Ectopic gene conversion causing quantitative trait variation

Authors: Marina Pfalz[1], Seïf-Eddine Naadja[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: Why is there so much non-neutral genetic variation segregating in natural populations? We dissect function and evolution of a near-cryptic quantitative trait locus (QTL) for defense metabolites in Arabidopsis using the CRISPR/Cas9 system and nucleotide polymorphism patterns. The QTL is explained by genetic variation in a family of four tightly linked indole-glucosinolate *O*-methyltransferase genes. Some of this variation appears to be maintained by balancing selection, some appears to be generated by non-reciprocal transfer of sequence, also known as ectopic gene conversion (EGC), between functionally diverged gene copies. Here we elucidate how EGC, as an inevitable consequence of gene duplication, could be a general mechanism for generating genetic variation for fitness traits.

Introduction

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, natural populations harbor too much non-neutral genetic variation for mutation-selection balance alone to explain (Charlesworth 2015). Indeed, additional processes like migration and drift introduce or maintain variation in natural populations, and some variation is maintained by balancing selection (Charlesworth 2015; Kroymann and Mitchell-Olds 2005). Here we propose an additional, general mechanism for generating non-neutral variation. We elucidate how ectopic gene conversion (EGC) generates fitness variation in populations, illustrated by dissection of a quantitative trait locus for glucosinolates in Arabidopsis.

Gene duplication generates evolutionary novelty via specialization of different copies (Ohno 1970; Ohta 2000), optimizing related functions within a complex fitness landscape of similar yet distinct tasks (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000; He and Zhang 2005). However, because these copies retain sequence similarity despite specialization, they are prone to unequal crossover and ectopic gene conversion (EGC). Unequal crossover will lead to variation in copy number and EGC, the non-reciprocal transfer of sequence between paralogous genes in the context of DNA repair, will homogenize variation among copies. Initially, in young gene copies, EGC is likely but will be largely neutral. As gene copies diverge in sequence and function, EGC will have a greater impact on phenotype but becomes less likely. Thus, it is easy to conclude that EGC can play only a very minor role in the evolution of gene families (Nei and Rooney 2006). Nonetheless, EGC has been documented across a wide range of organisms, from bacteria to humans (Ohta 1982; Innan 2003; Kroymann et al. 2003; Morris and Drouin 2004; Osada and Innan 2008; Benovoy and Drouin 2009; Arguello and Connallon 2011; Hanikenne et al. 2013; Harpak et al. 2017; Lamping et al. 2017) and is often cited as a mechanism for concerted evolution (Zimmer et al. 1980; Ohta 1991; Mano and Innan 2008).

We contend that EGC can also generate fitness variation that will be targeted by selection. When EGC places a specialized allele adapted to a particular cellular role into another, inappropriate context, this

should produce maladapted variants. Here we describe this process, detected while dissecting a QTL for glucosinolate metabolism in Arabidopsis. Sequence alignments revealed clear traces of EGC among four tandemly arranged members of a gene family for defense metabolites. Using CRISPR technology to knock out different copies of the gene family we showed which allele in which context was responsible for phenotypic variation, confirming that EGC was the mechanism that generated functional variation.

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 (Kliebenstein et al. 2001; Kliebenstein et al. 2005; Chan et al. 2011). Indole glucosinolates (IGs) are derived from tryptophan, are inducible by microbial pathogens or phytophagous insects, and play an important role in plant defense (Kim and Jander 2007; Bednarek et al. 2009; Clay et al. 2009; Pfalz et al. 2009; Pfalz et al. 2016). We had previously mapped leaf QTL for these IGs in Arabidopsis and cloned the *Indole Glucosinolate Modifier 1* (*IGM1*) QTL on chromosome 5 (Pfalz et al. 2007; Pfalz et al. 2009), 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, was thought to control variation for 4OHI3M but not 4MOI3M (Pfalz et al. 2007).

A tightly linked gene family of indole glucosinolate *O*-methyltransferase genes (*IGMT*s) underlies *IGM2*, via a complex determination of metabolite levels. The gene family presents multiple cases of shared polymorphisms that are unlikely to have arisen by independent point mutations. Instead, we contend that these occurred through EGC, and that the counterplay between selection and EGC shaped the QTL.

Results

Mapping and dissection of *IGM2*

We fine-mapped *IGM2* using F_2 progeny from a cross between two near-isogenic lines (NILs) of the Arabidopsis Da(1)-12 × Ei-2 mapping population, DE096 and DE155 (Pfalz et al. 2007). In leaves, DE096 had more 4OHI3M than DE155, but the NILs did not differ in their 4MOI3M content. In roots, DE096 had more 4OHI3M and less 4MOI3M than DE155 (**Fig. S1a**). DE096 and DE155 differed only in a ~ 20 Mbp region containing *IGM2*, while their marker genotypes were identical for the rest of the genome (**Fig. S1b**). We phenotyped for the fraction of 4OHI3M in the combined amount of 4OHI3M and 4MOI3M. *IGM2* mapped to a ca. 1.6 Mbp interval, with no evidence of additional QTL nearby (**Figs. S1c and S1d**). Statistical support for a QTL centered near a family of four tandemly arranged methyltransferase genes. These genes were previously identified as coding for indole glucosinolate *O*-methyltransferases (IGMTs) that convert 4OHI3M to 4MOI3M (Pfalz et al. 2011), making them plausible candidates for the QTL (**Figs. S1e and S1f**).

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

86 lower. Thus, genotypic differences in *IGMT* transcript abundance alone could not explain the observed
87 quantitative metabolite pattern.

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

95 We obtained quadruple knockouts, *igmt1-4*$_{DE155}$ and *igmt1-4*$_{DE096}$, for both NILs. While the leaves of
96 these mutants lacked 4MOI3M entirely, the roots still exhibited a background level of 4MOI3M,
97 suggesting the presence of another enzyme capable of converting 4OHI3M to 4MOI3M in roots. Our
98 prime candidate was IGMT5, encoded on chromosome 5, which normally converts 1-hydroxy-indol-
99 3ylmethyl to 1-methoxy-indole-3ylmethyl glucosinolate and has high root but low leaf activity (Pfalz et
100 al. 2016). We therefore generated an *igmt1-4* quadruple knockout in the *igmt5* mutant background.
101 As expected, these plants showed no detectable 4MOI3M in either leaves or roots (**Table 1**).

102 In addition to the quadruple mutant, referred to as *igmt1-4*$_{DE155}$, we obtained the genotypes *igmt3*$_{DE155}$,
103 *igmt1/3*$_{DE155}$, and *igmt1-3*$_{DE155}$ for DE155, and *igmt1/4*$_{DE096}$, *igmt1/3/4*$_{DE096}$, and *igmt1/2/4*$_{DE096}$ for
104 DE096 **(Fig. S2)**.

105 In leaves, we observed the most substantial differences in the levels of 4OHI3M, 4MOI3M and/or the
106 fraction of 4OHI3M when comparing *igmt1-3*$_{DE155}$ with *igmt1-4*$_{DE155}$, and DE096 wildtype with
107 *igmt1/4*$_{DE096}$ (**Fig. 2**). Thus, *IGMT4* had the largest impact on 4MOI3M generation in the leaves of both
108 NILs. Additionally, we found noticeable differences between *igmt1/2/4*$_{DE096}$ and *igmt1/4*$_{DE096}$, and
109 between *igmt1/3/4*$_{DE096}$ and *igmt1-4*$_{DE096}$ (**Fig. 2**), which highlighted the role of *IGMT2*$_{DE096}$. In stark
110 contrast, we only found minor differences for *IGMT2*$_{DE155}$ when comparing *igmt1/3*$_{DE155}$ and *igmt1-*
111 *3*$_{DE155}$, indicating that genotypic variation in *IGMT2* contributed to the leaf QTL.

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

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

125 Based on these results, we could explain the leaf QTL as the cumulative effects of two distinct
126 quantitative trait genes (QTGs), namely *IGMT2* and *IGMT4*. Genotypic variation in *IGMT4* led to an
127 excess of 4MOI3M in DE155 without altering 4OHI3M levels. This should result in a QTL for 4MOI3M,
128 with DE155 having the high allele for *IGMT4*. However, genotypic variation in *IGMT2* resulted in an

129 elevation of 4MOI3M in DE096, which counterbalanced the excess of 4MOI3M in DE155, and
130 concurrently caused a reduction in 4OHI3M. Consequently, the collective impact of genotypic variation
131 in both QTGs manifested as a disparity in 4OHI3M levels, with no noticeable difference in the levels of
132 4MOI3M.

133 In the roots of DE155, *IGMT1* was the primary contributor to 4MOI3M biosynthesis, as evidenced by
134 the contrast between *igmt3*$_{DE155}$ and *igmt1/3*$_{DE155}$ (**Fig. 2**). *IGMT4* and, notably, *IGMT2* also played roles,
135 but *IGMT3* had no discernible impact. In the roots of DE096, we noted the largest contrast between
136 *igmt1/4*$_{DE096}$ and DE096 wildtype. However, the *igmt1/4*$_{DE096}$ double mutant alone did not allow us to
137 discern whether the observed effects were attributable to *IGMT1*$_{DE096}$, *IGMT4*$_{DE096}$, or their
138 combination. Furthermore, *IGMT2*$_{DE096}$ and *IGMT3*$_{DE096}$ were implicated in the conversion of 4OHI3M
139 to 4MOI3M, primarily observed through the impact of the corresponding mutants on the fraction of
140 4OHI3M.

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

145 This root QTL pattern corresponded well with the abundance of *IGMT* transcripts in both NILs (**Fig. 1**),
146 suggesting that quantitative variation in the roots was mainly governed by differences in the
147 expression of individual genes. The transcript level of *IGMT1* was higher and that of *IGMT3* lower in
148 DE155 than in DE096, while the transcript level of *IGMT2* was similar between NILs. Interestingly,
149 *IGMT4*$_{DE096}$ exhibited a higher transcript abundance than *IGMT4*$_{DE155}$, indicating that *IGMT4* could be
150 another cryptic root QTG, further counteracting the effect of genotypic variation in *IGMT1*.

151

152 **Patterns of variation in the *IGMT* cluster**

153 In the comprehensive assessment of QTL effects and transcript level variation among NILs, *IGMT2*$_{DE155}$
154 stood out. The DE155 allele was obviously functional in roots, contributing to 4MOI3M at a level
155 comparable to its DE096 counterpart (**Fig. 2**). However, *IGMT2*$_{DE155}$ activity in leaves was relatively
156 minor, despite having a transcript level that was similar, if not slightly higher, than that of *IGMT2*$_{DE096}$
157 (**Fig. 1**).

158 To investigate the cause of this peculiar phenomenon, we built a gene tree with all four genes from
159 both NILs. We expected alleles of the same gene copies to cluster together, which was indeed the case
160 for *IGMT1* and *IGMT4*. Much to our surprise, however, the relationship between *IGMT2* and *IGMT3*
161 genes and alleles remained unresolved (**Fig. S4**), suggesting that *IGMT2*$_{DE155}$ and *IGMT3*$_{DE155}$ did not
162 evolve independently.

163 To examine this further, we procured and manually aligned sequences of the *IGMT* region from the
164 1001 Arabidopsis Genomes project (1001genomes.org), concentrating on *de novo*, reference-quality
165 assembles (Jiao and Schneeberger 2020), and incorporated additional short-read assemblies with near-
166 complete *IGMT* sequences (Gan et al. 2011) (**Table S2**). Our alignment contained data from 28
167 accessions, including several relict lineages (Toledo et al. 2020). All assemblies of the *IGMT* region,
168 particularly those from long reads, showed exactly four *IGMT* copies, suggesting that this configuration
169 exists since more than 200,000 generations (Durvasula et al. 2017).

170 We compared the coding sequences of all four *IGMT* genes and found numerous polymorphisms
171 (**10.6084/m9.figshare.28093955**). To ensure these were not sequencing errors, we amplified and
172 sequenced *IGMT4* coding sequence from 16 Arabidopsis accessions, confirming our findings. Strikingly,
173 many of these polymorphisms were not copy-specific, but shared by two or even more gene copies,
174 suggesting EGC (Mansai and Innan 2010).

175 We tested whether point mutations alone could explain the high number of cases of 'shared
176 polymorphisms', where the same polymorphism segregated at the same site in two or more copies.
177 We focused on instances where two independent point mutations affected the same site, either within
178 a single gene copy or across two distinct copies. The relative probabilities of these events depended
179 on the number of gene copies (n), being (n^{-1}) for two mutations at the same site within a single gene
180 copy and ($1 - n^{-1}$) for two mutations at the same site across two distinct copies.

181 We assumed a large population with initially no polymorphisms within copies but allowing for fixed
182 differences between gene copies. When the same gene copy mutates twice at the same site, two
183 outcomes are possible: either both mutations lead to the same polymorphism, appearing as a single,
184 copy-specific polymorphism, or they result in three different nucleotides segregating at the same site.
185 When two mutations occur at the same site in two different copies, this can lead to a shared
186 polymorphism or to two different copy-specific polymorphisms. The probabilities of these outcomes
187 depend on whether the copies initially had identical or different nucleotides at the side in question,
188 and in the latter case, whether the nucleotides differed by a transition or a transversion (see
189 **Supplementary text** for details). From our data, we estimated a transition/transversion ratio of 2.5935,
190 consistent with other studies (Ossowski et al. 2010; Weng et al. 2019). However, we were uncertain
191 about the initial count of identical *versus* different positions in gene copy comparisons. Therefore, we
192 examined the entire frequency spectrum of identical *versus* different positions, ranging from zero to
193 100%, considering their mutual exclusivity.

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

201 Thus, observed polymorphism patterns revealed far too many shared polymorphisms across gene
202 copies than would be expected from independent point mutations, making EGC the most likely
203 explanation. Indeed, visual inspection identified numerous linked specific variants at polymorphic
204 sites, sometimes spanning hundreds of positions, indicative of gene conversion tracts. These included
205 untranslated regions, coding sequences, and introns (**Tables S3** and **S4**). Reexamining the 13 cases of
206 three segregating nucleotides in the same copy or with two different segregating polymorphisms
207 across multiple gene copies revealed that most of these events were better explained by EGC than by
208 independent point mutations. Hence, shared polymorphisms resulting from two independent point
209 mutations were even more rare than initially suspected. Thus, shared polymorphisms overall indicate
210 EGC.

211 *IGMT4*, the QTG where the knockouts varied for 4MOI3M but not for 4OHI3M, displayed high levels of
212 intermediate frequency polymorphisms within a region of about 200 bp (**Fig. 3b**). This sequence stretch
213 encompassed codons encoding functionally important amino acids, including S-adenosyl-L-methionine

binding sites and the active site, corresponding to amino acids D240, R275 and H278, respectively (**Fig. 3c**). A sliding window analysis, using a window size of 100 and a step size of 25, found significantly elevated Tajima's *D* (Tajima 1989) in this region of *IGMT4* but not in the other three gene copies. Similarly, the HKA test (Hudson et al. 1987) revealed a significant excess of intermediate frequency polymorphisms in the C-terminal half of *IGMT4* ($\chi^2 = 6.16$, $p < 0.05$) compared to *IGMT*s from *A. lyrata*. The results of both tests support the presence of balancing selection.

EGC and subfunctionalization

What are the consequences of EGC for the *IGMT* gene cluster? All four gene products performed the same reaction – converting 4OHI3M to 4MOI3M – but showed organ-specific differences in transcript abundance (**Fig. 1**). CRISPR/Cas9-induced mutations in one gene were not effectively offset by the remaining gene copies, which implied that different *IGMT*s may be expressed in distinct cells. Indeed, *IGMT2* and *IGMT3* are expressed in different root tissues (Cao et al. 2024). Together, these findings suggest subfunctionalization, a process that preserves gene duplicates in a genome by partitioning the functions of the ancestral gene (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000; He and Zhang 2005). Subfunctionalization allows different gene copies to express in different contexts and to acquire mutations that optimize their function within these contexts, such as specific cellular environments. However, when EGC transfers these mutations to a gene copy specialized for a different cellular environment, they may impair the function of this copy, as exemplified by $IGMT2_{\mathrm{DE155}}$ in leaves (**Fig. 2**). Consequently, amino acids that enhance the environment-specific function of gene 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* (Hu et al. 2011) 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* ($t_{\mathrm{df}=28} = -2.42$, $p < 0.05$) (**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

257 gene conversion rates. Independent point mutations are highly unlikely to generate polymorphisms
258 shared across two or more gene copies, whereas EGC generates such shared polymorphisms,
259 especially visible when conversion tracts encompass long sequence stretches with multiple nucleotide
260 substitutions. Quantitative differences in IGMT performance were partially attributable to EGC, as
261 shown by comparing IGMT2 between two near-isogenic Arabidopsis lines. Both alleles were equally
262 active in roots, but IGMT2 from DE155, which had a clear signature of EGC, performed worse in leaves
263 compared to its DE096 counterpart. Furthermore, EGC events have shaped polymorphism patterns
264 across the entire *IGMT* gene cluster (**Fig. S7**), evidenced by sequence comparison of 28 Arabidopsis
265 accessions, suggesting that other *IGMT* gene copies in other accessions may also have experienced
266 performance alterations due to EGC.

267 Duplicated genes can persist longer than expected when higher gene dosage is advantageous (Sugino
268 and Innan 2006; Hanikenne et al. 2013; Heidel-Fischer et al. 2019) or when they functionally diverge.
269 Subfunctionalization is initially a neutral process during which gene duplicates partition the roles of
270 the ancestral gene (Force et al. 1999, Stolzfus 1999). Consequently, if the presence of
271 subfunctionalized copies is fixed by drift, selection preserves this configuration. Subfunctionalized
272 copies may acquire advantageous mutations that enhance copy-specific function or confer novel
273 function.

274 In all of these cases selection stabilizes the presence of multiple gene copies. For dosage effects, EGC
275 leads to homogenization, and perhaps rapid spread of advantageous mutations between copies
276 (Thomas 2006; Mano and Innan 2008; Hanikenne et al. 2013). Maladapted variants arise by mutation
277 but are removed by selection. However, if gene copies have functionally diverged, EGC, as an
278 unavoidable consequence of having multiple gene copies, introduces sequence from a copy adapted
279 to a specific context into another copy adapted to a different context, as exemplified here. Thus, EGC
280 generates inferior variants, which, within a population, manifest as QTL segregating fitness variation
281 (**Fig. 5**). By stabilizing the presence of multiple copies and favoring their specialization, natural selection
282 generates the opportunity for maladaptations to arise by EGC, until copies have sufficiently diverged
283 to prevent EGC. Thus, natural selection itself sets the stage for the generation of more fitness variation
284 than is explained by mutation-selection balance.

285

286 **Materials and Methods**

287 **Culture conditions**

288 Arabidopsis plants were cultivated in growth chambers with 11.5 hours of light at 22°C and 12.5 hours
289 of darkness at 16°C, maintaining approximately around 70 – 80% relative humidity. Seeds were sown
290 on soil and stratified for three days at 6°C. For glucosinolate analysis, one week after germination,
291 plants were transferred to sand and grown randomized in 96-celled trays, fertilized weekly with
292 Hydrocani C2 liquid fertilizer (Hydro Agri), as described by (Pfalz et al. 2016).

293 **Glucosinolate extraction and analysis**

294 Four weeks after germination, approximately 100 g leaf material and the entire root were harvested,
295 weighed and snap-frozen in liquid nitrogen. Glucosinolate extracts were separated in their desulfo-
296 form on a Vanquish HPLC system (Thermo), utilizing a LiChrospher 100 RP-18e LiChroCART column (250
297 x 4, 100A, 5 µm) as described by (Pfalz et al. 2011). Glucosinolate identification was based on retention
298 time and UV spectra, and quantification was based on integrated absorption peak area at 229 nm,
299 using sinigrin as an external standard and published response factors (Buchner 1987) to correct for
300 different UV absorption capacities of various IGs.

301 **RNA extraction and quantitative real-time quantitative PCR (RT-qPCR)**
302 Frozen leaf or root material was finely ground using a TissueLyser II (Qiagen). Total RNA was extracted
303 with TriSure (Bioline) or the NucleoSpin RNA kit (Macherey & Nagel), treated with Turbo DNase
304 (Ambion), and purified using RNeasy MinElute columns (Qiagen) or RNA Clean & Concentrator 5 (Zymo
305 Research). RNA quality and quantity were assessed by agarose gel electrophoresis and a Nanodrop
306 2000 (Thermo). First strand cDNA synthesis was performed with the Maxima First Strand cDNA
307 Synthesis Kit (Thermo), using 500 ng of total RNA. RT-qPCR was conducted with SYBR Green qPCR
308 Master Mix (Thermo) on a StepOnePlus Real Time PCR System (Applied Biosystems), using Arabidopsis
309 *UBQ10* (At4g05320) as the reference gene. Each genotype had three to six biological replicates. Primer
310 sequences are listed in **Table S1**.

311 **Generation of mutant plants**
312 Coding sequences of *IGMT1 – 4* from Arabidopsis accessions Col-0, Da(1)-12 and Ei-2 were used to
313 design three single guide RNA (sgRNA) sequences targeting all four genes simultaneously. The design
314 was performed using the CRISPOR software (crispor.tefor.net), with a preference for sites close to the
315 5'-end of the target genes and low off-target probability. A 1026 bp cassette flanked by attB sites,
316 including all three sgRNAs, each driven by an *At*U6-26 promoter and containing the tracrRNA scaffold
317 as well as the U6 terminator, was synthesized by Genewiz (Leipzig, Germany) and cloned into the
318 pDONR207 vector (Thermo). The final cloning step into pDE-Cas9 DsRed (Fauser et al. 2014; Morineau
319 et al. 2016), permitting selection by DsRed marker fluorescence, was achieved using Gateway™ LR
320 recombination. This vector was then used to transform *Agrobacterium tumefaciens* GV3101. The floral
321 dip method (Clough and Bent 1998) was employed to generate mutants in Arabidopsis DE096, DE155,
322 Col-0 and *igmt5*. After DNA extraction, mutations were identified by Sanger sequencing, using gene-
323 specific primers (**Table S1**) covering the target regions of all three sgRNAs.

324 **Statistical and computational analyses**
325 All measurements of were taken from distinct samples. Fine-mapping of *IGM2* was performed using a
326 general linear model in Systat V9, with markers and growth trays as fixed factors. Glucosinolate
327 quantities were analyzed via ANCOVA in jamovi Version 2.3.28.0 (www.jamovi.org), using the weight
328 of the harvested plant tissue as a covariate and genotype as a fixed factor. When pooling data from
329 multiple experiments, experiment was included as an additional fixed factor, along with an interaction
330 term for genotype × experiment. The leaf effect of *IGMT4* was assessed by ANCOVA, with leaf weight
331 as a covariate and genetic background, genotype at *IGMT4* and experiment as fixed factors. The data
332 for the fraction of 4OHI3M in the combined amount of 4OHI3M and 4MOI3M were arcsine square
333 root-transformed before ANOVA, with genotype as a fixed factor. Statistical differences between
334 genotypes were assessed using post-hoc two-sided *t*-tests. PAML (Yang 2007) was used to estimate
335 the transition/transversion ratio in *IGMT* genes. SplitsTree (Huson and Bryant 2006) was employed to
336 construct a genealogical network for Arabidopsis *IGMT* sequences. DnaSP v6 (Rozas et al. 2017) was
337 used for sliding window analyses of nucleotide polymorphisms and of Tajima's *D* and for the HKA test.
338 Functionally important sites in IGMT4 were identified using ScanProsite (De Castro et al. 2006) on the
339 Expasy webserver (www.expasy.org).

340 **References**

341 Arguello JR, Connallon T. 2011. Gene duplication and ectopic gene conversion in *Drosophila*. *Genes* 2:131-151.

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

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

346 Buchner R. 1987. Approach to determination of HPLC response factors for glucosinolates. In: Wathelet JP, editor.
347 Glucosinolates in Rapeseed: Analytical Aspects. Dordrecht (The Netherlands): Martinus Nijhoff Publishers. p. 50-
348 58.

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

351 Chan EKF, Rowe HC, Corwin JA, Joseph B, Kliebenstein DJ. 2011. Combining genome-wide association mapping
352 and transcriptional networks to identify novel genes controlling glucosinolates in *Arabidopsis thaliana*. *PLOS Biol.*
353 9:e1001125.

354 Charlesworth B. 2015. Causes of natural variation in fitness: Evidence from studies of *Drosophila* populations.
355 *Proc. Natl. Acad. Sci.* 112:1662-1669.

356 Clay NK, Adio AM, Denoux C, Jander G, Ausubel FM. 2009. Glucosinolate metabolites required for an Arabidopsis
357 innate immune response. *Science* 323:95-101.

358 Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of
359 *Arabidopsis thaliana*. *Plant J.* 16:735-743.

360 De Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, Gasteiger E, Bairoch A, Hulo N. 2016.
361 ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues
362 in proteins. *Nucleic Acids Res.* 34:W362-365.

363 Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, Tsuchimatsu T, Burbano HA, Picó FX,
364 Alonso-Blanco C, Hancock AM. 2017. African genomes illuminate the early history and transition to selfing in
365 *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. USA* 114:5213-5218.

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

368 Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by
369 complementary, degenerative mutations. *Genetics* 151:1531–1545.

370 Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan
371 VT, Kahles A, Bohnert R, Jean G, Derwent P, Kersey P, Belfield EJ, Harberd NP, Kemen E, Toomajian C, Kover PX,
372 Clark RM, Rätsch G, Mott R. 2011. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*.
373 *Nature* 477:419-423.

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

376 Harpak A, Lan X, Gao Z, Pritchard JK. 2017. Frequent nonallelic gene conversion on the human lineage and its
377 effect on the divergence of gene duplicates. *Proc. Natl. Acad. Sci. USA* 114:12779-12784.

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

380 Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer
381 G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME,
382 Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KFX, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D,
383 Guo YL. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat. Genet.*
384 43:476-481.

385 Hudson PR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data.
386 *Genetics* 116:153-159.

387 Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254-
388 267.

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

391 Jiao WB, Schneeberger K. 2020. Chromosome-level assemblies of multiple *Arabidopsis* genomes reveal hotspots
392 of rearrangements with altered evolutionary dynamics. *Nat. Commun.* 11:989.

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

395 Kliebenstein DJ, Kroymann J, Brown PD, Figuth A, Pedersen D, Gershenzon J, Mitchell-Olds T. 2001. Genetic
396 control of natural variation in *Arabidopsis thaliana* glucosinolate accumulation. *Plant Physiol.* 126:811-825.

397 Kliebenstein DJ., Kroymann J, Mitchell-Olds T. 2005. The glucosinolate-myrosinase system in an ecological and
398 evolutionary context. *Curr. Opin. Plant Biol.* 8:264-271.

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

401 Kroymann J, Mitchell-Olds T. 2005. Epistasis and balance polymorphism influencing complex trait variation.
402 *Nature* 435:95-98.

403 Lamping E, Zhu JY, Niimi M, Cannon RD. 2017. Role of ectopic gene conversion in the evolution of a *Candida*
404 *krusei* pleiotropic drug resistance transporter family. *Genetics* 205:1619-1639.

405 Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* 290:1151-1155.

406 Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics*
407 154:459–473.

408 Mano S, Innan H. 2008. The evolutionary rate of duplicated genes under concerted evolution. *Genetics* 180:493-
409 505.

410 Mansai SP, Innan H. 2010. The power of the methods for detecting interlocus gene conversion. *Genetics* 184:517-
411 527.

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

414 Morris RT, Drouin G. 2004. Ectopic gene conversions in four *Escherichia coli* genomes: increased recombination
415 in pathogenic strains. *J. Mol. Evol.* 58:596–605.

416 Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.*
417 39:121-152.

418 Ohno S. 1970. Evolution by Gene Duplication. Berlin: Springer-Verlag.

419 Ohta T. 1982. Allelic and nonallelic homology of a supergene family. *Proc. Natl. Acad. Sci. USA* 79:3251–3254.

420 Ohta T. 1991. Role of diversifying selection and gene conversion in evolution of major histocompatibility complex
421 loci. *Proc. Natl. Acad. Sci. USA* 88:6716-6720.

422 Ohta T. 2000. Evolution of gene families. *Gene* 259:45-52.

423 Osada N, Innan H. 2008. Duplication and gene conversion in the Drosophila melanogaster genome. *PLOS Genet.*
424 4:e1000305.

425 Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, Weigel D, Lynch M. 2010. The
426 rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science* 327:92-94.

427 Pfalz M, Mikkelsen M D, Bednarek P, Olsen CE, Halkier BA, Kroymann J. 2011. Metabolic engineering in *Nicotiana*
428 *benthamiana* reveals key enzyme functions in Arabidopsis indole glucosinolate modification. *Plant Cell* 23:716-
429 729.

430 Pfalz M, Mukhaimar M, Perreau F, Kirk J, Hansen CIC, Olsen CE, Agerbirk N, Kroymann J. 2016. Methyl transfer in
431 glucosinolate biosynthesis mediated by indole glucosinolate *O*-methyltransferase 5. *Plant Physiol.* 172:2190–
432 2203.

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

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

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. 2017. DnaSP v6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* 34:3299-3302.

Stoltzfus A. 1999. On the possibility of constructive neutral evolution. *J. Mol. Evol.* 49:169-181.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585-595.

Thomas JH. 2006. Concerted evolution of two novel protein families in *Caenorhabditis* species. *Genetics* 172:2269–2281.

Toledo B, Marcer A, Méndez-Vigo B, Alonso-Blanco C, Picó X. 2020. An ecological history of the relict genetic lineage of *Arabidopsis thaliana*. *Environ. Exp. Bot.* 170:103800.

Weng ML, Becker C, Hildebrandt J, Neumann M, Rutter MT, Shaw RG, Weigel D, Fenster CB. 2019 Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics* 211:703-714.

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

Zimmer E, Martin S, Beverley S, Kann Y, Wilson AC. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. *Proc. Natl. Acad. Sci. USA.* 77:2158-2162.

Acknowledgements

We are grateful for funding from the Agence Nationale de la Recherche (ANR-10-GENM-005, ANR-20-CE92-0042) (J.K.). We thank Marine Paupiére for her contribution to QTL mapping and George Sandler for helpful comments.

Author contributions: M.P. and S.-e. N. performed experiments and analyzed data. J.K. conceived the study and analyzed data. J.K. and J.S. wrote the manuscript with contributions from M.P.

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

Supplementary Text

Figs. S1 to S7

Tables S1 to S4

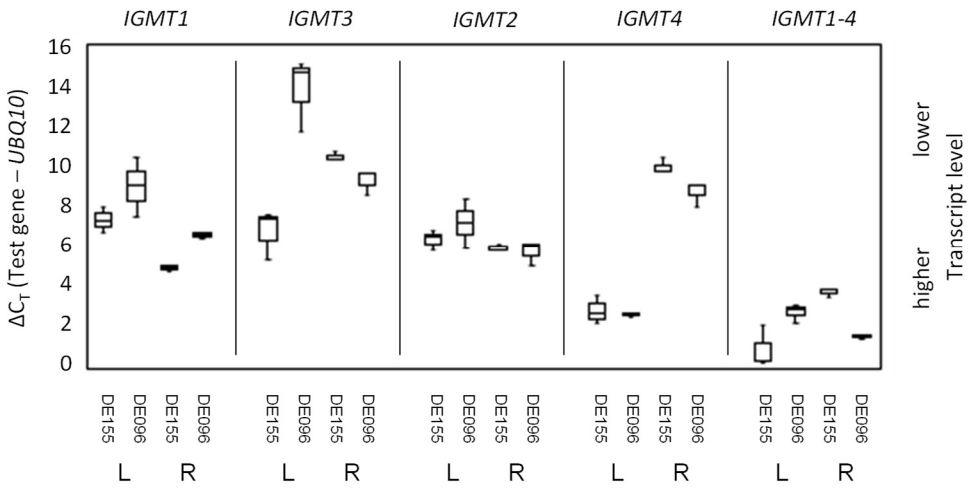


Fig. 1. *IGMT* transcript levels in DE155 and DE096. Shown are boxplots with maximum, median and minimum of ΔC_T values for leaves (L) and roots (R) in comparison to a control gene, *UBQ10*. Each genotype/transcript combination had n = 3 biological samples. *IGMT1* to *IGMT*4 represent individual genes,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

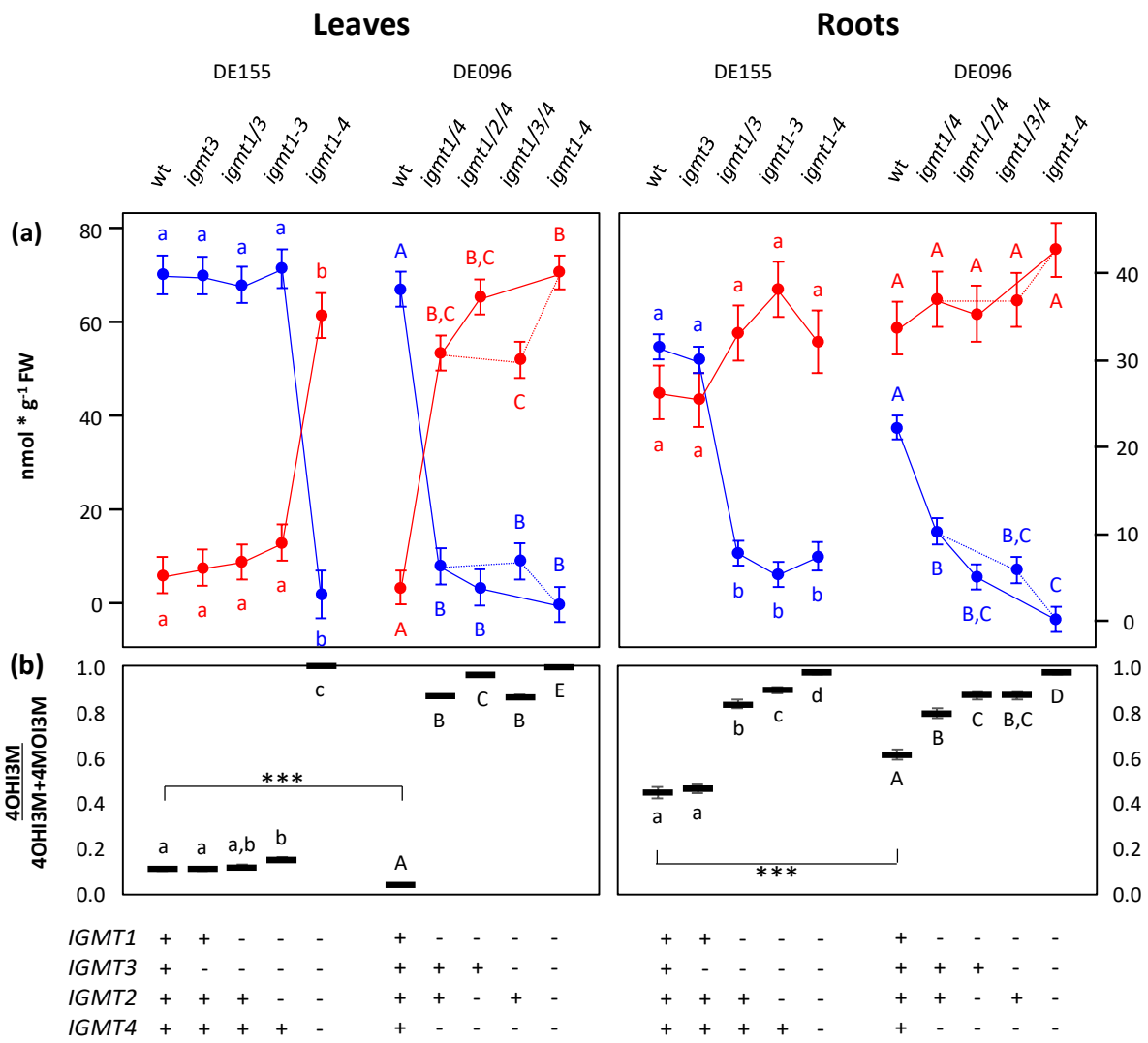


Fig. 2. Quantitative differences in 4MOI3M and 4OHI3M in leaves and roots. Shown are estimated marginal means of metabolite concentration in nmol per gram fresh weight (± SEM) for Arabidopsis DE155, DE096 and *igmt* mutants. Different upper and lower case letters indicate statistically significant differences.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

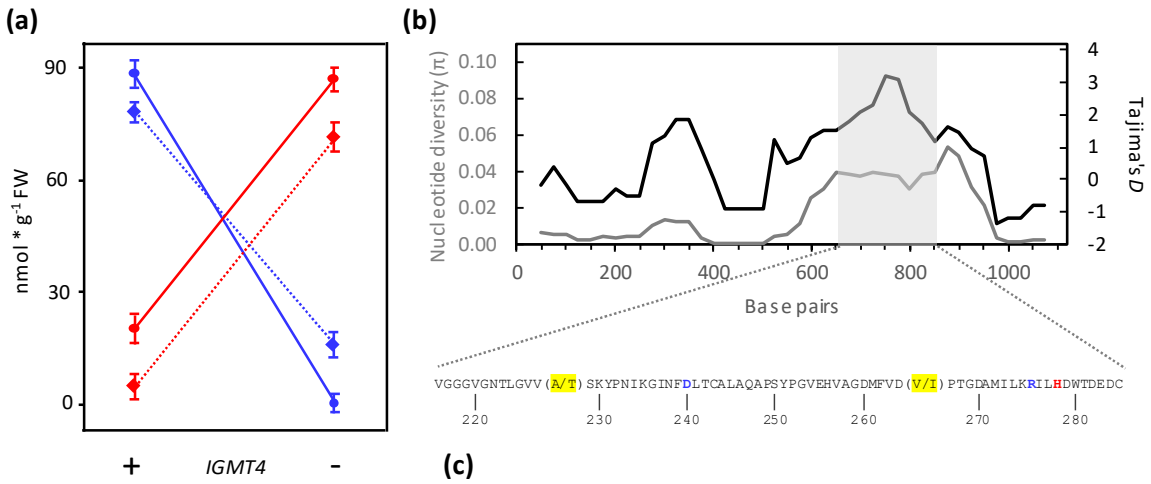


Fig. 3. Analysis of *IGMT4*. a) Quantitative effects of *IGMT4* on leaf IGs in DE155 and DE096. Data pooled from four independent experiments show estimated marginal means of metabolite concentration in nmol per gram fresh weight (± SEM) of 4OHI3M (red) and 4MOI3M (blue) with *IGMT4* intact (left) or defective (right) in DE155 (circles and solid lines) and DE096 (diamonds and dashed lines). Each genotype had n = 35 – 61 samples. The interaction of *IGMT4* by background is highly significant for 4MOI3M ($F_{1,164} = 18.7$; $p < 0.0001$) but not for 4OHI3M ($F_{1,164} = 0.0$; $p = 0.994$). There were no significant genotype or genotype-by-background effects for I3M, 1MOI3M, and total IG. b) Sliding window analyses of nucleotide polymorphisms (gray) and Tajima's *D* (black) with a window size of 100 and a step size of 25 nucleotides. Tajima's *D* is significantly elevated ($D > 2.07$; $p < 0.05$) from ca. 650 to ca. 850 bp (shaded gray). c) Functionally important amino acids. S-adenosyl-L-methionine binding sites are shown in blue, and the active site in red. Two polymorphic amino acids are highlighted yellow.

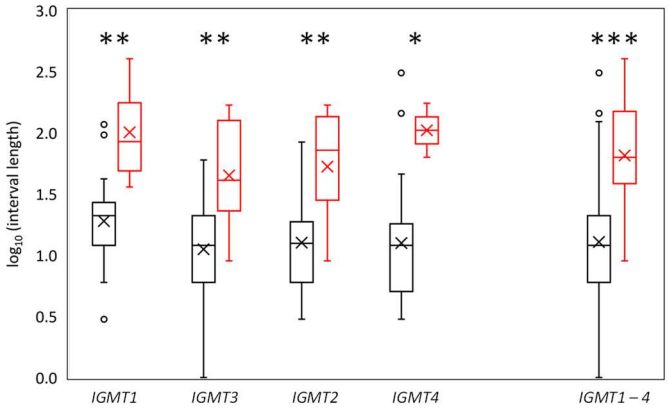


Fig. 4. Boxplot comparing the log10-transformed interval sizes without (black) and with (red) fixed derived amino acids. These derived amino acids are nonsynonymous substitutions that distinguish Arabidopsis IGMTs from those in *Arabidopsis lyrata*. Each interval is flanked by neighboring shared polymorphisms,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

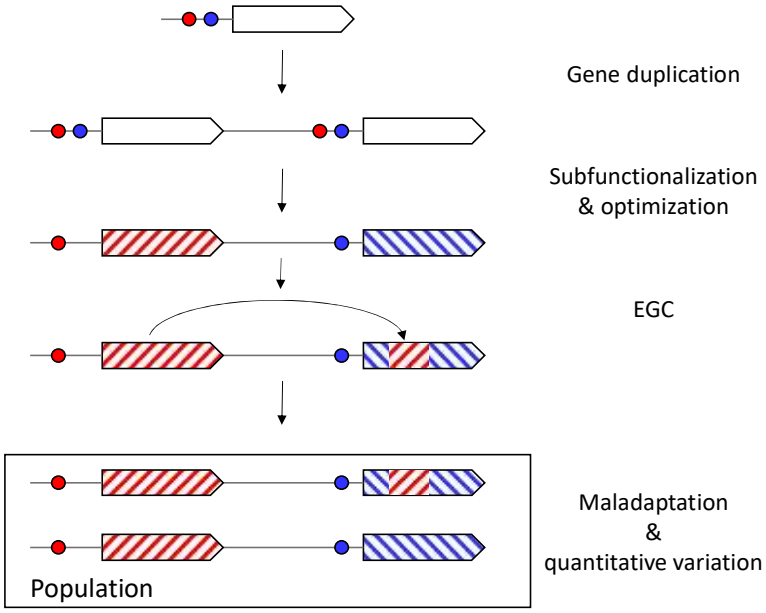


Fig. 5. Maladaptation caused by EGC in gene families. After gene duplication, the paralogs acquire specific expression patterns and optimize accordingly. EGC introduces sequence from one copy optimized for one context into another copy optimized for a different context, leading to maladaptation. This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

506

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

507 $8 \leq n \leq 12$

508 **Supplementary Materials**

509

510 **Supplementary text**

511 **Calculations of probabilities for different types of nucleotide polymorphisms**

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

516 (2 copies: AA*, AB, BA, BB*

517 3 copies: AA*, AB, AC, BA, BB*, BC, CA, CB, CC*

518 4 copies: AA*, AB, AC, AD, BA, BB*, BC, BD, CA, CB, CC*, CD, DA, DB, DC, DD*)

519

520 Each nucleotide can change into any of the three other nucleotides. There is one possibility for a
521 transition, t_i (*e.g.*, A -> G) and two possibilities for a transversion, t_v (*e.g.*, A -> C, A -> T). The probability
522 for a transition is $P_{ti} = x$, the probability for each transversion is $\frac{1}{2} P_{tv} = \frac{1}{2} (1 - x)$. The
523 transition/transversion ratio t_i/t_v can be estimated from the data.

524

525 We first determine the probability that the same mutation occurs twice at the same site, either in one
526 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	$P(X=\mu_2)$	$P(X=\mu_1)*P(X=\mu_2)$
528	A->C	$\frac{1}{2}(1-x)$	A->C	$\frac{1}{2}(1-x)$	$\frac{1}{4}(1-x)^2$
529	A->G	x	A->C	$\frac{1}{2}(1-x)$	$\frac{1}{2}x*(1-x)$
530	A->T	$\frac{1}{2}(1-x)$	A->C	$\frac{1}{2}(1-x)$	$\frac{1}{4}(1-x)^2$
531					
532	A->C	$\frac{1}{2}(1-x)$	A->G	x	$\frac{1}{2}x*(1-x)$
533	A->G	x	A->G	x	x^2
534	A->T	$\frac{1}{2}(1-x)$	A->G	x	$\frac{1}{2}x*(1-x)$
535					
536	A->C	$\frac{1}{2}(1-x)$	A->T	$\frac{1}{2}(1-x)$	$\frac{1}{4}(1-x)^2$
537	A->G	x/(n*l)	A->T	$\frac{1}{2}(1-x)$	$\frac{1}{2}x*(1-x)$
538	A->T	$\frac{1}{2}(1-x)$	A->T	$\frac{1}{2}(1-x)$	$\frac{1}{4}(1-x)^2$

539

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

542 $x^2 + 2 * \frac{1}{4}(1 - x)^2 = x^2 + \frac{1}{2}(1 - 2x + x^2) = 0.5 - x + 1.5\,x^2$

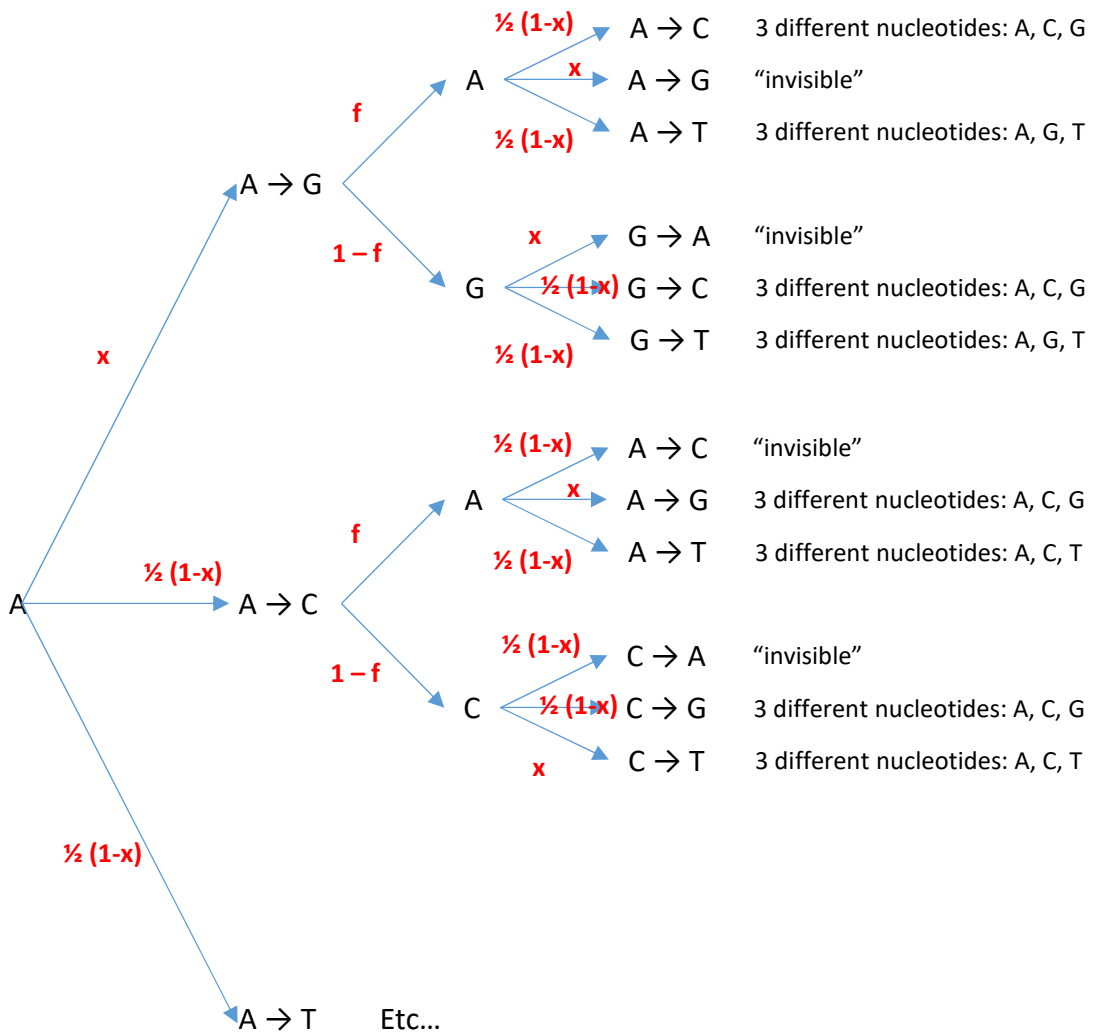
543 Note that identical mutations occurring independently at the same site in different individuals within
544 a large population go unnoticed.

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

546 $1 - (0.5 - x + 1.5\,x^2) = 0.5 + x - 1.5\,x^2$

547 If two distinct point mutations occur at the same site in the same copy in different individuals within a
548 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.

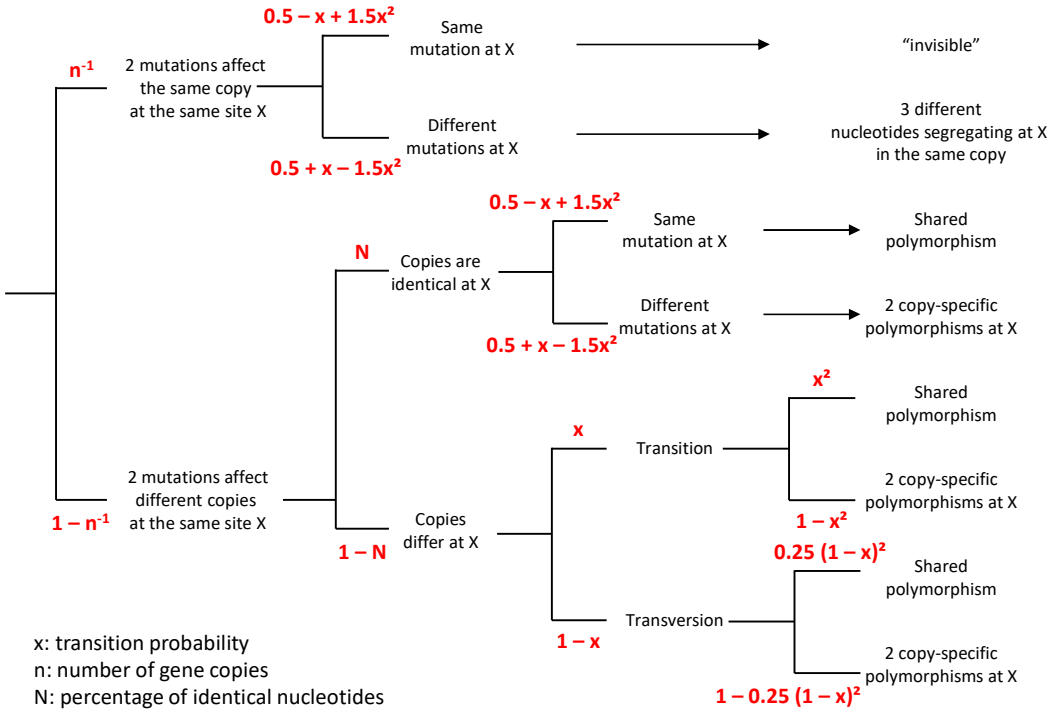
The relative frequencies of scenarios a) and b) are themselves determined by the probabilities of transitions and transversions. Specifically, in x instances, genes with a transition difference serve as the starting point. Conversely, in (1 – x) instances, genes with a transversion difference are the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

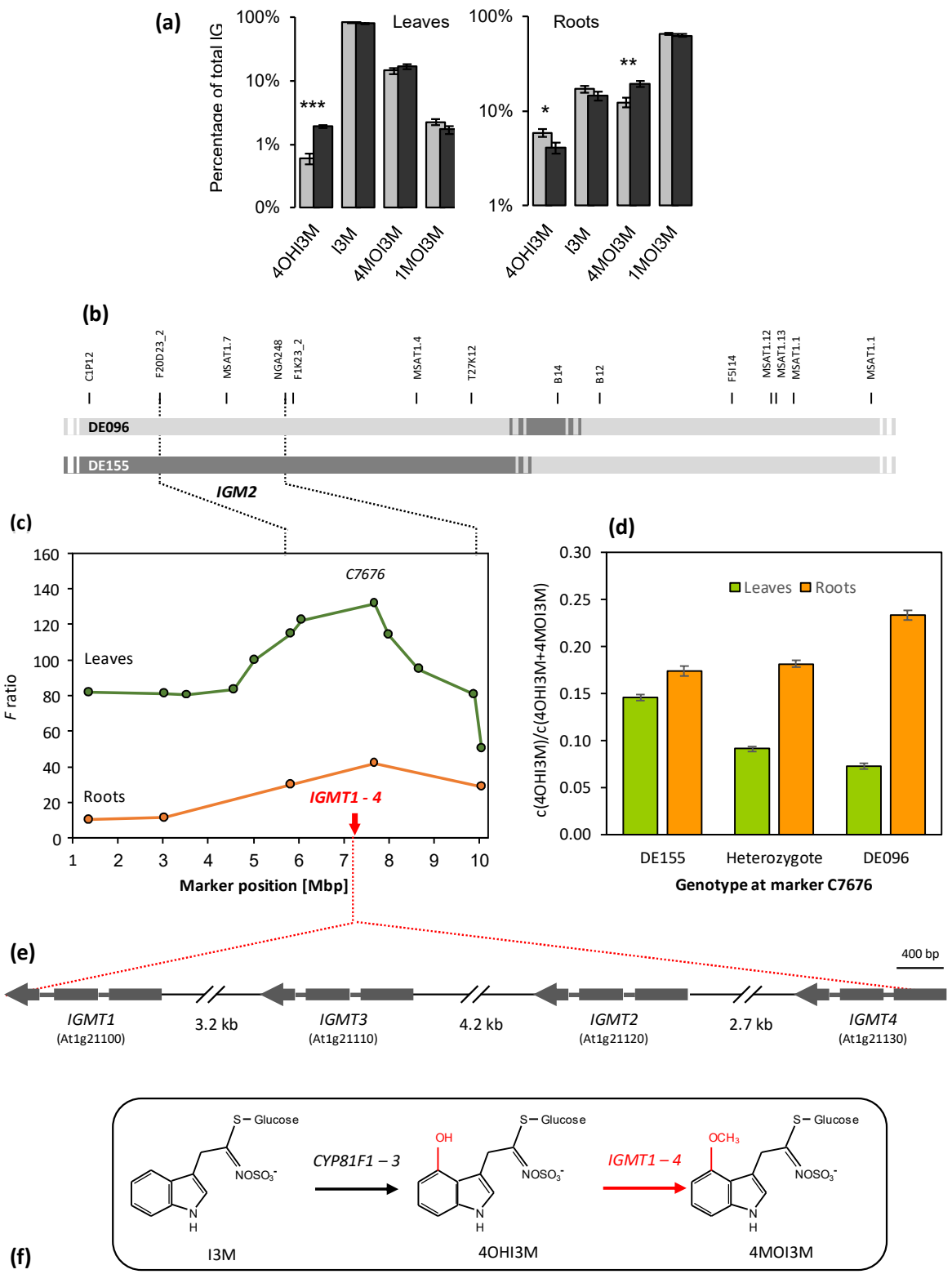
To address this, we have explored scenarios ranging from zero to 100% identical nucleotides, considering that the events of 'two mutations of identical nucleotides at the same position' and 'two mutations of distinct nucleotides at the same position' are mutually exclusive.

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



630 **Supplementary Figures and Tables**

631



632

633

634 **Fig. S1.** Fine mapping of *Indole Glucosinolate Modifier 2*. a) Indole glucosinolate phenotypes of near-
635 isogenic lines DE096 (light gray) and DE155, shown as percentages of total indole glucosinolate.

636 b) Genotypes of DE096 and DE155 on chromosome 1. DE096 carries parental alleles from the
637 Arabidopsis accession Ei-2 (light gray) in the *IGM2* candidate (dashed lines) region, whereas DE155
638 carries alleles from Da(1)-12 (dark gray). DE096 and DE155 marker genotypes do not differ on
639 chromosomes 2 – 5. c) Fine mapping. We fine-mapped the leaf QTL as the arcsine square root-
640 transformed proportion of 4OHI3M among modified indole glucosinolates [c4OHI3M x c(4OHI3M +
641 4MOI3M)$^{-1}$] in 184 F_2 progeny of a cross between near isogenic lines DE096 and DE155 with twelve
642 markers. Marker positions refer to their physical position in the Col-0 reference genome. Statistical
643 support (green line) for the presence of a QTL peaked close to *IGMT1-4*. Subsequently, we confirmed
644 the presence of a similar QTL in roots (orange line) with a subset of markers in an independent
645 experiment with 184 F_2 progeny. d) Estimated proportion (± SEM) of 4OHI3M by genotype at marker
646 C7676 in leaves (green) and roots (orange). e) Organization of IGMT1 – 4. Exon-intron structures are
647 based on the Col-0 accession and drawn to scale, with the intervening sequence shortened. The gene
648 order is the same in all investigated accessions. f) Chemical structures of indole glucosinolates and
649 reactions catalyzed by CYP81F1 – 3 and IGMT 1 – 4. *CYP81F1 – 3* gene products hydroxylate indol-
650 3ylmethyl glucosinolate (I3M) at the 4-position of the indole ring, producing 4-hydroxy-I3M (4OHI3M).
651 This compound then serves as the substrate for *IGMT1 – 4* gene products, which generate 4-methoxy-
652 I3M (4MOI3M).

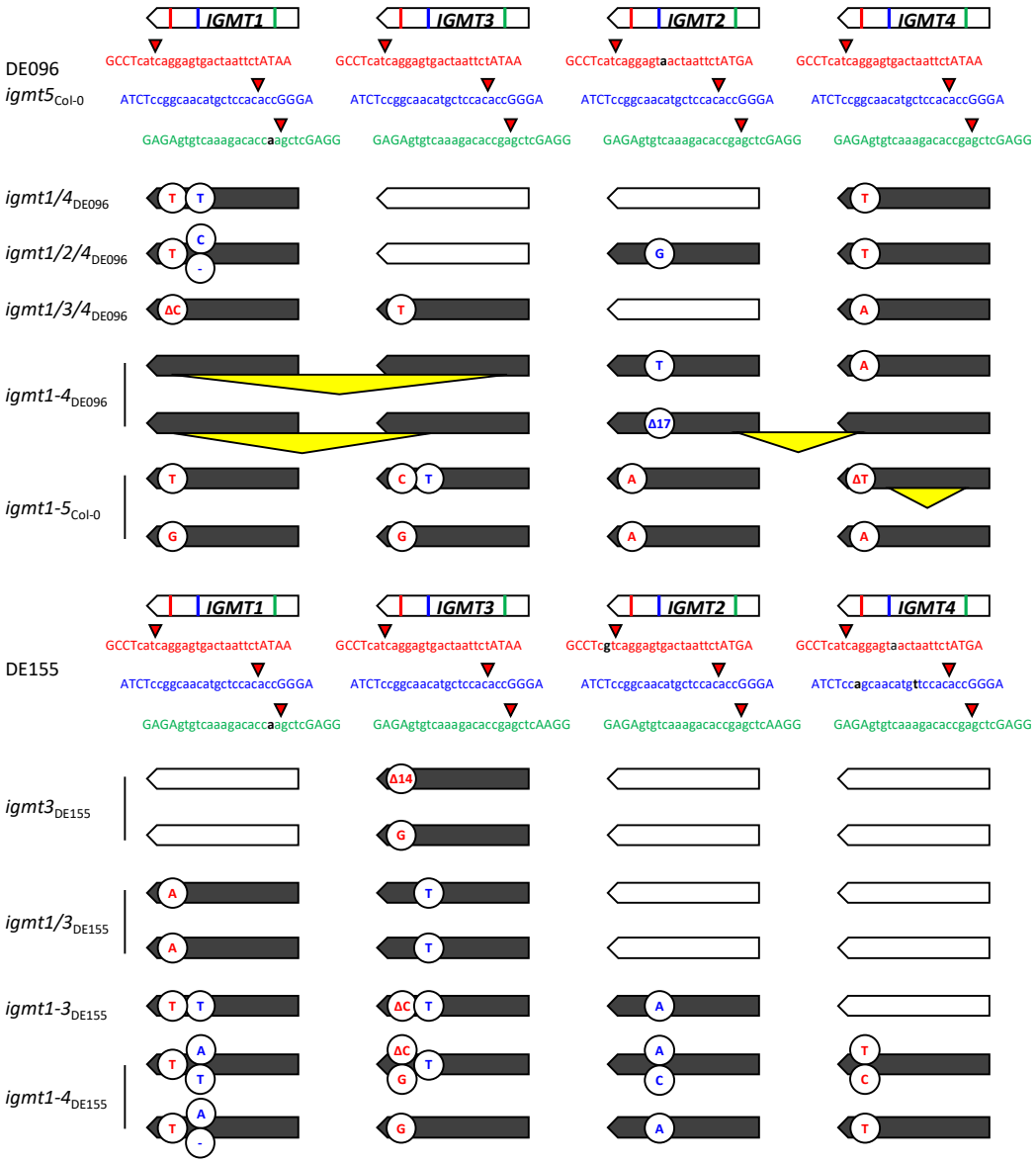


Fig. S2. Schematic representation of *IGMT* wild types and mutants. Locations of sgRNAs (red, blue and green colors) in *IGMT1 – 4*, corresponding sequences and flanking nucleotides (capital letters) in DE096, *igmt5*$_{\text{Col-0}}$, and DE155. Sequence deviations are shown in black. Mutations rendering 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.

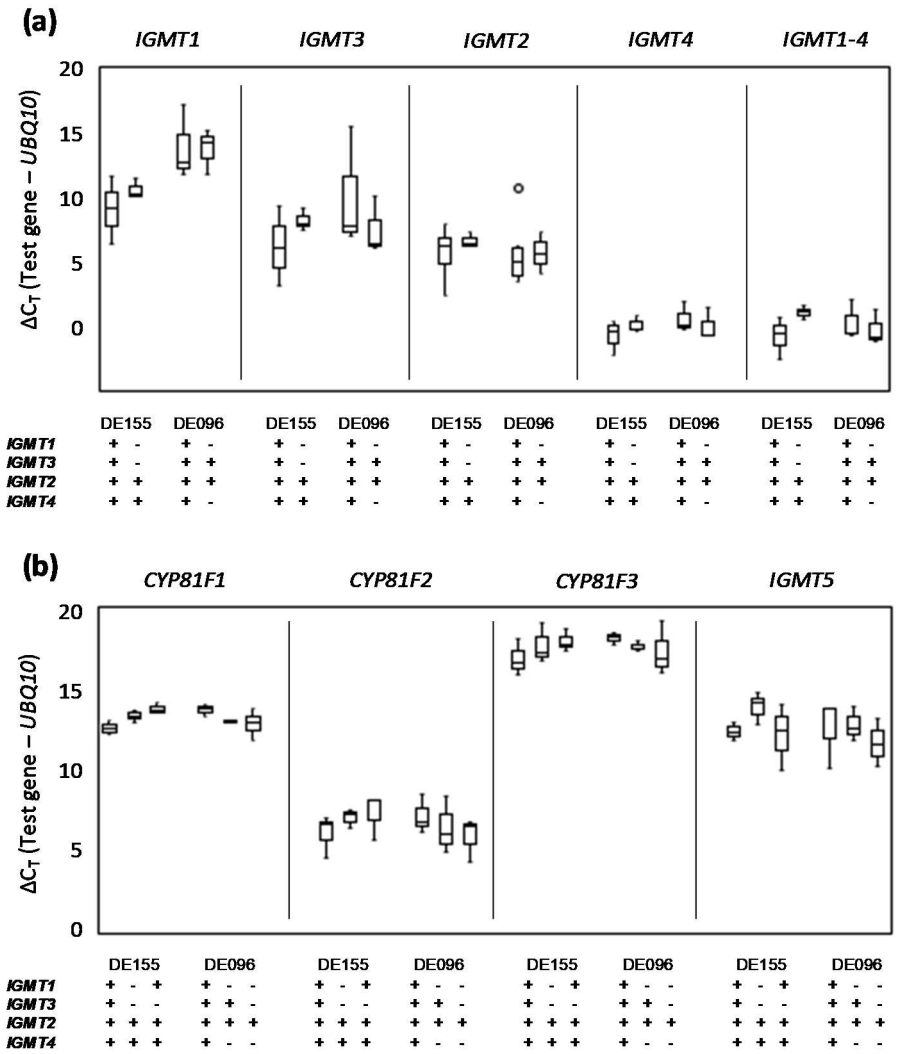


Fig. S3. *IGMT* and *CYP81F* transcript levels in wild type and mutant leaves. a) Boxplots showing *IGMT1 – 4* transcript levels in NILs and selected mutant leaves. b) *CYP81F1 – F3* and *IGMT5* transcript levels. Data are presented as median and range of ΔC_T values in comparison to the control gene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

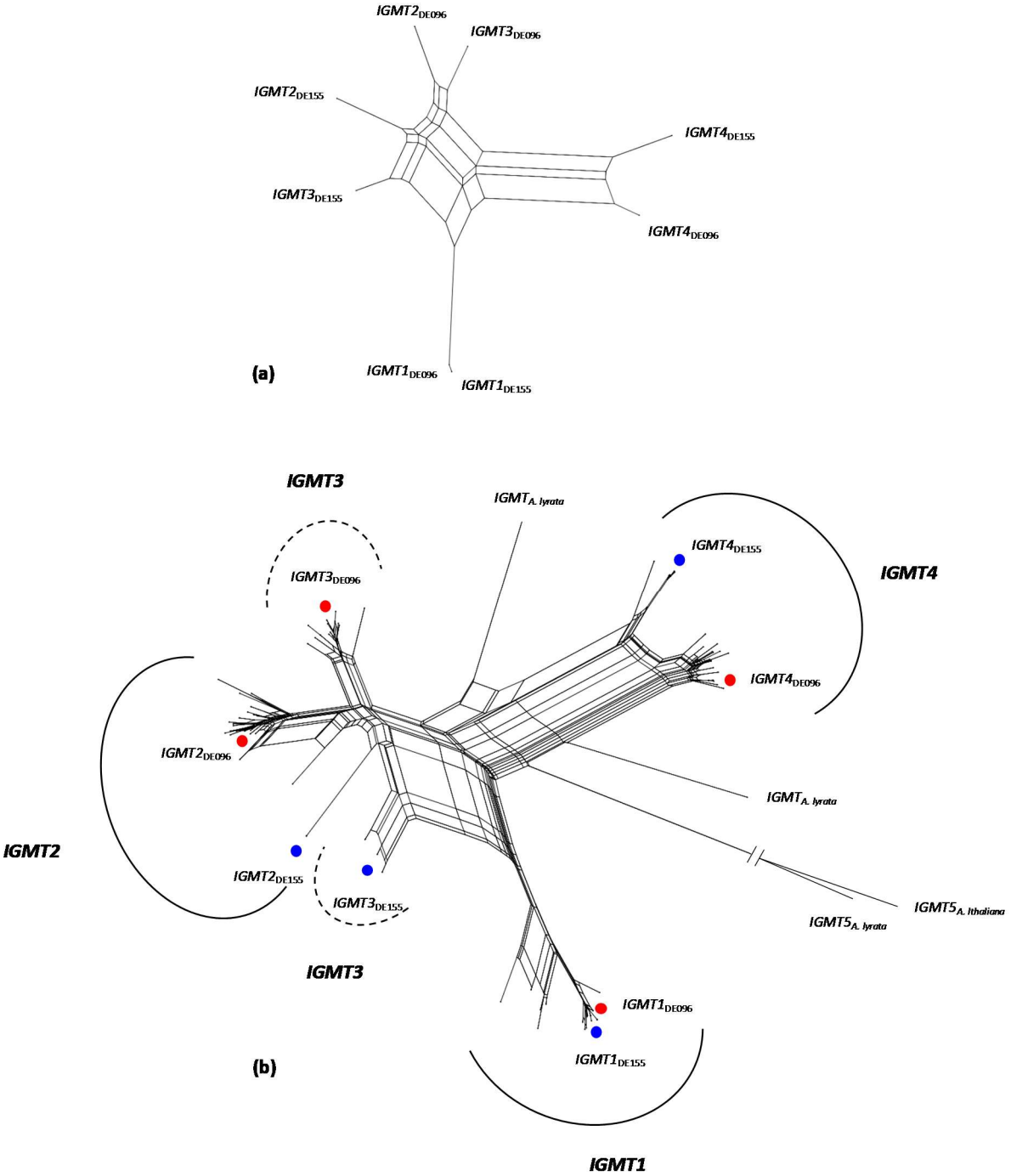
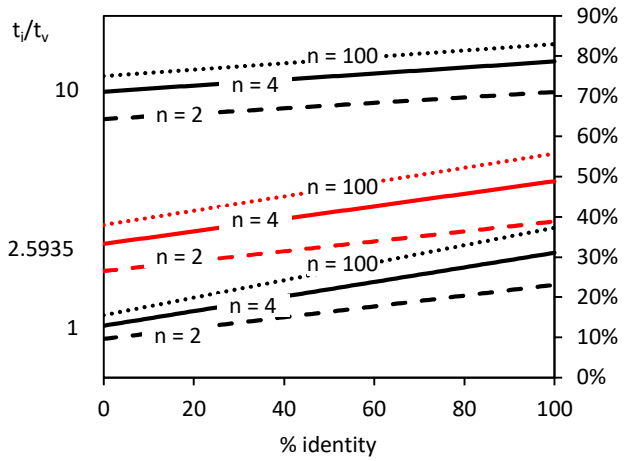
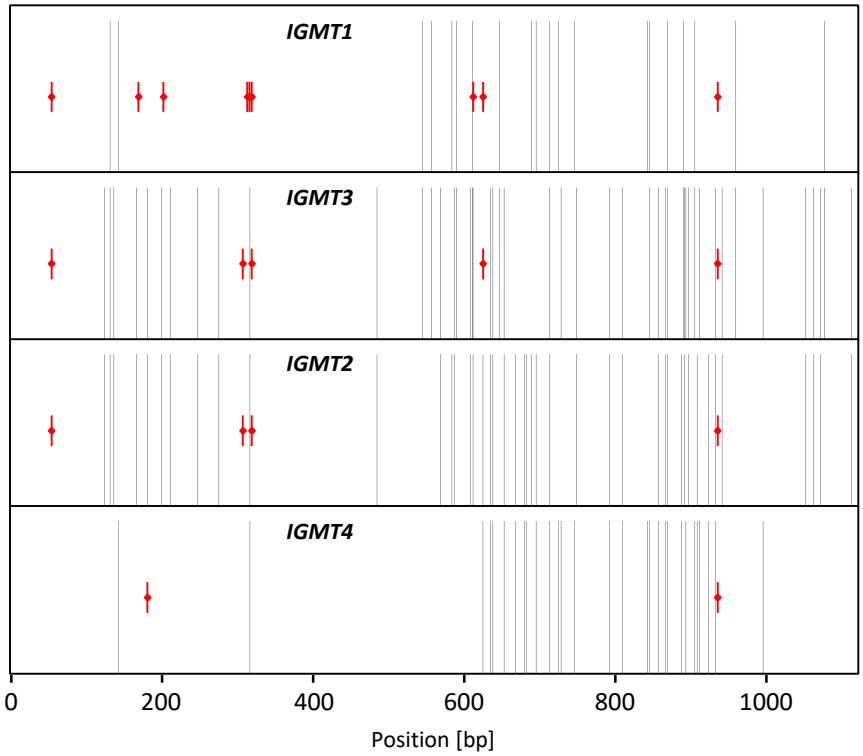


Fig. S4. Genealogical networks of *IGMT* genes. a) Network based on DE155 and DE096 *IGMT* genes. b) Network based on *IGMT1 – 4* genes from 28 Arabidopsis accessions, IGMT5, and corresponding sequences from *A. lyrata*. Colored dots indicate the positions of DE155 (blue) and DE096 (red). The *IGMT5*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



674

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



679

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

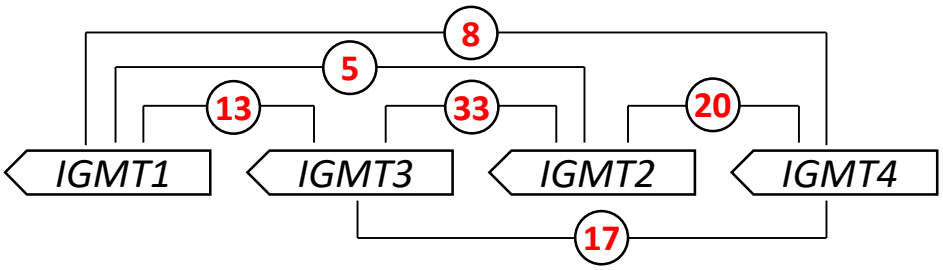


Fig. S7. Number of shared polymorphisms in pairwise comparisons of *IGMT* genes. Note that shared polymorphisms are more frequent when genes are in close proximity. Also, the two central genes exhibit the highest number of shared polymorphisms.

687 **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

688

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

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	

690

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

IGMT1	An-1	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Bur-0	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	C24	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Can-0	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Col-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Ct-1	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Cvi	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Da(1)-12	
	Edi-0	TTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Ei-2	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Eri	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Hi-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Kn-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Kyo	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Ler	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Ler-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAMACTGAT
	Mt-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	No	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Oy-0	TTTTSTTTCTTCTCCTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Po-0	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCAAACTGAT
	Rsch-4	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Sf-2	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Sha	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Tsu-0	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Wil-2	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Ws-0	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Wu-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Zu-0	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
IGMT2	An-1	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Bur-0	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	C24	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Can-0	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Col-0	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Ct-1	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Cvi	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Da(1)-12	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Edi-0	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Ei-2	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Eri	GTTTCTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Hi-0	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Kn-0	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Kyo	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Ler	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Ler-0	GTTACTTTCTTCTCCTTCAAGGCACTCGATCTACGTGATCAAACTGAT
	Mt-0	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	No	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Oy-0	GTTACTTTCTTCTCCTTCAAGGCACTCGATCTACGTGATCAAACTGAT
	Po-0	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Rsch-4	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Sf-2	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Sha	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Tsu-0	GTTATTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
	Wil-2	GTTACTTTCTTCTCCTTCAAGGCACTCGATCTACGTGATCAAACTGAT
	Ws-0	GTTACTTTCTTCTCCTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Wu-0	GTTTCTTTCTTCTCTTTCAAGGCACTCGATCTACGTGATCACACTGAT
	Zu-0	GTTACTTTCTTCTCCTTCAAGGCACTC----TACGTGATCAAACTGAT
Shared polymorphisms		↑↑ ↑ ↑↑↑↑ ↑

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

```
IGMT3 (At1g21110)
An-1    CTGTAAAAACATATTTAAA---ATTAGTATAAAGTTCAAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Bur-0   CTGTAAAAACATATTTAAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
C24     CTGTAAAAACATATTTAAA---ATTAGTATAAAGTTCAAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Can-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Col-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ct-1    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Cvi     CTGTAAAAACATATTTAAA---ATTAGTATAAAATTCTAA-TAGCATGTAGACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Da(1)-12CTGTAAAAACATATTTAAA---ATTAGTATAAAGTTCAAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Edi-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ei-2    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Eri     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Hi-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Kn-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Kyo     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ler     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ler-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAG-TAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Mt-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
No      CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Oy-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Po-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Rsch-4  CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Sf-2    CTGTAAAAACATATTTAA---ATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Sha     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Tsu-0   CTGTAAAAACATATTTAA---ATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Wil-2   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ws-0    CTGTAAAAACATATTTAA---ATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Wu-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTWCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Zu-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTWCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC

IGMT2 (At1g21120)
An-1    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Bur-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
C24     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Can-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Col-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Ct-1    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Cvi     CTGTAAAAACATGTTTAACACATAAGTATAAAGTACTAG-TAACATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Da(1)-12CTGTAAAAACATGTTTAACACATAAGTATAAAGTACTAG-TAACATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Edi-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTT--ATTGAGCTTAC
Ei-2    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Eri     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Hi-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Kn-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Kyo     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ler     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAG-TAACATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ler-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Mt-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
No      CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Oy-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Po-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Rsch-4  CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Sf-2    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Sha     CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Tsu-0   CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTTCTATTGAGCTTAC
Wil-2   CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAG-TAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Ws-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTT--ATTGAGCTTAC
Wu-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGTAAGTTCTATTGAGCTTAC
Zu-0    CTGTAAAAACATATTTAACACATAAGTATAAAGTTCTAGCTAGCATGTATACACTTCCAAAACATATAAGTGATAAGTT--ATTGAGCTTAC

IGMT4 (At1g21130)
An-1    CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Bur-0   CTATAAAAACATATTAAAA----TTAGTATGAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
C24     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Can-0   CTATAAAAACATAATAMAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Col-0   CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ct-1    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Cvi     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Da(1)-12CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Edi-0   CTATAAAAACATATTAAAA----TTAGTATGAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Ei-2    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Eri     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Hi-0    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Kn-0    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Kyo     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Ler     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAG-TAACATGTATACACTTCCAAAACATATAAATAATAAGTT--GTTGAGCTTAC
Ler-0   CTATAAAAACATATTAAACACATTAGTATAAAGTACTAG-TAACATGTATACACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Mt-0    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAGCTACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
No      CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Oy-0    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Po-0    CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Rsch-4  CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Sf-2    CTATAAAAACATATAAAA----ATTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Sha     CTGTAAAAACATATTTAACACATAAGTATAAAGTACTAGCTAGCATGTATACACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Tsu-0   CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAGCTACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Wil-2   CTATAAAAACATAATAAAA----TTAGTATAAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTTCTATTGAGCTTAC
Ws-0    CTATAAAAACATATTAAA----ATTAGTATAAAGTTCTAGCTACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC
Wu-0    CTATAAAAACATATTAAA----TTAGTATGAAGTTCTAG-TACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--GTTGAGCTTAC
Zu-0    CTATAAAAACATAATAAA----TTAGTATAAAGTTCTAGCTACCATGTATAAACTTCCAAAACATATAAGTAATAAGTT--ATTGAGCTTAC

Shared polymorphisms
```