Title: Ectopic gene conversion causing quantitative trait variation

Authors: Marina Pfalz[1], Seïf-Eddine Naàdja[1], Jacqui Anne Shykoff[1], Juergen Kroymann[1*]

Affiliations: [1]Ecologie Société Evolution, CNRS/Université Paris-Saclay/AgroParisTech, Gif-sur-Yvette, France

*juergen.kroymann@universite-paris-saclay.fr

Abstract: Mutation-selection balance explains the occurrence of genetic variation in populations: mutations arise continuously, and natural selection removes or fixes them at varying rates. However, there is more genetic variation segregating in populations than can be explained by mutation-selection balance alone. Here we elucidate a general mechanism for generating this segregating variation using the example of a gene family involved in plant defense.

Main Text:

Why is there so much genetic variation segregating in populations (*1*) when we expect natural selection to purge deleterious mutants rapidly? Clearly, other processes than mutation must be generating this variation. Here we propose a scenario whereby ectopic gene conversion (EGC) generates fitness variation in populations, illustrated by dissection of a quantitative trait locus for glucosinolates in Arabidopsis.

Glucosinolates are amino-acid derived defense metabolites found in Brassicales. Due to their extensive intra- and interspecific genetic variation, they have become model compounds for understanding the genomic architecture of adaptive quantitative traits (*2-4*). Indole glucosinolates (IGs) are derived from tryptophan, are inducible by microbial pathogens or phytophagous insects, and play an important role in plant defense (*5-9*). We mapped leaf QTL for these IGs in Arabidopsis and cloned the *Indole Glucosinolate Modifier 1* (*IGM1*) QTL on chromosome 5 (*9,10*), which controls variation for 4-hydroxy-indol-3ylmethyl glucosinolate (4OHI3M) and 4-methoxy-indol-3ylmethyl glucosinolate (4MOI3M). A second QTL on chromosome 1, termed *Indole Glucosinolate Modifier 2* (*IGM2*), which we dissect here, controls variation for 4OHI3M but not 4MOI3M (*10*).

A tightly linked gene family of indole glucosinolate *O*-methyltransferase genes (*IGMT*s) underlies *IGM2*, via a complex determination of metabolite levels. While dissecting the QTL, we discovered patterns of nucleotide polymorphism consistent with ectopic gene conversion (EGC), a process of non-reciprocal transfer of genetic material between paralogous loci during DNA repair. This was interesting because sequence transfer between specialized gene copies places sequence information into an inappropriate expression context, which should produce maladapted variants. Here we demonstrate that the QTL genes are shaped by EGC, but that selection acts against EGC near codons specifying fixed derived amino acids, *i.e.*, mutations that differentiate *Arabidopsis thaliana* gene copies from those of *Arabidopsis lyrata*. This allows us to conclude that gene copies have specialized upon subfunctionalization and that EGC indeed generates maladapted variants under these conditions.

38

39 **Results**

40 **Mapping and dissection of *IGM2***

41 We fine-mapped *IGM2* using F_2 progeny from a cross between two near-isogenic lines (NILs) of the
42 Arabidopsis Da(1)-12 × Ei-2 mapping population, DE096 and DE155 (*10*). In leaves, DE096 had more
43 4OHI3M than DE155, but the NILs did not differ in their 4MOI3M content. In roots, DE096 had more
44 4OHI3M and less 4MOI3M than DE155 (**fig. S1, A**). DE096 and DE155 differed only in a ~ 20 Mbp region
45 containing *IGM2*, while their marker genotypes were identical for the rest of the genome (**fig. S1, B**). We
46 phenotyped for the fraction of 4OHI3M in the combined amount of 4OHI3M and 4MOI3M. *IGM2*
47 mapped to a ca. 1.6 Mbp interval, with no evidence of additional QTL nearby (**fig. S1, C and D**). Statistical
48 support for a QTL centered near a family of four tandemly arranged methyltransferase genes. These
49 genes were previously identified as coding for indole glucosinolate *O*-methyltransferases (IGMTs) that
50 convert 4OHI3M to 4MOI3M (*11*), making them plausible candidates for the QTL (**fig. S1, E and F**).

51 We conducted quantitative real-time PCR to assess whether differences in *IGMT* expression could
52 explain the QTL. We used primers specific to the target genes, ensuring substantial mismatches with
53 other gene family members while avoiding polymorphisms at primer binding sites between accessions
54 (**table S1**). Additionally, we employed a primer pair that could simultaneously amplify all genes. Leaf
55 *IGMT1* and *3* transcript levels were higher in DE155 than in DE096, as was the overall transcript level,
56 while *IGMT2* and *4* transcript levels did not differ between the two NILs (**Fig. 1**). In roots, *IGMT1*
57 transcript levels were higher in DE155, but *IGMT3*, *4*, and particularly total *IGMT* transcripts were lower.
58 Thus, genotypic differences in *IGMT* transcript abundance alone could not explain the observed
59 quantitative metabolite pattern.

60 Therefore, we employed CRISPR/Cas9 to generate *igmt* mutants in both NILs. Our approach involved a
61 construct that produced three single guide RNAs (sgRNAs), each targeting distinct sites within the four
62 *IGMT* sequences. However, due to sequence variation, the sgRNAs did not match precisely in every
63 instance. We Sanger-sequenced all Cas9 target sites within the four *IGMT* genes in our mutant lines.
64 Typically, we detected small insertions or deletions that caused frameshifts. However, in certain cases,
65 we observed larger deletions resulting in the fusion of mismatched sequence from neighboring genes
66 (**fig. S2**).

67 We obtained quadruple knockouts, $igmt1\text{-}4_{DE155}$ and $igmt1\text{-}4_{DE096}$, for both NILs. While the leaves of
68 these mutants lacked 4MOI3M entirely, the roots still exhibited a background level of 4MOI3M,
69 suggesting the presence of another enzyme capable of converting 4OHI3M to 4MOI3M in roots. Our
70 prime candidate was IGMT5, encoded on chromosome 5, which normally converts 1-hydroxy-indol-
71 3ylmethyl to 1-methoxy-indole-3ylmethyl glucosinolate and has high root but low leaf activity (*5*). We
72 therefore generated an *igmt1-4* quadruple knockout in the *igmt5* mutant background. As expected,
73 these plants showed no detectable 4MOI3M in either leaves or roots (**Table 1**).

74 In addition to the quadruple mutant, referred to as $igmt1\text{-}4_{DE155}$, we obtained the genotypes $igmt3_{DE155}$,
75 $igmt1/3_{DE155}$, and $igmt1\text{-}3_{DE155}$ for DE155, and $igmt1/4_{DE096}$, $igmt1/3/4_{DE096}$, and $igmt1/2/4_{DE096}$ for DE096
76 **(fig. S2)**.

77 In leaves, we observed the most substantial differences in the levels of 4OHI3M, 4MOI3M and/or the
78 fraction of 4OHI3M when comparing *igmt1-3*$_{DE155}$ with *igmt1-4*$_{DE155}$, and DE096 wildtype with
79 *igmt1/4*$_{DE096}$ (**Fig. 2**). Thus, *IGMT4* had the largest impact on 4MOI3M generation in the leaves of both
80 NILs. Additionally, we found noticeable differences between *igmt1/2/4*$_{DE096}$ and *igmt1/4*$_{DE096}$, and
81 between *igmt1/3/4*$_{DE096}$ and *igmt1-4*$_{DE096}$ (**Fig. 2**), which highlighted the role of *IGMT2*$_{DE096}$. In stark
82 contrast, we only found minor differences for *IGMT2*$_{DE155}$ when comparing *igmt1/3*$_{DE155}$ and *igmt1-3*$_{DE155}$,
83 indicating that genotypic variation in *IGMT2* contributed to the leaf QTL.

84 However, given that the generation of each 4MOI3M molecule consumes one 4OHI3M molecule, the
85 contrast between an active IGMT2 in DE096 and a largely inactive IGMT2 in DE155 alone could not
86 explain the leaf QTL pattern. We therefore inspected the effects of IGMT4 in both NILs more closely,
87 based on the combined analysis of data for *igmt1-3*$_{DE155}$, *igmt1-4*$_{DE155}$, *igmt1/4*$_{DE096}$, and DE096 from four
88 independent experiments. Indeed, this analysis unveiled an additional quantitative difference between
89 the NILs (**Fig. 3**). While the increase in 4OHI3M was consistent in the comparisons from *igmt1-3*$_{DE155}$ to
90 *igmt1-4*$_{DE155}$ and from DE096 to *igmt1/4*$_{DE096}$, the reduction in 4MOI3M was significantly more
91 pronounced when *IGMT4* was knocked out in the DE155 background. This indicated that IGMT4 had a
92 higher level of activity in DE155 compared to DE096, thereby increasing the flux through the pathway.

93 We also conducted tests to ascertain if there were differences in the levels of *CYP81F* transcripts
94 between DE155 and DE096, or if mutations in *IGMT* genes influenced the transcript abundance of
95 wildtype *IGMT*s. However, we observed no substantial alterations (**fig. S3**).

96 Based on these results, we could explain the leaf QTL as the cumulative effects of two distinct
97 quantitative trait genes (QTGs), namely *IGMT2* and *IGMT4*. Genotypic variation in *IGMT4* led to an
98 excess of 4MOI3M in DE155 without altering 4OHI3M levels. This should result in a QTL for 4MOI3M,
99 with DE155 having the high allele for *IGMT4*. However, genotypic variation in *IGMT2* resulted in an
100 elevation of 4MOI3M in DE096, which counterbalanced the excess of 4MOI3M in DE155, and
101 concurrently caused a reduction in 4OHI3M. Consequently, the collective impact of genotypic variation
102 in both QTGs manifested as a disparity in 4OHI3M levels, with no noticeable difference in the levels of
103 4MOI3M.

104 In the roots of DE155, *IGMT1* was the primary contributor to 4MOI3M biosynthesis, as evidenced by the
105 contrast between *igmt3*$_{DE155}$ and *igmt1/3*$_{DE155}$ (**Fig. 2**). *IGMT4* and, notably, *IGMT2* also played roles, but
106 *IGMT3* had no discernible impact. In the roots of DE096, we noted the largest contrast between
107 *igmt1/4*$_{DE096}$ and DE096 wildtype. However, the *igmt1/4*$_{DE096}$ double mutant alone did not allow us to
108 discern whether the observed effects were attributable to *IGMT1*$_{DE096}$, *IGMT4*$_{DE096}$, or their combination.
109 Furthermore, *IGMT2*$_{DE096}$ and *IGMT3*$_{DE096}$ were implicated in the conversion of 4OHI3M to 4MOI3M,
110 primarily observed through the impact of the corresponding mutants on the fraction of 4OHI3M.

111 The influence of *IGMT1*$_{DE155}$ was substantially more pronounced than the combined effect of *IGMT1*$_{DE096}$
112 and *IGMT4*$_{DE096}$. Hence, *IGMT1* was a root QTG. *IGMT3* was a second, but cryptic, root QTG, with DE096
113 possessing the high allele that mitigated the impact of *IGMT1*$_{DE155}$. In contrast to leaves, genotypic
114 variation in *IGMT2* had no noticeable quantitative effect on roots.

115 This root QTL pattern corresponded well with the abundance of *IGMT* transcripts in both NILs (**Fig. 1**),
116 suggesting that quantitative variation in the roots was mainly governed by differences in the expression

117 of individual genes. The transcript level of *IGMT1* was higher and that of *IGMT3* lower in DE155 than in
118 DE096, while the transcript level of *IGMT2* was similar between NILs. Interestingly, $IGMT4_{DE096}$ exhibited
119 a higher transcript abundance than $IGMT4_{DE155}$, indicating that *IGMT4* could be another cryptic root
120 QTG, further counteracting the effect of genotypic variation in *IGMT1*.

121

122 **Ectopic gene conversion in the *IGMT* cluster**

123 In the comprehensive assessment of QTL effects and transcript level variation among NILs, $IGMT2_{DE155}$
124 stood out. The DE155 allele was obviously functional in roots, contributing to 4MOI3M at a level
125 comparable to its DE096 counterpart (**Fig. 2**). However, $IGMT2_{DE155}$ activity in leaves was relatively
126 minor, despite having a transcript level that was similar, if not slightly higher, than that of $IGMT2_{DE096}$
127 (**Fig. 1**).

128 To investigate the cause of this peculiar phenomenon, we built a gene tree with all four genes from both
129 NILs. We expected alleles of the same gene copies to cluster together, which was indeed the case for
130 *IGMT1* and *IGMT4*. Much to our surprise, however, the relationship between *IGMT2* and *IGMT3* genes
131 and alleles remained unresolved (**fig. S4**), suggesting that $IGMT2_{DE155}$ and $IGMT3_{DE155}$ did not evolve
132 independently. Such non-independence could arise from non-homologous recombination between gene
133 copies, generating chimeras from two or more copies. Alternatively, it could stem from the non-
134 reciprocal transfer of genetic material between paralogous loci, known as ectopic gene conversion
135 (EGC).

136 To examine this further, we procured and manually aligned sequences of the *IGMT* region from the 1001
137 Arabidopsis Genomes project (1001genomes.org), concentrating on *de novo*, reference-quality
138 assembles (*12*), and incorporated additional short-read assemblies with near-complete *IGMT* sequences
139 (*13*) (**table S2**). Our alignment contained data from 28 accessions. All assemblies of the *IGMT* region,
140 particularly those from long reads, showed exactly four *IGMT* copies. Because non-homologous
141 recombination usually generates copy number variation, we rule this out for explaining the apparent
142 non-independent evolution of *IGMT* genes.

143 We compared the coding sequences of all four *IGMT* genes and found numerous polymorphisms
144 (**10.6084/m9.figshare.28093955**). To ensure these were not sequencing errors, we amplified and
145 sequenced *IGMT4* coding sequence from 16 Arabidopsis accessions, confirming our findings. Strikingly,
146 many of these polymorphisms were not copy-specific, but shared by two or even more gene copies.

147 We tested whether point mutations could explain the high number of cases where the same
148 polymorphism segregated at the same site in two or more copies, which we refer to as 'shared
149 polymorphisms'. We focused on instances with two independent mutations affecting the same site,
150 either within a single gene copy or across two distinct copies. The relative probabilities of both events
151 depended on the number of gene copies (n), and were n^{-1} for two mutations at the same site in the
152 same gene copy and $1 - n^{-1}$ for two mutations at the same site in two distinct copies.

153 Assuming a large population and no initial polymorphisms (but still allowing fixed differences between
154 gene copies), two outcomes can occur when the same gene copy mutates twice at the same site. Either

155 both mutations lead to the same polymorphism, appearing as a single, copy-specific polymorphism, or
156 they result in three nucleotides segregating at the same site. When mutations hit the same site in two
157 different copies, this can lead to a shared polymorphism or to two copy-specific polymorphisms. The
158 probabilities of these outcomes depend on whether the copies initially had identical or different
159 nucleotides at the side in question, and in the latter case, it also depends on whether the nucleotides
160 differed by a transition or a transversion (see **Supplementary text** for details). From our data, we
161 estimated a transition/transversion ratio of 2.5935, which was in line with other studies (*14,15*).
162 However, we were uncertain about the initial count of identical *versus* different positions in gene copy
163 comparisons. Therefore, we examined the entire frequency spectrum of identical *versus* different
164 positions, ranging from zero to 100%, taking into account their mutual exclusivity.

165 For n = 4 copies and a transition rate of 2.5935, as in the case of the Arabidopsis *IGMT* genes, the
166 proportion of shared among visible types of polymorphisms can never exceed 50% (**fig. S5**). At higher
167 gene copy numbers, this proportion plateaus at 56%. In stark contrast, our entire dataset had only three
168 cases with three different segregating nucleotides in the same gene copy and ten cases with two
169 different segregating polymorphisms across multiple gene copies. However, shared polymorphisms
170 across copies occurred at 61 sites, making up over 80% of the total visible polymorphisms that affected
171 either the same copy twice or two different copies once each.

172 Thus, observed polymorphism patterns revealed far too many shared polymorphisms across gene copies
173 than would be expected from independent point mutations, making EGC the most likely explanation.
174 Indeed, visual inspection identified numerous linked specific variants at polymorphic sites, sometimes
175 spanning hundreds of positions, indicative of gene conversion tracts.

176 Reexamining the 13 cases of three segregating nucleotides in the same copy or with two different
177 segregating polymorphisms across multiple gene copies revealed that most of these events were better
178 explained by EGC than by independent point mutations. Hence, shared polymorphisms resulting from
179 two independent point mutations were even more rare than initially suspected. Thus, shared
180 polymorphisms overall indicate EGC.

181

182 **EGC and subfunctionalization**

183 What are the consequences of EGC for the *IGMT* gene cluster? All four gene products performed the
184 same reaction – converting 4OHI3M to 4MOI3M – but showed organ-specific differences in transcript
185 abundance (**Fig. 1**). CRISPR/Cas9-induced mutations in one gene were not effectively offset by the
186 remaining gene copies, which implied that different *IGMT*s may be expressed in distinct cells. Indeed,
187 *IGMT2* and *IGMT3* are expressed in different root tissues (*16*). Together, these findings suggest
188 subfunctionalization, a process that preserves gene duplicates in a genome by partitioning the functions
189 of the ancestral gene (*17-20*). Subfunctionalization allows different gene copies to express in different
190 contexts and to acquire mutations that optimize their function within these contexts, such as specific
191 cellular environments. However, when EGC transfers these mutations to a gene copy specialized for a
192 different cellular environment, they may impair the function of this copy, as exemplified by $IGMT2_{DE155}$
193 in leaves (**Fig. 2**). Consequently, amino acids that enhance the environment-specific function of gene
194 copies should be subject to purifying selection.

To test this, we focused on fixed derived amino acids identified by comparing *A. thaliana* and *Arabidopsis lyrata* (*21*) IGMT sequences. We divided each *A. thaliana IGMT* coding sequence into intervals delimited by adjacent polymorphisms shared with any of the other three gene copies. We then compared $\log_{10}$-transformed interval lengths with and without codons specifying fixed derived amino acids, using two-sided *t* tests. Intervals with one or more fixed derived amino acids were significantly larger than those without for *IGMT1* ($t_{df=19}$ = -3.12, $p < 0.01$), *IGMT3* ($t_{df=46}$ = -2.77, $p < 0.01$), *IGMT2* ($t_{df=42}$ = -2.88, $p < 0.01$), and *IGMT4* ($t_{df=28}$ = -2.42, $p < 0.05$) (**Fig. 4 and fig. S6**). Across all four genes, results were highly significant ($t_{df=141}$ = -5.72, $p < 0.0001$), suggesting that EGC was indeed selected against around those amino acids.

Discussion

We used the CRISPR/Cas9 system to dissect a near-cryptic QTL for defense metabolites. We obtained a set of different combinations of *IGMT* mutants in three genetic backgrounds but we did not obtain all possible knockout combinations. Consequently, we cannot entirely rule out that some observed effects result from interactions between gene products, such as heterodimers. However, the tissue-specific expression of these genes, coupled with the lack of coordinated expression and compensation for defective genes at the transcript level, argue against this possibility. Our analyses show that *IGM2* comprises a cluster of QTGs, with DE155 and DE096 both having high and low alleles, but at different gene copies. Thus, even small-effect QTL can have a complex genetic architecture.

To demonstrate the existence of EGC among *IGMT* copies, we counted sites with multiple polymorphisms, either within the same gene copy or across different copies. This approach provided a convenient means to assess EGC without requiring prior knowledge of mutation, recombination, or gene conversion rates. Independent point mutations are highly unlikely to generate polymorphisms shared across two or more gene copies, whereas EGC generates such shared polymorphisms, especially visible when conversion tracts encompass long sequence stretches with multiple nucleotide substitutions. Quantitative differences in IGMT performance were partially attributable to EGC, as shown by comparing IGMT2 between two near-isogenic Arabidopsis lines. Both alleles were equally active in roots, but IGMT2 from DE155, which had a clear signature of EGC, performed worse in leaves compared to its DE096 counterpart. Furthermore, EGC events have shaped polymorphism patterns across the entire *IGMT* gene cluster (**fig. S7**), evidenced by sequence comparison of 28 Arabidopsis accessions, suggesting that other *IGMT* gene copies may also have experienced performance alterations due to EGC.

Ectopic gene conversion (EGC) has been documented across a wide range of organisms, from bacteria to humans (*22-30*), indicating its universal influence on gene family evolution. Conversion affects not only coding sequences but also promoter regions, UTRs and introns (**tables S3 and S4**), potentially altering regulatory sequences and expression patterns (28). EGC tends to increase variation among alleles while reducing it between paralogs (*31,32*). EGC can spread advantageous mutations across gene copies (*28,33,34*), but will also place locally adapted variants in a new context where their performance may be impaired.

Gene duplication is essential for optimizing related functions within a complex fitness landscape of similar yet distinct tasks (*17-20*). However, EGC, as an unavoidable consequence of having multiple gene

235 copies, introduces sequence from a gene copy adapted to a specific context into another copy adapted
236 to a different context, as exemplified here. Thereby, EGC generates inferior variants, which, within a
237 population, manifest as QTL segregating fitness variation (**Fig. 5**). Thus, the adaptive process that
238 enables the evolution of specialized gene copies also generates temporary maladaptations until the
239 gene copies have sufficiently diverged to prevent EGC. This can explain why there is more segregating
240 fitness variation within populations than predicted under mutation selection balance (*1*).

241

242 **References and Notes**

243 1. B. Charlesworth, Causes of natural variation in fitness: Evidence from studies of *Drosophila* populations. *Proc.*
244 *Natl. Acad. Sci.* **112**, 1662-1669 (2015).

245 2. D. J. Kliebenstein, J. Kroymann, P. D. Brown, A. Figuth, D. Pedersen, J. Gershenzon, T. Mitchell-Olds, Genetic
246 control of natural variation in *Arabidopsis thaliana* glucosinolate accumulation. *Plant Physiol.* **126**, 811-825 (2001).

247 3. D. J. Kliebenstein, J. Kroymann, T. Mitchell-Olds, The glucosinolate-myrosinase system in an ecological and
248 evolutionary context. *Curr. Opin. Plant Biol.* **8**, 264-271 (2005).

249 4. E. K. F. Chan, H. C. Rowe, J. A. Corwin, B. Joseph, D. J. Kliebenstein, Combining genome-wide association
250 mapping and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*.
251 *PLOS Biol.* **9**, e1001125 (2011).

252 5. M. Pfalz, M. Mukhaimar, F. Perreau, J. Kirk, C. I. C. Hansen, C. E. Olsen, N. Agerbirk, J. Kroymann, Methyl transfer
253 in glucosinolate biosynthesis mediated by indole glucosinolate *O*-methyltransferase 5. *Plant Physiol.* **172**, 2190–
254 2203 (2016).

255 6. P. Bednarek, M. Piślewska-Bednarek, A. Svatoš, B. Schneider, J. Doubskỳ, M. Mansurova, M. Humphry, C.
256 Consonni, R. Panstruga, A. Sanchez-Vallet, A. Molina, P. Schulze-Lefert, A glucosinolate metabolism pathway in
257 living plant cells mediates broad-spectrum antifungal defense. *Science* **323**, 101-106 (2009).

258 7. N. K. Clay, A. M. Adio, C. Denoux, G. Jander, F. M. Ausubel, Glucosinolate metabolites required for an
259 Arabidopsis innate immune response. *Science* **323**, 95-101 (2009).

260 8. J. Kim, G. Jander, *Myzus persicae* (Green Peach Aphid) feeding on Arabidopsis induces the formation of a
261 deterrent indole glucosinolate. *Plant J.* **49**, 1008-1019 (2007).

262 9. M. Pfalz, H. Vogel, J. Kroymann, The gene controlling the *Indole Glucosinolate Modifier 1* QTL alters indole
263 glucosinolate structures and aphid resistance in Arabidopsis. *Plant Cell* **21**, 985-999 (2009).

264 10. M. Pfalz, H. Vogel, T. Mitchell-Olds, J. Kroymann, Mapping of QTL for resistance against the crucifer specialist
265 herbivore *Pieris brassicae* in a new Arabidopsis inbred line population, Da(1)-12 × Ei-2. *PLOS ONE* **6**, e578 (2007).

266 11. M. Pfalz, M. D. Mikkelsen, P. Bednarek, C. E. Olsen, B. A. Halkier, J. Kroymann, Metabolic engineering in
267 *Nicotiana benthamiana* reveals key enzyme functions in Arabidopsis indole glucosinolate modification. *Plant Cell*
268 **23**, 716-729 (2011).

269 12. W. B. Jiao, K. Schneeberger, Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots of
270 rearrangements with altered evolutionary dynamics. *Nat. Commun.* **11**, 989 (2020).

13. X. Gan, O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Rätsch, R. Mott, Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature* **477**, 419-423 (2011).

14. S. Ossowski, K. Schneeberger, J. I. Lucas-Lledó, N. Warthmann, R. M. Clark, R. G. Shaw, D. Weigel, M. Lynch, The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* **327**, 92-94 (2010).

15. M. L. Weng, C. Becker, J. Hildebrandt, M. Neumann, M. T. Rutter, R. G. Shaw, D. Weigel,C. B. Fenster, Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* **211**, 703-714 (2019).

16. H. Cao, X. Zhang, F. Li, Z. Han, X. Guo, Y. Zhang, Y., Glucosinolate *O*-methyltransferase mediated callus formation and affected ROS homeostasis in *Arabidopsis thaliana*. *Physiol. Mol. Biol. Pla.* **30**, 109-111 (2024).

17. A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan,J. Postlethwait, Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).

18. M. Lynch, A. Force, The probability of duplicate gene preservation by subfunctionalization. *Genetics* **154**, 459–473 (2000).

19. M. Lynch, J. S. Conery, The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151-1155 (2000).

20. X. He, J. Zhang, Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* **169**, 1157–1164 (2005).

21. T. T. Hu, P. Pattyn, E. G. Bakker, J. Cao, J. F. Cheng, R. M. Clark, N. Fahlgren, J. A. Fawcett, J. Grimwood, H. Gundlach, G. Haberer, J. D. Hollister,S. Ossowski, R. P. Ottilar, A. A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M. E. Nasrallah, J. Bergelson, J. C. Carrington, B. S. Gaut, J. Schmutz, K. F. X. Mayer, Y. Van de Peer, I. V. Grigoriev, M. Nordborg, D. Weigel, Y. L. Guo, The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.* **43**, 476-481 (2011).

22. H. Innan, A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc. Natl. Acad. Sci. USA* **100**, 8793-8798 (2003).

23. J. Kroymann, S. Donnerhacke, D. Schnabelrauch, T. Mitchell-Olds, Evolutionary dynamics of an Arabidopsis insect resistance quantitative trait locus. *Proc. Natl. Acad. Sci. USA* **100**, 14587-14592 (2003).

24. R. T. Morris, G. Drouin, Ectopic gene conversions in four *Escherichia coli* genomes: increased recombination in pathogenic strains, *J. Mol. Evol.* **58**, 596–605 (2004).

25. N. Osada, H. Innan, Duplication and gene conversion in the Drosophila melanogaster genome. *PLOS Genet.* **4**, e1000305 (2008).

26. D. Benovoy, G. Drouin, Ectopic gene conversions in the human genome. *Genomics* **93**, 27-32 (2009).

27. J. R. Arguello, T. Connallon, Gene duplication and ectopic gene conversion in *Drosophila*. *Genes* **2**, 131-151 (2011).

28. M. Hanikenne, J. Kroymann, A. Trampczynska, M. Bernal, P. Motte, S. Clemens, U. Krämer, Hard selective sweep and ectopic gene conversion in a gene cluster affording environmental adaptation. *PLOS Genetics* **9**, e1003707 (2013).

29. A. Harpak, X. Lan, Z. Gao, J. K. Pritchard, Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc. Natl. Acad. Sci. USA* **114**, 12779-12784 (2017).

30. E. Lamping, J. Y. Zhu, M. Niimi, R. D. Cannon, Role of ectopic gene conversion in the evolution of a *Candida krusei* pleiotropic drug resistance transporter family. *Genetics* **205**, 1619-1639 (2017).

31. H. Innan, A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics* **161**, 865-872 (2002).

32. T. Ohta, Allelic and nonallelic homology of a supergene family. *Proc. Natl. Acad. Sci. USA* **79**, 3251–3254 (1982).

33. J. H. Thomas, Concerted evolution of two novel protein families in *Caenorhabditis* species. *Genetics* **172**, 2269–2281 (2006).

34. S. Mano, H. Innan, The evolutionary rate of duplicated genes under concerted evolution. *Genetics* **180**, 493-505 (2008).

35. R. Buchner, in *Glucosinolates in Rapeseed*, J. P. Wathelet, Ed. (Martinus Nijhoff Publishers, 1987), pp. 50-58.

36. F. Fauser, S. Schiml, H. Puchta, Both CRISPR/Cas-based nucleases and nickases can be used efficiently for genome engineering in *Arabidopsis thaliana*. *Plant J.* **79**, 348-359 (2014).

37. C. Morineau, Y. Bellec, F. Tellier, L. Gissot, Z. Kelemen, F. Nogué, J. D. Faur, Selective gene dosage by CRISPR-Cas9 genome editing in hexaploid *Camelina sativa*. *Plant Biotechnol. J.* **15**, 729-739 (2016).

38. S. J. Clough, A. F. Bent, Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735-743 (1998).

39. Z. Yang, PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).

40. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254-267 (2006).

Acknowledgements

We thank Marine Paupiére for her contribution to QTL mapping.

Funding: Agence Nationale de la Recherche (ANR-10-GENM-005, ANR-20-CE92-0042)

Author contributions: Conceptualization: JK; Methodology: MP, SEN; Investigation: MP, SEN, JK; Funding acquisition: JK; Supervision: JK, MP; Writing: JK, JAS, MP

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or the supplementary materials.

Supplementary Materials

Materials and Methods

Supplementary Text

Figs. S1 to S7

Tables S1 to S4

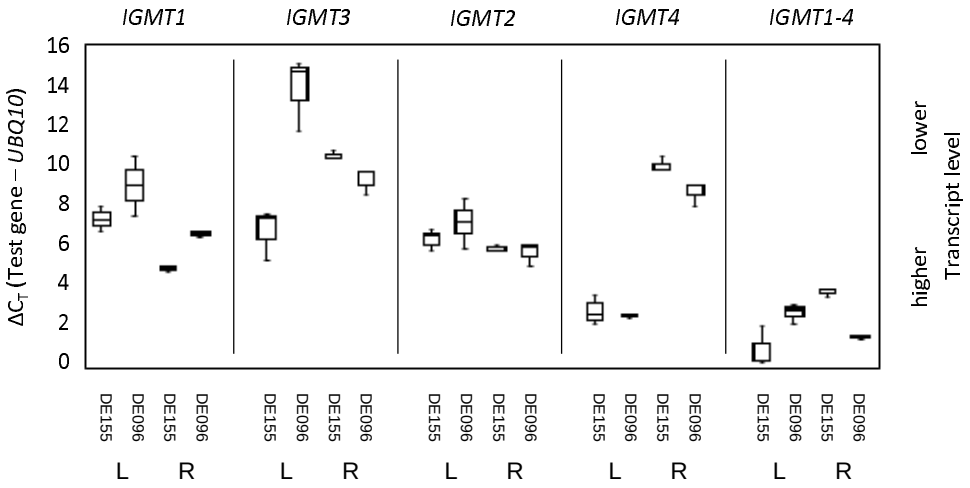
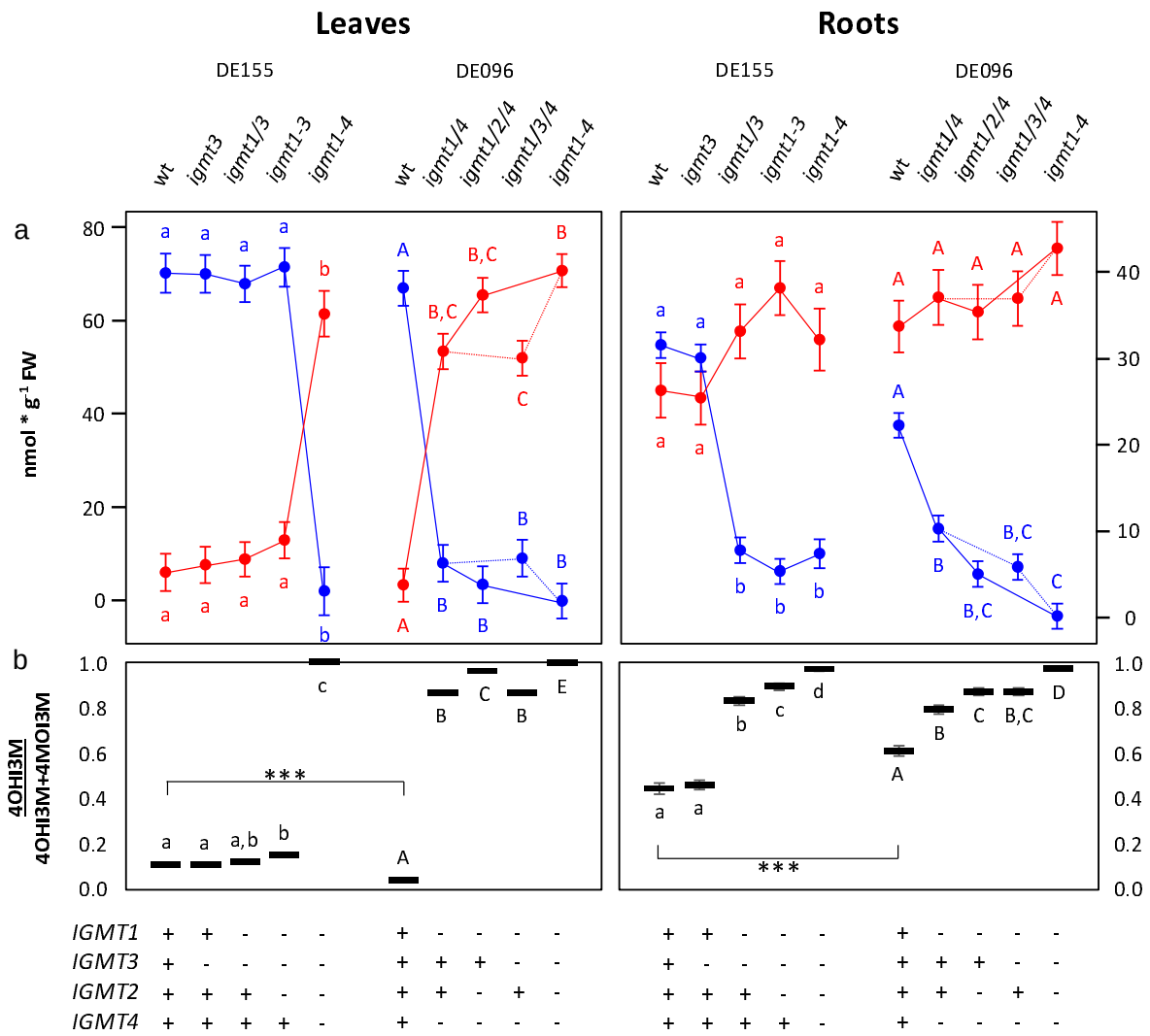


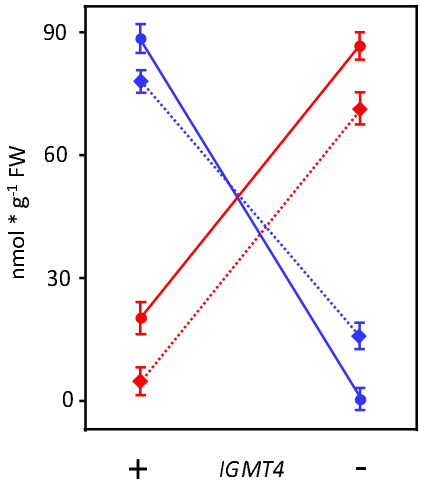
Fig. 1. *IGMT* transcript levels in DE155 and DE096. Shown are boxplots with maximum, median and minimum of ΔC_T values for leaves (L) and roots (R) in comparison to a control gene, *UBQ10*. Each genotype/transcript combination had n = 3 biological samples. *IGMT1* to *IGMT*4 represent individual genes, while *IGMT1-4* indicates total *IGMT* transcript levels. Note that higher ΔC_T values correspond to lower expression levels, and *vice versa*.



354

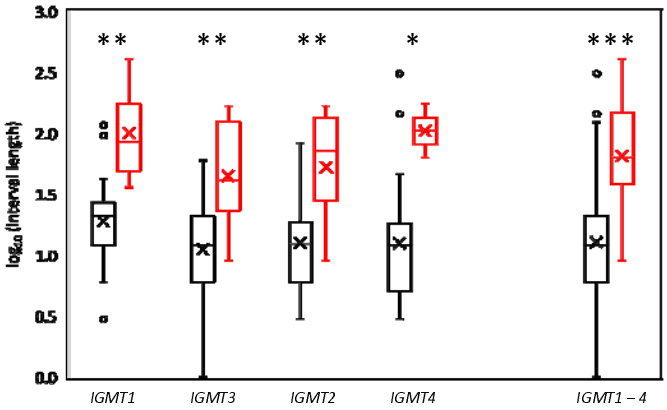
355 **Fig. 2. Quantitative differences in 4MOI3M and 4OHI3M in leaves and roots.** Shown are estimated
356 marginal means of metabolite concentration in nmol per gram fresh weight (± SEM) for Arabidopsis
357 DE155, DE096 and *igmt* mutants. Different upper and lower case letters indicate statistically significant
358 differences. Each genotype had n = 10 or 11 samples. (A) Concentration of 4MOI3M (blue) and 4OHI3M
359 (red). Lines connect meaningful comparisons. (B) Retransformed data for the fraction of 4OHI3M in the
360 combined amount of 4OHI3M and 4MOI3M. Genotypes at *IGMT 1 – 4* are indicated, with "-" for
361 defective and "+" for functional genes. NILs differ significantly for the fraction of 4OHI3M, indicated by
362 asterisks.

363



364 **Fig. 3. Quantitative effects of *IGMT4* on leaf IGs in DE155 and DE096.** Data pooled from four
365 independent experiments show estimated marginal means of metabolite concentration in nmol per
366 gram fresh weight (± SEM) of 4OHI3M (red) and 4MOI3M (blue) with *IGMT4* intact (left) or defective
367 (right) in DE155 (circles and solid lines) and DE096 (diamonds and dashed lines). Each genotype had n =
368 35 – 61 samples. The interaction of *IGMT4* by background is highly significant for 4MOI3M ($F_{1,164}$ = 18.7;
369 p < 0.0001) but not for 4OHI3M ($F_{1,164}$ = 0.0; p = 0.994). There were no significant genotype or genotype-
370 by-background effects for I3M, 1MOI3M, and total IG.

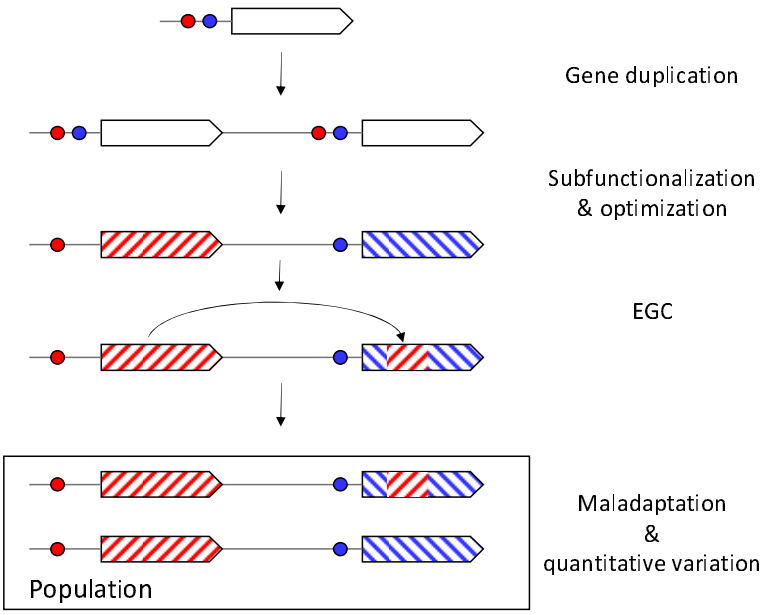
371



372

373 **Fig. 4. Boxplot comparing the log10-transformed interval sizes without (black) and with (red) fixed**
374 **derived amino acids.** These derived amino acids are nonsynonymous substitutions that distinguish
375 Arabidopsis IGMTs from those in *Arabidopsis lyrata*. Each interval is flanked by neighboring shared
376 polymorphisms, defined as the same nucleotides segregating at the same position in two or more gene
377 copies. Shown are mean (crosses), median (black horizontal line), and range of the data including
378 outliers (open circles). Statistical significance is indicated by asterisks (*: $p < 0.05$; **: $p < 0.01$; ***: $p <$
379 0.001).

380



381

382 **Fig. 5. Maladaptation caused by EGC in gene families.** After gene duplication, the paralogs acquire
383 specific expression patterns and optimize accordingly. EGC introduces sequence from one copy
384 optimized for one context into another copy optimized for a different context, leading to maladaptation.
385 This results in quantitative genetic variation in the population.

386

387

Table 1 Estimated Marginal Means - 4MOI3M/(4OHI3M + 4MOI3M) in Roots

			95% Confidence Interval	
Genotype	Mean	SEM	Lower	Upper
DE155	0.533	0.019	0.495	0.570
igmt1-4$_{DE155}$	0.050	0.015	0.020	0.080
DE096	0.361	0.017	0.326	0.396
igmt1-4$_{DE096}$	0.035	0.014	0.006	0.063
Col-0	0.297	0.017	0.264	0.330
igmt5$_{Col-0}$	0.260	0.019	0.223	0.297
igmt1-5$_{Col-0}$	0.000	0.015	-0.030	0.030

388 $8 \leq n \leq 12$

389

390 **Supplementary Materials**

391

392 **Materials and Methods**

393 **Culture conditions**

394 Arabidopsis plants were cultivated in growth chambers with 11.5 hours of light at 22°C and 12.5 hours of
395 darkness at 16°C, maintaining approximately around 70 – 80% relative humidity. Seeds were sown on
396 soil and stratified for three days at 6°C. For glucosinolate analysis, one week after germination, plants
397 were transferred to sand and grown randomized in 96-celled trays, fertilized weekly with Hydrocani C2
398 liquid fertilizer (Hydro Agri), as described by (*5*).

399 **Glucosinolate extraction and analysis**

400 Four weeks after germination, approximately 100 g leaf material and the entire root were harvested,
401 weighed and snap-frozen in liquid nitrogen. Glucosinolate extracts were separated in their desulfo-form
402 on a Vanquish HPLC system (Thermo), utilizing a LiChrospher 100 RP-18e LiChroCART column (250 x 4,
403 100A, 5 µm) as described by (*11*). Glucosinolate identification was based on retention time and UV
404 spectra, and quantification was based on integrated absorption peak area at 229 nm, using sinigrin as an
405 external standard and published response factors (*35*) to correct for different UV absorption capacities
406 of various IGs.

407 **RNA extraction and quantitative real-time quantitative PCR (RT-qPCR)**

408 Frozen leaf or root material was finely ground using a TissueLyser II (Qiagen). Total RNA was extracted
409 with TriSure (Bioline) or the NucleoSpin RNA kit (Macherey & Nagel), treated with Turbo DNase
410 (Ambion), and purified using RNeasy MinElute columns (Qiagen) or RNA Clean & Concentrator 5 (Zymo
411 Research). RNA quality and quantity were assessed by agarose gel electrophoresis and a Nanodrop 2000
412 (Thermo). First strand cDNA synthesis was performed with the Maxima First Strand cDNA Synthesis Kit
413 (Thermo), using 500 ng of total RNA. RT-qPCR was conducted with SYBR Green qPCR Master Mix
414 (Thermo) on a StepOnePlus Real Time PCR System (Applied Biosystems), using Arabidopsis *UBQ10*
415 (At4g05320) as the reference gene. Each genotype had three to six biological replicates. Primer
416 sequences are listed in **table S1**.

417 **Generation of mutant plants**

418 Coding sequences of *IGMT1 – 4* from Arabidopsis accessions Col-0, Da(1)-12 and Ei-2 were used to
419 design three single guide RNA (sgRNA) sequences targeting all four genes simultaneously. The design
420 was performed using the CRISPOR software (crispor.tefor.net), with a preference for sites close to the
421 5'-end of the target genes and low off-target probability. A 1026 bp cassette flanked by attB sites,
422 including all three sgRNAs, each driven by an *At*U6-26 promoter and containing the tracrRNA scaffold as
423 well as the U6 terminator, was synthesized by Genewiz (Leipzig, Germany) and cloned into the
424 pDONR207 vector (Thermo). The final cloning step into pDE-Cas9 DsRed (*36,37*), permitting selection by
425 DsRed marker fluorescence, was achieved using Gateway™ LR recombination. This vector was then used
426 to transform *Agrobacterium tumefaciens* GV3101. The floral dip method (*38*) was employed to generate
427 mutants in Arabidopsis DE096, DE155, Col-0 and *igmt5*. After DNA extraction, mutations were identified
428 by Sanger sequencing, using gene-specific primers (**table S1**) covering the target regions of all three
429 sgRNAs.

430 **Statistical analyses**

431 All measurements of were taken from distinct samples. Fine-mapping of *IGM2* was performed using a
432 general linear model in Systat V9, with markers and growth trays as fixed factors. Glucosinolate
433 quantities were analyzed via ANCOVA in jamovi Version 2.3.28.0 (www.jamovi.org), using the weight of
434 the harvested plant tissue as a covariate and genotype as a fixed factor. When pooling data from
435 multiple experiments, experiment was included as an additional fixed factor, along with an interaction
436 term for genotype × experiment. The leaf effect of *IGMT4* was assessed by ANCOVA, with leaf weight as
437 a covariate and genetic background, genotype at *IGMT4* and experiment as fixed factors. The data for
438 the fraction of 4OHI3M in the combined amount of 4OHI3M and 4MOI3M were arcsine square root-
439 transformed before ANOVA, with genotype as a fixed factor. Statistical differences between genotypes
440 were assessed using post-hoc two-sided *t*-tests. PAML (*39*) was used to estimate the
441 transition/transversion ratio in *IGMT* genes. SplitsTree (*40*) was employed to construct a genealogical
442 network for Arabidopsis *IGMT* sequences.

443

444

Supplementary text

Calculations of probabilities for different types of nucleotide polymorphisms

We consider a large population with n copies of a gene. Initially, there is no polymorphism within copies but fixed differences among copies are allowed. We examine cases where two independent point mutations affect the same site, either twice in the same copy or once in two different copies. The probabilities of these events are $P_{same\ copy} = n^{-1}$ and $P_{different\ copies} = (1 - n^{-1})$, respectively.

(2 copies: AA*, AB, BA, BB*

3 copies: AA*, AB, AC, BA, BB*, BC, CA, CB, CC*

4 copies: AA*, AB, AC, AD, BA, BB*, BC, BD, CA, CB, CC*, CD, DA, DB, DC, DD*)

Each nucleotide can change into any of the three other nucleotides. There is one possibility for a transition, t_i (*e.g.*, A -> G) and two possibilities for a transversion, t_v (*e.g.*, A -> C, A -> T). The probability for a transition is $P_{ti} = x$, the probability for each transversion is ½ P_{tv} = ½ (1 – x). The transition/transversion ratio t_i/t_v can be estimated from the data.

We first determine the probability that the same mutation occurs twice at the same site, either in one copy or in two different copies, with μ_1 being the first and μ_2 being the second mutation:

μ_1	$P(X=\mu_1)$	μ_2	$P(X=\mu_2)$	$P(X=\mu_1)*P(X=\mu_2)$
A->C	½ (1-x)	A->C	½ (1-x)	¼ $(1-x)^2$
A->G	x	A->C	½ (1-x)	½ x*(1-x)
A->T	½ (1-x)	A->C	½ (1-x)	¼ $(1-x)^2$
A->C	½ (1-x)	A->G	x	½ x*(1-x)
A->G	x	A->G	x	x^2
A->T	½ (1-x)	A->G	x	½ x*(1-x)
A->C	½ (1-x)	A->T	½ (1-x)	¼ $(1-x)^2$
A->G	x/(n*l)	A->T	½ (1-x)	½ x*(1-x)
A->T	½ (1-x)	A->T	½ (1-x)	¼ $(1-x)^2$

The pattern shown above for possible outcomes starting from 'A' is also applicable to the three other nucleotides. The probability that the same point mutation occurs twice at the same site (red cases) is:

$x^2 + 2 * ¼ (1 - x)^2 = x^2 + ½ (1 - 2x + x^2) = 0.5 – x + 1.5 x^2$

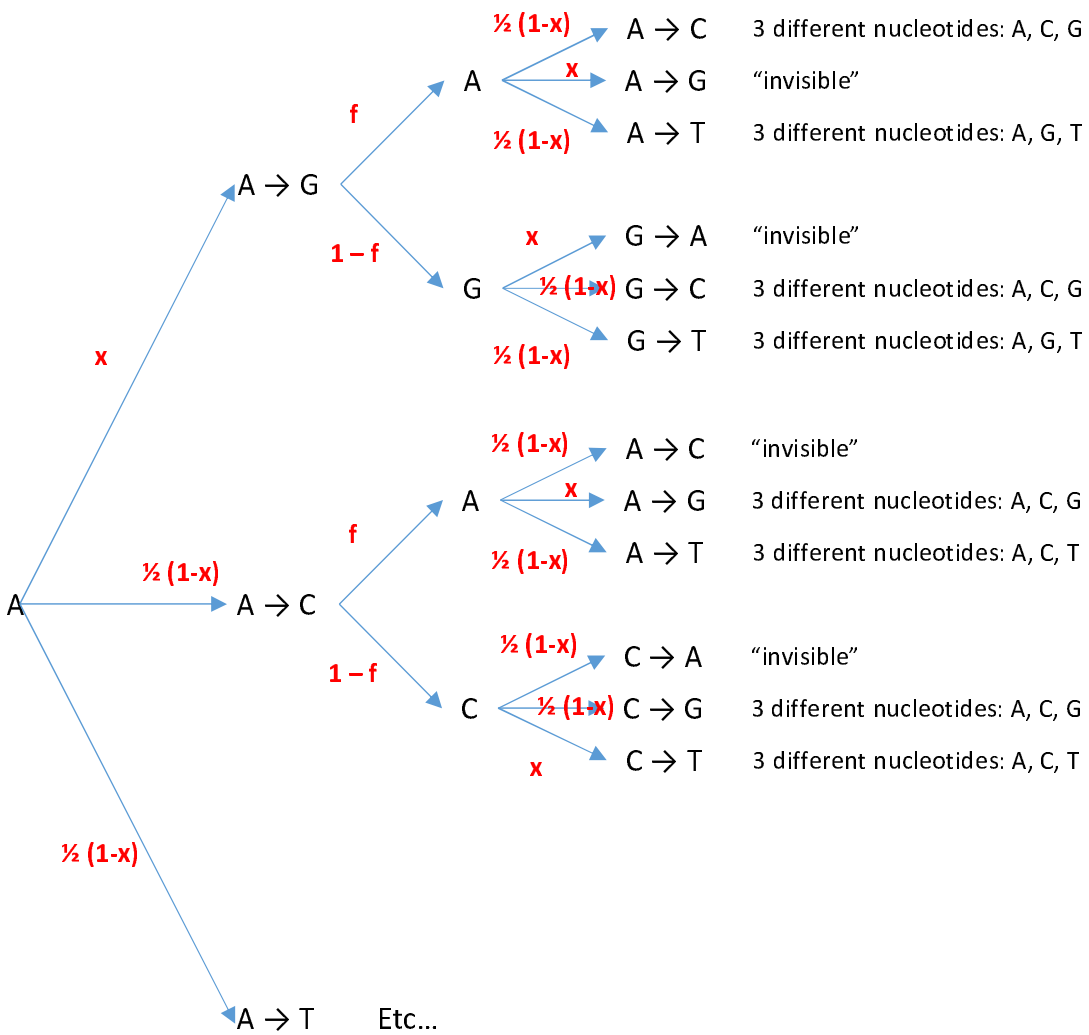
Note that identical mutations occurring independently at the same site in different individuals within a large population go unnoticed.

The probability of different point mutations arising at the same site (blue cases) is:

$1 – (0.5 – x + 1.5 x^2) = 0.5 + x – 1.5 x^2$

If two distinct point mutations occur at the same site in the same copy in different individuals within a large population, three different nucleotides will segregate at that site.

The first mutation creates a polymorphism at the affected site, and the mutant nucleotide may increase in frequency (f) within the population. Does this alter the probabilities of the outcomes following a second mutation at the same site in the same copy?



P (X = 'invisible') = (x * f * x) + (x * (1 – f) * x) + 2 (½ (1 – x) * f * ½ (1 – x)) + 2 (½ (1 – x) * (1 – f) * ½ (1 – x))

$= f * x^2 + (1 - f) * x^2 + 2 (¼ f * (1 - x)^2) + 2 (¼ (1 - f) * (1 - x)^2)$

$= (f + 1 - f) * x^2 + 2 (¼ f * (1 - x)^2 + ¼ (1 - f) * (1 - x)^2) = x^2 + 2 * ¼ (f + 1 - f) * (1 - x)^2$

$= x^2 + ½ (1 - x)^2 = x^2 + ½ (1 - 2x + x^2) = x^2 + 0.5 - x + 0.5 x^2 = 0.5 - x + 1.5 x^2$

Thus, the probabilities for 'invisible' events and those where three different nucleotides segregate at the same site after the second mutation are independent of the mutant allele frequency.

If two mutations hit the same site in the same copy, the probabilities of the outcomes are:

1. **Invisible**: $(0.5 - x + 1.5\ x^2)$

2. **Three nucleotides segregating at the same site** in one copy: $(0.5 + x - 1.5\ x^2)$

If two mutations occur at the same site in different copies, assuming <u>both copies initially have the same nucleotide</u>, two outcomes are possible:

1. **Shared polymorphism**: Identical mutations result in the same nucleotides segregating at the same site in different copies. Probability: $(0.5 - x + 1.5\ x^2)$

2. **Copy-specific polymorphisms**: Different mutations differ result in unique polymorphisms at the same site in each copy. Probability: $(0.5 + x - 1.5\ x^2)$

If the two copies initially had different nucleotides due to an ancient mutation, the outcomes depend on whether the differences are transitions or transversions.

a) The copies differ by a transition: A & G or C & T

μ_1	$P(X=\mu_1)$	μ_2	$P(X=\mu_2)$	$P(X=\mu_1)*P(X=\mu_2)$
A->C	½ (1-x)	G->A	x	½ x*(1-x)
A->G	x	G->A	x	x^2
A->T	½ (1-x)	G->A	x	½ x*(1-x)
A->C	½ (1-x)	G->C	½ (1-x)	¼ $(1-x)^2$
A->G	x	G->C	½ (1-x)	½ x*(1-x)
A->T	½ (1-x)	G->C	½ (1-x)	¼ $(1-x)^2$
A->C	½ (1-x)	G->T	½ (1-x)	¼ $(1-x)^2$
A->G	x	G->T	½ (1-x)	½ x*(1-x)
A->T	½ (1-x)	G->T	½ (1-x)	¼ $(1-x)^2$

A 'shared polymorphism' (red) arises with probability $P = x^2$. With probability $P = (1 - x^2)$, two copy-specific polymorphisms occur at the same site in both copies.

b) The copies differ by a transversion: A & C, A & T, G & C, or G & T

μ_1	$P(X=\mu_1)$	μ_2	$P(X=\mu_2)$	$P(X=\mu_1)*P(X=\mu_2)$
A->C	½ (1-x)	C->A	½ (1-x)	¼ $(1-x)^2$
A->G	x	C->A	½ (1-x)	½ x*(1-x)
A->T	½ (1-x)	C->A	½ (1-x)	¼ $(1-x)^2$

A->C	½ (1-x)	C->G	½ (1-x)	¼ $(1-x)^2$
A->G	x	C->G	½ (1-x)	½ x*(1-x)
A->T	½ (1-x)	C->G	½ (1-x)	¼ $(1-x)^2$
A->C	½ (1-x)	C->T	x	½ x*(1-x)
A->G	x	C->T	x	x^2
A->T	½ (1-x)	C->T	x	½ x*(1-x)

A 'shared polymorphism' (red) occurs with a probability of $P = ¼ (1 – x)^2$. With a probability of $P = 1 – (¼ (1 – x)^2)$, we obtain two copy-specific polymorphisms at the same site in both copies.

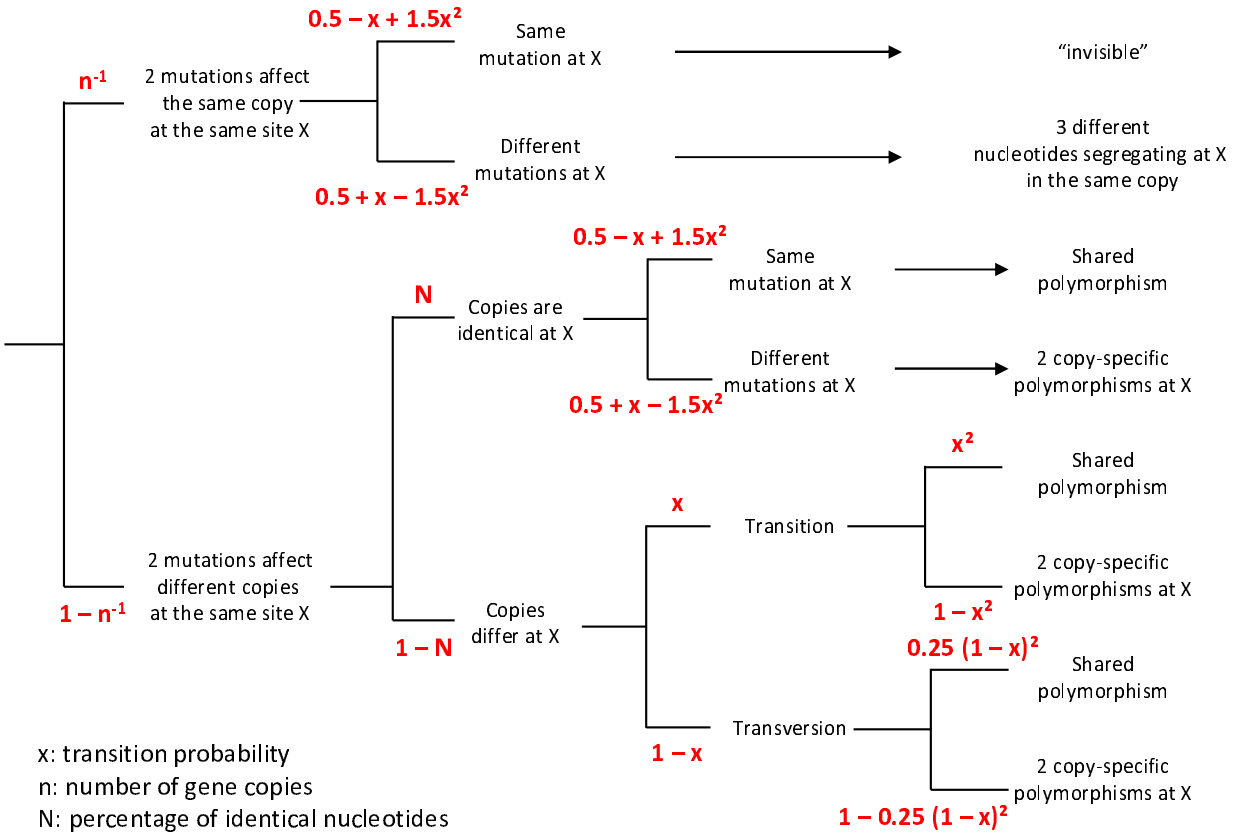
The relative frequencies of scenarios a) and b) are themselves determined by the probabilities of transitions and transversions. Specifically, in x instances, genes with a transition difference serve as the starting point. Conversely, in (1 – x) instances, genes with a transversion difference are the starting point.

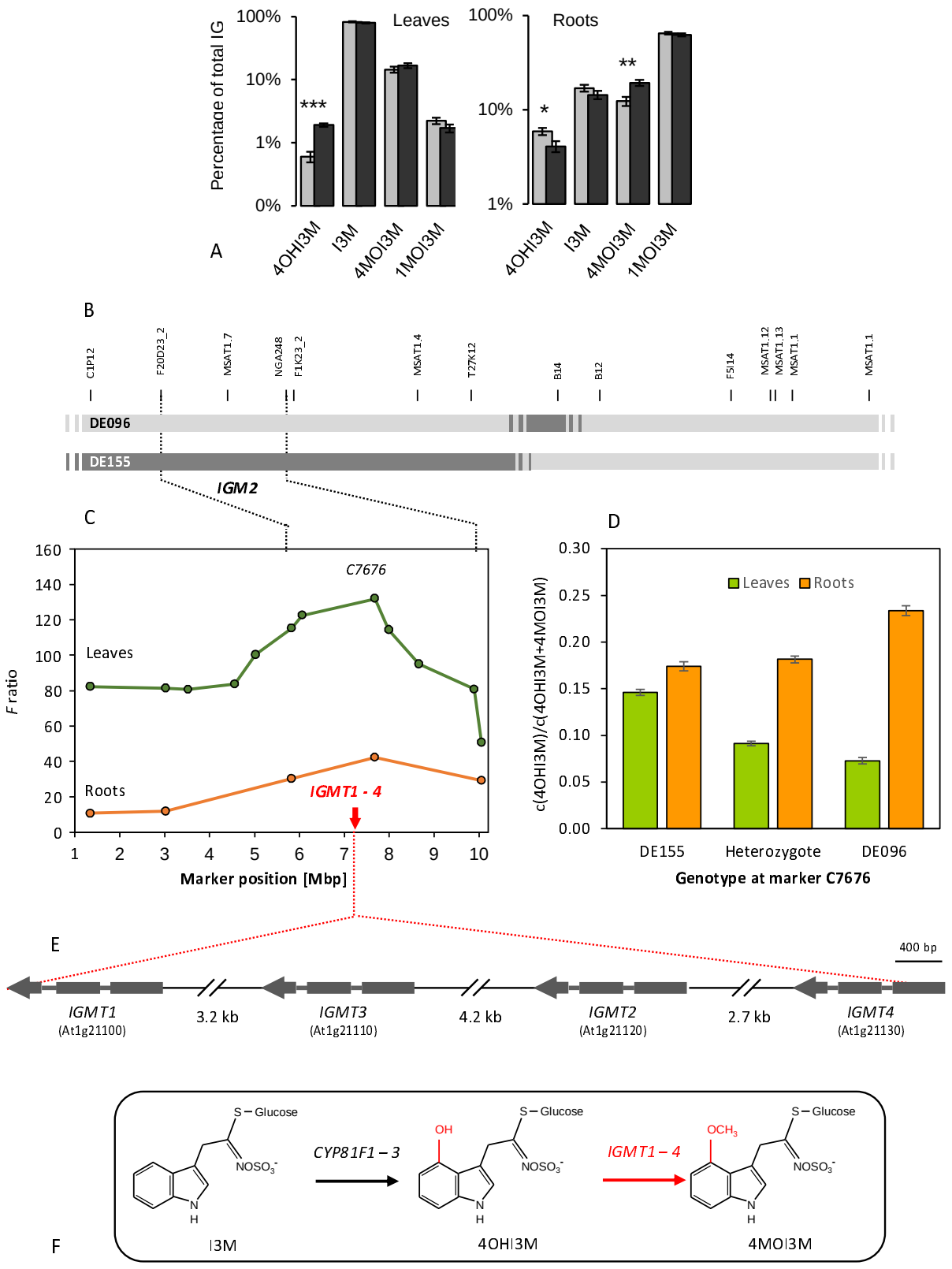
Therefore, the probability of obtaining a shared polymorphism starting with a transition difference is $P = x^3$, and the probability when starting with a transversion difference is $P = ¼ (1-x)^3$. This sums up to a total probability of $P = ¼ (1 – 3x + 3x^2 + 3x^3)$ for copies that initially had different nucleotides at the relevant position. Conversely, the total probability of ending up with two different copy-specific polymorphisms at the same position in both copies is $P = ¾ (1 + x - x^2 - x^3)$.

The exact number of mutable sites in our dataset is uncertain. An initial approximation can be based on the number of synonymous sites, though many synonymous sites in *IGMT*s remain unmutated. Conversely, some non-synonymous sites exhibit polymorphisms or substitutions.

To address this, we have explored scenarios ranging from zero to 100% identical nucleotides, considering that the events of ‘two mutations of identical nucleotides at the same position’ and ‘two mutations of distinct nucleotides at the same position’ are mutually exclusive.

Probabilities of different outcomes for two mutations occurring at the same site X:




A
Leaves
Roots
Percentage of total IG
100%
10%
1%
0%
***
*
**
4OHI3M
I3M
4MOI3M
1MOI3M
B
C1P12
F20D23_2
MSAT1.7
NGA248
F1K23_2
MSAT1.4
T27K12
B14
B12
F5I14
MSAT1.12
MSAT1.13
MSAT1.1
MSAT1.1
DE096
DE155
IGM2
C
C7676
Leaves
Roots
IGMT1 - 4
F ratio
Marker position [Mbp]
D
Leaves
Roots
c(4OHI3M)/c(4OHI3M+4MOI3M)
DE155
Heterozygote
DE096
Genotype at marker C7676
E
400 bp
IGMT1
(At1g21100)
3.2 kb
IGMT3
(At1g21110)
4.2 kb
IGMT2
(At1g21120)
2.7 kb
IGMT4
(At1g21130)
F
S – Glucose
NOSO3-
CYP81F1 – 3
OH
IGMT1 – 4
OCH3
I3M
4OHI3M
4MOI3M

575

576

577 **Fig. S1. Fine mapping of *Indole Glucosinolate Modifier 2***. (A**)** Indole glucosinolate phenotypes of near-
578 isogenic lines DE096 (light gray) and DE155, shown as percentages of total indole glucosinolate. (B)
579 Genotypes of DE096 and DE155 on chromosome 1. DE096 carries parental alleles from the Arabidopsis
580 accession Ei-2 (light gray) in the *IGM2* candidate (dashed lines) region, whereas DE155 carries alleles
581 from Da(1)-12 (dark gray). DE096 and DE155 marker genotypes do not differ on chromosomes 2 – 5. (C)
582 Fine mapping. We fine-mapped the leaf QTL as the arcsine square root-transformed proportion of
583 4OHI3M among modified indole glucosinolates [c4OHI3M x $c(4OHI3M + 4MOI3M)^{-1}$] in 184 F_2 progeny of
584 a cross between near isogenic lines DE096 and DE155 with twelve markers. Marker positions refer to
585 their physical position in the Col-0 reference genome. Statistical support (green line) for the presence of
586 a QTL peaked close to *IGMT1-4*. Subsequently, we confirmed the presence of a similar QTL in roots
587 (orange line) with a subset of markers in an independent experiment with 184 F_2 progeny. (D) Estimated
588 proportion (± SEM) of 4OHI3M by genotype at marker C7676 in leaves (green) and roots (orange). (E)
589 Organization of IGMT1 – 4. Exon-intron structures are based on the Col-0 accession and drawn to scale,
590 with the intervening sequence shortened. The gene order is the same in all investigated accessions. (F)
591 Chemical structures of indole glucosinolates and reactions catalyzed by CYP81F1 – 3 and IGMT 1 – 4.
592 *CYP81F1 – 3* gene products hydroxylate indol-3ylmethyl glucosinolate (I3M) at the 4-position of the
593 indole ring, producing 4-hydroxy-I3M (4OHI3M). This compound then serves as the substrate for *IGMT1*
594 *– 4* gene products, which generate 4-methoxy-I3M (4MOI3M).

595

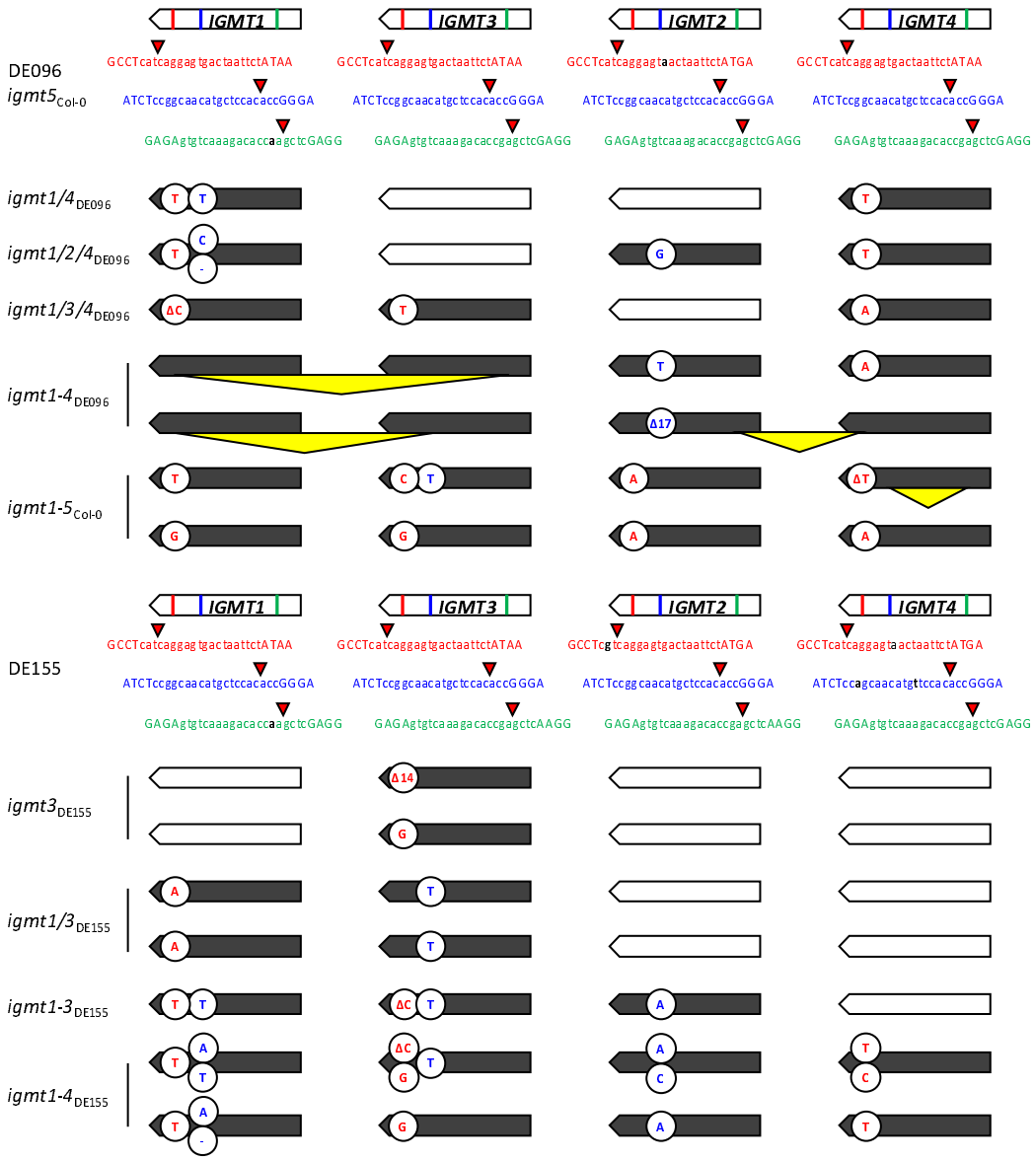
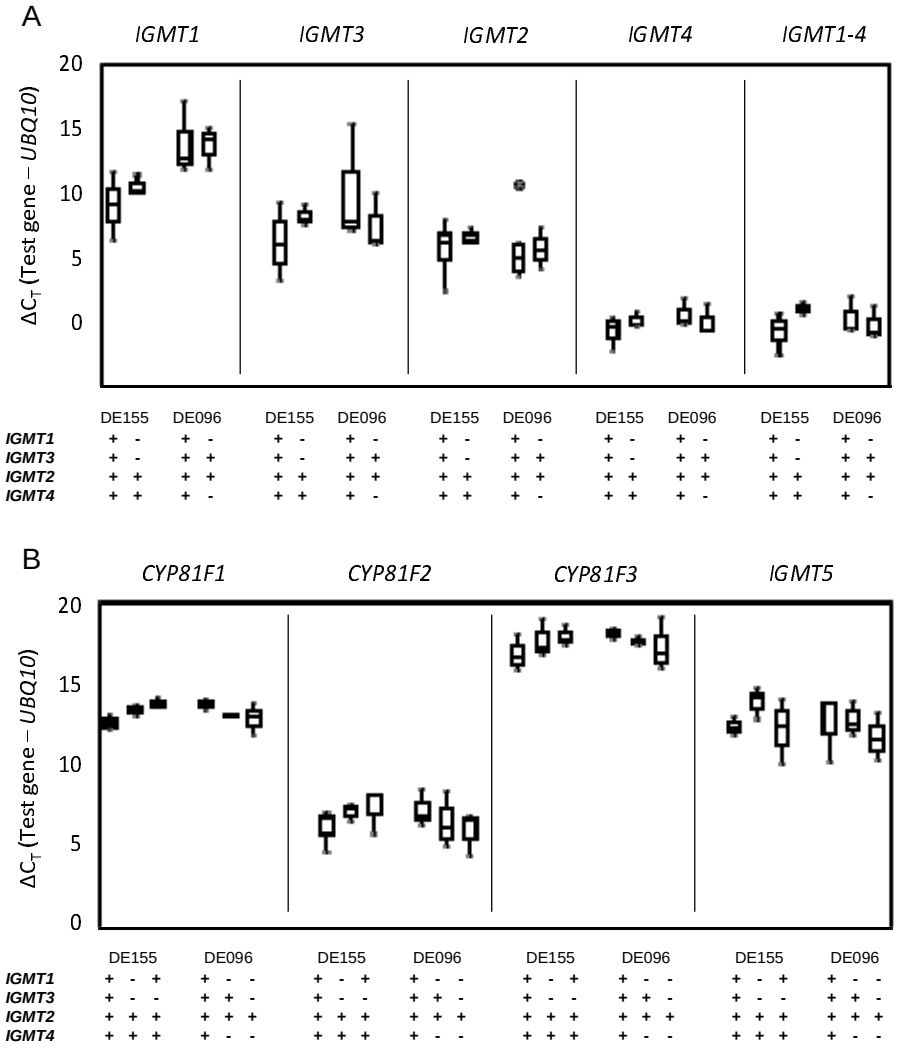


Fig. S2. Schematic representation of *IGMT* wild types and mutants. Locations of sgRNAs (red, blue and green colors) in *IGMT1 – 4*, corresponding sequences and flanking nucleotides (capital letters) in DE096, *igmt5*$_{\text{Col-0}}$, and DE155. Sequence deviations are shown in black. Mutations rendering genes non-functional are shown in colors corresponding to the different target sites. Insertions are denoted by letters, deletions by a Δ proceeding letters (in case of single-nucleotide deletions) or numbers (in case of multiple nucleotide deletions). Larger deletions, some resulting in fusions between parts of adjacent gene copies, are marked by yellow triangles. Note that some of the target sites were still segregating for polymorphisms when analyzed, leading to heterozygous configurations; however, additional mutations at other target sites ensured non-functionality of the gene products.

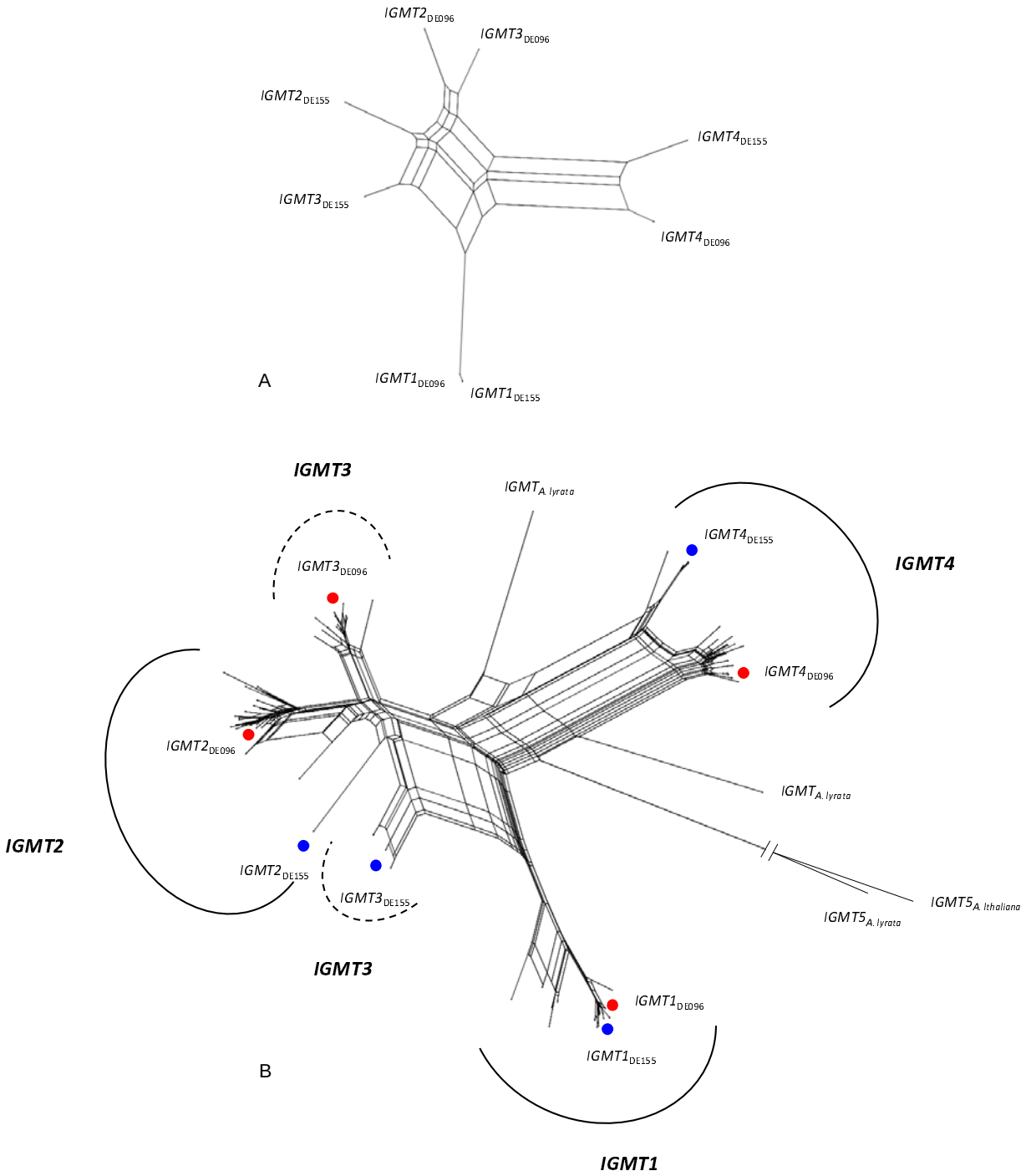
607



608

609 **Fig. S3. *IGMT* and *CYP81F* transcript levels in wild type and mutant leaves.** (A) Boxplots showing *IGMT1*
610 *– 4* transcript levels in NILs and selected mutant leaves. (B) *CYP81F1 – F3* and *IGMT5* transcript levels.
611 Data are presented as median and range of ΔC_T values in comparison to the control gene *UBQ10*.
612 Genotypes at *IGMT1 – 4* are indicated, with "-" for defective and "+" for functional genes.

613

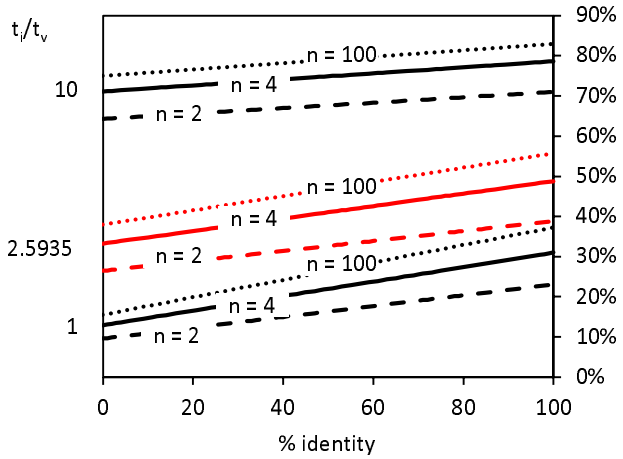


614

615

616 **Fig. S4. Genealogical networks of *IGMT* genes. (**A) Network based on DE155 and DE096 *IGMT* genes. **(**B)
617 Network based on *IGMT1 – 4* genes from 28 Arabidopsis accessions, IGMT5, and corresponding
618 sequences from *A. lyrata*. Colored dots indicate the positions of DE155 (blue) and DE096 (red). The
619 *IGMT5* branch was shortened to 1/5 of its original length. Note that there are two clusters with *IGMT3*
620 genes.

621

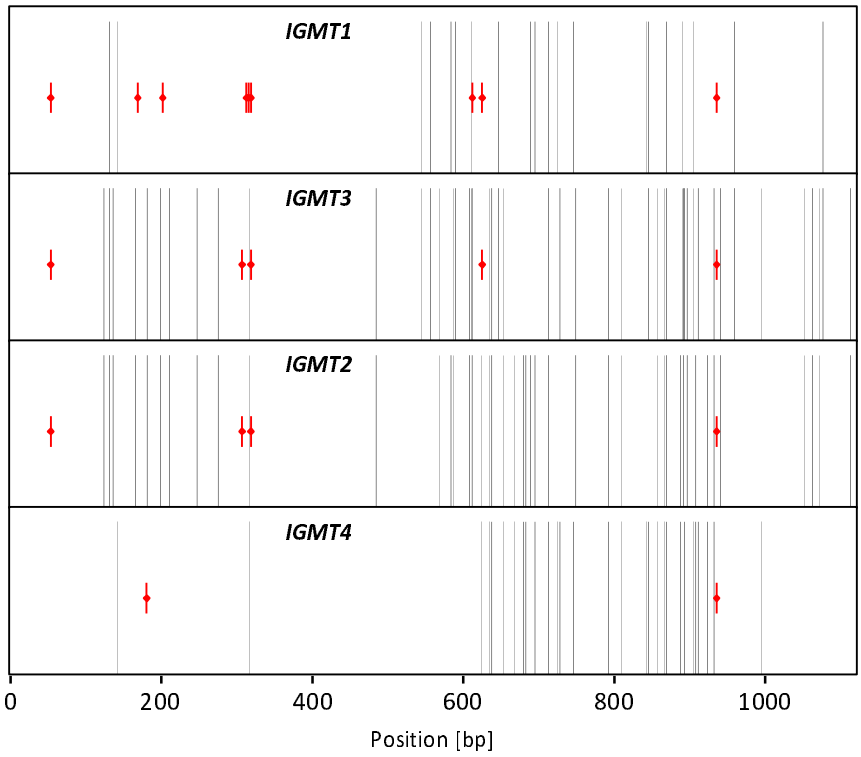


622

623 **Fig. S5. Expected proportion of shared among visible types of polymorphisms for different**
624 **transition/transversion (t_i/t_v) ratios and copy numbers (n).** % identity refers to cases where two
625 mutations hit different copies at the same site, which may carry either identical or different nucleotides.
626 Data for Arabidopsis *IGMT1 – 4* are indicated by a solid red line.

627

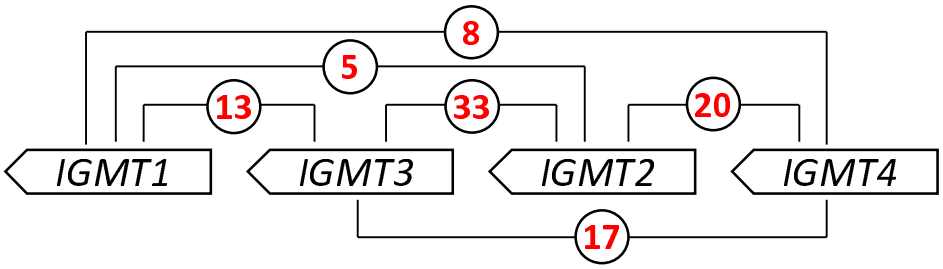
628



629

630 **Fig. S6. Position of derived amino acids and shared polymorphisms in *IGMT1 – 4* coding sequences.** A
631 polymorphism is considered 'shared' when the same nucleotides segregate at the same site in two or
632 more gene copies. Intervals flanked by two adjacent shared polymorphisms (black) are significantly
633 larger when they contain derived amino acids (red). Positions are shown in base pairs.

634



635 **Fig. S7. Number of shared polymorphisms in pairwise comparisons of *IGMT* genes.** Note that shared
636 polymorphisms are more frequent when genes are in close proximity. Also, the two central genes
637 exhibit the highest number of shared polymorphisms.

638

639 **Table S1. Primer sequences used in this study.**

Fine mapping of *IGM2*			
Marker	**Sequence (5' → 3')**	**Sequence (5' → 3')**	**Physical position [Mbp] on Chr 1 in Col-0**
C1P42	GACATAGTTGGCGATGATTG	CAAACAATGGGAGAAGTTGGTC	1.34
C1P43	CTTTTGGAGACACTCAAAGTTG	TAGAGTAAGGTACAAGACCTC	3.02
c3519	CAGTTTGGTAGTTGGATCTGG	GTATAACACTAATCCGAGAACC	3.52
c4556	CGCATGATTCGTCTTGTAAC	CTTTAGAGAGAGGGATCAATAC	4.56
c5020	AGTTTGCTGGTGGATTCGTG	CATTGCATGGCCAAATTCCTC	5.02
F20D23_2	CCGTCACACCATTCACAATC	CCAACCCCTTATATATCGTTC	5.82
c6061	ACTAAAGGAAGAGATCAATCAC	GCTGATGCTTAGTCCTTTC	6.06
c7676	GTGATTTAGTTAGGCATTAAGTTTC	CTCTATATTCTCAGCTTTCTTTC	7.68
M3	GCTTTTATCAGCTCAAACAT	ACTCTTACGTTTGGAGTTCA	7.99
c8654	CACCTGATCTAACGAGTGTTTC	GTTAGCATTTTTATGCCACTATTAG	8.65
M5	TCTGTATCTCGGTGAATTCTCC	TACCGAACCAAAACACAAAG	9.89
F1K23	GAACCAATAAGGAGGCTCAAC	CCATACGGAGAAACCTTCTTC	10.00

Amplification and sequencing of *IGMT1-4* promoters & genes		
Target	**Primer name**	**Sequence (5' → 3')**
IGMT1 (At1g21100)	PromMT 100_1F	CATTTTCCTTTGTTTTTGTTTC
	PromMT 100_F	TTTCCTTTGTTTTTGTTTCTTTCTTC
	PromMT 100_R	CTATAACAATTGAGATGGGAAAC
IGMT3 (At1g21110)	PromMT 110_F	TTCTCTTTCTTTTTTGTTTCTTTCTTC
	PromMT 110_1F	CATTTCTCTTTCTTTTTTGTTTC
	PromMT 110_R	CAACTGTTTCTAAATGGCTATC
	PromMT 110_2R	GTAGCGGTGTAATCATTACTTC
	PromMT 110_3R	GGTATAAGAGATATTGCGTGTG
IGMT2 (At1g21120)	PromMT 120_F	TTCTCTTTCTTTTTTGTTACTTTCTTC
	PromMT 120_1F	CATTTCTCTTTCTTTTTTGTTAC
	PromMT 120_R	GAATCAAACCGACGATGAAC
	PromMT 120_2R	GACTAACGGCCGAAAGGGAC
IGMT4 (At1g21130)	PromMT 130_1F	CATTTTTATTTCTTTTTGTTTCTTT
	PromMT 130_R	GAAACTTAACTGGGCCAAGC
	PromMT 130_2R	GCATAACGTAACAGACCATG
	RIK-3F	GGGATACCTTCTAGAAGAAAC
	RIK-3R	CAGAACTCAATAATCCAGCAG

Detection of changes in sgRNA sequences (PCR & Sanger sequencing)		
Target	**Primer name**	**Sequence (5' → 3')**
IGMT1 (At1g21100)	gRNA21100F	GAAAGTAGCCGACCGTGGAG
	gRNA21100R	GCACAAGAGGCTCTTACTTTTC
IGMT2 (At1g21120)	gRNA21120F	CAGAACTCAATAATCCAGCAGT
	gRNA21120R	CTCTTACCTTTCTAAACAGAGAC
IGMT3 (At1g21110)	gRNA21110F	CAAATGAAAGCAGCTGGTCTTG
	gRNA21110R	CTACACAGAGACTACTAACTCAG
IGMT4 (At1g21130)	gRNA21130F	CAGAACTCAATAATCCAGCAGT
	gRNA21130R	CTACGGTACATTACATAGAGAC

Real Time quantitative PCR		
Target	**Primer name**	**Sequence (5' → 3')**
IGMT1 (At1g21100)	RT-MT100for	CTCTCACCCTATGAAATAGCAA
	RT-MT100rev	GCTCTGTAGACCCTCTCGC
IGMT2 (At1g21120)	120new_f	CATGGTCAAGTGTGAGAAGGTC
	120new_r	GAGGCAAGGGATCCAATATCC
IGMT3 (At1g21110)	110new_f	CATGGTCAAGTGTGAGAAG
	110new_r	GTTTGCGTTGATGTCACC
IGMT4 (At1g21130)	130new_f	GCGAGAGGCTTCAGCATAGAG
	130new_r	GGGATACCTTCTAGAAGAAACC
IGMT5 (At5g76790)	RT-76790F	CAGTCCAGTCATGAAGTATAC
	RT-76790R	GTGCTTTAACACAAGCTCCTTC
CYP81F1 (At4g37430)	RT-CYP430F	GAACCGCTCTTATCTGATTG
	RT-CYP430R	GGAAATAGCCGGAATGTTTC
CYP81F2 (At5g57220)	Cyt2F	GTCACAGGGAGACGCTACTAC
	Cyt2R	CACCACTGTTGTCATTGATGTC
CYP81F3 (At4g37400)	RT-CYP400F	GCCTAATGCTGGGCATGATG
	RT-CYP400R	CTTTAATATCCTCTGTTGGTG
CYP81F4 (At4g37410)	37410seq2	CCCTGAGTACTACACTGAC
	RT-410-(09)R	CAAAACTTTTGGGTGGTTCAAC
IGMT1-4	Universal (IGMT)-for	ACGTATACTTCATGATTGGAC
	Universal (IGMT)-rev	CAATAATCCAGCAGTGATAAG
UBQ10 (At4g05320)	UBQ10f	AAAGAGATAACAGGAACGGAAACATAGT
	UBQ10r	GGCCTTGTATAATCCCTGATGAATAAG

640

641 **Table S2. Arabidopsis accessions used in this study.**

642

Accession	Country	Note
Col-0	Poland	reference genome
An-1	Belgium	
C24		
Cvi-0	Cape Verdi Islands	
Eri-1	Sweden	Jiao & Schneeberger, 2020
Kyo	Japan	
Ler	Poland	
Sha	Tadjikistan	
Bur-0	Ireland	
Can-0	Canary Isles	
Ct-1	Italy	
Edi-0	Scotland	
Hi-0	Netherlands	
Kn-0	Lithuania	
Ler-0	Poland	
Mt-0	Libya	
No-0	Germany	Gan et al., 2011
Oy-0	Norway	
Po-0	Germany	
Rsch-4	Russia	
Sf-2	Spain	
Tsu-0	Japan	
Wil-2	Russia	
Ws-0	Russia	
Wu-0	Germany	
Zu-0	Germany	
Da(1)-12	Czech Republic	
Ei-2	Germany	*IGMT1* , *2* , *3* , *4* ; this study
No-0	Germany	
Aa-0	Germany	
Ag-0	France	
Bl-1	Italy	
Cal-0	United Kingdom	
Can-0	Canary Islands	
Cnt-1	Canterbury	
Cvi-0	Cape Verdi Islands	
Ema-1	United Kingdom	near-complete *IGMT4* ; this study
Hodja	Tadjikistan	
Kas-0	India	
Ler-0	Germany	
Lip-0	Poland	
Ma		
Mr-0	Italy	
Oy-0	Norway	
Tsu-0	Japan	

643

Table S3. Shared polymorphisms in the 5'UTR of *IGMT1* and *2*.

| | | |

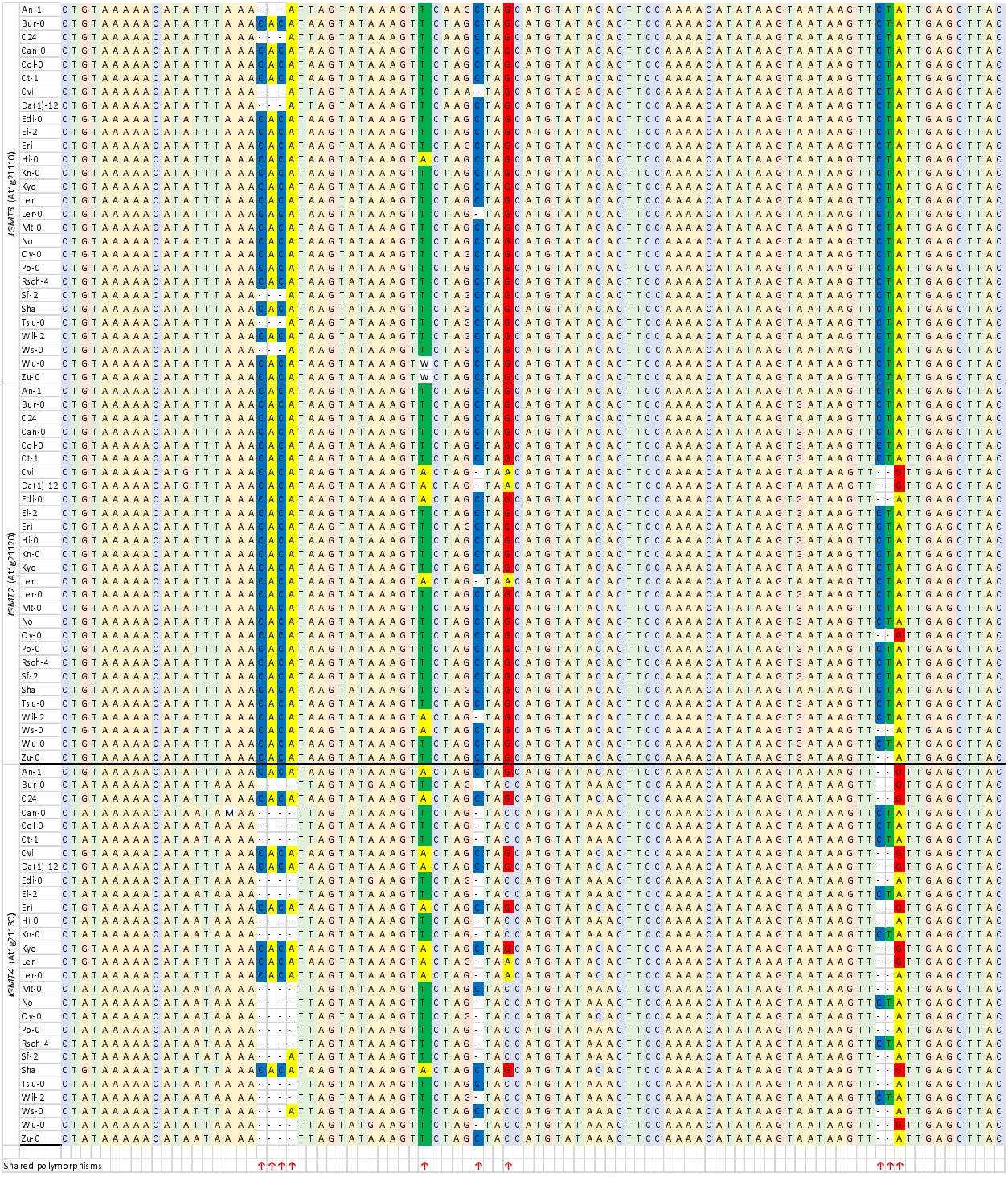



Table S4. Shared polymorphisms in an *IGMT* intron.