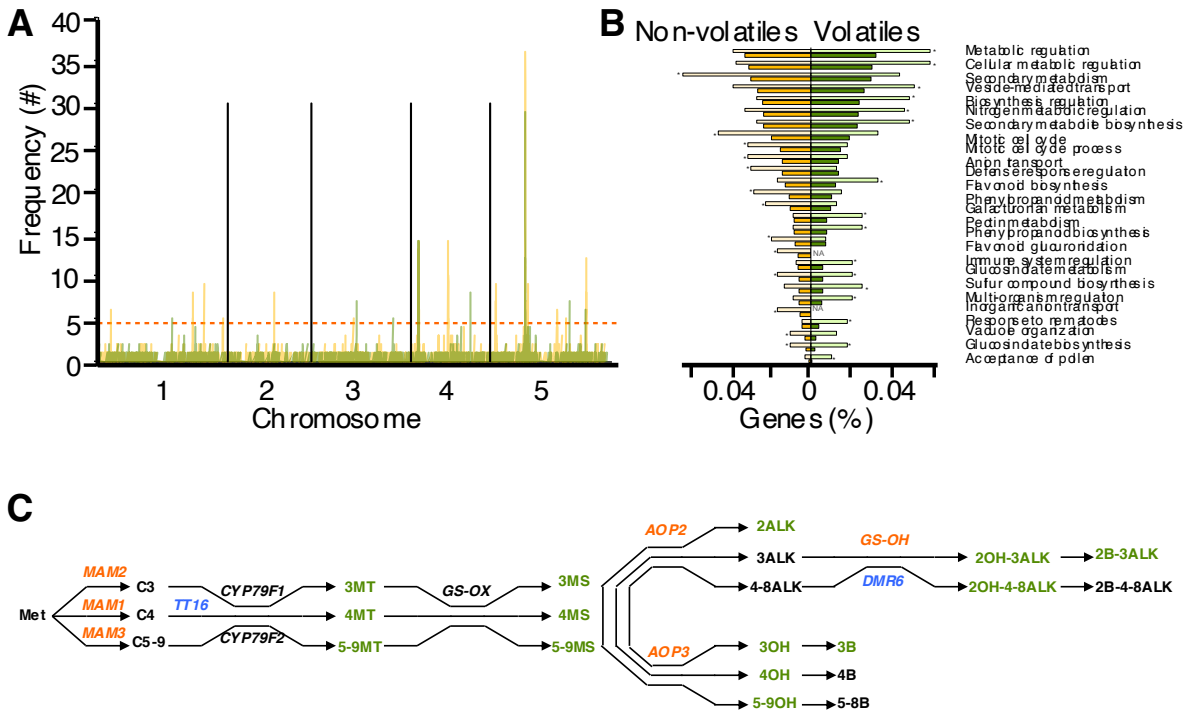


Figure 2: Genetic mapping of metabolic variation.

(**A**) Frequency distribution of QTLs over the five chromosomes of Arabidopsis. A QTL hotspot is defined as a locus associated with variation in at least five metabolites. Yellow and green bars indicate the number of non-volatile and volatile metabolites for which a QTL was detected, respectively. The red dotted horizontal line indicates the QTL hotspot threshold. (**B**) Enrichment analysis of candidate genes associated with variation in metabolic content. Dark yellow and green bars indicate the reference genome-wide percentage of genes involved in the processes plotted on the right. Light yellow and green bars indicate this percentage for candidate genes, associated with variation in non-volatile and volatile metabolites, respectively. Asterisks indicate significant enrichment ($P_{BENJAMINI} < 0.01$), while *NA* denotes less than two candidate genes involved in the respective process. (**C**) Biosynthesis pathway of glucosinolates. Genes involved in biosynthesis are plotted above arrows. Red genes were assigned as candidate genes explaining the observed variation in glucosinolates. Blue genes have not previously been annotated as biosynthesis genes but were strongly associated with the observed variation in glucosinolates. Synthesised glucosinolates are plotted following arrows. Green glucosinolates were detected and putatively annotated in the current study. C3, C4 and C5-9 indicate glucosinolate carbon-chain-length; MT is Methylthioalkyl; MS is Methylsulfinylalkyl; ALK is Alkenyl; OH is Hydroxyalkyl; OH-ALK is Hydroxyalkenyl; B is Benzoyloxyalkyl; B-ALK is Benzoyloxyalkenyl.

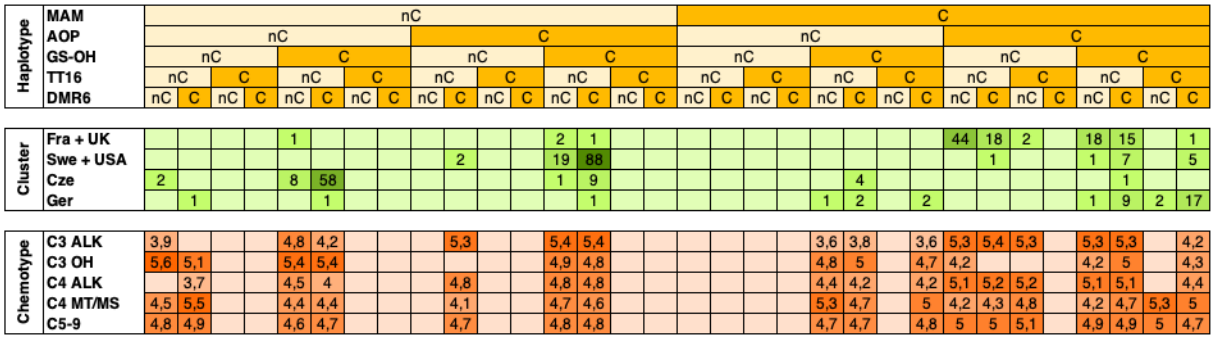


Figure 3: Genotypic and geographic patterns of adaptation of glucosinolate profiles.

The upper panel divides haplotypes based on the genotypes of five loci involved in the biosynthesis of glucosinolates. C denotes Col-0 alleles, while nC denotes non-Col-0 alleles. The second panel divides the collection of Arabidopsis accessions analyzed in three geographical clusters based on hierarchical clustering of metabolite profiles. Columns align with the observed haplotype indicated in the upper panel. For each cluster and haplotype the number of accessions is given. The lower panel divides the detected glucosinolates in five classes. Columns align again with the observed haplotype indicated in the upper panel. For each haplotype the relative average abundance of glucosinolates assigned to a specific chemotype is given.