

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

2 Rik Kooke^{1,2}, Willem Kruijer², Henriette D.L.M. van Eekelen³, Frank F.M. Becker¹, Ron Wehrens², Robert D. Hall^{3,4},
3 Roland Mumm³, Ric C.H. de Vos³, Fred A. van Eeuwijk² and Joost J.B. Keurentjes¹

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

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

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

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

8

9 **ORCID Identifier**

10 R. Kooke

11 W. Kruijer

12 H.D.L.M. van Eekelen

13 F.F.M. Becker

14 R. Wehrens

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

16 R. Mumm

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

18 F.A. van Eeuwijk

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

20

21 **Corresponding Author**

22 Joost J.B. Keurentjes

23 Droevendaalsesteeg 1

24 NL-6708 PB Wageningen

25 The Netherlands

26 T: 0317 483149

27 E: Joost.Keurentjes@wur.nl

28

29 **Classification**

30 Biological Sciences - Genetics

31

32 **Keywords**

33 *Arabidopsis thaliana*, Metabolomics, Secondary metabolism, Genetics, Adaptation

34

35

36

37

38 **Abstract**

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

51

52

53 **Introduction**

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

65 *Arabidopsis thaliana* is no exception and is deservedly considered the reference plant species for studies on natural
66 variation (Alonso-Blanco et al., 2009). *A. thaliana* diverged five to six million years ago from other species and despite
67 relatively recent bottleneck events, due to glacial-interglacial climate changes, has adapted to a wide range of
68 environmental settings spanning nearly all terrestrial habitats across the globe (Lyu, 2017, Shimizu and Purugganan,
69 2005). The species displays a wide diversity in the manifestation of traits and the abundant genetic resources
70 available enabled the analysis of linkage between sequence diversity and natural variation in adaptive properties
71 (Bergelson and Roux, 2010, Kover and Mott, 2012, Trontin et al., 2011). Being sessile organisms, unable to escape

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

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

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

102

103 **Results**

104 **Natural variation in plant secondary metabolism is driven by pathway diversity**

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

107 in four-week-old rosettes of plants grown under standard conditions in short days. This population has previously
108 been analyzed extensively for numerous morphological, metabolic and stress-related traits (Bac-Molenaar et al.,
109 2015b, Davila Olivas et al., 2017, Fusari et al., 2017, Thoen et al., 2017, Wu et al., 2018, Kooke et al., 2016), although
110 the systematic effect-size of genomic variation and genetic architecture was hardly established in these studies. Here,
111 we investigated plant metabolic content in two independent replicate samples of each accession through both
112 untargeted UPLC-Orbitrap-FTMS and SPME-GCMS based profiling, enabling the detection and relative quantification
113 of non-volatile and volatile secondary metabolites, respectively (Salem et al., 2020, Kooke et al., 2019, Wehrens et
114 al., 2016). A total number of, respectively, 567 and 603 non-volatile and volatile compounds were detected of which
115 many could be putatively annotated using public and in-house databases (Perez de Souza et al., 2017, Tikunov et al.,
116 2012, Kooke et al., 2019, Witjes et al., 2019). Widespread quantitative and qualitative variation in chemical profiles
117 was observed among the genotypes of the population (Figure 1A; Supplemental Table 1). A large number of volatile
118 and non-volatile metabolites were commonly detected in all accessions analyzed (32% and 17%, respectively), while
119 9% and 16% of the compounds were detected in less than 50 accessions, respectively (Figure 1C). For instance, each
120 accession contained on average eight of the 94 rare non-volatile metabolites, but the accessions Zdr-6 and Mt-0
121 contained no less than 38 and 39 of these compounds, respectively. The lowest number (228) of non-volatile
122 metabolites was detected in Cvi-0, although this accession contained the highest level of 3-butenyl glucosinolate.
123 This pattern of variation in phytochemical profiles is consistent with earlier findings in a much smaller set of
124 accessions (Keurentjes et al., 2006). Moreover, the variation in abundance of detected metabolites was substantial,
125 with similar median coefficients of variation of 101% and 82% for volatile and non-volatile compounds, respectively.
126 However, the broad-sense heritability (H^2) of non-volatile metabolites was on average much higher than that of the
127 volatile metabolites, possibly due to stronger stochasticity in the release of volatile compounds. Nonetheless, median
128 heritabilities of 40% (volatiles) and 64% (non-volatiles) demonstrate that a large part of the variation in secondary
129 metabolites is heritable (Figure 1B).

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

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

142 one or more SNPs was substantially lower (WGR: 157 metabolites (26%, $|\beta| > 0.01$); GWA: 134 metabolites (22%,
143 $P_{BONF} < 0.05$)) (Supplemental Table 2). The difference in mapping power between the two types of metabolites reflects
144 their contrast in heritability, while the discrepancy between the WGR and GWA approach might be an effect of low
145 minor allele frequencies. Rare large-effect loci are better detected using WGR while the power to detect true
146 positives with LMM GWAS declines with decreasing MAFs. The latter would suggest that variation in volatile
147 metabolites, in contrast to variation in non-volatile metabolites, is more frequently driven by rare alleles. The overlap
148 in metabolites for which variation could be associated to genetic variation by both WGR and GWA was 83% for the
149 non-volatiles and 57% for the volatiles, emphasizing the differences between metabolite types and mapping
150 approach.

151
152 To assign genomic functions causal for the observed variation in metabolic profiles, candidate genes in linkage
153 disequilibrium (LD) with the large-effect SNPs detected by WGR ($|\beta| > 0.01$) were identified. For many metabolites
154 multiple QTLs were detected, often represented by series of large-effect SNPs in LD. In total, 1,097 unique candidate
155 genes explaining variation in the 301 non-volatile metabolites were assigned, whereas 651 candidate genes were
156 assigned explaining variation in the 154 volatile metabolites (Supplemental Table 2). Only 120 of the assigned genes
157 explain variation in both non-volatiles and volatiles, illustrating that there is little overlap in the genetic regulation of
158 volatile and non-volatile secondary metabolites.

159 At a number of loci, variation in several metabolites mapped to the same candidate gene. A QTL hotspot was defined
160 as a gene locus substantially associated with variation in at least five metabolites by a minimum of two large-effect
161 SNPs ($|\beta| > 0.01$). This criterion was met by nine loci after WGR analysis of non-volatile metabolites and three loci
162 for volatile metabolites, with a small overlap of two hotspots detected with both platforms (Supplemental Table 2,
163 Figure 2A). With the exception of *ACD6* and *GLABRA1*, which are involved in trade-offs between growth and defence
164 (Fusari et al., 2017, Todesco et al., 2010) and trichome development (Wang et al., 2019, Herman and Marks, 1989,
165 Marks and Feldmann, 1989), respectively, all assigned hotspot candidate genes are directly related to the
166 biosynthesis of secondary metabolites. Most prominent is the glucosinolate metabolism (Figure 2C), which is
167 represented by three QTL hotspots, explaining variation in 82 metabolites. While variation in various non-volatile
168 intact glucosinolates maps to all three QTL hotspots, variation in the well-known volatile glucosinolate breakdown
169 products, such as nitriles and isothiocyanates (Hansen et al., 2008, Textor et al., 2004, Kliebenstein et al., 2001) maps
170 predominantly to the *AOP* and *MAM* loci. Other QTL hotspots mainly represent phenolic compounds, such as
171 phenylpropanoids and flavonoids (Goujon et al., 2003, Ishihara et al., 2016, Kim et al., 2004, Ross et al., 1999). The
172 analysis of QTL hotspots illustrates that many secondary metabolites are regulated simultaneously by natural
173 variation in a small number of biosynthesis genes upstream of branching points in metabolic pathways. However,
174 enrichment analysis demonstrates that many more genes involved in secondary metabolic processes are assigned as
175 candidate genes explaining variation in non-volatile metabolites than can be expected by chance (43 genes, $P_{BENJAMINI}$
176 < 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

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

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

229
230 An important topic in the evolution of species is whether adaptive traits have spread through natural populations by
231 drift, recombination and selection of a single alteration or by convergent evolution, in which similar effects are
232 repeatedly obtained by multiple independent mutation events. We addressed this issue by investigating genetic and
233 allelic heterogeneity in the regulation of metabolic content. For this, we performed a detailed analysis of the major-
234 effect WGR QTLs, which are represented by multiple large-effect SNPs ($|\beta| > 0.01$). Notably, collocating SNPs often
235 differ markedly in effect-size and MAF, and sometimes are not even in LD with each other. This strongly suggests that
236 different haplotypes of particular loci exist within the investigated population, with unequal selection forces acting
237 on the distinctive SNPs. Exemplary, at a number of hotspot QTLs the genetic variation explaining the abundance of a
238 specific metabolite is often determined predominantly by a single major-effect SNP and several smaller-effect SNPs.
239 However, this major-effect SNP is not necessarily the same for other metabolites controlled by the identical hotspot
240 QTL (Supplemental Table 2). So, where independent mutation events at a single locus might lead to allelic
241 heterogeneity in the regulation of one metabolite it might cause haplotype diversity in the regulation of another and,
242 depending on the adaptive value of one or the other metabolite, results in different selective forces on sequence
243 variation.

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

246 large-effect SNPs in the candidate gene *ACD6*, which are not in LD with each other ($R^2=0.11$) (Supplemental Figure
247 3A). However, the first SNP is in strong LD with various polymorphisms in the *ACD6* promoter, while the second SNP
248 is in strong LD with non-synonymous SNPs in the last exon of the coding region, resulting in the existence of four
249 different haplotypes. Strikingly, 68% of the accessions belonged to haplotype IV, whereas 25%, 6%, and 1% of the
250 investigated accessions belonged to haplotypes III, I, and II, respectively. Haplotype analysis of variation in a
251 representative metabolite controlled by this locus revealed that this metabolite was detected in 90% (18/20) of the
252 accessions of haplotype I, whereas this was only the case for 40% (35/88) and 14% (33/237) of the accessions of the
253 haplotype classes III and IV, respectively. This metabolite could not be detected in any of the accessions (0/2)
254 belonging to haplotype class II (Supplemental Figure 3B). Finally, the abundance of this metabolite was substantially
255 higher in haplotype class I than in any of the other haplotype classes (Supplemental Figure 3C). This indicates an
256 intragenic epistatic interaction effect between the two gene domains. The non-synonymous gene body SNPs might
257 determine the functionality of the gene product and are epistatic over the promoter SNPs, which, in turn, control the
258 level of transcription and, hence, the quantity of the gene product. The variation between haplotypes in the efficiency
259 to produce and accumulate specific metabolites strongly suggests that selective forces have established the observed
260 differences in the global population haplotype frequencies.

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

269
270 In other instances, multiple large-effect SNPs at a single locus are associated to more than one likely candidate gene,
271 suggesting genetic heterogeneity or even channelling of biosynthesis genes (Witjes et al., 2019). The QTL hotspot at
272 the top of chromosome five, explaining variation in the highest number of metabolites, predominantly
273 glucosinolates, links to three genes (*MAM*, *TT16* and *DMR6*), which all might be involved in metabolite biosynthesis
274 (Falcone Ferreyra et al., 2015, Kroymann et al., 2001, Xu et al., 2017). Variation in the majority of metabolites that
275 map to this locus is best explained by large-effect SNPs that are most strongly associated with *MAM*, although
276 quantitative variation in a minor number of metabolites is stronger linked to large-effect SNPs at the position of *TT16*
277 or *DMR6* (Supplemental Figure 4). Moreover, both *TT16* and *DMR6* are not in strong LD with *MAM* ($R^2 < 0.23$). A
278 similar observation can be made for metabolite variation mapping to the QTL hotspot at the top of chromosome four
279 (Supplemental Figure 5). Here, variation in metabolite abundance is most strongly associated to large-effect SNPs at
280 the position of either the *AOP* gene or *DAAR1*, both of which have been implicated in metabolite biosynthesis

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

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

294

295 **Polygenic adaptation of plant secondary metabolism to local settings**

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

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

317 The high F_{ST} values observed for some SNPs suggests the existence of population structure and a relationship
318 between metabolic profiles and the geographic origin of accessions. To investigate this further a two-dimensional
319 hierarchical clustering was performed on the geographic distribution of analyzed accessions and the accumulation of
320 detected compounds therein. Not surprisingly, given their strong enrichment in the phytochemical profile, aliphatic
321 glucosinolates and their volatile derivatives were among the strongest determinants of genotype clustering
322 (Supplemental Figure 8A). The aliphatic glucosinolates were further categorized on chain length and side-chain
323 modification and based on this classification four main groups of genotypes, each with a different organization of
324 glucosinolate biosynthesis alleles and enriched in accessions from specific geographic regions, could be distinguished
325 (Supplemental Figure 8B). For instance, alkenyl and long-chain glucosinolates group closely together and are, on
326 average, detected at high levels in the majority of genotypes from the UK and France (group 1). From the 105
327 accessions within this group, 45% is of French origin (*i.e.*, 77% of all French accessions) while 27% originates from the
328 UK (*i.e.*, 60% of all accessions from the UK) (Supplemental Table 1). The qualitative and quantitative variation in
329 glucosinolate content is largely explained by allelic variation in five biosynthesis genes (*viz.* *MAM*, *AOP*, *TT16*, *DMR6*,
330 and *GS-OH*). Both the *TT16* locus and the *DMR6* locus have not been reported earlier in glucosinolate studies,
331 probably due to the close vicinity to *MAM*. Both loci are, however, not in very strong LD with *MAM* ($R^2 < 0.23$). A
332 candidate gene for *TT16* might be coding for a succinyl CoA ligase involved in the TCA cycle (AT5G23250), or a
333 nicotinamidase involved in coniferin metabolism (AT5G23220/30). The *DMR6* gene is co-expressed with *GS-OH*,
334 shares protein domains with *AOP* and *GS-OH* and exhibits, like *AOP* and *GS-OH*, oxidoreductase activity (Warde-Farley
335 et al., 2010). Given the variation in C4-8 alkenyl glucosinolates explained by *DMR6*, the loss of an active *GS-OH* allele
336 might be compensated by a similar function of *DMR6*. To test this hypothesis, a population of doubled haploid (DH)
337 lines was generated from a cross between the accessions UKNW06-460 and Tamm-2 and subjected to metabolic
338 analyses. Unlike the reference accession Col-0, which lacks expression of functional *AOP* (Kliebenstein et al., 2001),
339 these accessions share functional *AOP* and *MAM* genes, necessary for the production of 3-butenyl, which is the
340 precursor for the synthesis of 2-hydroxy-3-butenyl. However, UKNW06-460 and Tamm-2 differ in functionality of the
341 *GSOH* and *DMR6* loci. Tamm-2 lacks functional alleles of both *GSOH* and *DMR6* and is unable to synthesise 2-hydroxy-
342 3-butenyl despite the fair amount of available 3-butenyl, whereas UKNW06-460 contains both a functional *GSOH* and
343 *DMR6* allele, which allows it to produce 2-hydroxy-3-butenyl in addition to 3-butenyl (Supplemental figure 9A).

344 Recombinant DH lines were divided in four haplotype classes based on functionality of the *GSOH* and *DMR6* alleles.
345 With a few exceptions, haplotypes lacking a functional *DMR6* allele (haplotype I and II) did not produce detectable
346 levels of 3-butenyl and 2-hydroxy-3-butenyl. Haplotypes containing a functional allele of *DMR6* (haplotype III and IV),
347 however, were able to synthesise high levels of both compounds irrespective of the presence of a functional *GSOH*
348 allele (haplotype IV) or not (Haplotype III), although the production of 2-hydroxy-3-butenyl is much higher when both
349 genes are functional (Supplemental figure 9B). While the model of redundant functions of the *GSOH* and *DMR6* gene

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

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

368 Additive effects increase the selection coefficient of specific loci, which drives the co-evolution of certain allele
369 combinations in particular environments, as was previously observed for glucosinolate metabolism (Brachi et al.,
370 2015). In line with this, the occurrence of some haplotypes deviates significantly from estimates based on population
371 allele frequencies. For instance, the non-Col-0 *GS-OH* allele almost exclusively occurs together with the Col-0 *MAM*
372 and *AOP* allele, while the combination of the nonCol-0 *AOP* allele and the Col-0 *MAM* allele is extremely rare (Figure
373 3). Interestingly, a sixth locus, encoding an epithiospecifier protein (*ESP*) involved in glucosinolate breakdown and
374 explaining a large fraction of the variation in C4-8 alkenyl glucosinolates acts hypostatic to the *GS-OH* and *MAM* loci
375 (Supplemental Table 11). A co-evolution of *ESP* with *GS-OH* and *MAM*, and to a lesser extent *DMR6*, is further
376 suggested by moderate to high levels of LD between SNPs representing the various loci, even though located on
377 different chromosomes (Supplemental Figure 10). The active nonCol-0 *ESP* allele is found in all, but one accessions
378 from France and the UK (Supplemental Table 12), suggesting a favourable production of nitriles over isothiocyanates
379 in these regions. In contrast, the vast majority of accessions from the Czech Republic carry the weak Col-0 allele,
380 further confirming strong population structure for the accumulation of these compounds.

381 Analogous to the glucosinolates and their volatile derivatives many other metabolites, including phenylpropanoids
382 and flavonoids, display strong correlations with geographical clines and climate parameters (Supplemental Table 13).
383 In addition, the loci explaining most of this variation, like *BGLU6* and *UGT78D2*, exhibit high F_{ST} values and a strong
384 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

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

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

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

439

440

441 **Material and Methods**

442 **Plant growth conditions**

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

452

453 **Metabolomics**

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

457 For the detection of non-volatiles, aqueous-methanol extracts were prepared from 50 mg frozen ground material to
458 which 200 μ l of 94 % MeOH containing 0.125 % formic acid was added (De Vos et al., 2007). Batch sizes ranged from
459 78 to 80 samples, with the exception of the last batch, batch 10, containing 48 samples (Wehrens et al., 2016). After
460 sonication and filtering, the crude extracts were analysed as described previously (van Duynhoven et al., 2014) using
461 UPLC (Waters Aquity) coupled to a high-resolution Orbitrap FTMS (Thermo). A 20 min gradient of 5–35 % acetonitril,
462 acidified with 0.1 % formic acid, at a flow rate of 400 μ l/min was used to separate compounds on a 2.1 x 150 mm²
463 C18-BEH column (1.7 μ m particle size) at 40 °C. Metabolites were detected using a LTQ-Orbitrap hybrid MS system
464 operating in negative electrospray ionization mode heated at 300 °C with a source voltage of 4.5 kV. The transfer
465 tube in the ion source was replaced and the FTMS recalibrated after each sample batch, without stopping the UPLC
466 system. After pre-processing raw data files in an untargeted manner using a Metalign-MSClust based workflow
467 (Kooke et al., 2019, Lommen and Kools, 2012, Tikunov et al., 2012), metabolites occurring in fewer than 20 different
468 genotypes were removed, leading to a data matrix containing relative intensities of 567 reconstructed metabolites
469 in 761 samples (including 51 QCs). The percentage of non-detects (*i.e.*, an intensity value below the arbitrarily set
470 detection threshold) in this matrix is 48 %. For individual metabolites, the fraction of non-detects can be much larger,
471 and in this data set is up to 97 %.

472 The detection of volatiles is based on aliquots of the same Arabidopsis material as described for the non-volatiles.
473 The aim here was to analyse volatile organic compounds (VOCs) present in the leaf material using solid phase
474 microextraction (SPME) of the headspace. Extracts of 50 mg from frozen ground material were analysed on a GC-MS
475 system (Agilent GC7890A with a quadrupole MSD Agilent 5978C) in fifteen batches of 34-99 samples, with, on
476 average, 15 study samples per QC, as described earlier (Cordovez et al., 2015, Mumm et al., 2016, Verhoeven et al.,
477 2012). In contrast to these studies, the temperature program of the GC oven started at 45 °C (2 min hold) and rose
478 first with 8–190 °C.min⁻¹, followed by 25–280 °C (2 min hold). This data set contains information on 753 samples
479 (including 50 QCs) with, in total, 40 % non-detects, similar to what was found for the non-volatiles. For individual
480 metabolites, the percentage of non-detects goes up to 97 %. As for the non-volatiles, only those volatile metabolites
481 were retained that were present in at least 20 different genotypes, in this case 603 metabolites.

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

485

486 **Descriptive statistics**

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

489 where r is the number of replicates and $MS(G)$ and $MS(E)$ are the mean sums of squares for genotype and residual
490 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
491 genetic effects (σ^2_A) in account. Marker-based estimates of narrow-sense heritability were obtained from the mixed
492 model, $y_{ij} = \mu + G_i + E_{ij}$, ($i = 1, \dots, n$, $j = 1, \dots, r$), $G \sim N(0, \sigma^2_A K)$, $E_{i,j} \sim N(0, \sigma^2_E)$, where y_{ij} is the phenotypic response of
493 replicate j of genotype i , μ is the intercept, $G = (G_1, \dots, G_n)$ is the vector of random genetic effects, and the errors E_{ij}
494 have independent normal distributions with variance σ^2_E which is the residual variance for a single individual (Kruijer
495 et al., 2014). The vector G has a multivariate $N(0, \sigma^2_A K)$ distribution, and the genetic relatedness matrix K is estimated
496 from standardized SNP-scores. The coefficient of variation (CV_G) was calculated as $CV_G = (\sigma_G / \bar{X}) * 100\%$.

497

498 **Genome-wide association mapping**

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

501 All traits were analysed with a Bayesian statistical model, BayesR, which uses a Gibbs sampling approach to estimate
502 variant effects that are modelled as a mixture distribution of four normal distributions. SNPs were assigned a prior
503 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
504 genetic variance. This allows many uninformative variants to be dropped from the model and permits remaining
505 variants to have moderate to large effects. Gibbs sampling was performed for 50,000 iterations, after 20,000 burn-
506 in iterations. Conventional GWAS was performed using the R-package statgenGWAS
507 (<https://github.com/Biometris/statgenGWAS/>), following the approach of previous studies (Kang et al., 2010, Kruijer
508 et al., 2014).

509

510 **Enrichment analysis**

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

515

516 **Fixation index**

517 A large subset of the global Arabidopsis population was divided into five sub-populations based on the countries of
518 origin that were best represented in the collection, viz. Germany (57 accessions), France (56 accessions), Sweden (48
519 accessions), UK (47 accessions), and the Czech Republic (29 accessions). These accessions together comprised 68%
520 (237 out of 350 genotypes) of the entire population. F_{ST} values were calculated using the following formula: $F_{ST} =$
521 $(\pi_{Between} - \pi_{Within}) / \pi_{between}$, where $\pi_{Between}$ and π_{Within} represent the expected heterozygosity (i.e., genetic diversity)
522 between populations and the expected heterozygosity within populations, respectively.

523

524 **Validation of *GSOH*/*DMR6* redundancy**

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

547

548 **Climate Data**

549 Climate data for the collection origin of each accession was obtained from the Climate Research Unit at the University
550 of East Anglia. Data were extracted for nine climate variables giving the average per month over a 30-year (1961-
551 1990) period (New et al., 2002). From these 9 variables, most other variables were extracted. Day length (spring) and
552 relative humidity (spring) from the site of 306 accessions were obtained from the NCEP-NCAR climate reanalysis
553 project (Hancock et al., 2011, Kistler et al., 2001) and the FAO GeoNetwork
554 (<http://fao.org.geonetworks/srv/en/main.home>).

555

556

557 **Funding**

558 This research was part of the “Learning from Nature” program, supported by the Dutch Technology Foundation
559 (STWGrant 1099), which is part of the Netherlands Organization for Scientific Research (NWO). Further support was
560 obtained from a Centre of BioSystems Genomics (CBSG) Metabolomics Hotel Project (TD16-5), A CBSG Arabidopsis
561 project (AA3-WU-PL), a Netherlands Metabolic Centre (NMC) project (Population Metabolomics – Arabidopsis
562 3370046800) and a booster grant (050-040-213) of the Netherlands Genomics Initiative (NGI) to the Consortium for
563 Improving Plant Yield (CIPY).

564
565

566 Acknowledgements

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

568
569

570 References

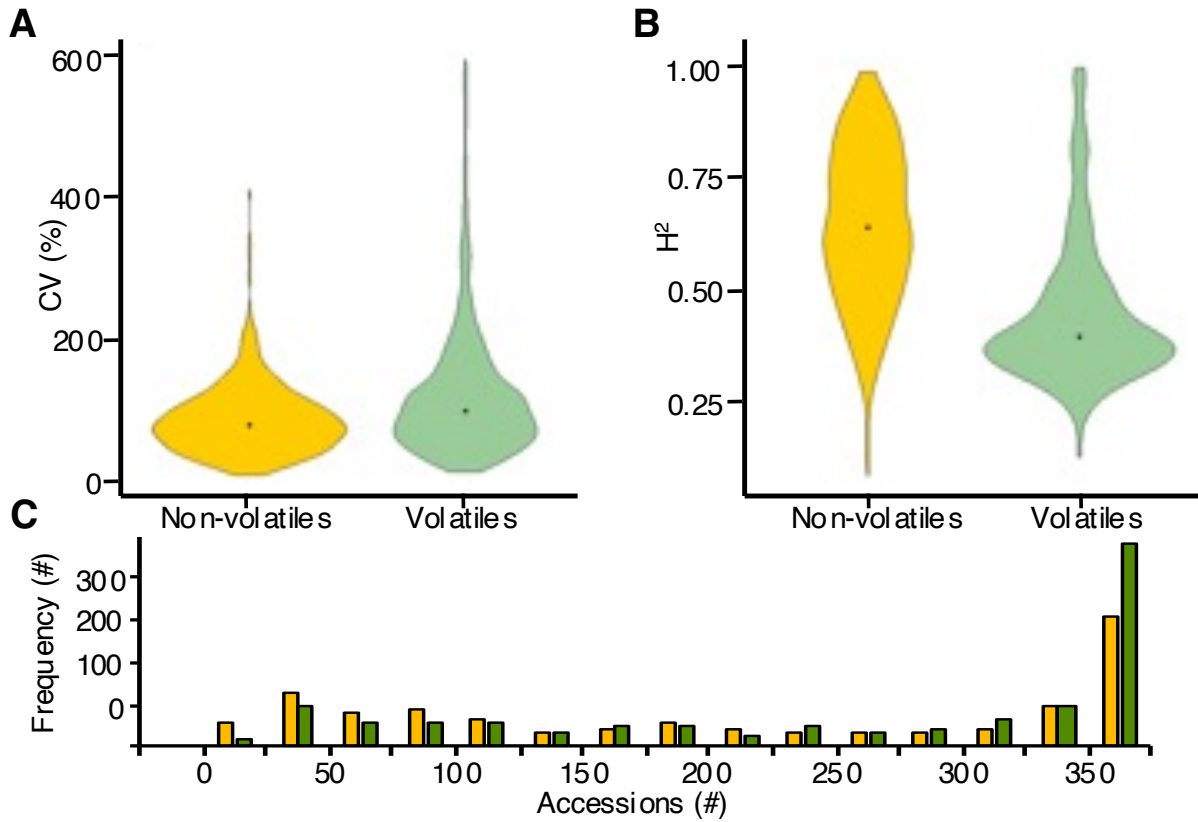
- 571 ALONSO-BLANCO, C., AARTS, M. G., BENTSINK, L., KEURENTJES, J. J., REYMOND, M., VREUGDENHIL, D. &
572 KOORNNEEF, M. 2009. What has natural variation taught us about plant development, physiology,
573 and adaptation? *Plant Cell*, 21, 1877-96.
- 574 BAC-MOLENAAR, J. A., FRADIN, E. F., RIENSTRA, J. A., VREUGDENHIL, D. & KEURENTJES, J. J. B. 2015a. GWA
575 Mapping of Anthocyanin Accumulation Reveals Balancing Selection of MYB90 in *Arabidopsis thaliana*.
576 *PLoS One*, 10, e0143212.
- 577 BAC-MOLENAAR, J. A., VREUGDENHIL, D., GRANIER, C. & KEURENTJES, J. J. B. 2015b. Genome-wide
578 association mapping of growth dynamics detects time-specific and general quantitative trait loci.
579 *Journal of Experimental Botany*, 66, 5567-5580.
- 580 BAXTER, I., BRAZELTON, J. N., YU, D., HUANG, Y. S., LAHNER, B., YAKUBOVA, E., LI, Y., BERGELSON, J.,
581 BOREVITZ, J. O., NORDBORG, M., VITEK, O. & SALT, D. E. 2010. A coastal cline in sodium accumulation
582 in *Arabidopsis thaliana* is driven by natural variation of the sodium transporter *AtHKT1;1*. *PLoS*
583 *Genet*, 6, e1001193.
- 584 BENDEROTH, M., TEXTOR, S., WINDSOR, A. J., MITCHELL-OLDS, T., GERSHENZON, J. & KROYMANN, J. 2006.
585 Positive selection driving diversification in plant secondary metabolism. *Proc Natl Acad Sci U S A*, 103,
586 9118-23.
- 587 BERGELSON, J. & ROUX, F. 2010. Towards identifying genes underlying ecologically relevant traits in
588 *Arabidopsis thaliana*. *Nat Rev Genet*, 11, 867-79.
- 589 BOYLE, E. A., LI, Y. I. & PRITCHARD, J. K. 2017. An Expanded View of Complex Traits: From Polygenic to
590 Omnigenic. *Cell*, 169, 1177-1186.
- 591 BRACHI, B., MEYER, C. G., VILLOUTREIX, R., PLATT, A., MORTON, T. C., ROUX, F. & BERGELSON, J. 2015.
592 Coselected genes determine adaptive variation in herbivore resistance throughout the native range of
593 *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*, 112, 4032-4037.
- 594 CANNELL, N., EMMS, D. M., HETHERINGTON, A. J., MACKAY, J., KELLY, S., DOLAN, L. & SWEETLOVE, L. J. 2020.
595 Multiple Metabolic Innovations and Losses Are Associated with Major Transitions in Land Plant
596 Evolution. *Curr Biol*, 30, 1783-1800.e11.
- 597 CHAE, L., LEE, I., SHIN, J. & RHEE, S. Y. 2012. Towards understanding how molecular networks evolve in
598 plants. *Curr Opin Plant Biol*, 15, 177-84.
- 599 CHAN, E. K., ROWE, H. C., HANSEN, B. G. & KLIEBENSTEIN, D. J. 2010. The complex genetic architecture of the
600 metabolome. *PLoS Genet*, 6, e1001198.
- 601 CHATEIGNER, A., LESAGE-DESCAUSES, M. C., ROGIER, O., JORGE, V., LEPLÉ, J. C., BRUNAUD, V., ROUX, C. P.,
602 SOUBIGOU-TACONNAT, L., MARTIN-MAGNIETTE, M. L., SANCHEZ, L. & SEGURA, V. 2020. Gene
603 expression predictions and networks in natural populations supports the omnigenic theory. *BMC*
604 *Genomics*, 21, 416.

- 605 CORDOVEZ, V., CARRION, V. J., ETALO, D. W., MUMM, R., ZHU, H., VAN WEZEL, G. P. & RAAIJMAKERS, J. M.
606 2015. Diversity and functions of volatile organic compounds produced by *Streptomyces* from a
607 disease-suppressive soil. *Front Microbiol*, 6, 1081.
- 608 DAVILA OLIVAS, N. H., KRUIJER, W., GORT, G., WIJNEN, C. L., VAN LOON, J. J. A. & DICKE, M. C. 2017. Genome-
609 wide association analysis reveals distinct genetic architectures for single and combined stress
610 responses in *Arabidopsis thaliana*. *New Phytologist*, 213, 838-851.
- 611 DE LOS CAMPOS, G., PÉREZ, P., VAZQUEZ, A. I. & CROSSA, J. 2013. Genome-enabled prediction using the BLR
612 (Bayesian Linear Regression) R-package. *Methods Mol Biol*, 1019, 299-320.
- 613 DE VOS, R. C., MOCO, S., LOMMEN, A., KEURENTJES, J. J. B., BINO, R. J. & HALL, R. D. 2007. Untargeted large-
614 scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat Protoc*, 2,
615 778-91.
- 616 DUBIN, M. J., ZHANG, P., MENG, D., REMIGEREAU, M.-S., OSBORNE, E. J., PAOLO CASALE, F., DREWE, P.,
617 KAHLES, A., JEAN, G., VILHJÁLMSSON, B., JAGODA, J., IREZ, S., VORONIN, V., SONG, Q., LONG, Q.,
618 RÄTSCH, G., STEGLE, O., CLARK, R. M. & NORDBORG, M. 2015. DNA methylation in *Arabidopsis* has a
619 genetic basis and shows evidence of local adaptation. *eLife*, 4, e05255.
- 620 FALCONE FERREYRA, M. L., EMILIANI, J., RODRIGUEZ, E. J., CAMPOS-BERMUDEZ, V. A., GROTEWOLD, E. &
621 CASATI, P. 2015. The Identification of Maize and *Arabidopsis* Type I FLAVONE SYNTHASEs Links
622 Flavones with Hormones and Biotic Interactions. *Plant Physiol*, 169, 1090-107.
- 623 FANG, C., FERNIE, A. R. & LUO, J. 2019. Exploring the Diversity of Plant Metabolism. *Trends Plant Sci*, 24, 83-98.
- 624 FILIAULT, D. L., SEYMOUR, D. K., MARUTHACHALAM, R. & MALOOF, J. N. 2017. The Generation of Doubled
625 Haploid Lines for QTL Mapping. *Methods Mol Biol*, 1610, 39-57.
- 626 FU, J., KEURENTJES, J. J. B., BOUWMEESTER, H., AMERICA, T., VERSTAPPEN, F. W., WARD, J. L., BEALE, M. H.,
627 DE VOS, R. C., DIJKSTRA, M., SCHELTEMA, R. A., JOHANNES, F., KOORNNEEF, M., VREUGDENHIL, D.,
628 BREITLING, R. & JANSEN, R. C. 2009. System-wide molecular evidence for phenotypic buffering in
629 *Arabidopsis*. *Nat Genet*, 41, 166-7.
- 630 FUSARI, C. M., KOOKE, R., LAUXMANN, M. A., ANNUNZIATA, M. G., ENKE, B., HOEHNE, M., KROHN, N., BECKER,
631 F. F. M., SCHLERETH, A., Sulpice, R., STITT, M. & KEURENTJES, J. J. B. 2017. Genome-Wide
632 Association Mapping Reveals That Specific and Pleiotropic Regulatory Mechanisms Fine-Tune Central
633 Metabolism and Growth in *Arabidopsis*. *The Plant Cell*, 29, 2349-2373.
- 634 GOUJON, T., SIBOUT, R., POLLET, B., MABA, B., NUSSAUME, L., BECHTOLD, N., LU, F., RALPH, J., MILA, I.,
635 BARRIÈRE, Y., LAPIERRE, C. & JOUANIN, L. 2003. A new *Arabidopsis thaliana* mutant deficient in the
636 expression of O-methyltransferase impacts lignins and sinapoyl esters. *Plant Molecular Biology*, 51,
637 973-989.
- 638 HANCOCK, A. M., BRACHI, B., FAURE, N., HORTON, M. W., JARYMOWYCZ, L. B., SPERONE, F. G., TOOMAJIAN, C.,
639 ROUX, F. & BERGELSON, J. 2011. Adaptation to climate across the *Arabidopsis thaliana* genome.
640 *Science*, 334, 83-6.
- 641 HANSEN, B. G., KERWIN, R. E., OBER, J. A., LAMBRIX, V. M., MITCHELL-OLDS, T., GERSHENZON, J., HALKIER, B.
642 A. & KLIEBENSTEIN, D. J. 2008. A Novel 2-Oxoacid-Dependent Dioxygenase Involved in the Formation
643 of the Goiterogenic 2-Hydroxybut-3-enyl Glucosinolate and Generalist Insect Resistance in
644 *Arabidopsis*. *Plant Physiology*, 148, 2096-2108.
- 645 HERMAN, P. L. & MARKS, M. D. 1989. Trichome Development in *Arabidopsis thaliana*. II. Isolation and
646 Complementation of the GLABROUS1 Gene. *Plant Cell*, 1, 1051-1055.
- 647 HOLSINGER, K. E. & WEIR, B. S. 2009. Genetics in geographically structured populations: defining, estimating
648 and interpreting F(ST). *Nat Rev Genet*, 10, 639-50.
- 649 HORTON, M. W., HANCOCK, A. M., HUANG, Y. S., TOOMAJIAN, C., ATWELL, S., AUTON, A., MULIYATI, N. W.,
650 PLATT, A., SPERONE, F. G., VILHJÁLMSSON, B. J., NORDBORG, M., BOREVITZ, J. O. & BERGELSON, J.
651 2012. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from
652 the RegMap panel. *Nat Genet*, 44, 212-6.
- 653 HUANG, D. W., SHERMAN, B. T. & LEMPICKI, R. A. 2008. Systematic and integrative analysis of large gene lists
654 using DAVID bioinformatics resources. 4, 44.
- 655 ISHIHARA, H., TOHGE, T., VIEHÖVER, P., FERNIE, A. R., WEISSHAAR, B. & STRACKE, R. 2016. Natural variation
656 in flavonol accumulation in *Arabidopsis* is determined by the flavonol glucosyltransferase BGLU6.
657 *Journal of Experimental Botany*, 67, 1505-1517.

- 658 JOSEPH, B., CORWIN, J. A. & KLIEBENSTEIN, D. J. 2015. Genetic variation in the nuclear and organellar
659 genomes modulates stochastic variation in the metabolome, growth, and defense. *PLoS Genet*, 11,
660 e1004779.
- 661 JOSEPHS, E. B., STINCHCOMBE, J. R. & WRIGHT, S. I. C. 2017. What can genome-wide association studies tell us
662 about the evolutionary forces maintaining genetic variation for quantitative traits? *New Phytologist*,
663 214, 21-33.
- 664 KANG, H. M., SUL, J. H., SERVICE, S. K., ZAITLEN, N. A., KONG, S. Y., FREIMER, N. B., SABATTI, C. & ESKIN, E.
665 2010. Variance component model to account for sample structure in genome-wide association
666 studies. *Nat Genet*, 42, 348-54.
- 667 KERWIN, R., FEUSIER, J., CORWIN, J., RUBIN, M., LIN, C., MUOK, A., LARSON, B., LI, B., JOSEPH, B., FRANCISCO,
668 M., COPELAND, D., WEINIG, C. & KLIEBENSTEIN, D. J. 2015. Natural genetic variation in *Arabidopsis*
669 *thaliana* defense metabolism genes modulates field fitness. *eLife*, 4, e05604.
- 670 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.
671 C., VREUGDENHIL, D. & KOORNNEEF, M. 2006. The genetics of plant metabolism. *Nat Genet*, 38, 842-
672 9.
- 673 KEURENTJES, J. J. B., SULPICE, R., GIBON, Y., STEINHAUSER, M. C., FU, J., KOORNNEEF, M., STITT, M. &
674 VREUGDENHIL, D. 2008. Integrative analyses of genetic variation in enzyme activities of primary
675 carbohydrate metabolism reveal distinct modes of regulation in *Arabidopsis thaliana*. *Genome Biol*, 9,
676 R129.
- 677 KIM, S.-J., KIM, M.-R., BEDGAR, D. L., MOINUDDIN, S. G. A., CARDENAS, C. L., DAVIN, L. B., KANG, C. & LEWIS, N.
678 G. 2004. Functional reclassification of the putative cinnamyl alcohol dehydrogenase multigene family
679 in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 101,
680 1455-1460.
- 681 KISTLER, R., KALNAY, E., COLLINS, W., SAHA, S., WHITE, G., WOOLLEN, J., CHELLIAH, M., EBISUZAKI, W.,
682 KANAMITSU, M., KOUSKY, V., VAN DEN DOOL, H., JENNE, R. & FIORINO, M. 2001. The NCEP-NCAR 50-
683 year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological*
684 *Society*, 82, 247-267.
- 685 KLIEBENSTEIN, D. 2009. Advancing genetic theory and application by metabolic quantitative trait loci
686 analysis. *Plant Cell*, 21, 1637-46.
- 687 KLIEBENSTEIN, D. J., CACHO, N. I. & STANISLAV, K. 2016. Chapter Three - Nonlinear Selection and a Blend
688 of Convergent, Divergent and Parallel Evolution Shapes Natural Variation in Glucosinolates. *Advances*
689 *in Botanical Research*. Academic Press.
- 690 KLIEBENSTEIN, D. J., LAMBRIX, V. M., REICHEL, M., GERSHENZON, J. & MITCHELL-OLDS, T. 2001. Gene
691 duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent
692 dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *Plant Cell*, 13, 681-93.
- 693 KOOKE, R. & KEURENTJES, J. J. B. 2012. Multi-dimensional regulation of metabolic networks shaping plant
694 development and performance. *Journal of Experimental Botany*, 63, 3353-3365.
- 695 KOOKE, R., KRUIJER, W., BOURS, R., BECKER, F. F. M., KUHN, A., GEEST, H. V. D., BUNTJER, J., DOESWIJK, T.,
696 GUERRA, J., BOUWMEESTER, H. J., VREUGDENHIL, D. & KEURENTJES, J. J. B. 2016. Genome-wide
697 association mapping and genomic prediction elucidate the genetic architecture of morphological
698 traits in *Arabidopsis thaliana*. *Plant Physiology*.
- 699 KOOKE, R., MORGADO, L., BECKER, F., VAN EEKELLEN, H., HAZARIKA, R., ZHENG, Q., DE VOS, R. C. H.,
700 JOHANNES, F. & KEURENTJES, J. J. B. 2019. Epigenetic mapping of the *Arabidopsis* metabolome
701 reveals mediators of the epigenotype-phenotype map. *Genome Res*, 29, 96-106.
- 702 KORTE, A. & FARLOW, A. 2013. The advantages and limitations of trait analysis with GWAS: a review. *Plant*
703 *Methods*, 9, 29.
- 704 KOVER, P. X. & MOTT, R. 2012. Mapping the genetic basis of ecologically and evolutionarily relevant traits in
705 *Arabidopsis thaliana*. *Curr Opin Plant Biol*, 15, 212-7.
- 706 KROYMANN, J., TEXTOR, S., TOKUHISA, J. G., FALK, K. L., BARTRAM, S., GERSHENZON, J. & MITCHELL-OLDS, T.
707 2001. A gene controlling variation in *Arabidopsis* glucosinolate composition is part of the methionine
708 chain elongation pathway. *Plant Physiol*, 127, 1077-88.
- 709 KRUIJER, W., BOER, M., MALOSETTI, M., FLOOD, P. J., ENGEL, B., KOOKE, R., KEURENTJES, J. J. B. & EEUWIJK, F.
710 V. 2014. Marker-Based Estimation of Heritability in Immortal Populations. *Genetics*.

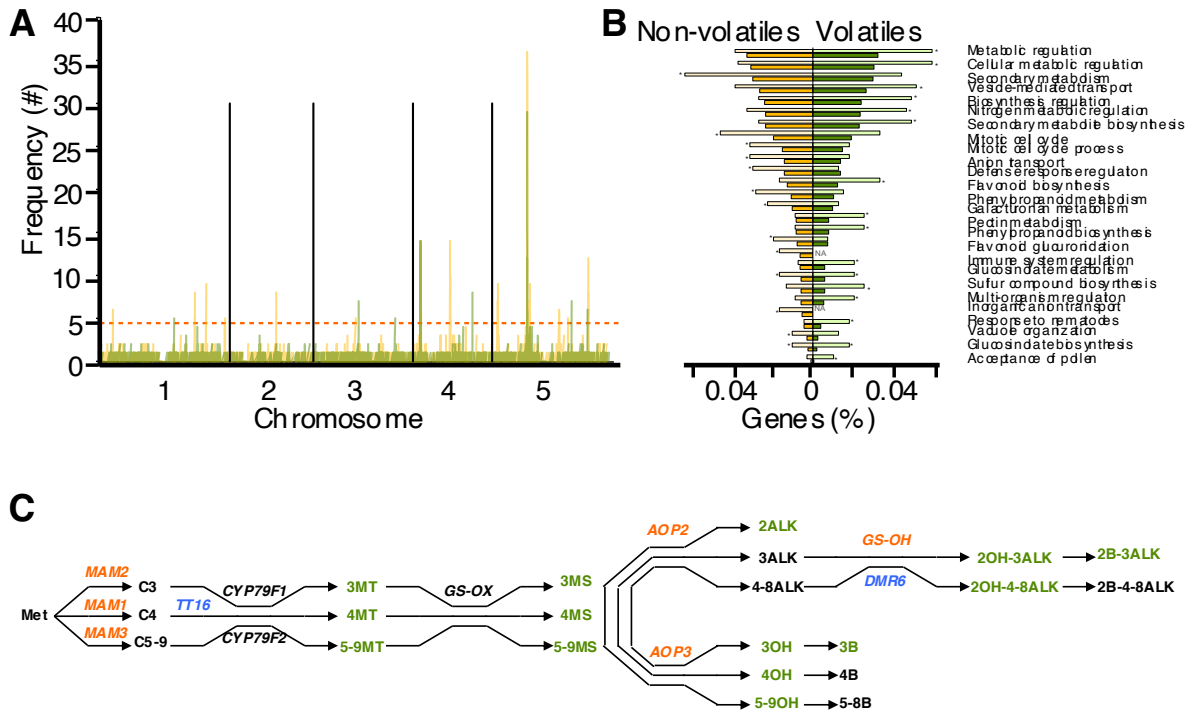
- 711 LI, X., TIEMAN, D., LIU, Z., CHEN, K. & KLEE, H. J. 2020. Identification of a lipase gene with a role in tomato fruit
712 short-chain fatty acid-derived flavor volatiles by genome-wide association. *The Plant Journal*, 104,
713 631-644.
- 714 LI, Y., HUANG, Y., BERGELSON, J., NORDBORG, M. & BOREVITZ, J. O. 2010. Association mapping of local
715 climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*, 107, 21199-
716 204.
- 717 LIPPERT, C., LISTGARTEN, J., LIU, Y., KADIE, C. M., DAVIDSON, R. I. & HECKERMAN, D. 2011. FaST linear mixed
718 models for genome-wide association studies. *Nat Methods*, 8, 833-5.
- 719 LISTGARTEN, J., KADIE, C., SCHADT, E. E. & HECKERMAN, D. 2010. Correction for hidden confounders in the
720 genetic analysis of gene expression. *Proc Natl Acad Sci U S A*, 107, 16465-70.
- 721 LOMMEN, A. & KOOLS, H. J. 2012. MetAlign 3.0: performance enhancement by efficient use of advances in
722 computer hardware. *Metabolomics*, 8, 719-726.
- 723 LYU, J. 2017. *Arabidopsis* evolution: Roots in Africa. *Nat Plants*, 3, 17091.
- 724 MARKS, M. D. & FELDMANN, K. A. 1989. Trichome Development in *Arabidopsis thaliana*. I. T-DNA Tagging of
725 the GLABROUS1 Gene. *Plant Cell*, 1, 1043-1050.
- 726 MOSER, G., LEE, S. H., HAYES, B. J., GODDARD, M. E., WRAY, N. R. & VISSCHER, P. M. 2015. Simultaneous
727 discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. *PLoS*
728 *genetics*, 11, e1004969.
- 729 MUMM, R., HAGEMAN, J. A., CALINGACION, M. N., DE VOS, R. C. H., JONKER, H. H., ERBAN, A., KOPKA, J.,
730 HANSEN, T. H., LAURSEN, K. H., SCHJOERRING, J. K., WARD, J. L., BEALE, M. H., JONGEE, S., RAUF, A.,
731 HABIBI, F., INDRASARI, S. D., SAKHAN, S., RAMLI, A., ROMERO, M., REINKE, R. F., OHTSUBO, K.,
732 BOUALAPHANH, C., FITZGERALD, M. A. & HALL, R. D. 2016. Multi-platform metabolomics analyses of
733 a broad collection of fragrant and non-fragrant rice varieties reveals the high complexity of grain
734 quality characteristics. *Metabolomics*, 12, 38.
- 735 NEW, M., LISTER, D., HULME, M. & MAKIN, I. 2002. A high-resolution data set of surface climate over global
736 land areas. *Climate Research*, 21, 1-25.
- 737 OLSON-MANNING, C. F., WAGNER, M. R. & MITCHELL-OLDS, T. 2012. Adaptive evolution: evaluating empirical
738 support for theoretical predictions. 13, 867.
- 739 PAYNE, J. L. & WAGNER, A. 2019. The causes of evolvability and their evolution. *Nat Rev Genet*, 20, 24-38.
- 740 PEREZ DE SOUZA, L., NAAKE, T., TOHGE, T. & FERNIE, A. R. 2017. From chromatogram to analyte to
741 metabolite. How to pick horses for courses from the massive web resources for mass spectral plant
742 metabolomics. *Gigascience*, 6, 1-20.
- 743 ROCKMAN, M. V. 2012. THE QTN PROGRAM AND THE ALLELES THAT MATTER FOR EVOLUTION: ALL THAT'S
744 GOLD DOES NOT GLITTER. *Evolution*, 66, 1-17.
- 745 ROSS, J. R., NAM, K. H., D'AURIA, J. C. & PICHERSKY, E. 1999. S-Adenosyl-l-Methionine:Salicylic Acid Carboxyl
746 Methyltransferase, an Enzyme Involved in Floral Scent Production and Plant Defense, Represents a
747 New Class of Plant Methyltransferases. *Archives of Biochemistry and Biophysics*, 367, 9-16.
- 748 SALEM, M. A., PEREZ DE SOUZA, L., SERAG, A., FERNIE, A. R., FARAG, M. A., EZZAT, S. M. & ALSEEKH, S. 2020.
749 Metabolomics in the Context of Plant Natural Products Research: From Sample Preparation to
750 Metabolite Analysis. *Metabolites*, 10.
- 751 SEGRÈ, D., VITKUP, D. & CHURCH, G. M. 2002. Analysis of optimality in natural and perturbed metabolic
752 networks. *Proc Natl Acad Sci U S A*, 99, 15112-7.
- 753 SHIMIZU, K. K. & PURUGGANAN, M. D. 2005. Evolutionary and ecological genomics of *Arabidopsis*. *Plant*
754 *Physiol*, 138, 578-84.
- 755 SMITH, S. M. & MAUGHAN, P. J. 2015. SNP genotyping using KASPar assays. *Methods Mol Biol*, 1245, 243-56.
- 756 STRAUCH, R. C., SVEDIN, E., DILKES, B., CHAPPLE, C. & LI, X. 2015. Discovery of a novel amino acid racemase
757 through exploration of natural variation in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A*, 112, 11726-
758 31.
- 759 TEXTOR, S., BARTRAM, S., KROYMANN, J., FALK, K. L., HICK, A., PICKETT, J. A. & GERSHENZON, J. 2004.
760 Biosynthesis of methionine-derived glucosinolates in *Arabidopsis thaliana*: recombinant expression
761 and characterization of methylthioalkylmalate synthase, the condensing enzyme of the chain-
762 elongation cycle. *Planta*, 218, 1026-1035.
- 763 THOEN, M. P. M., DAVILA OLIVAS, N. H., KLOTH, K. J., COOLEN, S., HUANG, P.-P., AARTS, M. G. M., BAC-
764 MOLENAAR, J. A., BAKKER, J., BOUWMEESTER, H. J., BROEKGAARDEN, C., BUCHER, J., BUSSCHER-
765 LANGE, J., CHENG, X., FRADIN, E. F., JONGSMA, M. A., JULKOWSKA, M. M., KEURENTJES, J. J. B.,

- 766 LIGTERINK, W., PIETERSE, C. M. J., RUYTER-SPIRA, C., SMANT, G., TESTERINK, C., USADEL, B., VAN
767 LOON, J. J. A., VAN PELT, J. A., VAN SCHAİK, C. C., VAN WEES, S. C. M., VISSER, R. G. F., VOORRIPS, R.,
768 VOSMAN, B., VREUGDENHIL, D., WARMERDAM, S., WIEGERS, G. L., VAN HEERWAARDEN, J., KRUIJER,
769 W., VAN EEUWIJK, F. A. & DICKE, M. C. 2017. Genetic architecture of plant stress resistance: multi-
770 trait genome-wide association mapping. *New Phytologist*, 213, 1346-1362.
- 771 TIKUNOV, Y. M., LAPTENOK, S., HALL, R. D., BOVY, A. & DE VOS, R. C. H. 2012. MSclust: a tool for unsupervised
772 mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. *Metabolomics*,
773 8, 714-8.
- 774 TODESCO, M., BALASUBRAMANIAN, S., HU, T. T., TRAW, M. B., HORTON, M., EPPLE, P., KUHNS, C.,
775 SURESHKUMAR, S., SCHWARTZ, C., LANZ, C., LAITINEN, R. A., HUANG, Y., CHORY, J., LIPKA, V.,
776 BOREVITZ, J. O., DANGL, J. L., BERGELSON, J., NORDBORG, M. & WEIGEL, D. 2010. Natural allelic
777 variation underlying a major fitness trade-off in *Arabidopsis thaliana*. *Nature*, 465, 632-6.
- 778 TRONTIN, C., TISNÉ, S., BACH, L. & LOUDET, O. 2011. What does *Arabidopsis* natural variation teach us (and
779 does not teach us) about adaptation in plants? *Curr Opin Plant Biol*, 14, 225-31.
- 780 VAN DUYNHOVEN, J., VAN DER HOOFT, J. J., VAN DORSTEN, F. A., PETERS, S., FOLTZ, M., GOMEZ-ROLDAN, V.,
781 VERVOORT, J., DE VOS, R. C. & JACOBS, D. M. 2014. Rapid and sustained systemic circulation of
782 conjugated gut microbial catabolites after single-dose black tea extract consumption. *J Proteome Res*,
783 13, 2668-78.
- 784 VERHOEVEN, H. A., JONKER, H., DE VOS, R. C. & HALL, R. D. 2012. Solid phase micro-extraction GC-MS analysis
785 of natural volatile components in melon and rice. *Methods Mol Biol*, 860, 85-99.
- 786 WALSH, B. & LYNCH, M. 2018. *Evolution and Selection of Quantitative Traits*, New York, U.S.A., Oxford
787 University Press.
- 788 WANG, Z., YANG, Z. & LI, F. 2019. Updates on molecular mechanisms in the development of branched trichome
789 in *Arabidopsis* and nonbranched in cotton. *Plant Biotechnol J*, 17, 1706-1722.
- 790 WARDE-FARLEY, D., DONALDSON, S. L., COMES, O., ZUBERI, K., BADRAWI, R., CHAO, P., FRANZ, M., GROUIOS,
791 C., KAZI, F., LOPES, C. T., MAITLAND, A., MOSTAFAVI, S., MONTOJO, J., SHAO, Q., WRIGHT, G., BADER, G.
792 D. & MORRIS, Q. 2010. The GeneMANIA prediction server: biological network integration for gene
793 prioritization and predicting gene function. *Nucleic Acids Res*, 38, W214-20.
- 794 WEHRENS, R., HAGEMAN, J. A., VAN EEUWIJK, F., KOOKE, R., FLOOD, P. J., WIJNKER, E., KEURENTJES, J. J. B.,
795 LOMMEN, A., VAN EEKELEN, H. D., HALL, R. D., MUMM, R. & DE VOS, R. C. 2016. Improved batch
796 correction in untargeted MS-based metabolomics. *Metabolomics*, 12, 88.
- 797 WITJES, L., KOOKE, R., VAN DER HOOFT, J. J. J., DE VOS, R. C. H., KEURENTJES, J. J. B., MEDEMA, M. H. &
798 NIJVEEN, H. 2019. A genetical metabolomics approach for bioprospecting plant biosynthetic gene
799 clusters. *BMC Res Notes*, 12, 194.
- 800 WU, S., ALSEEKH, S., CUADROS-INOSTROZA, Á., FUSARI, C. M., MUTWIL, M., KOOKE, R., KEURENTJES, J. J. B.,
801 FERNIE, A. R., WILLMITZER, L. & BROTMAN, Y. 2016. Combined Use of Genome-Wide Association
802 Data and Correlation Networks Unravels Key Regulators of Primary Metabolism in *Arabidopsis*
803 *thaliana*. *PLoS Genet*, 12, e1006363.
- 804 WU, S., TOHGE, T., CUADROS-INOSTROZA, Á., TONG, H., TENENBOIM, H., KOOKE, R., MÉRET, M., KEURENTJES,
805 J. J. B., NIKOLOSKI, Z., FERNIE, A. R., WILLMITZER, L. & BROTMAN, Y. 2018. Mapping the *Arabidopsis*
806 Metabolic Landscape by Untargeted Metabolomics at Different Environmental Conditions. *Mol Plant*,
807 11, 118-134.
- 808 XU, W., BOBET, S., LE GOURRIEREC, J., GRAIN, D., DE VOS, D., BERGER, A., SALSAC, F., KELEMEN, Z.,
809 BOUCHEREZ, J., ROLLAND, A., MOUILLE, G., ROUTABOUL, J. M., LEPINIEC, L. & DUBOS, C. 2017.
810 TRANSPARENT TESTA 16 and 15 act through different mechanisms to control proanthocyanidin
811 accumulation in *Arabidopsis* testa. *J Exp Bot*, 68, 2859-2870.
- 812 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,
813 R. 2020. Metabolite signatures of diverse *Camellia sinensis* tea populations. *Nat Commun*, 11, 5586.
- 814 ZHANG, W., ALSEEKH, S., ZHU, X., ZHANG, Q., FERNIE, A. R., KUANG, H. & WEN, W. 2020. Dissection of the
815 domestication-shaped genetic architecture of lettuce primary metabolism. *The Plant Journal*, 104,
816 613-630.
- 817 ZHOU, X. & STEPHENS, M. 2012. Genome-wide efficient mixed-model analysis for association studies. *Nat*
818 *Genet*, 44, 821-4.
- 819
- 820



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



829
 830 **Figure 2:** Genetic mapping of metabolic variation.
 831 (A) Frequency distribution of QTLs over the five chromosomes of Arabidopsis. A QTL hotspot is defined as a locus
 832 associated with variation in at least five metabolites. Yellow and green bars indicate the number of non-volatile and
 833 volatile metabolites for which a QTL was detected, respectively. The red dotted horizontal line indicates the QTL
 834 hotspot threshold. (B) Enrichment analysis of candidate genes associated with variation in metabolic content. Dark
 835 yellow and green bars indicate the reference genome-wide percentage of genes involved in the processes plotted on
 836 the right. Light yellow and green bars indicate this percentage for candidate genes, associated with variation in non-
 837 volatile and volatile metabolites, respectively. Asterisks indicate significant enrichment ($P_{BENJAMINI} < 0.01$), while NA
 838 denotes less than two candidate genes involved in the respective process. (C) Biosynthesis pathway of glucosinolates.
 839 Genes involved in biosynthesis are plotted above arrows. Red genes were assigned as candidate genes explaining the
 840 observed variation in glucosinolates. Blue genes have not previously been annotated as biosynthesis genes but were
 841 strongly associated with the observed variation in glucosinolates. Synthesised glucosinolates are plotted following
 842 arrows. Green glucosinolates were detected and putatively annotated in the current study. C3, C4 and C5-9 indicate
 843 glucosinolate carbon-chain-length; MT is Methylthioalkyl; MS is Methylsulfinylalkyl; ALK is Alkenyl; OH is
 844 Hydroxyalkyl; OH-ALK is Hydroxyalkenyl; B is Benzoyloxyalkyl; B-ALK is Benzoyloxyalkenyl.

845
 846

Haplotype	nC														C													
	nC							C							nC							C						
	nC		C		nC		C		nC		C		nC		C		nC		C		nC		C					
	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C				
MAM																												
AOP																												
GS-OH																												
TT16																												
DMR6																												

Cluster	nC														C													
	nC							C							nC							C						
	nC		C		nC		C		nC		C		nC		C		nC		C		nC		C					
	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C				
Fra + UK					1																							
Swe + USA									2																			
Cze	2				8	58																						
Ger		1				1																						

Chemotype	nC														C													
	nC							C							nC							C						
	nC		C		nC		C		nC		C		nC		C		nC		C		nC		C					
	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C	nC	C				
C3 ALK	3,9				4,8	4,2																						
C3 OH	5,6	5,1			5,4	5,4																						
C4 ALK		3,7			4,5	4																						
C4 MT/MS	4,5	5,5			4,4	4,4																						
C5-9	4,8	4,9			4,6	4,7																						

847
848 **Figure 3:** Genotypic and geographic patterns of adaptation of glucosinolate profiles.
849 The upper panel divides haplotypes based on the genotypes of five loci involved in the biosynthesis of glucosinolates.
850 C denotes Col-0 alleles, while nC denotes non-Col-0 alleles. The second panel divides the collection of Arabidopsis
851 accessions analyzed in three geographical clusters based on hierarchical clustering of metabolite profiles. Columns
852 align with the observed haplotype indicated in the upper panel. For each cluster and haplotype the number of
853 accessions is given. The lower panel divides the detected glucosinolates in five classes. Columns align again with the
854 observed haplotype indicated in the upper panel. For each haplotype the relative average abundance of
855 glucosinolates assigned to a specific chemotype is given.
856
857