Using single-cell genomics to explore transcriptional divergence and *cis*-regulatory dynamics of duplicated genes

Xiang Li[1], Xuan Zhang[1], Robert J. Schmitz[1]

[1]Department of Genetics, University of Georgia, Athens, GA, 30602, USA.

Address for correspondence: schmitz@uga.edu

Abstract

Gene duplication is a major source of evolutionary innovation, enabling the emergence of novel expression patterns and functions. Leveraging single-cell genomics, we investigated the transcriptional dynamics and *cis*-regulatory evolution of duplicated genes in cultivated soybean (*Glycine max*), a species that has undergone two rounds of whole-genome duplication. Our analysis revealed extensive transcriptional divergence among these duplicated gene pairs, including dosage balanced, paralogue dominance, divergence, specialization, and widespread mono-expressed pattern. These expression divergences are correlated with sequence variation in their associated accessible chromatin regions (ACRs), where *cis*-regulatory elements reside. Moreover, different duplication mechanisms, including whole-genome duplication and small-scale duplication, likely give rise to distinct types of *cis*-regulatory variants, leading to varying effects on transcriptional divergence. To further explore the evolution of *cis*-regulatory elements and their impact on gene expression, we focused on gene sets in which duplicated genes are derived from two rounds of whole-genome duplication and share a common ancestral gene, as *cis*-regulatory elements are typically co-duplicated with genes during whole-genome duplication. Most ACRs were retained after duplication whereas a subset likely arose *de novo*. Both mutations in duplicated elements and the emergence of *de novo* evolved elements contribute to widespread

expression divergence within these gene sets. We also characterized the evolution of cell-type-specific expression and cell-type-specific ACRs among these sets. Collectively, our findings highlight the important role of *cis*-regulatory evolution in shaping transcriptional divergence and facilitating the retention of duplicated genes, offering a high-resolution view through the lens of single-cell genomics.

Introduction

Gene duplication is a crucial driver of evolution, providing the raw materials for new gene functions and expression patterns (Lynch and Conery 2000; Birchler and Yang 2022). It arises through small-scale duplication, such as tandem or proximal duplications, or through whole-genome duplication that replicates entire genomic segments simultaneously (Zhang 2003; Kuzmin et al. 2022). These mechanisms generate duplicated genes with varying evolutionary rates, essentiality, and functions (Freeling 2009; Qiao et al. 2019; Shi et al. 2020; Cai and Des Marais 2024; Benoit et al. 2025). In plants, approximately 65% of annotated genes have duplicates, mostly derived from whole-genome duplication, consistent with the prevalence of palaeopolyploidy in land plants (Wendel 2000; Blanc and Wolfe 2004; Panchy et al. 2016; Tiley et al. 2016). The retention of these duplicates has been instrumental in shaping plant genomes and contributing to important agricultural traits (Panchy et al. 2016; Salman-Minkov et al. 2016; Benoit et al. 2025).

Following gene duplication, one or both copies may be subject to relaxed selective constraints and, in some instances, acquire new functions that contribute to adaptation (neofunctionalization) or partition ancestral functions (subfunctionalization). However, functional divergence is not always necessary for their retention and typically unfolds over extended evolutionary timescales (Panchy et al. 2016; Birchler and Yang 2022). Multiple mechanisms have been proposed to account for the dispensability of retained duplicates (Innan and Kondrashov 2010; Panchy et al. 2016; Qiu et al. 2020). Moreover, factors such as selection, genetic drift and genetic redundancy (or genetic buffering) can also

influence gene retention (Panchy et al. 2016; Iohannes and Jackson 2023). On the other hand, this redundancy invariably complicates genotype-phenotype associations and limits the transferability of genetic findings across closely related species (Iohannes and Jackson 2023; Benoit et al. 2025). Despite their evolutionary and functional significance, the dynamics of gene duplication and diversification over short evolutionary timescales remain insufficiently explored (Benoit et al. 2025).

Divergent expression among duplicate genes is a widespread phenomenon and provides crucial insight into the evolution of duplicated genes. In Arabidopsis, 70% of duplicated gene pairs exhibit significant transcriptional differences (Ganko et al. 2007). Similarly, approximately 50% of soybean duplicated genes, derived from whole-genome duplication, show transcriptional divergence (Roulin et al. 2013). This divergence is likely driven by genetic variants within regulatory regions (Wray et al. 2003), particularly *cis*-regulatory elements that control temporospatial expression patterns (Wittkopp and Kalay 2012; Schmitz et al. 2022). In cotton, divergence in *cis*-regulatory sequences between its diploid progenitors contributes to transcriptional divergence in 40% of the duplicated genes (Chaudhary et al. 2009). Similar patterns have been observed in Arabidopsis (Haberer et al. 2004), soybean (Fang et al. 2023), strawberry (Fang et al. 2024) and additional cotton studies (Han et al. 2022; Hu et al. 2024). Despite these valuable insights, a key underexplored question remains: how do changes in gene expression and *cis*-regulatory control influence the retention of duplicated genes? Specifically, what types of transcriptional divergence and genetic variations in *cis*-regulatory regions persist in duplicated genes to ensure their retention over relatively short evolutionary timescales?

Soybean (*Glycine max*) has gone through two rounds of whole-genome duplication, approximately 59 and 13 million years ago (Schmutz et al. 2010; Koenen et al. 2021; Zhuang et al. 2022), which make it an excellent model for exploring this question. Gene pairs resulting from both whole-genome and small-scale duplications have been identified in soybean (Qiao et al. 2019). Previous studies have also explored the evolutionary fates of duplicated genes in soybean, highlighting divergence in gene expression and

71 the role of *cis*-regulatory control (Du et al. 2012; Zhao et al. 2017; Fang et al. 2023). However, these
72 studies primarily assessed transcriptional divergence at the tissue level, often relying on differential
73 expression as the sole metric (Fang et al. 2023), which likely oversimplifies "divergence" and overlooks
74 differences at the cellular level. For instance, even when two duplicated genes exhibit similar overall
75 expression levels within a given tissue, they may display distinct expression patterns across different cell
76 types within that tissue. The recently generated single-cell assay for transposase-accessible chromatin
77 sequencing (scATAC-seq) and single-nucleus RNA sequencing (snRNA-seq) data in soybean (Zhang et al.
78 2024) provide a valuable opportunity to study transcriptional and *cis*-regulatory sequence divergence of
79 duplicated genes at a single-cell resolution. snRNA-seq enables precise measurement of gene expression
80 at the cellular level, whereas scATAC-seq profiles chromatin accessibility, offering deeper insights into
81 how *cis*-regulatory activity, particularly cell-type-specific ones, contributes to gene expression
82 divergence.

83 In this study, we investigated duplicated genes by integrating single-cell expression and chromatin
84 accessibility profiles to examine transcriptional and *cis*-regulatory sequence divergence. This approach
85 provides a more comprehensive understanding of transcriptional divergence and its correlation with *cis*-
86 regulatory divergence, shedding light on how these processes are intertwined with the distinct fates of
87 duplicated genes arising from different duplication mechanisms. Given that *cis*-regulatory elements are
88 typically duplicated along with genes during whole genome duplication, we further examined their
89 evolution and impacts on expression divergence by focusing on gene sets in which genes are derived
90 from two rounds of whole-genome duplications and share the same ancestral gene. Our analysis
91 revealed that most *cis*-regulatory elements were retained after whole-genome duplication, whereas a
92 subset evolved *de novo*. Mutations within these duplicated elements and the emergence of novel ones
93 collectively drive transcriptional divergence. Lastly, we explored the evolution of cell-type-specific
94 expression and cell-type-specific *cis*-regulatory sequences among these gene sets. Overall, our findings

95 underscore the power of single-cell genomics in revealing the divergence of expression and chromatin
96 accessibility in retained duplicated genes, offering key insights into the evolution of *cis*-regulatory
97 elements and gene duplication.
98

99 Results

100 Single-cell transcriptomics reveal the expression relationships of retained duplicated genes

101 A previous study identified 73,983 duplicated gene pairs in soybean, classifying them into two main
102 categories: whole-genome duplication and small-scale duplication gene pairs (Qiao et al. 2019). The
103 small-scale duplicated gene pairs were further subdivided into four types based on their duplication
104 mechanisms: tandem duplication (closely adjacent duplicated genes on the same chromosome),
105 proximal duplication (duplicated genes located nearby on the same chromosome but separated by up to
106 10 other genes), transposed duplication (duplicated genes transposed to distant genomic positions), and
107 dispersed duplication (duplicated genes scattered across the genome, neither neighboring nor colinear)
108 (**Fig. 1a and supplementary table S1**) (Qiao et al. 2019). Importantly, individual genes can undergo
109 multiple duplication events via distinct duplication mechanisms. For instance, a gene duplicated through
110 whole-genome duplication may later experience tandem duplication. Among all identified duplicated
111 gene pairs, whole-genome duplication derived gene pairs represented the largest proportion (47.1%),
112 followed by dispersed duplication (40.9%) (**Fig. 1b and supplementary table S2**). The remaining 12% of
113 duplicated gene pairs arose from tandem, proximal, and transposed duplications, with transposed
114 duplication being the most prevalent among them (**Fig. 1b and supplementary table S2**).

115 The expression dynamics of duplicated gene pairs play a crucial role in their retention (Thompson
116 et al. 2016; Birchler and Yang 2022; Iohannes and Jackson 2023). To investigate this, we integrated the
117 gene pair information with the single-cell gene expression profiles from seven tissues - root, hypocotyl,
118 nodule, globular stage seed, heart stage seed, cotyledon stage seed and early maturation stage seed

119 (Zhang et al. 2024). Building on the methods developed by (Benoit et al. 2025), we refined the approach
120 to assess gene expression at the level of individual cells within each tissue, rather than solely across
121 different tissues (Benoit et al. 2025). This cell-level resolution allows for a more precise and
122 comprehensive examination of expression relationships of duplicated genes, providing insight into the
123 expression-based mechanisms that contribute to their retention (Coate et al. 2020). In addition to the
124 four previously described biologically meaningful expression patterns (Benoit et al. 2025) – (1) dosage
125 balanced, where duplicates retain similar expression profiles and levels across cell types within a tissue
126 (**Fig. 1c**); (2) paralogue dominance, where duplicates share similar expression profiles but exhibit
127 divergent expression levels (**Fig. 1d**); (3) specialization, where duplicates display distinct expression
128 profiles, with one acquiring cell-type-specific expression (**Fig. 1e**); (4) divergence, where duplicates differ
129 in both expression profiles and levels – we introduced a fifth category (**Fig. 1f**): (5) mono-expression, in
130 which only one gene in the pair is expressed whereas the other remains inactive in a given tissue (**Fig.**
131 **1g).** Additionally, gene pairs in which neither gene is expressed in the examined tissues are classified as
132 non-expression, whereas those that do not fit any of the above categories are grouped as "others" (**Fig.**
133 **1h**). Among the five biologically meaningful expression patterns mentioned above, the mono-expression
134 category consistently contains the most gene pairs across different tissue types (**Fig. 2a and**
135 **supplementary table S3**). This is followed by dosage balanced, divergence, paralogue dominance, and
136 specialization patterns, except in the hypocotyl, where divergence gene pairs outnumber dosage
137 balanced ones, and in the early maturation seed stage, where paralogue dominance exceeds divergence
138 (**Fig. 2a and supplementary table S3**).
139 When examining the expression pattern of gene pairs based on their duplication mechanisms, we
140 found that those derived from whole-genome duplication and dispersed duplication exhibited similar
141 patterns, whereas gene pairs from tandem, proximal, and transposed duplications showed greater
142 similarity to each other (**Fig. 2b and supplementary fig. S1 and table S3**). The unpredictable, seemingly

143 random nature of dispersed duplication events (Ganko et al. 2007; Qiao et al. 2019) may explain why the
144 expression pattern of gene pairs derived from this mechanism differ from those derived from other
145 small-scale duplications. Based on this distinction, we focused on gene pairs arising from tandem,
146 proximal, and transposed duplications—collectively referred to as regional duplications due to their
147 tendency to remain within specific chromosomal regions—to compare their expression patterns with
148 those of gene pairs derived from whole-genome duplication.

149 First, the proportion of expressed gene pairs—where at least one gene is active in a given tissue—
150 varied by duplication mechanisms. For instance, in early nodule tissue, over 60% of gene pairs from
151 regional duplications (excluding transposed duplication) were not expressed, whereas only 37% of gene
152 pairs derived from whole-genome duplication remained inactive (**supplementary table S3**). Second,
153 among expressed gene pairs, most regional duplications exhibited mono-expression (**Fig. 2b and**
154 **supplementary fig. S1**). Specifically, up to 72% of expressed gene pairs from regional duplications
155 showed mono-expression in early nodule tissue, compared to 43% of those from whole-genome
156 duplication (**supplementary table S3**). Lastly, a few gene pairs from regional duplications displayed other
157 biologically meaningful expression patterns, such as dosage balanced, paralogue dominance,
158 specialization, or divergence expression (**Fig. 2b and supplementary fig. S1**). For example, only 11% of
159 expressed gene pairs from proximal duplication exhibited these patterns in early nodule tissue,
160 compared to approximately 20% of expressed gene pairs from whole-genome duplication
161 (**supplementary table S3**). These results suggested that gene pairs exhibiting correlated expression
162 patterns, such as dosage balanced, paralogue dominance and specialization expression, are more likely
163 to originate from whole-genome duplication than from regional duplications. Similar trends have been
164 previously reported in Arabidopsis, where duplicated genes originating from large-scale duplications and
165 retained in duplicated segments exhibit more correlated expression patterns than those arising from
166 small-scale duplications or those not located within duplicated segments (Casneuf et al. 2006).

By assigning an expression pattern, including non-expressed, “others”, mono-expressed, dosage balanced, paralogue dominance, specialization and divergence, to each gene pair across the seven examined tissue types, we were able to analyze their expression dynamics across tissues. In total, we identified 1,344 distinct expression patterns (**supplementary table S4**). To simplify the results, we focused on the top 25 expression patterns across tissue types, each comprising at least 0.5% of the gene pairs (**Fig. 2c and supplementary table S4)**. The most prevalent expression patterns involved gene pairs exhibiting mono-expression in at least one tissue and non-expression in the remaining tissues (**Fig. 2c**). Together, these patterns accounted for more than 20% of the gene pairs analyzed (**supplementary table S4**). Additionally, in gene pairs exhibiting a mono-expression across tissues, the same gene within the pair was always expressed across different tissue types (**supplementary fig. S2 and table S5**). A previous study also found that differential expression within gene pairs tend to occur in the same direction across tissues (Roulin et al. 2013). The non-expression in all examined tissues was the second most common expression pattern across tissue types (**Fig. 2c**), which was expected given that our study included only seven tissue types. A lot of gene pairs also exhibited mono-expression across seven tissues, ranking the third most common expression pattern (**Fig. 2c**). The remaining top 25 expression patterns across tissue types often involved “others” expression pattern along with biologically meaningful expression patterns such as mono-expression, dosage balanced and divergence (**Fig. 2c**). Notably, dosage balanced expression was the most frequently observed biologically meaningful pattern when combined with “others” expression pattern (**Fig. 2c**). Together, the single-cell expression profiles revealed a wide variety of expression patterns exhibited by duplicated gene pairs both within and across tissue types.

Correlations between genetic variants within ACRs and expression patterns of duplicated gene pairs

Cis-regulatory variants are thought to drive transcriptional divergence in duplicated gene pairs and contribute to their retention (Iohannes and Jackson 2023); to explore this, we analyzed Accessible

191 Chromatin Regions (ACRs), where *cis*-regulatory elements reside, from scATAC-seq data in each of the
192 seven tissue types corresponding to the snRNA-seq data (Zhang et al. 2024). For each gene pair, we
193 associated ACRs with duplicated gene based on their proximity and counted the number of ACRs
194 associated with each gene separately. In some cases, both genes had at least one associated ACR,
195 whereas in others, only one gene was linked to ACR whereas the other one had no associated ACRs. The
196 latter one was more frequently observed in gene pairs exhibiting mono-expression than in those
197 exhibiting other biologically meaningful expression patterns (**Fig. 3a and supplementary fig. S3**). Gene
198 pairs where only one gene had associated ACR also showed a higher proportion of mono-expression in
199 the respective tissue compared to those where both genes had ACR (**supplementary table S6**). These
200 findings suggest that the presence of *cis*-regulatory elements in only one duplicated gene, may be a key
201 factor, although not the sole mechanism, driving mono-expression in duplicated gene pairs.
202 To assess genetic variation in ACRs associated with duplicated gene pairs, we used a BLAST-based
203 approach following the method of (Mendieta et al. 2024). For each gene pair, the ACR sequence
204 associated with one duplicated gene was used as the query, whereas the reference was defined as the
205 genomic region spanning from the upstream to the downstream gene of the corresponding duplicate
206 (**Fig. 3b**). This approach accounts for the possibility that the ACR linked to the corresponding duplicate
207 may be positioned between it and either its upstream or downstream gene (**Fig. 3b**). Conserved ACRs
208 were defined based on BLASTed ratio, calculated as the length of the BLASTed sequence in the reference
209 region divided by the length of the query ACR. ACRs with BLASTed ratios greater than 0.1 were
210 considered conserved (**Fig. 3c**). We found that gene pairs exhibiting dosage balanced had the highest
211 percentage of conserved ACRs, whereas divergence gene pairs had the lowest (**Fig. 3d, supplementary**
212 **fig. S3 and table S7**). Dosage balanced gene pairs exhibit the most similar expression profiles and levels
213 (**Fig. 1c**), whereas divergence gene pairs show great divergence (**Fig. 1f**). Similarly, gene pairs displaying
214 paralogue dominance - characterized by similar expression profiles but varied expression level (**Fig. 1d**) -

215 also had a high percentage of conserved ACRs (**Fig. 3d, supplementary fig. S3 and table S7**). These
216 results suggest that proportion of conserved ACRs correlates with the degree of expression similarity
217 between duplicated genes.

218 Further analysis of genetic variation between conserved ACRs and their corresponding BLASTed
219 sequences - including average BLASTed ratio and average mismatch rate (calculated as the ratio of
220 mismatches between the BLASTed sequence and the ACR query to the BLASTed sequence length) -
221 revealed that ACRs and their BLASTed sequences were more divergent in gene pairs with divergence
222 expression pattern compared to those exhibiting other expression patterns (**Fig. 3e, supplementary fig.
223 S3, table S7**). Interestingly, gene pairs exhibiting mono-expression had a higher percentage of conserved
224 ACRs, greater BLASTed ratio, and lower mismatch rate than those with divergence expression pattern
225 (**Fig. 3d-3e, supplementary fig. S3, table S7**), despite mono-expression being a more extreme form of
226 divergent expression. One possible explanation is that these measures were based on conserved ACRs,
227 indicating the presence of both ACRs and their corresponding BLASTed sequences in the reference
228 region. However, a higher percentage of unBLASTed ACRs - those without a corresponding sequence in
229 the reference region - was found in mono-expressed gene pairs compared to those with other
230 biologically meaningful expression patterns (**Fig. 3f, supplementary fig. S3, table S7**). This suggests that
231 ACRs associated with mono-expressed gene pairs are more likely to be unique, lacking counterparts in
232 their associated duplicated genes. This uniqueness may result from the *de novo* emergence or loss of *cis*-
233 regulatory elements, which could play a key role in driving mono-expression. Overall, our analysis
234 highlights the correlation between expression divergence and ACR sequence variation, considering both
235 the presence or absence of corresponding ACRs and the sequence divergence between conserved ACR
236 and its corresponding counterpart.

237 When further examining these BLAST results in the context of duplication mechanisms, distinct
238 patterns emerged among gene pairs derived from different duplication mechanisms (**supplementary fig.**

239 **S4 and table S8**). Gene pairs originating from tandem duplication exhibited the highest percentage of
240 conserved ACRs (**supplementary fig. S4**). This observation may be influenced by a methodological
241 limitation of our BLAST approach, which tends to map ACRs back to themselves for tandem-duplication-
242 associated ACRs. Specifically, the reference region of one gene encompasses both the tandemly
243 duplicated gene and its adjacent upstream or downstream gene, often the gene itself, as tandemly
244 duplicated genes are positioned next to each other (**Fig. 3b**). Tandemly duplicated genes can sometimes
245 exhibit similar expression profiles, with increased copy number leading to higher overall expression
246 levels, suggesting that either the shared *cis*-regulatory elements regulate both copies, or that *cis*-
247 regulatory elements were co-duplicated along with the genes. However, relatively fewer tandem-
248 duplicated gene pairs exhibited dosage balanced expression pattern (**Fig. 2b and supplementary fig. S1**),
249 indicating that genetic variation in ACRs may be disproportionately masked or skewed by the current
250 method. Alternatively, the transcriptional activities of these genes could be impacted by positional
251 differences of *cis*-regulatory elements, the dependence of expression on DNA folding or competition for
252 transcription factors at duplicated *cis*-regulatory elements.

253 For gene pairs derived from other duplications—particularly whole-genome duplication, proximal
254 duplication, and transposed duplication—ACRs and their corresponding BLASTed sequences displayed
255 similar levels of divergence (**supplementary fig. S4 and table S8**), despite differing gene expression
256 patterns (**Fig. 2b and supplementary fig. S1**). This pattern may be explained by the difference in the ratio
257 of unBLASTed ACRs. Gene pairs derived from whole-genome duplication had a much lower percentage of
258 unBLASTed ACRs compared to those from proximal and transposed duplications (**supplementary fig. S4**
259 **and table S8**). *Cis*-regulatory elements of duplicated genes derived from whole-genome duplication were
260 more likely to be entirely duplicated (Shoemaker et al. 2006), whereas the variation in the presence or
261 absence of duplicated *cis*-regulatory elements was more likely common in gene pairs derived from other

262 duplication mechanisms. Together, different duplication mechanisms likely give rise to different types of
263 *cis*-regulatory variants, leading to varying effects on transcriptional divergence.
264
265 ## The evolution of duplicated *cis*-regulatory regions and their impact on expression dynamics and gene
266 retention following two rounds of whole-genome duplications
267 Soybean is a paleopolyploid species that has undergone two rounds of whole-genome duplication,
268 theoretically resulting in four duplicated genes originating from a single ancestral gene. Although
269 extensive gene loss followed these duplications (Schmutz et al. 2010; Du et al. 2012; Zhao et al. 2017),
270 the retained complete gene sets provide a valuable opportunity to examine the evolution of *cis*-
271 regulatory elements - duplicated alongside their associated genes (Li and Schmitz 2025) - and their
272 impact on gene expression and retention. Specifically, we focused on gene sets containing four
273 duplicated genes derived from whole-genome duplication and sharing a common ancestral gene across
274 all bean-like (papilionoid) legume species.
275 A total of 2,383 gene sets, comprising 9,532 genes, were identified for analysis (**supplementary**
276 **table S9**). Within individual tissue, these duplicated gene sets showed transcriptional divergence, both in
277 the number of genes expressed and in the expression correlation among the expressed genes - an
278 indicator of expression similarity (**Fig. 4a and 4b, supplementary fig. S5**). For instance, in cotyledon stage
279 seed, ~ 40% of gene sets contained at least one duplicated gene being not expressed (**Fig. 4a and**
280 **supplementary fig. S5**). Given that a coexpression value of 0.9 had been established as the threshold for
281 defining co-expression for duplicated gene pairs (Benoit et al. 2025), most gene sets with more than two
282 expressed genes had mean coexpression value below this threshold, indicating their low expression
283 correlations (**Fig. 4b and supplementary fig. S5**). Based on these observations, we categorized gene sets
284 into mono-, dual-, tri-, and quad-expressed groups, and further classified them (except mono-expressed)
285 as correlated if their mean coexpression value exceeded 0.9, and uncorrelated otherwise (**Fig. 4c and**

286 **supplementary table S10**). Among these, the quad-uncorrelated category contained the largest number
287 of gene sets (**Fig. 4c, supplementary fig. S5 and table S10**). Additionally, quad-expressed gene sets
288 displayed higher expression levels than those in other categories, with the difference most pronounced
289 when compared to mono-expressed gene sets, which consistently exhibited the lowest expression levels
290 (**supplementary fig. S6**). Across tissues, a variety of expression pattern combinations were observed,
291 with quad-uncorrelated expression across all seven examined tissues being the most prevalent (**Fig. 4d,**
292 **supplementary tables S11 and S12**). Collectively, these results reveal widespread transcriptional
293 divergence among the four duplicated genes within the set, reflected in variation in expression presence
294 or absence, expression correlation, and overall expression level.

295 To investigate genetic variation in ACRs underlying transcriptional divergence within duplicated
296 gene sets, we BLASTed each ACR associated with a duplicated gene against the reference regions of the
297 other three duplicated genes in the same set. To account for varying levels of sequence divergence
298 among ACRs arising from the first and second rounds of whole-genome duplication, we assigned each
299 ACR a conservation score, defined as the number of times it was conserved (BLASTed ratio > 0.1) across
300 the three BLASTs. Based on this metric, ACRs were classified into three categories: unique ACRs, which
301 did not BLAST to any regions near the other duplicated genes (conservation score = 0); two-copy
302 conserved ACRs, which appeared to be conserved only between the two genes duplicated from the most
303 recent whole-genome duplication (conservation score = 1/3); and multiple-copy conserved ACRs, which
304 were conserved across three or four duplicated genes from both rounds of whole-genome duplications
305 (conservation score = 1 or 2/3, with the latter suggesting possible loss of ACR in one duplicate) (**Fig. 5a**).
306 For each gene set, we then calculated the average BLASTed ratio and mismatch rate for two-copy and
307 multi-copy conserved ACRs to quantify ACR sequence divergence.

308 The proportion of ACRs associated with gene set that were two-copy or multi-copy conserved was
309 generally similar, and both were substantially higher than the proportion of unique ACRs—a pattern

310 consistent across gene sets with varying expression profiles (**Fig. 5b, supplementary fig. S7 and table**
311 **S13**). These findings suggest that most ACRs associated with duplicated gene sets are conserved
312 elements, likely retained from the first or second rounds of whole-genome duplication, whereas only a
313 small fraction are unique ACRs, potentially representing *de novo* elements that evolved at a low rate.
314 Notably, multi-copy conserved ACRs showed a lower average BLASTed ratio and higher average mismatch
315 rate than two-copy conserved ACRs (**Fig. 5c, supplementary fig. S8, and table S13**), indicating greater
316 sequence divergence in multi-copy conserved ACRs. This pattern supports a temporal model of
317 regulatory sequence divergence, where ACRs gradually accumulate sequence divergence over time
318 following duplication.

319 To examine the relationship between genetic variants associated with ACRs and expression
320 divergence within duplicated gene sets, we compared the proportions of different ACR categories and
321 their BLAST sequence similarity across gene sets with varying expression patterns. First, in gene sets with
322 correlated expression—particularly those showing tri- or quad-correlated patterns —the proportion of
323 multi-copy conserved ACRs was higher than that of two-copy conserved ACRs (**Fig. 5b, supplementary**
324 **fig. S7 and table S13**), suggesting that multi-copy conserved ACRs may play a more prominent role in
325 maintaining coordinated expression, especially in sets with more than two co-expressed genes. Second,
326 sequence variation among ACRs was associated with expression divergence in corresponding gene
327 duplicates. For example, in cotyledon tissue, tri- or quad-uncorrelated groups exhibited a lower average
328 BLASTed ratio in multi-copy conserved ACRs compared to correlated groups, as highlighted by red
329 squares in **Fig. 5c**. A similar pattern was observed for two-copy conserved ACRs in some other tissues
330 and when using mismatch rate as an alternative measure of sequence divergence (**supplementary fig. S8**
331 **and table S13**). These findings suggested that genetic variation in either multi-copy or two-copy
332 conserved ACRs can contribute to expression divergence. Lastly, we found no consistent patterns in the
333 proportions of different ACR types, average BLASTed ratios or average mismatch rates across gene sets

334 grouped by the number of expressed genes (**Fig. 5b – 5c, supplementary fig. S8 and table S13**),
335 indicating that regulation of gene expression presence or absence may be more case-dependent or
336 influenced by mechanisms beyond *cis*-regulatory elements.
337

338 ## The evolution of cell-type-specific expression and cell-type-specific ACRs

339 Some genes exhibiting cell-type-specific expression are essential for determining or maintaining
340 cell identity, with *cis*-regulatory elements controlling at which cell types, during what developmental
341 stages and at what levels these genes are expressed (Wittkopp and Kalay 2012; Schmitz et al. 2022;
342 Marand et al. 2023). This specificity can differ among duplicated gene sets - some retain expression in
343 the same cell type across all four copies (**Fig. 6a**), whereas others exhibit varied cell-type-specific
344 expression (**Fig. 6b - 6c**). These patterns, in turn, enable the identification of candidate *cis*-regulatory
345 regions responsible for expression specificity and provides insights into the mechanisms driving its
346 evolution.
347 Here, we continued to focus on gene sets originating from two rounds of whole-genome
348 duplication that share a common ancestral origin, exclusively selecting sets in which all four genes are
349 expressed and at least one exhibit cell-type-specific expression. This filtering step allowed us to focus on
350 changes in cell-type-specific expression, rather than the confounding effects of loss of expression in
351 some duplicates. To assess the evolutionary dynamics of cell-type-specific expression, we clustered
352 duplicated genes based on their expression contributions and then counted the number of shared
353 clusters within each set, which reflects the degree of conservation in cell-type-specific expression (**Fig.
354 6**). Only a few sets retained a single shared cluster (**Fig. 7, supplementary fig. S9 and table S14**), implying
355 the potential conservation of cell-type-specific expression from the ancestral gene (**Fig. 6a**), whereas
356 most sets had multiple clusters (**Fig. 7, supplementary fig. S9 and table S14**), particularly in uncorrelated

357 sets, likely reflecting subfunctionalization of ancestral gene expression or the emergence of novel cell-
358 type-specific expression (**Fig. 6b -6c**).
359 To explore the role of *cis*-regulatory elements in shaping these expression patterns, we examined
360 cell-type-specific ACRs (ctACRs) associated with these gene sets (Zhang et al., 2024). Gene sets that
361 retained a single shared cluster always contained at least one conserved ctACR; notably, in some tissue
362 types, no gene sets composed exclusively of unique ctACRs contained a single shared cluster (**Fig. 7,**
363 **supplementary fig. S10**). In contrast, gene sets with multiple shared clusters showed a variety of ctACR
364 configurations—some with only conserved ctACRs, some with only unique ctACRs, and others with a mix
365 of both (**Fig. 7, supplementary fig. S10**). These patterns suggest that changes in cell-type-specific
366 expression can arise from either sequence divergence within conserved ctACRs or from the emergence
367 of *de novo* evolved ctACRs. Moreover, multi-copy conserved ctACRs in gene sets with a single shared
368 cluster exhibited much higher average BLASTed ratios and lower average mismatch rates compared to
369 those in gene sets with multiple shared clusters (**supplementary fig. S11 and table S15**), highlighting the
370 important role of multi-copy conserved ctACRs in maintaining conserved cell-type-specific expression
371 within the set.
372

373 ## Discussion

374 This study investigates the expression patterns of duplicated gene pairs and gene sets across
375 diverse cell types and tissues using single-cell genomic data, offering a comprehensive understanding of
376 the diversity and prevalence of their transcriptional profiles. Further analysis of ACRs associated with
377 these duplicated gene pairs or sets revealed correlations between the sequence variation in ACRs and
378 expression divergence. Additionally, different duplication mechanisms likely give rise to distinct types of
379 *cis*-regulatory variants, resulting in varying effects on transcriptional divergence. To gain deeper insights
380 into the evolution of *cis*-regulatory elements, we analyzed the expression patterns and associated ACRs

381 of four genes derived from whole-genome duplication that share a common evolutionary origin. Most
382 ACRs associated with these gene sets were retained from the first or second whole-genome duplication
383 events and have accumulated mutations over time, whereas a small subset of these ACRs likely evolved
384 *de novo*. The mutations in conserved ACRs and the emergence of novel ACRs collectively contribute to
385 widespread expression divergence—reflected in the number of genes expressed, expression level, and
386 expression correlation. Furthermore, we examined the evolution of cell-type-specific expression and
387 ctACRs, revealing dynamics of cell-type-specific gene expression within duplicated gene sets. These
388 changes were driven by mutations in conserved ctACRs as well as the emergence of novel regulatory
389 elements, highlighting the regulatory complexity underlying cell-type-specific expression patterns.
390

391 ## Single-cell genomics: a key to understanding transcriptional divergence and genetic variants in chromatin
392 accessibility

393 The transcriptional divergence of duplicated genes has been extensively studied in polyploid plants
394 (Du et al. 2012; Roulin et al. 2013; Zhao et al. 2017; Bird et al. 2020; Wang et al. 2021; Han et al. 2022;
395 Fang et al. 2023; Hu et al. 2024). However, due to technological limitations, most previous studies have
396 relied on bulk RNA-seq, assessing divergence primarily through differential expression analyses at tissue
397 level, with only few focusing on the single-cell level (Coate et al. 2020). Yet, such tissue-level differential
398 expression—whether high or low across entire tissues—can result from diverse underlying patterns,
399 including paralog dominance, specialization, or divergence (**Fig. 1**), which cannot be resolved by using
400 bulk RNA-seq. In contrast, single-cell genomics enables gene expression analysis at the cellular level,
401 providing a more detailed view of expression divergence, including shifts in expression across specific cell
402 types, the presence or absence of expression in particular tissues, and altered expression patterns across
403 tissue types (**Fig. 1 and 2**). Additionally, this improved resolution of transcriptional divergence,

404 particularly across different cell types within individual tissues, allows for a deeper understanding of its
405 relationship with ACR genetic variants (**Fig. 3**).
406 Admittedly, bulk ATAC-seq, which identifies ACRs within tissues, is sufficient to uncover broad
407 correlations. Combining bulk ATAC-seq with snRNA-seq could potentially pinpoint candidate regulatory
408 regions underlying specific transcriptional patterns, such as conserved ACRs with high BLASTed ratio or
409 low mismatch rate associated with gene pairs exhibiting dosage balanced expression or unique ACRs
410 linked to gene pairs with mono- or divergence expression patterns. However, integrating scATAC-seq with
411 snRNA-seq provides a more powerful approach to explore the evolution of cell-type-specific expression
412 and ctACRs. We did observe the correlations between genetic similarity in ctACRs and expression
413 similarity of their associated gene sets (**supplementary fig. S11**) and the contributions of both mutations
414 in conserved ctACRs and emergence of novel ACRs to the dynamics of the cell-type-specific expression
415 within the sets (**Fig. 7 and supplementary fig. S10**). Importantly, these results offer a valuable resource
416 to identify *cis*-regulatory elements responsible for the conservation (**e.g. Fig. 6a**) or shifts (**Fig. 6b - 6c**) in
417 cell-type-specific expression in specific case studies in the future.
418

419 ## The evolution of *cis*-regulatory elements

420 The duplication of *cis*-regulatory elements serves as a key source for the emergence of new
421 regulatory elements (McDonald and Reed 2024; Li and Schmitz 2025). Similar to protein-coding
422 sequences of duplicated genes, which often experience reduced selective constraints (Kondrashov et al.
423 2002), duplicated *cis*-regulatory regions may also undergo relaxed selection, allowing mutations to
424 accumulate through genetic drift and potentially driving phenotypic novelty. Moreover, this pattern may
425 be more pronounced in whole-genome duplications than small-scale duplications. Duplicates arising
426 from whole-genome duplication tend to retain shared *cis*-regulatory elements (Shoemaker et al. 2006),
427 although this is not necessarily the case for those evolved from small-scale duplications (Arsovski et al.

428 2015). For example, a higher number of conserved non-coding regions are more likely to be retained in
429 duplicated genes arising from whole-genome duplication (Schnable et al. 2011). Our observation of a
430 higher percentage of ACRs that can be BLASTed to corresponding duplicated regions in whole-genome
431 duplication-derived genes compared to small-scale duplication-derived genes (**supplementary fig. S4**),
432 also underscores this distinction. These differences, in turn, can influence expression pattern divergence,
433 with more pronounced divergence seen in duplicated genes from small-scale duplications, especially
434 those arising from regional duplications (**Fig. 2 and supplementary fig. S1**).
435 On the other hand, the evolution of *cis*-regulatory elements can involve modifications to
436 preexisting elements through mutations, deletions, insertions, or the *de novo* emergence of regulatory
437 sequences from previously non-regulatory regions (Katikaneni and Lowe 2025; Li and Schmitz 2025). Our
438 analysis of proportional of different categories of ACRs associated with the gene sets with four whole-
439 genome duplication derived genes with shared common ancestor (**Fig. 5a**), potentially provides insights
440 into the relative contributions of these mechanisms to *cis*-regulatory sequence evolution. Most ACRs
441 associated with these sets were retained from the first or second rounds of whole-genome duplication
442 whereas only a small subset of ACRs evolved *de novo* (**Fig. 5b and supplementary figs. S7 and S8**)
443 suggesting that both modifications of existing sequences and *de novo* emergence of *cis*-regulatory
444 elements exist but the latter one probably occur in a lower rate. Additionally, the obvious lower average
445 BLASTed ratio and higher average mismatch rate in multi-copy conserved ACRs than those in two-copy
446 conserved ACRs (**Fig. 5c**), suggest that mutations in ACRs have accumulated over time.
447

448 ## The retention of duplicated genes and the involvement of *cis*-regulatory control

449 Three canonical outcomes of duplicated genes over long evolutionary time include gene loss,
450 subfunctionalization and neofunctionalization (Ohno 1970; Lynch and Conery 2000; Birchler and Yang
451 2022; Kuzmin et al. 2022; Gout et al. 2023). However, before duplicated genes can undergo

452 subfunctionalization or neofunctionalization, they must first be maintained without being selected
453 against—a process that can persist for an extended period, though in some situations
454 subfunctionalization and neofunctionalization can arise quickly (Birchler et al. 2005; Birchler and Yang
455 2022). The prevailing consensus in the field suggests that the prolonged maintenance phase is primarily
456 driven by selective constraints on gene dosage (Iohannes and Jackson 2023). Duplicated genes can
457 impose negative fitness effects due to increased expression dosage, which can lead to stoichiometric
458 imbalances (Birchler et al. 2005; Freeling and Thomas 2006; Veitia et al. 2008; Veitia et al. 2013; Birchler
459 and Yang 2022). This issue is particularly critical for genes involved in protein-protein interactions, as
460 proper stoichiometries among interacting partners are essential for maintaining functional complexes
461 (Defoort et al. 2019; Birchler and Yang 2022; Iohannes and Jackson 2023). A key distinction between
462 whole-genome duplication and small-scale duplication lies in dosage balance: whereas whole-genome
463 duplication preserves the relative dosage of all interacting partners, small-scale duplication disrupts this
464 balance because only a subset of genes is duplicated (Ohno 1970; Defoort et al. 2019; Kuzmin et al.
465 2022). The greater prevalence of similar expression patterns—including dosage balanced, paralogue
466 dominant, and specialization expression—and the reduced incidence of mono-expression in gene pairs
467 derived from whole-genome duplication compared to those from small-scale duplication, particularly
468 regional ones, (**Fig. 2b and supplementary fig. S1**) likely reflects the impact of dosage-balance selection.
469 Although the initial similarity in expression among gene pairs derived from whole-genome duplication
470 may simply result from the simultaneous duplication of *cis*-regulatory elements alongside the genes, the
471 continued maintenance of these regulatory elements suggests the potential influence of selective
472 pressures.
473 On the other hand, the process of maintaining dosage balance is complex and can involve several
474 mechanisms: hypofunctionalization, where both duplicates reduce their expression; dosage balance,
475 where the duplicates remain reduced but similar expressions to preserve the original gene's function; or

476 compensatory drift, where one duplicate decreases in expression while the other increases to maintain
477 functionality of the progenitor gene; and these processes can occur gradually or bypass certain stages
478 (Birchler et al. 2005; Qian et al. 2010; Thompson et al. 2016; Veitia 2017; Birchler and Yang 2022;
479 Iohannes and Jackson 2023). These transcriptional dynamics correspond with the widespread expression
480 divergence we observed among four-gene sets derived (**Fig. 4 and supplementary fig. S5**), suggesting
481 that these genes are likely in the midst of dosage-balance processes and may ultimately evolve toward
482 subfunctionalization, neofunctionalization, or loss.
483 Our analyses of single-cell genomic data also reflect an important additional factor to consider in
484 gene retention - the influence of cell-type specificity. Duplicated genes may be in dosage balance or
485 compensatory drift in certain cell types, but not in other cell types (Birchler and Yang 2022; Iohannes and
486 Jackson 2023). This is reflected in our observation that different combinations of expression patterns
487 occur across tissues within the gene pairs (**Fig. 2c**) or gene sets (**Fig. 4d**). Another example is that the
488 gene set exhibiting consistent phloem-specific expression in the globular stage seed tissue but different
489 cell-type-specific expression in early-stage nodule tissue (**supplementary fig. S12**). Such variability
490 suggests that gene retention may be driven by expression in specific tissues while being dispensable in
491 others. This highlights the importance of understanding the context dependence of this phenomenon,
492 particularly through the lens of single-cell genomics.
493 In summary, understanding the evolution of paralogs is crucial, as many crops - including Solanum,
494 grasses, and legumes - have similarly complex gene families that undoubtedly influence agriculturally
495 important traits. Gaining insights into the retention of gene duplications and their underlying
496 mechanisms can inform strategies for fine-tuning gene dosage, facilitating the engineering of
497 quantitative variation, and overcoming redundancy barriers in functional studies.
498

499 # Materials and Methods

500 ## Duplicated gene pairs and gene sets in cultivated soybean

501 Duplicated gene pairs in the soybean genome (Williams 82), were categorized based on their
502 duplication mechanisms - whole-genome duplication, tandem duplication, proximal duplication,
503 transposed duplication and dispersed duplication (**Fig. 1**) - and obtained from the Plant Duplication Gene
504 Database (http://pdgd.njau.edu.cn:8080) (Qiao et al. 2019). Gene pairs located in scaffolds were
505 excluded from further analyses. To identify gene sets containing four whole-genome duplication derived
506 genes, which share a common ancestral gene across all bean-like (papilionoid) legume species, we
507 overlapped the legume gene family dataset (https://www.soybase.org/) with duplicated genes
508 originating from whole-genome duplication. Specifically, we first filtered the legume gene families to
509 retain only those containing four duplicated genes in soybean, and then further refined the selection by
510 ensuring that these four genes were derived from whole-genome duplication based on the gene pair
511 information obtained earlier.

512

513 ## Expression relationships of gene pairs or gene sets

514 Raw gene counts from seven tissues - root, hypocotyl, nodule, globular stage seed, heart stage
515 seed, cotyledon stage seed and early maturation stage seed - were obtained from snRNA-seq (Zhang et
516 al. 2024). Data from each tissue was analyzed separately. Only genes with two replicates per cell type
517 within the same tissue were included in subsequent analyses. For each tissue, we calculated the
518 Spearman correlation between cell type replicates and excluded cell types with low correlation (0.75 or
519 blow). Lowly expressed genes were filtered using “filterByExpr” function, and the remaining gene counts
520 were normalized using “cpm” function in edgeR (Chen et al. 2025). Genes with zero expression across all
521 cell types within a tissue were classified as “non-expressed” and excluded from further analysis. The

522 remaining genes were used to construct coexpression networks (Benoit et al. 2025) to assess whether
523 duplicated gene pairs exhibit similar expression profiles across cell types within a tissue.
524 To quantify expression relationships between duplicated genes within the same pair, we calculated
525 the absolute mean and standard deviation (SD) of log2-transformed fold change across cell types within
526 the same tissue (Benoit et al. 2025). Following (Benoit et al. 2025), we classified duplicated gene pairs
527 into six categories based on coexpression and variation in expression levels, introducing "mono-
528 expression" as an additional category:
529 1. Dosage-balanced: coexpression > 0.9, mean $\log_2$ fold change <1, SD $\log_2$ fold change <1
530 2. Paralogue dominance: coexpression > 0.9, mean $\log_2$ fold change >= 1, SD $\log_2$ fold change <1
531 3. Specialized: coexpression > 0.9, mean $\log_2$ fold change >= 1, SD $\log_2$ fold change >=1
532 4. Diverged: coexpression < 0.5, mean $\log_2$ fold change >= 1, SD $\log_2$ fold change >=1
533 5. Mono-expression: coexpression = NA.
534 6. Others: expressed gene pairs that did not fall into any of the above categories.
535 To investigate expression relationships among gene sets, we summarized the metric of
536 coexpression for all six gene pairs within each gene set. Gene sets were first grouped based on the
537 number of genes expressed. For sets with at least two expressed genes, we further classified them as
538 either correlated or uncorrelated based on their average pairwise coexpression values. Gene sets with a
539 mean coexpression value $>$ 0.9 were designated as correlated, while those below this threshold were
540 categorized as uncorrelated.
541
542 Identification of cell-type-specific genes in each tissue in cultivated soybean
543 To identify cell-type-specific genes, we performed differential expression analysis for each cell type
544 within its respective tissue. Gene counts were aggregated across nuclei of the same cell type and
545 compared to a reference panel, which uniformly sampled gene counts from other cell types within the

546 same tissue, using edgeR (Chen et al. 2025). Genes with a $\log_2$ fold change > 1 and an adjusted p-value <
547 0.05 were classified as highly expressed in a given cell type. A gene was defined as cell-type-specific if it
548 was classified as highly expressed in at least one cell type within its tissue.

549 To investigate the evolution of cell-type-specific expression among gene sets, we first filtered for
550 gene sets in which four duplicated genes were expressed and at least one duplicate exhibited cell-type-
551 specific expression in the relevant tissue. For each gene, we calculated the proportion of expression
552 attributable to each cell type by dividing the expression level of that cell type by the total expression
553 level within the tissue. The resulting matrix of cell-type contribution proportions was used for clustering
554 analysis. Pairwise distances between genes were computed as 1 minus the Spearman correlation
555 coefficient, and agglomerative hierarchical clustering was performed using the average linkage method
556 in R. The optimal number of clusters was determined using the gap statistic method with 50
557 permutations and a maximum of 60 clusters using cluster package. Genes were then assigned to clusters
558 by cutting the dendrogram at the selected value of k. The number of clusters observed within each gene
559 set reflects the conservation of cell-type-specific expression among the duplicated genes: a single shared
560 cluster indicates conserved expression across all four genes (e.g. **Fig. 6a**), whereas two or more clusters
561 indicate divergence in cell-type-specific expression among the duplicates (e.g. **Fig. 6b - 6c**).

562

563 ### ACR comparisons between the two duplicated genes within each pair

564 ACRs identified in cultivated soybean was obtained from (Zhang et al. 2024). The gene closest to
565 each ACR was defined as the gene-associated ACR, and this association was determined using bedtools
566 (Quinlan and Hall 2010). BLAST analysis was performed to examine the sequence divergence of ACRs
567 associated with the two genes within the same duplication gene pair. The sequence of the ACR
568 associated with one duplicated gene was used as the query, while the sequence of the two adjacent
569 genes from the other duplicated gene served as the reference (**Fig. 3b**). These sequences were isolated

using the “getfasta” function in bedtools (Quinlan and Hall 2010). Then sequence comparisons were performed using BLASTn with the following parameters “-task blastn-short -evalue 1e-3 -max_target_seqs 4-word_size 7 -gapopen 5 -gapextend 2 -penalty -1 -reward 1 -dust no -outfmt 6” following (Mendieta et al. 2024). To align ACR sequence comparisons with the expression relationships of duplicated genes in each tissue, we identified ACRs associated with each gene in each tissue separately and conducted BLAST analyses independently.

ACRs were classified as conserved if more than 10% (approximately 50 bp) were BLASTed to the reference region (**Fig. 3c**). The mean percentage of conserved ACRs were calculated for each gene pair by dividing the number of conserved ACRs by the total number of ACRs associated with that pair and then averaged across gene pairs with the same expression pattern (**Fig. 3d**). The BLASTed ratio was calculated by dividing the length of the BLASTed sequence in the reference region by the length of the query ACR. The mismatch rate was calculated as the number of mismatches between the ACR and the BLASTed sequence, divided by the length of the BLASTed sequence in the reference region. The percentage of unBLASTed ACRs was calculated by dividing the number of ACRs could not be BLASTed to the reference region by the total number of ACRs associated with the gene pair. The average BLASTed ratio, mismatch rate, and percentage of unBLASTed ACRs across all the ACRs associated with the gene pair were calculated to represent the level of sequence divergence of ACRs for each gene pair.

When measuring ACR sequence divergence across duplicated genes with the gene sets, we classified the ACRs based on conservation score, defined as the number of times an ACR was conserved (with a BLASTed ratio > 0.1) across the three BLASTs. Based on this metric, ACRs were categorized as: unique ACRs, two-copy conserved ACRs and multi-copy conserved ACRs (**Fig. 5a**). We examined the compositions of these ACRs for ACRs associated with each gene set. For each gene set, we calculated the average BLASTed ratio and mismatch rate for two-copy conserved ACRs, and multiple-copy conserved

ACRs associated with each gene set. When comparing the ACR explaining cell-type-specific expression, we specifically focused on cell-type-specific ACRs identified in (Zhang et al. 2024).

During the preparation of this manuscript, we used ChatGPT to refine code and improve readability and language. All content was subsequently reviewed and edited by the authors, who take full responsibility for the final version.

Acknowledgement

We sincerely appreciate the insightful feedback from Dr. Jeffrey Doyle from Cornell University, Dr. Scott Jackson from University of Georgia, Dr. Steven Cannon from USDA-ARS, and the members of the Schmitz lab on this manuscript, particularly Cullan Meyer, Kevin Sun, Dr. Hao Zhang, Dr. Shan Liang and Dr. Mark Minow. This research was supported by the United Soybean Board (2432-201-0102) and the National Science Foundation (IOS-1856627) to R.J.S.

Reference

Arsovski AA, Pradinuk J, Guo XQ, Wang S, Adams KL. 2015. Evolution of Cis-Regulatory Elements and Regulatory Networks in Duplicated Genes of Arabidopsis. *Plant Physiol.* 169:2982–2991.

Benoit M, Jenike KM, Satterlee JW, Ramakrishnan S, Gentile I, Hendelman A, Passalacqua MJ, Suresh H, Shohat H, Robitaille GM, et al. 2025. Solanum pan-genetics reveals paralogues as contingencies in crop engineering. *Nature*:1–11.

Birchler JA, Riddle NC, Auger DL, Veitia RA. 2005. Dosage balance in gene regulation: biological implications. *Trends Genet.* 21:219–226.

Birchler JA, Yang H. 2022. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* 34:2466–2474.

618 Bird KA, Niederhuth CE, Ou S, Gehan M, Pires JC, Xiong Z, VanBuren R, Edger PP. 2020. Replaying
619 the evolutionary tape to investigate subgenome dominance in allopolyploid Brassica napus.
620 *New Phytol.*

621 Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age
622 distributions of duplicate genes. *Plant Cell* 16:1667–1678.

623 Cai H, Des Marais DL. 2024. Expression variability following gene duplication facilitates gene
624 retention. *bioRxiv* [Internet]:2024.11.06.622370. Available from:
625 https://www.biorxiv.org/content/10.1101/2024.11.06.622370v2.abstract

626 Casneuf T, De Bodt S, Raes J, Maere S, Van de Peer Y. 2006. Nonrandom divergence of gene
627 expression following gene and genome duplications in the flowering plant Arabidopsis
628 thaliana. *Genome Biol.* 7:R13.

629 Chaudhary B, Flagel L, Stupar RM, Udall JA, Verma N, Springer NM, Wendel JF. 2009. Reciprocal
630 silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton
631 (gossypium). *Genetics* 182:503–517.

632 Chen Y, Chen L, Lun ATL, Baldoni PL, Smyth GK. 2025. edgeR v4: powerful differential analysis of
633 sequencing data with expanded functionality and improved support for small counts and
634 larger datasets. *Nucleic Acids Res.* 53:gkaf018.

635 Coate JE, Farmer AD, Schiefelbein JW, Doyle JJ. 2020. Expression partitioning of duplicate genes at
636 single cell resolution in Arabidopsis roots. *Front. Genet.* 11:596150.

637 Defoort J, Van de Peer Y, Carretero-Paulet L. 2019. The evolution of gene duplicates in angiosperms
638 and the impact of protein-protein interactions and the mechanism of duplication. *Genome*
639 *Biol. Evol.* 11:2292–2305.

640 Du J, Tian Z, Sui Y, Zhao M, Song Q, Cannon SB, Cregan P, Ma J. 2012. Pericentromeric effects shape
641 the patterns of divergence, retention, and expression of duplicated genes in the
642 paleopolyploid soybean. *Plant Cell* 24:21–32.

643 Fang C, Jiang N, Teresi SJ, Platts AE, Agarwal G, Niederhuth C, Edger PP, Jiang J. 2024. Dynamics of
644 accessible chromatin regions and subgenome dominance in octoploid strawberry. *Nat.*
645 *Commun.* 15:1–14.

646 Fang C, Yang M, Tang Y, Zhang L, Zhao H, Ni H, Chen Q, Meng F, Jiang J. 2023. Dynamics of *cis*-
647 regulatory sequences and transcriptional divergence of duplicated genes in soybean. *Proc.*
648 *Natl. Acad. Sci. U. S. A.* 120:e2303836120.

649 Freeling M. 2009. Bias in plant gene content following different sorts of duplication: tandem, whole-
650 genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* 60:433–453.

651 Freeling M, Thomas BC. 2006. Gene-balanced duplications, like tetraploidy, provide predictable
652 drive to increase morphological complexity. *Genome Res.* 16:805–814.

653 Ganko EW, Meyers BC, Vision TJ. 2007. Divergence in expression between duplicated genes in
654 Arabidopsis. *Mol. Biol. Evol.* 24:2298–2309.

655 Gout J-F, Hao Y, Johri P, Arnaiz O, Doak TG, Bhullar S, Couloux A, Guérin F, Malinsky S, Potekhin A, et
656 al. 2023. Dynamics of gene loss following ancient whole-genome duplication in the cryptic
657 Paramecium complex. *Mol. Biol. Evol.* 40:msad107.

658 Haberer G, Hindemitt T, Meyers BC, Mayer KFX. 2004. Transcriptional Similarities, Dissimilarities,
659 and Conservation of cis-Elements in Duplicated Genes of Arabidopsis. *Plant Physiol.*
660 136:3009–3022.

661 Han J, Lopez-Arredondo D, Yu G, Wang Y, Wang B, Wall SB, Zhang X, Fang H, Barragán-Rosillo AC,
662 Pan X, et al. 2022. Genome-wide chromatin accessibility analysis unveils open chromatin
663 convergent evolution during polyploidization in cotton. *Proc. Natl. Acad. Sci. U. S. A.*
664 119:e2209743119.

665 Hu G, Grover CE, Vera DL, Lung P-Y, Girimurugan SB, Miller ER, Conover JL, Ou S, Xiong X, Zhu D, et
666 al. 2024. Evolutionary Dynamics of Chromatin Structure and Duplicate Gene Expression in
667 Diploid and Allopolyploid Cotton. *Mol. Biol. Evol.* [Internet] 41. Available from:
668 http://dx.doi.org/10.1093/molbev/msae095

669 Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing
670 between models. *Nat. Rev. Genet.* 11:97–108.

671 Iohannes SD, Jackson D. 2023. Tackling redundancy: genetic mechanisms underlying paralog
672 compensation in plants. *New Phytol.* 240:1381–1389.

673 Katikaneni A, Lowe CB. 2025. Novelty versus innovation of gene regulatory elements in human
674 evolution and disease. *Curr. Opin. Genet. Dev.* 90:102279.

675 Koenen EJM, Ojeda DI, Bakker FT, Wieringa JJ, Kidner C, Hardy OJ, Pennington RT, Herendeen PS,
676 Bruneau A, Hughes CE. 2021. The origin of the legumes is a complex paleopolyploid
677 phylogenomic tangle closely associated with the Cretaceous-Paleogene (K-Pg) mass
678 extinction event. *Syst. Biol.* 70:508–526.

679 Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene
680 duplications. *Genome Biol.* 3:RESEARCH0008.

681 Kuzmin E, Taylor JS, Boone C. 2022. Retention of duplicated genes in evolution. *Trends Genet.*
682 38:59–72.

683 Li X, Schmitz RJ. 2025. Cis-regulatory dynamics in plant domestication. *Trends Genet.* [Internet] 0.
684 Available from: https://www.cell.com/action/showAbstract?pii=S0168952525000460

685 Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science*
686 290:1151–1155.

687 Marand AP, Eveland AL, Kaufmann K, Springer NM. 2023. cis-Regulatory Elements in Plant
688 Development, Adaptation, and Evolution. *Annu. Rev. Plant Biol.* 74:111–137.

689 McDonald JMC, Reed RD. 2024. Beyond modular enhancers: new questions in cis-regulatory
690 evolution. *Trends Ecol. Evol.* [Internet] 0. Available from:
691 http://www.cell.com/article/S0169534724001708/abstract

692 Mendieta JP, Tu X, Jiang D, Yan H, Zhang X, Marand AP, Zhong S, Schmitz RJ. 2024. Investigating the
693 cis-regulatory basis of C3 and C4 photosynthesis in grasses at single-cell resolution. *Proc.*
694 *Natl. Acad. Sci. U. S. A.* 121:e2402781121.

695 Ohno S. 1970. Evolution by gene duplication. 1970th ed. Berlin, Germany: Springer Available from:
696 https://link.springer.com/book/10.1007/978-3-642-86659-3

697 Panchy N, Lehti-Shiu MD, Shiu S-H. 2016. Evolution of gene duplication in Plants1[OPEN]. *Plant*
698 *Physiol.* 171:2294–2316.

699 Qian W, Liao B-Y, Chang AY-F, Zhang J. 2010. Maintenance of duplicate genes and their functional
700 redundancy by reduced expression. *Trends Genet.* 26:425–430.

701 Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution
702 in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* 20:38.

703 Qiu Y, Van Tay Y, Ruan Y, Adams KL. 2020. Divergence of duplicated genes by repeated partitioning
704 of splice forms and subcellular localization. *New Phytol.* 225:1011–1022.

705 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
706 *Bioinformatics* 26:841–842.

707 Roulin A, Auer PL, Libault M, Schlueter J, Farmer A, May G, Stacey G, Doerge RW, Jackson SA. 2013.
708 The fate of duplicated genes in a polyploid plant genome. *Plant J.* 73:143–153.

709 Salman-Minkov A, Sabath N, Mayrose I. 2016. Whole-genome duplication as a key factor in crop
710 domestication. *Nat Plants* 2:16115.

711 Schmitz RJ, Grotewold E, Stam M. 2022. Cis-regulatory sequences in plants: Their importance,
712 discovery, and future challenges. *Plant Cell* 34:718–741.

713 Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J,
714 et al. 2010. Genome sequence of the palaeopolyploid soybean. *Nature* 465:120–120.

715 Schnable JC, Pedersen BS, Subramaniam S, Freeling M. 2011. Dose-sensitivity, conserved non-
716 coding sequences, and duplicate gene retention through multiple tetraploidies in the
717 grasses. *Front. Plant Sci.* 2:2.

718 Shi T, Rahmani RS, Gugger PF, Wang M, Li H, Zhang Y, Li Z, Wang Q, Van de Peer Y, Marchal K, et al.
719 2020. Distinct Expression and Methylation Patterns for Genes with Different Fates following
720 a Single Whole-Genome Duplication in Flowering Plants. *Mol. Biol. Evol.* 37:2394–2413.

721 Shoemaker RC, Schlueter J, Doyle JJ. 2006. Paleopolyploidy and gene duplication in soybean and
722 other legumes. *Curr. Opin. Plant Biol.* 9:104–109.

723 Thompson A, Zakon HH, Kirkpatrick M. 2016. Compensatory drift and the evolutionary dynamics of
724 dosage-sensitive duplicate genes. *Genetics* 202:765–774.

725 Tiley GP, Ané C, Burleigh JG. 2016. Evaluating and characterizing ancient whole-genome
726 duplications in plants with gene count data. *Genome Biol. Evol.* 8:1023–1037.

727 Veitia RA. 2017. Gene duplicates: Agents of robustness or fragility? *Trends Genet.* 33:377–379.

728 Veitia RA, Bottani S, Birchler JA. 2008. Cellular reactions to gene dosage imbalance: genomic,
729 transcriptomic and proteomic effects. *Trends Genet.* 24:390–397.

730 Veitia RA, Bottani S, Birchler JA. 2013. Gene dosage effects: nonlinearities, genetic interactions,
731 and dosage compensation. *Trends Genet.* 29:385–393.

732 Wang L, Jia G, Jiang X, Cao S, Chen Z, Song Q. 2021. Altered chromatin architecture and gene
733 expression during polyploidization and domestication of soybean. *Plant Cell* [Internet].
734 Available from: https://academic.oup.com/plcell/article/33/5/1430/6175070

735 Wendel JF. 2000. Genome evolution in polyploids. *Plant Mol. Biol.* 42:225–249.

736 Wittkopp PJ, Kalay G. 2012. Cis-regulatory elements: molecular mechanisms and evolutionary
737 processes underlying divergence. *Nature Reviews Genetics* 13:59–69.

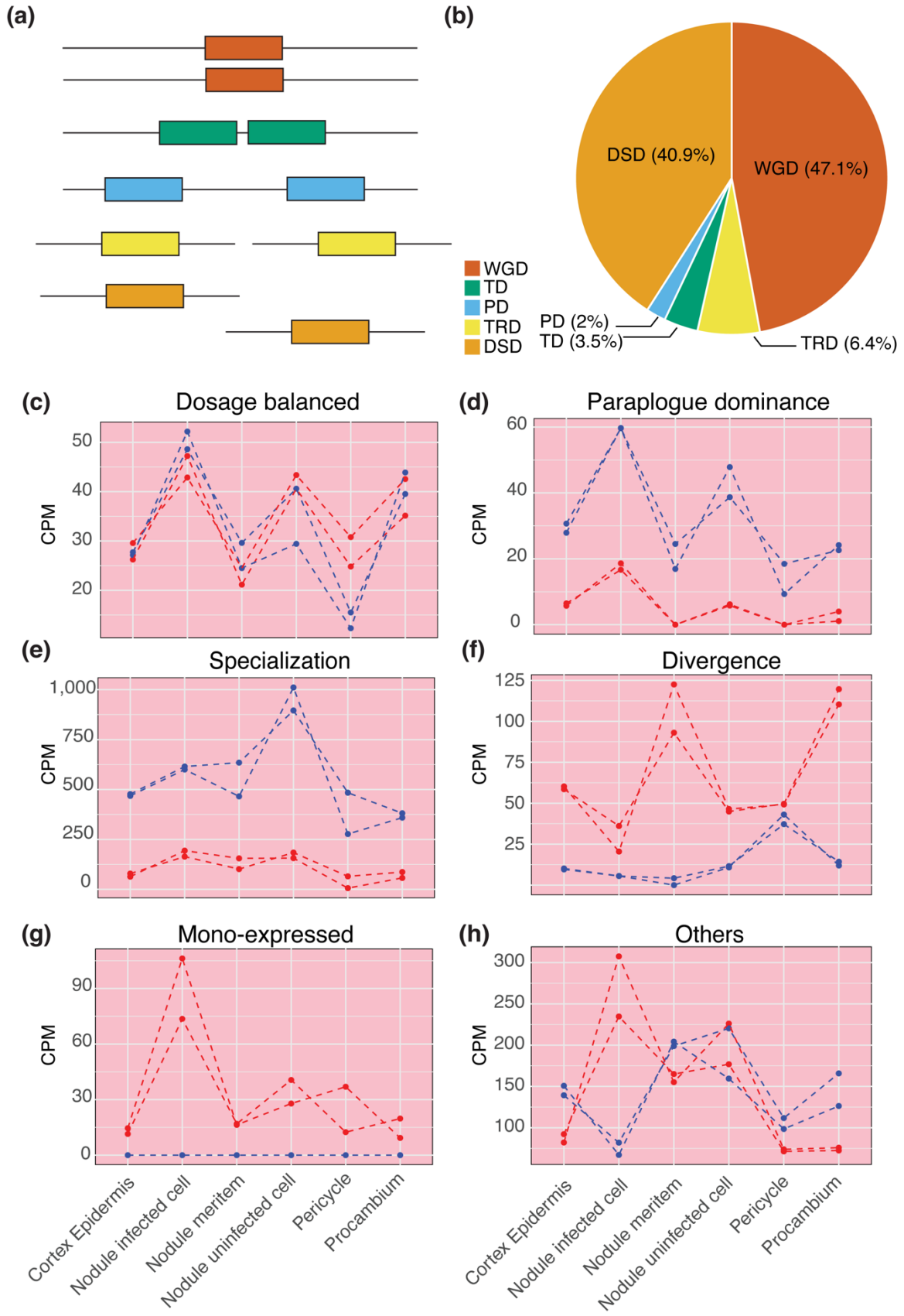
738 Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. 2003. The evolution
739 of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* 20:1377–1419.

740 Zhang J. 2003. Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18:292–298.

741 Zhang X, Luo Z, Marand AP, Yan H, Jang H, Bang S, Mendieta JP, Minow MAA, Schmitz RJ. 2024. A
742 spatially resolved multi-omic single-cell atlas of soybean development. *Cell* 188:550-
743 567.e19.

744 Zhao M, Zhang B, Lisch D, Ma J. 2017. Patterns and Consequences of Subgenome Differentiation
745 Provide Insights into the Nature of Paleopolyploidy in Plants. *Plant Cell* 29:2974–2994.

746 Zhuang Y, Wang X, Li X, Hu J, Fan L, Landis JB, Cannon SB, Grimwood J, Schmutz J, Jackson SA, et al.
747 2022. Phylogenomics of the genus Glycine sheds light on polyploid evolution and life-
748 strategy transition. *Nat. Plants* 8:233–244.


(a)
(b)
WGD
TD
PD
TRD
DSD
DSD (40.9%)
WGD (47.1%)
PD (2%)
TD (3.5%)
TRD (6.4%)
(c)
Dosage balanced
(d)
Paraplogue dominance
(e)
Specialization
(f)
Divergence
(g)
Mono-expressed
(h)
Others
CPM
Cortex Epidermis
Nodule infected cell
Nodule meritem
Nodule uninfected cell
Pericycle
Procambium

749

750 **Fig. 1. Different duplication mechanisms and distinct expression patterns.**

751 (a) Illustration of different gene duplication mechanisms, including whole-genome duplication (WGD),
752 dispersed duplication (DSD), proximal duplication (PD), tandem duplication (TD), and transposed
753 duplication (TRD); (b) Proportional contributions of each duplication mechanism to the duplicated gene
754 pairs identified in soybean; (c-h) Representative examples of duplicated gene pairs exhibiting distinct
755 expression patterns across cell types within early nodule tissue: (c) dosage balanced, (d) paralogue
756 dominance, (e) specialization, (f) divergence, (g) mono-expressed and (h) others. In each panel, blue and
757 red dots/lines represent the two genes within a duplicated pair, with two biological replicates included
758 for each gene.

759

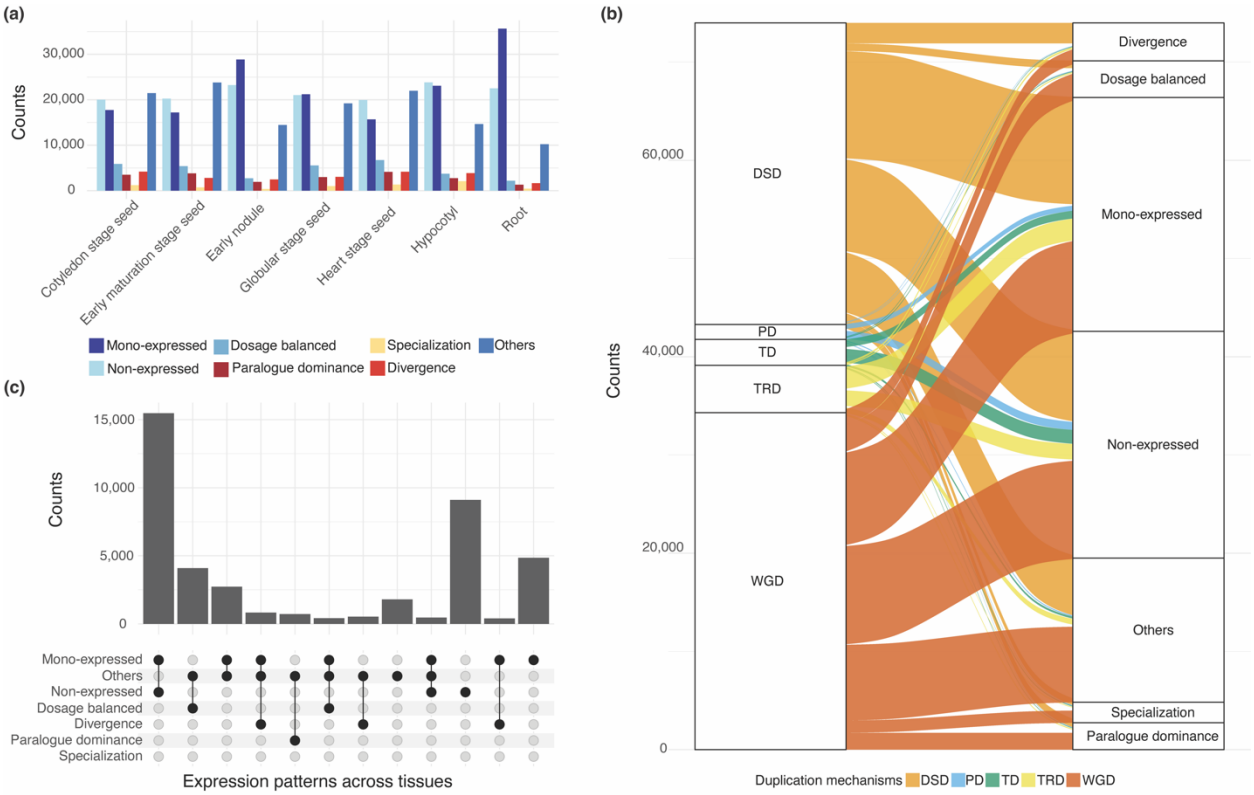
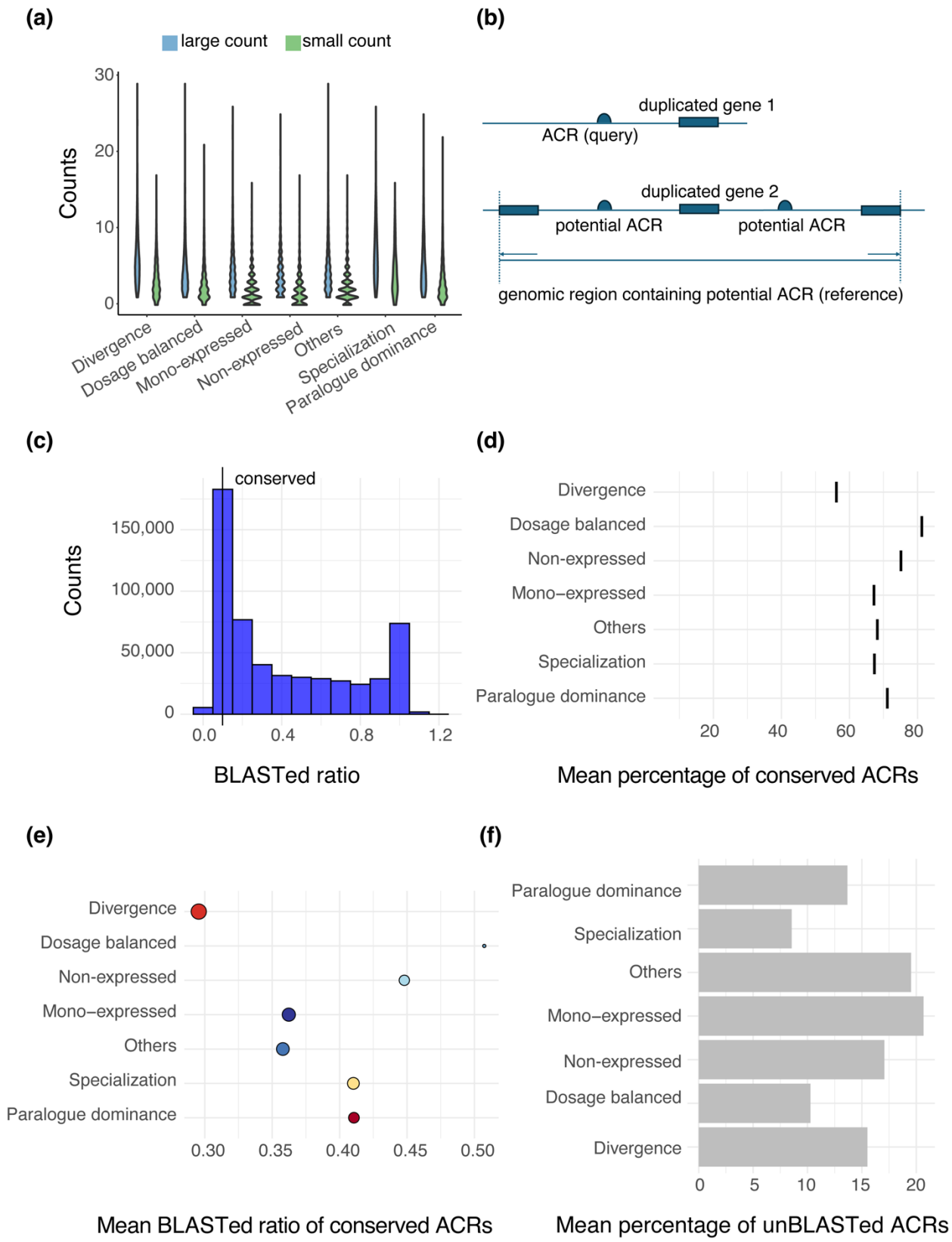


Fig. 2. Distributions of expression patterns within and across tissues of duplicated gene pairs derived from distinct duplication mechanisms.

(a) Number of duplicated gene pairs classified by distinct expression patterns within a tissue, including mono-expressed, dosage balanced, paralogue dominance, specialization, divergence and others (as described in Fig.1), as well as non-expressed category. (b) Relationship between gene duplication mechanisms and expression patterns. The left panel categorizes genes based on duplication mechanisms: whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), dispersed duplication (DSD), and transposed duplication (TRD). The right panel represents gene expression patterns. Flow width indicates the relative frequency of each transition, highlighting how different duplication mechanisms contribute to expression variation. (c) Number of duplicated gene pairs classified by a combination of expression patterns across seven examined tissues.


(a)
large count
small count
Counts
30
20
10
0
Divergence
Dosage balanced
Mono-expressed
Non-expressed
Others
Specialization
Paralogue dominance
(b)
duplicated gene 1
ACR (query)
duplicated gene 2
potential ACR
potential ACR
genomic region containing potential ACR (reference)
(c)
conserved
Counts
150,000
100,000
50,000
0
0.0
0.4
0.8
1.2
BLASTed ratio
(d)
Divergence
Dosage balanced
Non-expressed
Mono−expressed
Others
Specialization
Paralogue dominance
20
40
60
80
Mean percentage of conserved ACRs
(e)
Divergence
Dosage balanced
Non-expressed
Mono−expressed
Others
Specialization
Paralogue dominance
0.30
0.35
0.40
0.45
0.50
Mean BLASTed ratio of conserved ACRs
(f)
Paralogue dominance
Specialization
Others
Mono−expressed
Non-expressed
Dosage balanced
Divergence
0
5
10
15
20
Mean percentage of unBLASTed ACRs

774

775

776 **Fig. 3. Genetic variation within ACRs associated with gene pairs with different expression patterns.**
777 (a) Number of ACRs associated with each duplicated gene within gene pairs exhibiting different
778 expression patterns in the cotyledon stage seed. Within each gene pair, the gene associated with a
779 higher number of ACRs is classified into the large count group, whereas the one with fewer ACRs is
780 classified into the small count group. (b) Illustration of the query and reference region for a BLAST: an
781 ACR sequence associated with duplicated gene1 as a query, whereas the genomic region spanning from
782 the upstream to the downstream gene of duplicated gene2 as a reference region. (c) Number of ACRs
783 BLASTed to the reference region in the cotyledon stage seed. ACRs with a BLASTed ratio greater than 0.1
784 are classified as conserved. (d) Mean percentage of conserved ACRs (black bars) across gene pairs with
785 different expression patterns in the cotyledon stage seed. (e) Mean BLASTed ratio of conserved ACRs
786 across gene pairs with different expression patterns in the cotyledon stage seed. The size of dots
787 represents the mean mismatch rate of conserved ACRs across gene pairs with different expression
788 patterns in the cotyledon stage seed. The color of dots represents different expression patterns. (f) Mean
789 percentage of ACRs without a corresponding sequence in the reference region (unBLASTed ACRs) across
790 gene pairs with different expression patterns in the cotyledon stage seed.

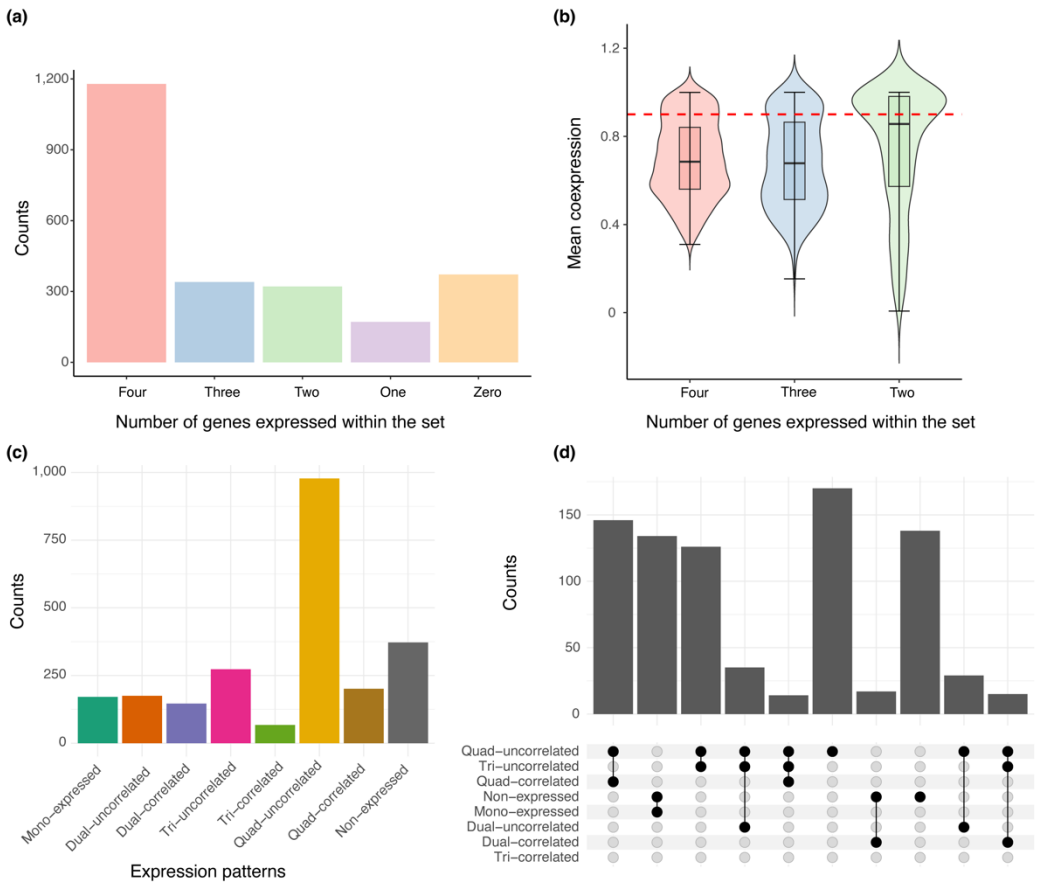


Fig. 4. Expression divergence of duplicated genes within the sets.

(a) Number of duplicated gene sets classified by the number of genes expressed with each set in cotyledon stage seed. (b) Mean coexpression among expressed genes within gene sets, grouped by the number of genes expressed per set in cotyledon stage seed. (c) Number of duplicated gene sets classified by distinct expression patterns in cotyledon stage seed, including mono-expressed, dual-uncorrelated, dual-correlated, tri-uncorrelated, tri-correlated, quad-uncorrelated, quad-correlated and non-expressed. (d) Number of duplicated gene sets classified by a combination of expression patterns across seven examined tissues.

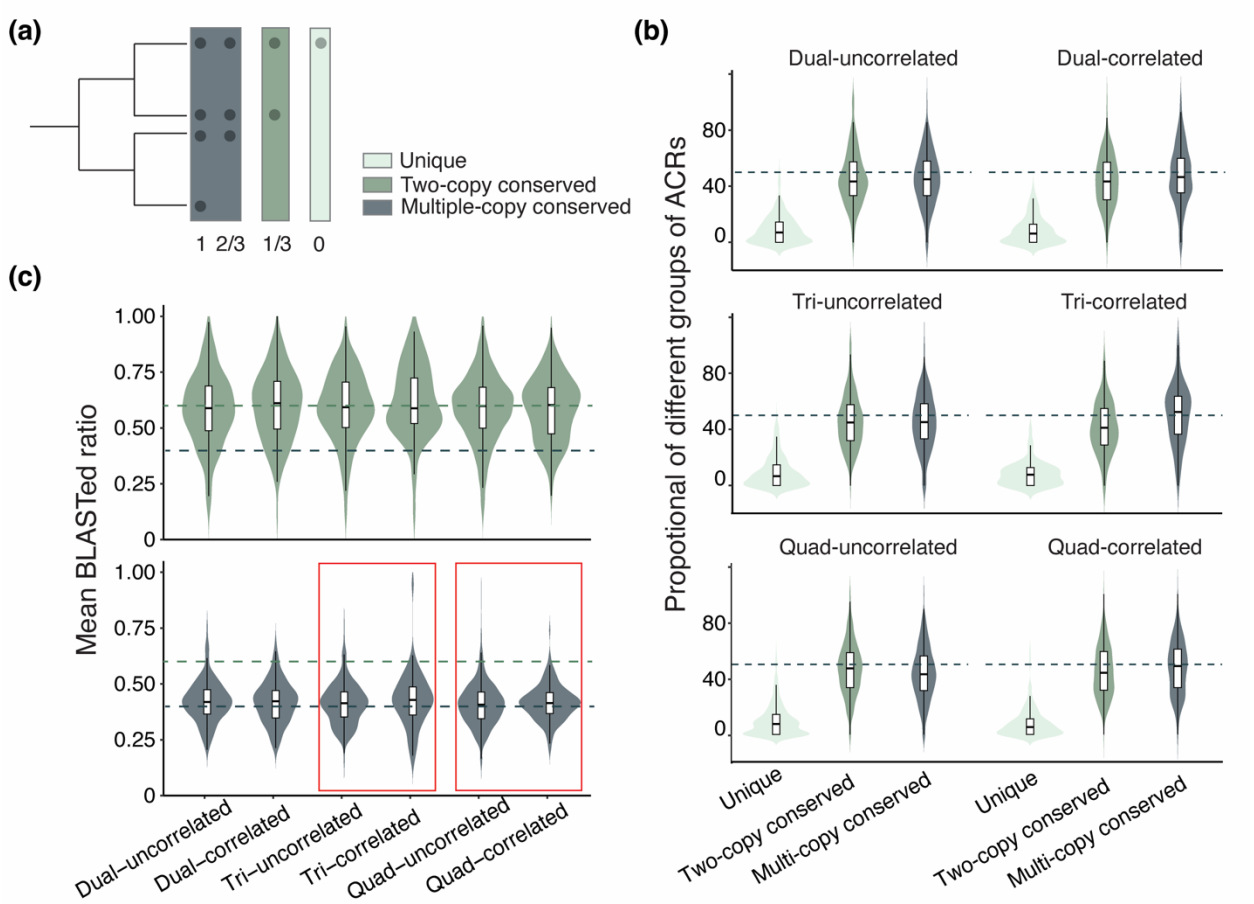


Fig. 5. Proportions and sequence divergence of ACRs with different levels of conservation across gene sets with distinct expression patterns.

(a) Classifications of ACRs based on their conservation score: unique ACR, two-copy conserved ACR and multi-copy conserved ACR. (b) Proportions of each ACR category identified in gene sets with different expression patterns in cotyledon stage seed. (c) Mean BLASTed ratio of two-copy conserved or multi-copy conserved ACRs across gene sets with different expression patterns in cotyledon stage seed. Comparisons where correlated groups had higher mean BLASTed ratios than uncorrelated ones are highlighted with red squares.

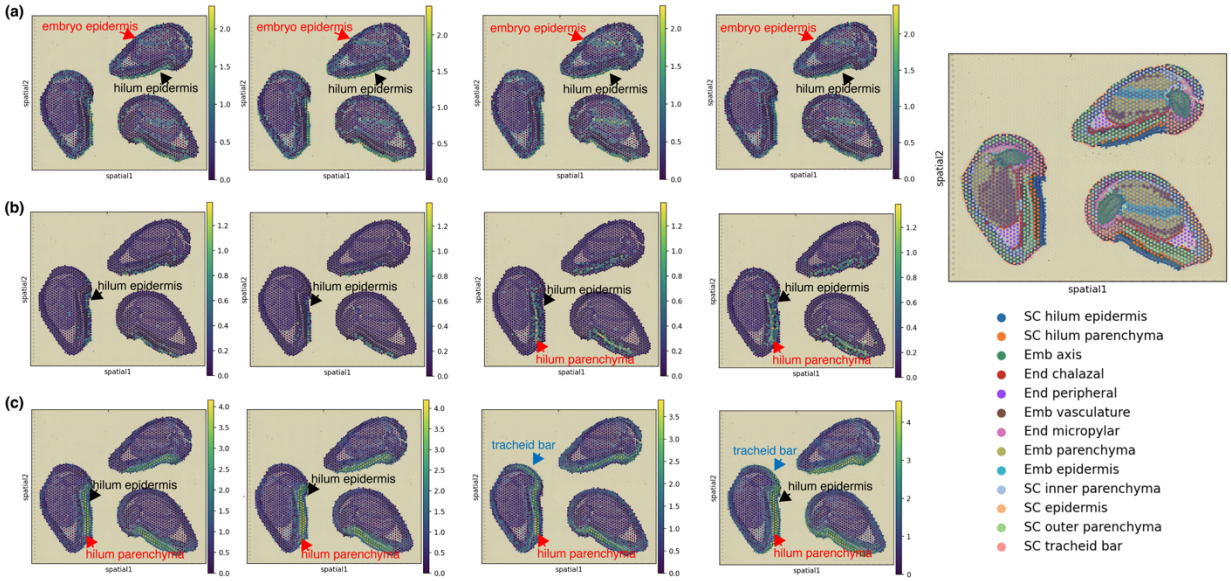
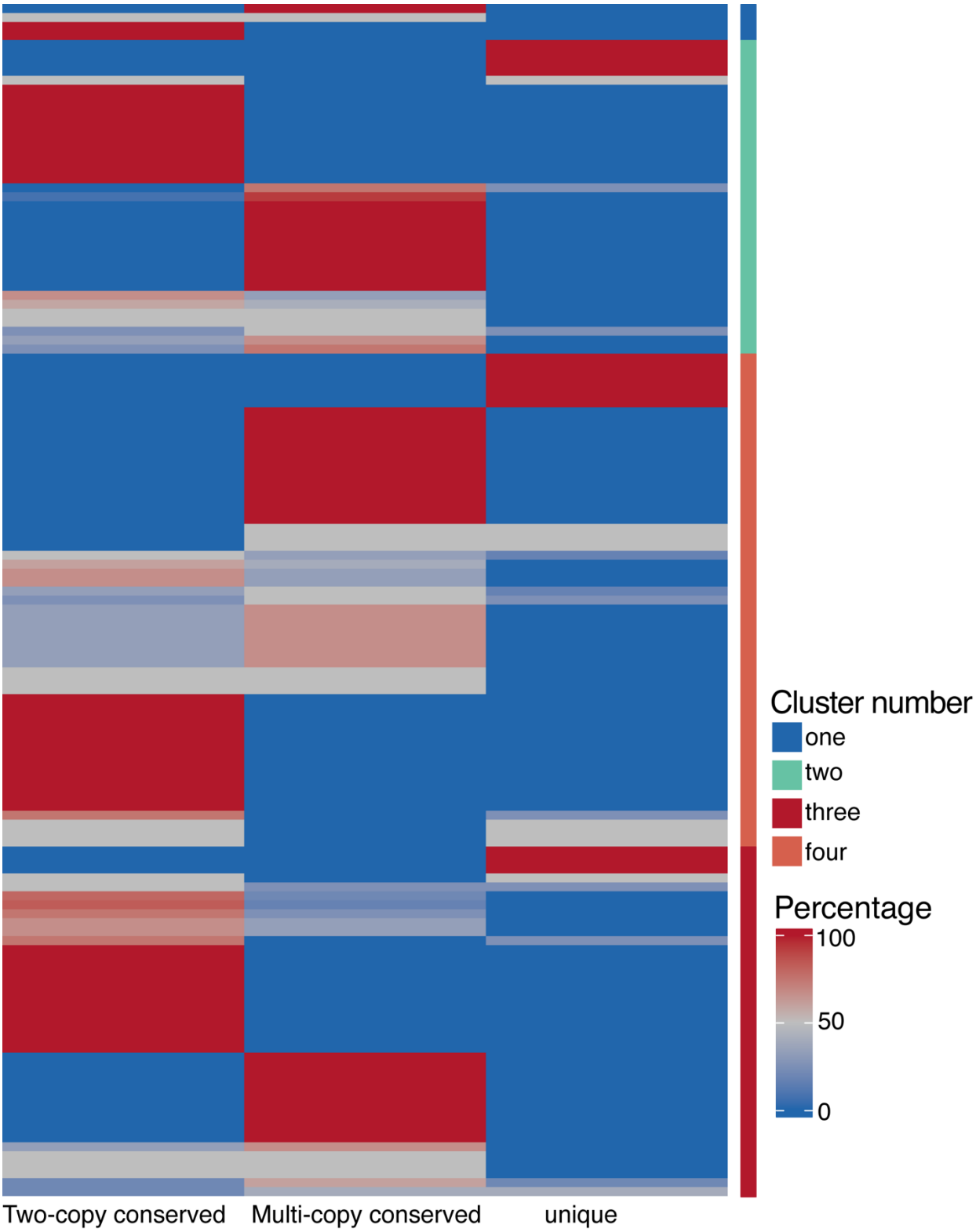


Fig. 6. The dynamics of cell-type-specific expression within gene sets in cotyledon stage seed. (a) Gene set with a single shared cluster: all four duplicated genes exhibited specific expression in the hilum epidermis and embryo epidermis. (b) Gene set with two shared clusters: the two genes on the left formed a cluster with expression specific to the hilum epidermis, while the two on the right formed a separate cluster with expression in both the hilum epidermis and hilum parenchyma. (c) Gene set with three shared clusters: the two genes on the left clustered together and were expressed in the hilum epidermis and hilum parenchyma; the third gene formed a second cluster and was expressed in the tracheid-bar and hilum parenchyma; the fourth gene formed a third cluster and was expressed in the tracheid-bar, hilum epidermis, and hilum parenchyma. Cell-type annotations for cotyledon stage seeds were shown on the right. Arrows indicated the cells where each gene was specifically expressed, with different arrow colors denoting distinct cell types in each panel. WGD: whole-genome duplication; SC: seed coat; Emb: embryo; End: endosperm.



821

822 **Fig. 7. Heatmap of proportions of cell-type-specific ACRs (ctACRs) with different levels of conservation**

823 **within each gene set with at least one gene exhibiting cell-type-specific expression in root tissue.**

824 Each row represents a gene set in which all four genes are expressed and at least one exhibits cell-type-
825 specific expression in the root. The three columns display the proportions of two-copy conserved,
826 multiple-copy conserved and unique ACRs associated with each gene set. Color intensity ranges from
827 blue (0%) to red (100%). The bar alongside the heatmap indicates the number of shared clusters
828 identified for each gene set.