Purifying selection provides buffering of the natural variation co-expression network in a forest tree species

Niklas Mähler[1, 2], Barbara K Terebieniec[2], Jing Wang[3], Pär K Ingvarsson[3], Nathaniel R Street[2¶*] and Torgeir R Hvidsten[1, 2¶]

[1]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Ås, Norway

[2]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, Umeå, Sweden

[3]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, Umeå, Sweden

[*] Corresponding author

E-mail: nathaniel.street@umu.se

[¶] These authors contributed equally to this work

Abstract

Several studies have investigated general properties of the genetic architecture of gene expression variation. Most of these used controlled crosses and it is unclear whether their findings extend to natural populations. Furthermore, systems biology has established that biological networks are buffered against large effect mutations, but there remains little data resolving this with natural variation of gene expression. Here we utilise RNA-Sequencing to assay gene expression in winter buds undergoing bud flush in a natural population of *Populus tremula*. We performed expression

Quantitative Trait Locus (eQTL) mapping and identified 164,290 significant eQTLs associating 6,241 unique genes (eGenes) with 147,419 unique SNPs (eSNPs). We found approximately four times as many local as distant eQTLs, which had significantly higher effect size. eQTLs were primarily located in regulatory regions of genes (UTRs or flanking regions), regardless of whether they were local or distant. We used the gene expression data to infer a co-expression network and investigated to what degree eQTLs could explain the structure of the network: eGenes were present in the core of 28 of 38 network modules, however, eGenes were overall underrepresented in cores and overrepresented in the periphery of the network, with a negative correlation between eQTL effect size and network connectivity. We also observed a negative correlation between eQTL effect size and allele frequency and found that core genes have experienced stronger selective constraint. Our integrated genetics and genomics results suggest that prevalent purifying selection is the primary mechanism underlying the genetic architecture of natural variation in gene expression in *P. tremula* and that highly connected network hubs are buffered against deleterious effects as a result of regulation by numerous eSNPs, each of minor effect.

Author summary

Numerous studies have shown that many genomic polymorphisms contributing to phenotypic variation are located outside of protein coding regions, suggesting that they act by modulating gene expression. Furthermore, phenotypes are seldom explained by individual genes, but rather emerge from networks of interacting genes. The effect of regulatory variants and the interaction of genes can be described by co-expression networks, which are known to contain a small number of highly connected nodes and many more lowly connected nodes, making them robust to random mutation. While previous studies have examined the genetic architecture of gene expression variation, few were performed in natural populations with fewer still integrating the co-expression network.

42 We undertook a study using a natural population of European aspen (*Populus tremula*), showing that
43 expression variance is substantially smaller among individuals than between tissues within the same
44 individual, suggesting that stabilizing selection may act to restrict the scale of expression variation.
45 We further show that highly connected genes within the co-expression network are associated with
46 polymorphisms of lower than average effect size, suggesting purifying selection. These genes are
47 therefore buffered against large expression modulation, providing a mechanistic explanation of how
48 network robustness is created and maintained at the population level.

49 Introduction

50 A central aim of biology is to understand how emergent phenotypes are encoded in the genome and
51 how genetic variation engenders phenotypic variation within populations. While the biological
52 function of many genes can be ascertained through use of mutant screens, such approaches tend to
53 identify genes essential to a biological process. There is no clear consensus as to whether variation in
54 these genes underlies natural variation in phenotypic traits, how this relates to the evolutionary
55 history of the traits or to the properties of the associated transcriptional networks. Additionally,
56 many traits are complex, being controlled by the action and interaction of numerous genes, often
57 rendering mutational approaches inadequate or infeasible. In such cases, turning to nature and
58 employing population genetics and systems biology approaches (systems genetics) to uncover genetic
59 and genomic co-variation between assayed traits (genomic markers, gene expression, protein
60 abundance, metabolites etc.) and 'classical' phenotypes (physiognomy, morphology, physiology)
61 represents an appealing alternative.

62 The availability of massively parallel sequencing technologies affords new possibilities for addressing
63 biological questions, for example enabling the generation of *de novo* genome assemblies and

64 population-wide resequencing data that can be used to perform genome-wide association studies
65 (GWAS), even in species with large genomes that harbour high levels of polymorphism or that
66 display rapid linkage disequilibrium (LD) decay [1]. The use of genome-wide resequencing data
67 allows the discovery of, in theory, all genetic polymorphisms within an individual. These genetic
68 markers, of which single nucleotide polymorphisms (SNPs) are currently the most commonly
69 considered, can then be used to perform association or linkage mapping to identify the subset of
70 polymorphisms engendering phenotypic variation among individuals. While the majority of SNPs are
71 expected to have no, or insignificant, consequence for a given trait, an implication of the infinitesimal
72 model [2] is the expectation that the control of variation for any given trait will be contributed to by
73 a large number of loci, each of small effect size. It is not clear how this holds across the scale of
74 molecular to complex, integrative phenotypes (*e.g.* from transcription to morphology) or whether this
75 pattern contrasts between adaptive and selectively neutral phenotypic trait variation.

76 Advances in sequencing technologies have concordantly revolutionised transcriptomics studies,
77 particularly in non-model organisms. Following seminal work [3,4], numerous early studies in a range
78 of species established that there is a significant heritable component underlying natural variation of
79 gene expression levels among individuals within populations [5–17] and that this variation underlies a
80 number of phenotypes [18–25]. Given these findings, it became apparent that gene expression values
81 could be considered in the same way as any other quantitative phenotype and be subjected to linkage
82 or association mapping to identify polymorphisms contributing to expression level variation among
83 individuals [26], as first reported in [20], with the identified loci termed expression Quantitative Trait
84 Locus (eQTL; [7]) or, less commonly, expression level polymorphisms (ELPs; [27]). eQTLs are
85 classified as either local or distant acting depending on the physical location of the associated
86 polymorphism in relation to the gene that the eQTL is mapped for: local eQTLs are usually defined
87 as being located within a specified physical distance of the gene location (typically up to 2 Mbp,

88 although this varies depending on species) while distant eQTLs represent polymorphisms that are
89 located beyond that threshold distance or on another chromosome. eQTLs can further be classified
90 as acting in *cis* or *trans*: *cis* eQTLs act in an allele-specific manner and are usually considered to be
91 local, although long-range *cis* interactions can occur, for example when a polymorphism is located in
92 an enhancer that is physically distant from the gene of interest; *trans* acting eQTLs affect both alleles
93 of a gene and are most commonly located distant to that gene. There continues to be strong interest
94 in eQTLs as they can identify mechanistic links between phenotype and genotype [28,29].
95 Importantly, the majority of polymorphisms that have been associated to phenotypes using GWAS
96 in a wide range of species are located outside of protein coding or transcribed regions [30–33],
97 suggesting that they influence expression rather than altering protein or transcript function.

98 A number of previous eQTL studies have been conducted using plant species including *Arabidopsis*
99 *thaliana* [34–39], *Zea mays* [40–42] and *Oryza sativa* [43,44], and in forest tree species [8,45–47]. These,
100 together with studies in other eukaryotic systems, have yielded generalities concerning the genetic
101 architecture of gene expression variation, including that a greater number of local eQTLs are
102 typically identified and that these individually explain a larger proportion of gene expression variance
103 than do distant eQTLs [35,48–51]. With the exception of human studies, the majority of previous
104 work was conducted using controlled, often inter-specific, crosses and it is not clear how generally
105 applicable their conclusions are for natural populations of unrelated individuals. Few studies have
106 considered whether observed, heritable variation is adaptive [52,53] or whether signatures of
107 population differentiation are observed at the transcriptome level [53–58], and there is a lack of
108 consensus as to whether natural variation in gene expression is selectively neutral [53].

109 Species in the *Populus* genus have been established as a powerful model system for forest tree
110 genomics due to their relatively small genome, rapid growth, propensity for clonal propagation and

111 ease of genetic transformation [59]. *P. tremula* (European aspen) has many features that render it a
112 particularly useful model for population genetics and speciation studies [60,61]. To facilitate use of *P.*
113 *tremula* as a model system we have produced a draft *de novo* genome assembly of the *P. tremula*
114 genome (available at [62]) and resequencing data [63] for all individuals comprising the Swedish
115 Aspen (SwAsp) collection [64].

116 Here, we utilise this resource in combination with population-wide RNA-Seq data, assaying gene
117 expression in winter buds at the point of bud break. We performed eQTL mapping and constructed
118 a co-expression network, with the results integrated to provide insight into the genetic architecture of
119 natural variation in gene expression levels. Our analyses revealed a scale free and modular co-
120 expression network, with highly connected genes being under-represented for eQTLs and with
121 eQTL effect size being negatively correlated with both gene connectivity and minor allele frequency.
122 Patterns of polymorphism and divergence within highly connected genes demonstrate that these
123 experience stronger selective constraint relative to genes in the network periphery. Our results
124 therefore highlight the importance of purifying selection for explaining observed standing variation
125 in gene expression in natural populations in addition to explaining how buffering of the
126 transcriptional network may be achieved. We propose that network modules can be explained by
127 eQTLs in one or a few highly connected module genes and that these are evolving under strong
128 functional constraint compared to genes in the network periphery.

129 Results

130 Population level gene expression similarity

131 We utilised the northern common garden (located at 63.9° N, near Umeå, Sweden) of the SwAsp
132 collection [64], which comprises 116 *P. tremula* genotypes sampled across the species distribution

133 range in Sweden (56.2° to 66.4° N, **Fig 1**A). The SwAsp collection was previously shown to contain
134 abundant genetic variation, with low linkage disequilibrium (LD) [61,65] and minimal population
135 structure [66]. Based on whole genome re-sequencing data that were aligned to a *de novo* assembly of
136 the *P. tremula* genome (available at PopGenIE.org; [62]), we called 4,509,654 SNPs after stringent
137 filtering, which we utilised to perform eQTL mapping.

138 To enable examination of the genetic architecture of natural variation in gene expression within the
139 SwAsp collection, we generated paired-end RNA-Seq expression data from winter buds at the point
140 of spring bud flush for 219 individuals (clonal replicates), representing 86 genotypes, and first
141 examined the distribution of broad-sense heritability (H^2) and population differentiation (Q_{ST}) (**Fig**
142 **1**B,C respectively) for all expressed, annotated genes. H^2 ranged from 0.0 to 1.0 with a mean (± s.d)
143 of 0.30 (0.22) and with 5,924 genes (17%) having $H^2 > 0.5$. There was a weak positive correlation
144 between H^2 and median expression (Pearson r = 0.09, $p < 2.2\times10^{-16}$), and a relatively strong positive
145 correlation to expression variance (Pearson r = 0.43, $p < 2.2\times10^{-16}$). Q_{ST} ranged from 0.0 to 1.0 with a
146 mean (± s.d) of 0.06 (0.12) andhad a weak negative correlation with expression variance (Pearson r =
147 -0.02, $p < 4.5\times10^{-4}$) and a positive correlation with median expression (Pearson r = 0.18, $p < 2.2\times10^{-}$
148 ^{16}). To further examine whether population structure was apparent on the basis of expression
149 variation among genotypes, we performed hierarchical clustering of all individuals (**Fig 1**D) or
150 genotypes (**S1 Fig**). We observed some evidence of clustering among genotypes but this could not
151 be explained by population structure.

152 We selected the 500 genes with the highest H^2 (0.88-1.0, 0.93 ± 0.03) and Q_{ST} (0.54-1.0, 0.71 ± 0.13)
153 and subjected these to Gene Ontology (GO) enrichment analysis to determine their biological
154 relevance (see **S1 File** for all results). Genes with high H^2 were enriched for categories including
155 protein phosphorylation (GO:0006468; $p = 3.7\times10^{-6}$), while high Q_{ST} genes were enriched in terms

156 including translation (GO:0006412; $p = 4.2\times10^{-20}$) and gene expression (GO:0010467; $p = 4.6\times10^{-12}$).
157 Likewise, we considered the genes with the lowest values, which revealed enrichment of terms
158 including cell wall modification (GO:0042545; $p = 2.6\times10^{-7}$) for the 2,289 genes with an H^2 of zero
159 and enrichment of terms including amino acid activation (GO:0043038; p = 0.0012) among the
160 11,895 genes with a Q_{ST} of zero.

161 We performed a regression analysis to ascertain whether a set of geographic (latitude, longitude,
162 elevation), climatic (temperature, precipitation) or other (time since sample collection) factors
163 significantly explained the global patterns of gene expression similarity among genotypes (**S2 Fig**), as
164 identified by performing a PCA of the expression data. None of the gene expression principal
165 components (PCs) were significantly explained by environmental factors, with the only significant
166 results found between PCs 2, 5 and 7 and the number of hours from collecting branches from the
167 field until bud sampling for RNA extraction, which explained 6.6%, 3.2%, and total 2.1% expression
168 variance, respectively.

169 We subsequently filtered expression values to remove unexpressed genes and uninformative
170 expression profiles with low variance. Of 35,154 annotated genes, 20,835 were expressed in all
171 samples, including biological replicates, and 23,183 were expressed in all genotypes when considering
172 genotype means. Filtering to remove uninformative expression retained 22,306 genes with the 12,848
173 removed genes representing those that were either not expressed in our bud samples (6,736 genes
174 with median expression of zero of which 2,385 had no detectable expression at all), or that were
175 weakly expressed (1,762 genes with variance < 0.05 and median expression < 2), together with genes
176 that had stable expression among genotypes (4,350 genes with expression variance < 0.05 and
177 median expression >= 2). The latter potentially represent genes with canalised gene expression.
178 Analysis of this set of stably expressed genes identified enrichment for GO categories including

179 protein transport (GO:0015031, $p = 6.8\times10^{-11}$) and protein localisation (GO:0008104, $p = 2.2\times10^{-10}$).
180 In contrast, the 500 genes with the highest variance across all samples were enriched for GO
181 categories related to protein phosphorylation (GO:0006468, $p < 10^{-6}$), chitin metabolic process
182 (GO:0006030, $p < 10^{-4}$), and cell wall macromolecule catabolic process (GO:0016998, $p < 10^{-4}$).
183 Comparing the variance of these 500 genes with mean F_{ST} calculated using SNPs within those genes
184 revealed no apparent relationship.

185 A recent reanalysis [67] of two existing datasets assaying gene expression among natural accessions
186 of *A. thaliana* [68,69] observed that thousands of genes displayed clear present/absent expression
187 among accessions. In contrast, we did not find any genes displaying such a pattern of expression
188 variation (**S3 Fig**), an observation that we also confirmed in an independent *P. tremula* dataset [70],
189 albeit containing substantially fewer genotypes.

190 eQTL mapping

191 We performed eQTL mapping, defining an eQTL as a significant association between a SNP
192 (termed an eSNP) and the expression of a gene (termed an eGene). Furthermore, we classified an
193 eQTL as *local* if the eSNP was located on the same chromosome and not more than 100 kbp from
194 the associated eGene, and as *distant* otherwise. We did not consider whether eQTLs acted in *cis* or
195 *trans*.

196 In total 164,290 eQTLs were identified at 5% empirical FDR: 131,103 local and 33,187 distant.
197 These eQTLs represented pairwise associations between 6,241 unique genes (eGenes) and 147,419
198 unique SNPs (eSNPs). The genomic context of eSNPs was determined by overlapping eSNP
199 positions with gene annotations. After normalising for feature length, the majority of local eSNPs
200 were located within untranslated regions (UTRs) and up- or down-stream (regulatory) regions of

201 genes, with distinctly lower representation within exons than introns (**Fig 2**B). The genomic context
202 distribution of local and distant eSNPs was largely similar, although there were distinctly more
203 eSNPs located within intergenic regions for distant eQTLs. Dividing the local eQTLs on the basis of
204 whether they were located within or near to the associated eGene or within or near any other gene
205 inside of our local distance threshold revealed that approximately half were located within or near to
206 the eGene itself. There was a clear tendency for a local eSNP to be located proximal to the
207 transcription start site (TSS) or the stop codon (**S4 Fig**). eGenes had significantly higher heritability
208 (difference in median of 0.24, Mann-Whitney $p < 2.2\times10^{-16}$) than non eGenes (**Fig 2**D), with this
209 trend being slightly higher for local than distant eQTLs (**S5 Fig**). There were no significant
210 differences in median expression between eGenes and non eGenes or between eGenes with local
211 and distant eQTLs. eGenes with at least one local eQTL were enriched for GO categories related to
212 tRNA metabolic process (GO:0006399, $p = 1.5\times10^{-5}$), ncRNA metabolic process (GO:0034660, p =
213 2.6×10^{-5}) and organonitrogen compound biosynthetic process (GO:1901566, $p = 2.2\times10^{-5}$), among
214 others, while eGenes with at least one distant eQTL were enriched for categories including protein
215 phosphorylation (GO:0006468, p = 0.0064; see **S1 File** for all results).

216 The vast majority of eSNPs were associated with a single eGene (132,258 eSNPs) with a maximum
217 of six eGenes associated with a single eSNP (S6A Fig). In contrast only 1,248 of the 6,241 eGenes
218 were associated with a single eSNP, while there were eGenes associated with up to 1,547 eSNPs (S6B
219 Fig). To partially account for linkage, we fitted linear models between the expression of each eGene
220 and all the significant eSNPs for that gene, both local and distant. The use of a linear model masks
221 eSNPs that contain identical/redundant information and thus effectively identifies haplotype blocks
222 present in all individuals (which we refer to as 'unique eSNPs'), while also producing a measure of
223 how well the combination of eSNPs explains the expression of the corresponding eGene (in terms of

224 percentage variance explained, %VE). Of the 4,993 eGenes associated with more than one eSNP,
225 4,703 were also associated to more than one unique eSNP. The adjusted %VE for the combination
226 of eQTLs was, in general, higher than for single eSNPs. Local eSNPs explained significantly more of
227 the variance than did distant eSNPs (local mean adjusted %VE = 51, distant mean adjusted %VE =
228 47, Mann-Whitney $p < 2.2\times10^{-16}$, **Fig 2**A) and also had higher statistical significance (**Fig 2**E). There
229 was a positive correlation between the maximum %VE of the eSNPs associated with an eGene and
230 gene expression H^2 (Pearson r = 0.47, $p < 2.2\times10^{-16}$).

231 A number of previous studies in both plants and a range of other organisms have identified distinct
232 eQTL hotspots [20,35,37,48,51], which represent loci where large numbers of *trans*-acting variants are
233 co-located. To detect possible genomic hotspots, we plotted eSNP positions against the positions of
234 the associated genes (**Fig 2**C), as well as the number of eSNPs and eGenes for 100 kbp genomic
235 windows along the chromosome (**S7 Fig**). We did not identify any clear hotspots in our data.

236 Co-expression network

237 We used genotype mean gene expression values to calculate a co-expression network and
238 subsequently considered the network characteristics for genes with and without a mapped eQTL
239 (eGenes). In common with other biological networks, the network had a good scale-free fit (R^2 =
240 0.97). We compared our network with the *P. tremula* expression atlas (exAtlas; [62]), which represents
241 different tissues collected from a single genotype, and observed that the correlation distribution for
242 the exAtlas samples was much wider (mean correlation 0.01 ± s.d. 0.36) than that of our population
243 expression data (mean correlation 0.00 ± s.d. 0.12; Kolmogorov-Smirnov D = 0.14, $p < 2.2\times10^{-16}$;
244 **Fig 3**A).

245 Clustering analysis of the co-expression network identified 38 co-expression modules (two examples
246 are shown in **Fig 3**C). These were enriched for a number of Gene Ontology (GO) categories
247 including translation (modules 9, 10, and 14), photosynthesis (module 22) and oxidation-reduction
248 process (module 29; for all results see **S1 File**). Despite the narrow distribution of correlation values,
249 the modules were reasonably well defined, as indicated by the normalised connectivity difference
250 (k_{diff}), *i.e.* the difference between intra- and inter-modular connectivity. All modules exhibited a
251 positive mean k_{diff}, with only 157 genes (0.7%) having a negative k_{diff}. This was in stark contrast to
252 genes assigned to the 'junk' module (*i.e.* all genes not assigned to any well-defined module), where
253 there were 480 genes with negative k_{diff} (29%).

254 ## eGenes are under-represented in network module cores

255 To test whether eQTLs could explain expression variation of highly connected genes, which are
256 those that drive the network structure, we examined the relationship of eGenes to network
257 connectivity. In general, eGenes had lower connectivity and betweenness centrality than non-eGenes
258 (Mann-Whitney $p < 2.2\times10^{-16}$ for both, **Fig 3**B). Moreover, genes with a positive k_{diff} were
259 significantly under-represented for eGenes (hypergeometric test, $p = 1.7\times10^{-27}$). We defined the core
260 of each module as the 10% of genes in a module with the highest normalised k_{diff} while also having
261 an intra-modular connectivity >1. Using this definition, all 38 modules contained at least one core
262 gene, with the percentage of core genes ranging from 2-10% (**S2 File**). Among these module cores,
263 28 contained at least one eGene, with 29 module cores being significantly under-represented for
264 eGenes (hypergeometric test, 5% FDR). This further emphasised that eGenes were not central in the
265 network. Subsequently, we tested whether the periphery of the network was overrepresented for
266 eGenes. Sixty-four of 142 peripheral genes were eGenes, representing a significant enrichment
267 (hypergeometric test $p = 2.6\times10^{-5}$). On the other hand, network module cores were enriched for

268 both transcription factors (hypergeometric test, $p = 5.6\times10^{-7}$) and phylogenetically conserved genes
269 (hypergeometric test, $p = 8.6\times10^{-5}$) defined as genes with orthologs in *P. tremuloides*, *P. trichocarpa* and
270 *A. thaliana* (20,318 genes; [71]). Furthermore, *P. tremula*-specific genes, *i.e.* genes without orthologs in
271 *P. tremuloides*, *P. trichocarpa* or *A. thaliana* (1,614 genes), were slightly under-represented in network
272 module cores (hypergeometric test, $p = 0.013$) while being slightly over-represented among eGenes
273 (hypergeometric test, $p = 0.010$).

274 In addition to eGenes having generally lower connectivity, there was also a negative relationship
275 between eQTL effect size and co-expression connectivity, with eGenes having the highest
276 connectivity also having lower effect size (Pearson $r = -0.15$, $p < 10^{-10}$, **Fig 3**D). As expected,
277 eGenes also had higher expression heritability (H^2) than non-eGenes (difference in median
278 heritability 0.24, Mann-Whitney $p < 2.2\times10^{-16}$). In addition, H^2 correlated positively with eQTL effect
279 size (Pearson $r = 0.59$, $p < 2.2\times10^{-16}$) and negatively with connectivity (Pearson $r = -0.30$, $p <$
280 2.2×10^{-16}).

281 eGenes had higher expression variance than non-eGenes (median variance of 0.075 versus 0.048;
282 Mann-Whitney $p < 2.2\times10^{-16}$). In particular, 60 of the 75 genes with variance > 2 (hypergeometric
283 test, $p < 2.2\times10^{-16}$) and 178 of the 250 genes with variance > 1 ($p < 2.2\times10^{-16}$), were eGenes (**S8 Fig**).
284 There was a weak negative relationship between network connectivity and gene expression variance
285 (Pearson $r = -0.08$, $p < 2.2\times10^{-16}$). Transcription factors had higher connectivity than non-
286 transcription factors (Mann-Whitney p-value $< 2.2\times10^{-16}$) and there was a slight overrepresentation
287 of TFs among core eGenes ($p=0.017$).

288 ## Paralogs with diverged expression are more likely to be eGenes

289 In *P. tremula* 3,910 paralog pairs were detected [71], with 2,140 of these (4,185 unique genes) passing
290 our gene expression and variance filtering criteria. These paralogs were significantly under-
291 represented for eGenes, with 1,078 of the 4,185 genes having at least one associated eSNP
292 (hypergeometric test, $p = 0.0004$). Comparing the expression correlation of paralog pairs to that of
293 random gene pairs showed that paralogs exhibited conserved regulation (permutation test, $p <$
294 0.001). We compared the expression correlation distributions of paralog pairs containing 0, 1, and 2
295 eGenes (**Fig 4**) and found that a higher number of eGenes in a pair was associated with lower
296 expression correlation (linear model $\hat{\beta} = -10.52$, $p < 10^{-16}$). Excluding paralogs did not alter the
297 fact that eGenes had significantly lower network connectivity than non-eGenes.

298 ## Population genetics signatures of eQTLs

299 We examined allele frequency distribution of eSNPs compared to all SNPs, finding that that eSNPs
300 included proportionally fewer low-frequency variants and proportionally more intermediate
301 frequency variants relative to the genome-wide level (S9A Fig). We detected a significant negative
302 correlation between minor allele frequency and eQTL effect size (Spearman's rho = -0.285, $p <$
303 0.001; S9B Fig), indicating that variants of large effect were present only at relatively low frequencies
304 with few large effect eQTLs detected. While purifying selection acting on eSNPs is a likely
305 explanation for this pattern, limited statistical power to detect associations between low frequency
306 variants and gene expression levels could also contribute to this observed pattern. However, as
307 shown in [28], even if the pattern is confounded by statistical artefacts, it likely represents a true
308 biological phenomenon, where large effect eQTLs are constrained to segregate at low frequencies.

309 We further compared patterns of genetic variation between genes with and without eQTLs (eGenes
310 vs. non-eGenes) and found that eGenes exhibited significantly higher genetic variation than non-
311 eGenes (**Fig 5**A). Furthermore, genes in network module cores harboured significantly lower genetic
312 variation compared to genes in the periphery of the network (**Fig 5**A and S10A Fig). To test whether
313 different gene categories were experiencing different levels of selective constraint, we also assessed
314 the site frequency spectrum across genes using Tajima's D [72]. Even if Tajima's D values are overall
315 negative in *P. tremula*, likely due to a historical range expansion [61], we would expect genes in
316 module cores to have lower values of Tajima's D compared to genome-wide averages if they are
317 indeed evolving under stronger purifying selection than peripheral genes. Likewise, if eGenes are
318 experiencing relaxed selective constraints, they should exhibit higher Tajima's D values relative to the
319 genome-wide average. Indeed, both of these expectations were supported by our data (**Fig 5**B, S10B
320 Fig): In accordance with stronger selective constraints for genes in network cores, we found
321 significantly reduced ratios of non-synonymous to synonymous polymorphisms ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and
322 substitutions (d_N/d_S) for these genes compared to peripheral genes in the network (**Fig 5**C,D). For
323 peripheral genes, eGenes had significantly higher $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and d_N/d_S compared to non-eGenes,
324 again suggesting that genes with eQTLs in the network periphery might be under relaxed purifying
325 selection compared non-eGenes. In comparison, there was no significant difference of $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ or
326 d_N/d_S between eGenes and non-eGenes located in network cores, suggesting that, within the
327 network core, both these classes of genes are evolving under comparably strong selective constraints.

328 # Discussion

329 ## Expression heritability and population differentiation

330 We present one of the first studies assaying gene expression in a natural outbreeding population. We
331 identified prevalent heritability in gene expression levels, in line with observations in a number of
332 species [5–17,53], with 5,924 genes (17%) having $H^2 > 0.5$. The majority of previous studies were
333 performed in controlled, often inter-specific, crosses, but our results confirm that there is also
334 substantial genetic variation contributing to gene expression in natural populations, demonstrating
335 the potential for selection to modulate expression levels.

336 In agreement with results reported for non-phenology related phenotypic traits in the SwAsp
337 collection [73], we observed no evidence of population structure, clinal variation or sub-population
338 differentiation (Q_{ST}) in our gene expression data, which is not surprising given the lack of population
339 structure present at the genome level [66]. Although some evidence of clustering was apparent (**Fig**
340 **1**D), this disappeared after hidden confounder removal (**S11 Fig**). Reassuringly, biological replicates
341 of a genotype clustered together for both the original and the adjusted data, indicating that our data
342 were reliable and that reproducible biological differences (corresponding to genetic differences) in
343 gene expression among genotypes could be detected. We were not able to identify the underlying
344 factors responsible for the clustering, but were able to exclude a set of climatic, environmental and
345 experimental factors that represented the most obvious candidates. However, it is of course possible
346 that environmental factors, especially latitude, may correlate with the expression of specific genes.
347 We did identify enrichment of GO categories in the set of the 500 most variable genes, suggesting
348 that this represents a signature with biological relevance. At the other extreme, for genes with
349 canalised expression, there was also enrichment for GO categories, with these potentially
350 representing genes having housekeeping functions.

eQTLs

In line with other published work, we found more local than distant eQTLs, with local eQTLs explaining significantly more of the variance in gene expression than distant eQTLs [35,48–51] (**Fig 2**A). Although each eSNP was typically associated with only a single gene, many genes were associated with more than one unique eSNP, with the combined variance explained by these eSNPs being higher than for the individual eSNPs, both for local and distant eQTLs. This suggests that the observed expression variation resulted from the combined effects of several eSNPs, consistent with the infinitesimal model. Looking specifically at only the subset of 1,100 eGenes with both local and distant eQTLs, we observed an average increase in the proportion of variance explained by all eSNPs compared to only local eSNPs (an increase of 3.4 percentage points after adjustment).

Local eQTLs were most frequently located in regulatory regions (**Fig 2**B), with UTRs having the highest density of local eSNPs (~1.5 eSNP per kbp sequence) followed by flanking regions (2 kbp up- and downstream, ~1 per kbp sequence) and introns (~0.75 per kbp sequence), which is consistent with previous findings in natural populations of humans, *Drosophila* and *Capsella grandiflora* [28]. The distributions of genomic context for local and distant eQTLs were broadly similar, however distant eQTLs were somewhat more common in exonic regions compared to local ones, as would be expected by indirectly acting eQTLs, and were more frequently located in intergenic regions, potentially representing distantly acting regulatory elements such as enhancers. It has previously been reported that distant eQTLs in humans are most frequently located in introns, followed by intergenic regions [74]. Our results only partially agree with this, with distant eQTLs being far more commonly associated with UTRs and flanking regions, followed by introns and intergenic regions. Finally, the global distribution of genomic contexts for all investigated SNPs was similar to that of local and distant eSNPs, suggesting that SNPs affecting expression do not have a characteristic genomic

374 signature and most likely representing the concentration of SNPs to regions of active recombination
375 [61].

376 Co-expression

377 The pairwise expression correlations underlying our co-expression network were low, at least
378 compared to those observed in typical systems biology studies, for example in the *P. tremula* exAtlas
379 network [62], where samples originated from different tissues of a single genotype (**Fig 3**A). Our
380 results thus demonstrate that natural variation clearly only operates by inducing a small proportion of
381 the known possible variation in expression for a gene, potentially reflecting a substantial influence by
382 stabilizing selection on natural variation in gene expression. Despite the low correlations, our natural
383 variation co-expression network displayed characteristics typical to biological networks [75]; the
384 network was scale-free with hubs and distinct modules. Specifically, we identified 38 co-expression
385 modules that were enriched for a number of functional categories, suggesting that they represent
386 biologically meaningful units.

387 Genes with the highest expression variance, and the highest heritability, were enriched for eGenes.
388 Perhaps counterintuitively given that genetic variation underlies heritable gene expression variation,
389 genes co-expressed with many other genes (*i.e.* genes with high connectivity) were under-represented
390 for eGenes, while genes with few co-expression partners (located in the network periphery) were
391 enriched for eGenes (**Fig 3**B). Furthermore, eGene connectivity was negatively correlated with
392 variance explained by the associated eSNPs (**Fig 3**D). It has similarly been reported that conserved
393 genes and hubs in human protein-protein interaction networks were less likely to be associated with a
394 detectable eQTL and that the effect size of eQTLs was negatively correlated with connectivity in the
395 protein-protein interaction network [76].

396 The simplest explanation for the presence of natural variation in gene expression would be that each
397 unique expression pattern is caused by a single eSNP: In network modules, where many genes share
398 the same expression pattern, a single eSNP modulating the expression of a central regulator would
399 be sufficient to induce similar expression variation (i.e. co-expression) in the set of connected genes
400 within that network module; in our data, we observed at least one eGene in 28 of the 38 modules,
401 thus potentially accounting for the co-expression structure of a majority of the modules. In
402 contrast, a gene located in the network periphery would have low connectivity and a relatively unique
403 expression profile. An eSNP modulating expression of such a gene would likely therefore affect only
404 the peripheral eGene that it is associated with: Consistent with this, we observed an
405 overrepresentation of eGenes in the network periphery.

406 Regulatory networks are known to be redundant, scale-free and to have an inherent ability to buffer
407 against single mutations of large negative effect [77–79]. In our case, genes with the potential to
408 cause such large negative effects would be observed as hubs in the co-expression network, and we
409 would therefore expect hubs to be buffered against single polymorphisms of large effect size and, in
410 particular, the persistence of large negative effect mutations. As such, we could expect the expression
411 of highly connected genes to be subjected to both strong stabilizing and purifying selection, which
412 would result in them being controlled largely by eSNPs of small additive effects. However, due to
413 statistical limitations, many of these small-effect eSNPs would represent false negatives (*i.e.* would be
414 unobserved), likely explaining why the core of 10 of the 38 modules within our network did not
415 contain any eGenes. Additionally, several other types of complex regulatory mechanisms can
416 preclude our ability to detect eSNPs: Epistatic interactions between eSNPs, such as AND-logics,
417 would result in very little correlation between the individual AND-interacting eSNPs and the
418 expression of the controlled gene. Likewise, the expression of a gene will frequently not only be a
419 function of the associated eSNPs but also of the expression of the regulators acting through those

420 eSNPs, which are themselves affected by additional eSNPs. Furthermore, the regulator(s) that
421 initiated an observed co-expression module may no longer be expressed at the sampling point used
422 for RNA isolation, and thus the eSNPs that determined the expression variation of that regulator will
423 no longer be visible to us (or a 'ghost signature' may remain, but be too weak to reach significance),
424 which again could explain the lack of eGenes in some module cores.

425 Our results suggest that natural variation in gene expression follows the infinitesimal model, with few
426 examples of a single eSNP explaining a large proportion of the expression variance of a gene. As we
427 argue, this may arise from selective processes that maintain biological robustness. However, we
428 cannot extrapolate these findings to suggest whether this is the rule for all cases. For example, we
429 observed a snapshot of expression during the process of bud flush, a trait that does not appear to be
430 under strong directional selection. To ascertain the general relevance of our findings, similar studies
431 should be performed assaying expression for phenotypic traits that do show evidence of selection,
432 such as the timing of autumn bud set within the SwAsp collection [73], although the specifics of how
433 to sample comparative material while controlling for confounding effects to enable expression
434 profiling presents significant challenges in such a case.

435 The Salicaceae lineage underwent a recent whole-genome duplication shared by all member species
436 and that remains represented by a large number of paralogous gene pairs in the genomes of *Populus*
437 species. If many of these duplicated genes are functionally redundant or in the process of diverging,
438 one would expect them to be overrepresented for eGenes as sub- or neo-functionalisation requires
439 derived SNPs to drive expression differences. However, we saw an under-representation of eGenes
440 in paralog pairs, suggesting that SNPs which initially induced expression divergence have reached
441 fixation. Interestingly, we found progressively lower expression correlation between paralog pairs

442 containing zero, one or two eGenes (**Fig 4**) indicating that a subset of paralogs are still undergoing
443 sub- or neo-functionalisation, with their associated eSNPs driving expression divergence.

444 Signatures of selection

445 We observed that both eGenes and non-eGenes in network module cores had significantly reduced
446 genetic diversity compared to more peripheral genes (**Fig 5**). This suggests that network core genes,
447 which represent genes of high potential effect, have evolved under stronger evolutionary constraint
448 than genes in the periphery of the network. As genes in module cores generally have higher levels of
449 pleiotropy and lower levels of dispensability [77], they are consequently more constrained against
450 changes in both gene expression and protein sequence compared to genes in the periphery of the
451 network [80]. Furthermore, we show that eGenes are overrepresented in the periphery of the
452 network and that they have experienced lower levels of purifying selection relative to non-eGenes,
453 corroborating a relaxation of both expression and coding constraints for these genes. Relaxed
454 selection of peripheral eGenes is thus expected to result in an accumulation of weakly deleterious
455 mutations that will segregate as intra-specific polymorphisms.

456 Consistent with the observation that network core genes are evolving under stronger functional
457 constraint than genes in the network periphery, we found that conserved genes (*i.e.* old genes) were
458 significantly over-represented in network cores while *P. tremula* specific genes (*i.e.* young genes) were
459 under-represented. This is in agreement with previous studies that have shown that old genes tend to
460 be more central in regulatory networks, with increased constraints on expression and fewer
461 associated eQTLs [81,82]. Finally, we detected a significant negative correlation between eQTL effect
462 size and eSNP allele frequency (S9B Fig), suggesting that prevalent purifying selection may be acting
463 on expression variation of the eQTLs detected in our study, which is consistent with recent results
464 from *Capsella grandiflora* [28].

465 # Conclusion

466 We present one of the first explorations of the genetic architecture of natural variation in a gene co-
467 expression network within a natural, out-breeding population. We identified substantial heritability
468 associated with gene expression, with eSNPs being most frequently located within regulatory regions,
469 suggesting that they act by inducing expression variance rather than affecting protein function. We
470 propose that the modular co-expression network is determined by a relatively small set of highly
471 connected regulators displaying expression variation due to eSNPs. Since these regulators will be
472 hubs, modulation of which would have potentially large negative effects, they are buffered by
473 complex regulatory mechanisms. Supporting this hypothesis, we find that eGenes are
474 underrepresented in module cores and that eSNP effect size is negatively correlated to network
475 connectivity. Our results point towards stronger selection acting on network core genes compared to
476 genes in the periphery of the co-expression network, with eSNP effect size being negatively
477 correlated to allele frequency. This suggests that the majority of segregating variation associated with
478 gene expression is slightly deleterious and that gene expression in general is under purifying selection.

479 # Materials and Methods

480 ## Samples

481 We collected branches form the SwAsp collection common garden in north of Sweden on 27th May
482 2012, before natural bud break but as close to the point of natural spring bud break as possible.
483 Branches were placed in the greenhouse facility at the Umeå Plant Science Centre under conditions
484 selected to induce rapid bud break (24 h light, temperature of 20 °C and humidity 50-70 %). At a
485 defined point of emergence (**S12 Fig**), buds were harvested, flash frozen in liquid nitrogen and

486 stored at -80 ° C until used for RNA isolation. Only terminal buds were sampled (i.e. no lateral
487 buds were included). The time from the day branches were placed in the greenhouse until bud flush
488 sampling ranged from one day to eight days (S13A Fig) and there was a high, positive correlation to
489 bud flush date recorded in the field for the same year (S13B Fig; $r=0.776$, $p < 2.2\times10^{-16}$). As has
490 previously been reported [73], there was no apparent Q_{ST} for bud flush, either in the field or for the
491 greenhouse material (Q_{ST} 0.13 and 0.07 respectively), however H^2 was high (H^2 = 0.82 and 0.71
492 respectively).

493 ## RNA isolation

494 One to two buds per clonal replicate were ground using one 3 mm stainless steel bead (Qiagen,
495 Redwood city, USA) in Corning® 96 well PP 1.2 ml cluster tubes (Sigma-Aldrich, St. Louis, USA)
496 using a Mixer Mill MM400 (Retsch, Haan, Germany) at 20 Hz for 2 x 15 sec. Total RNA was
497 extracted from all samples according to [83] with the omission of the L spermidine. Buffer volumes
498 were adjusted according to starting material (70 - 130 mg). RNA isolation was performed using one
499 extraction with CTAB buffer followed by one chloroform : isoamyl alcohol IAA (24:1) extraction.
500 All other steps were performed as in [83]. DNA contamination was removed using DNA-free™
501 DNA removal Kit (Life Technologies, Carlsbad, USA). RNA purity was measured using a
502 NanoDrop 2000 (Thermo Scientific, Wilmington, USA) and RNA integrity was assessed using the
503 Plant RNA Nano Kit for the Bioanalyzer (Agilent Technologies, Santa Clara, USA).

504 ## RNA-Sequencing and analysis

505 RNA-Sequencing was performed as in [70]. Briefly, paired-end (2 × 100 bp) RNA-Seq data were
506 generated using standard Illumina protocols and kits (TruSeq SBS KIT-HS v3, FC-401-3001; TruSeq
507 PE Cluster Kit v3, PE-401-3001) and all sequencing was performed using the Illumina HiSeq 2000

508 platform at the Science for Life Laboratory, Stockholm, Sweden. Raw data is available at the
509 European Nucleotide Archive (ENA, [84]) with accession number ERP014886.

510 RNA-Seq FASTQ-files were pre-processed and aligned to v1.0 of the *P. tremula* reference genome
511 [71] as in [85]. In short, reads were quality and adapter trimmed using Trimmomatic v0.32 [86],
512 rRNA matching reads were filtered using SortMeRNA v1.9 [87], reads were aligned to the v1.0 *P.*
513 *tremula* reference genome using STAR 2.4.0f1 [88] and read counts were obtained using htseq-count
514 from HTSeq [89]. FastQC [90] was used to track read quality throughout the process. Normalised
515 gene expression values were obtained by applying a variance stabilising transformation (VST) to the
516 raw counts from HTSeq, as implemented in the DESeq2 R package [91].

517 ## Gene expression H^2 and Q_{ST}

518 We calculated repeatability as an assumed upper bound estimate of broad sense heritability of gene
519 expression (see [92] for discussion) from the variance estimates in our data according to the equation

520 $$H^2 = \frac{V_G}{V_P}$$

521 where V_G is the genetic component of the variance calculated as the expression variance between
522 genotypes for a particular gene (*i.e.* variance among genotype means) and V_P is the total phenotypic
523 variance calculated as the sum of V_G and V_E, where V_E is the environment component of the
524 variance calculated as the expression variance within genotypes for a particular gene (i.e. the mean
525 variance among clonal replicates). Point estimates of H^2 were obtained using the repeatability
526 function from the heritability v1.1 R package [93].

527 Population differentiation (Q_{ST}; [94]) was calculated as

528 $$Q_{ST} = \frac{V_{between}}{V_{between} + 2V_{genetic}}$$

529 where V_{between} is the variance among populations and V_{genetic} is the genetic variance among genotypes
530 as computed using the lmer function from the lme4 v1.1.12 R package [95] using the formula

531 ```
expression ~ 1 + (1|population) + (1|clone)
```

532 where *expression* is the expression of a gene, *population* is a factor representing the population of each
533 sample, and *clone* is a factor representing genotype replicates. As we use repeatability as an upper
534 bound estimate of $H^2$, our $Q_{ST}$ estimates are conservative [96].

535 ## Hidden confounder removal

536 Gene expression data was adjusted for hidden confounders before mapping eQTLs and constructing
537 the co-expression network. Hidden confounders in the gene expression data was accounted for by
538 regressing out the 9 first principal components (PCs) of the gene expression data [97–99]. The
539 number of components to remove was determined by running the eQTL mapping with 0 to 20 PCs
540 removed and selecting the number of components that yielded the largest number of significant
541 eQTLs (Benjamini-Hochberg $p < 0.05$) (**S14 Fig**). This approach is based on the assumption that the
542 number of identified eQTLs will increase if the removed PCs are removing unwanted, systematic
543 variation (i.e. noise) rather than informative biological variation [97–99].

544 ## eQTL mapping

545 eQTL mapping was performed by associating gene expression with biallelic SNPs using the R
546 package Matrix eQTL v2.1.1 [100]. The corresponding raw resequencing data for all SwAsp
547 genotypes is available at the NCBI SRA resource as BioProject PRJNA297202 (SRA: SRP065057)
```

548 and the resultant Variant Call Format (VCF) file used for association mapping is available from [63].
549 Before performing association mapping, genes were filtered on variance so that only genes with a
550 gene expression variance above 0.05 were included. SNPs were also filtered on minor allele
551 frequency (MAF) and major genotype frequency (MGF); any SNPs with MAF < 0.1 or MGF > 0.9
552 were excluded to avoid spurious associations. A motivating example for the MGF filtering can be
553 seen in **S15 Fig**. We also generated an LD-trimmed SNP set by removing one SNP from each pair of
554 SNPs with a between SNP correlation coefficient (r^2) >0.2 in blocks of 50 SNPs using PLINK v1.9
555 [101], yielding 217,489 independent SNPs that were retained for analyses of population structure.
556 The first genotype principal component based on the set of independent SNPs was used as a
557 covariate in the linear model used by Matrix eQTL to account for the weak signature of population
558 structure. Permutation testing was used to determine eQTL significance whereby genotype sample
559 labels were permuted 1000 times and the maximum absolute t-statistic from Matrix eQTL was
560 recorded for each gene across all SNPs for each permutation. Empirical p-values were calculated
561 with the empPvals function in the qvalue v2.4.2 R package [102], and q-values (empirical FDR) were
562 calculated with the qvalue function in the same package.

563 When determining the genomic context of eSNPs, there were some cases where introns overlapped
564 exons as a result of overlapping gene model being present on the same strand. These 41 eSNPs were
565 discarded from the counting. Another type of overlap that was discarded were cases where an eSNP
566 overlapped a gene feature, but no sub-feature inside that gene (e.g. UTR, exon or intron). These
567 1961 eSNPs were excluded from the counting. Since many of the features overlap (e.g. exon and
568 untranslated regions), the priority for counting was untranslated region, exon/intron,
569 upstream/downstream and intergenic.

570 ## Co-expression

571 The R package WGCNA v1.51 [103] was used for constructing a co-expression network. The input
572 gene expression values were per-gene genotype means. We chose to use the unsigned network type
573 for this study with the motivation that we did not want to discard negative relationships. By looking
574 at this from an eQTL perspective, an eSNP can be positively associated with one gene while
575 negatively associated with another. The relationship between these two genes would be missed if we
576 used the signed or the signed-hybrid approaches. Using the unsigned approach, we assure that genes
577 with strong negative correlation end up in the same network modules. A soft thresholding power of
578 5 was used to calculate adjacencies. The topological overlap matrix (TOM) was generated using the
579 TOMsimilarity function with the signed approach in order to take negative edges into account (see
580 [103] for details). In order to identify network modules, hierarchical clustering was applied to the
581 TOM dissimilarity matrix (1 – TOM) and the resulting dendrogram was divided into modules using
582 the cutreedynamic function. The connectivity of the network was then defined as the adjacency sum
583 for each node, i.e. the weights of the edges that are connected to this node. This concept was applied
584 to modules as well to obtain measures of intra- and inter-modular connectivity, i.e. the connectivity
585 based on edges connecting the gene with other genes inside the same module, and connectivity based
586 on edges connecting the gene with genes outside of the module.

587 To define the periphery of the network we applied a hard edge-threshold to the network where only
588 gene-pairs with an absolute Pearson correlation > 0.22 were linked, which corresponded to the top
589 1% most correlated gene pairs. Genes were then classed as peripheral if they linked to only one other
590 gene.

591 Conserved genes and *P. tremula*-specific genes were inferred from gene family analysis based on
592 protein sequence similarity available on the PopGenIE FTP server [104].

593 ## Population genetic analysis

594 The minor allele frequency (MAF) was compared between eSNPs (both local and distant) and all
595 SNPs throughout the genome. To test for the signature of selection on eQTLs, we also estimate the
596 correlation between minor allele frequency and eQTL effect size. Furthermore, we estimated and
597 compared nucleotide diversity (θ_π) and Tajima's D [72] between eGenes and non-eGenes, as well as
598 genes locating at central or peripheral in the gene network (core and peripheral genes). We only used
599 the reads with mapping quality above 30 and the bases with quality score higher than 20 for the
600 estimation of θ_π and Tajima's D by ANGSD [105]. Genes with less than 50 covered sites left from
601 previous quality filtering steps (Wang et al. In prep) were excluded. In the end, there were 131 and
602 1,563 eGenes and non-eGenes that were located in module cores, respectively. Similarly there were
603 5,759 eGenes and 12,893 non-eGenes located the periphery of the gene expression network. In
604 addition, θ_π and Tajima's D were also calculated for 0-fold and 4-fold degenerate sites within genes,
605 representing non-synonymous and synonymous sites, respectively. We further compared the
606 estimates of $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ between different categories of genes. We used the transcript with the
607 highest content of protein-coding sites to categorize synonymous and non-synonymous sites within
608 each gene. Finally, a total of 33,039 orthologous genes were determined by best BLASTp sequence
609 homology searches between *P. tremula* and *P. trichocarpa* [71,104]. We estimated the ratio of non-
610 synonymous substitution and synonymous substitution rate (d_N/d_S) using gKaKs pipeline [106] with
611 codeml method [107] for 133 and 1,599 orthologous eGenes and non-eGenes within module cores
612 of gene expression network, and for 5,758 and 12,828 orthologous eGenes and non-eGenes locating
613 at peripheral in the gene network. Significance for the above statistical measurements between each
614 pair of gene categories was evaluated using Mann-Whitney tests.

615 # Acknowledgements

616 PKI has been supported by a grant from Vetenskapsrådet. JW was supported by a scholarship from
617 the Chinese Scholarship Council. NRS is supported by the Trees and Crops for the Future (TC4F)
618 project. The authors also would like to acknowledge support from the Science for Life Laboratory,
619 the National Genomics Infrastructure (NGI), and the Uppsala Multidisciplinary Center for
620 Advanced Computational Science (UPPMAX) for providing assistance in massive parallel sequencing
621 and computational infrastructure. We thank Dr Kathryn Robinson for her help sampling material,
622 for providing environmental data and for informative discussion.

623 # References

624 1. Ingvarsson PK, Hvidsten TR, Street NR. Towards integration of population and comparative
625 genomics in forest trees. New Phytol. 2016; doi:10.1111/nph.14153

626 2. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian
627 Inheritance. Trans R Soc Edinburgh. 1919;52: 399–433. doi:10.1017/S0080456800012163

628 3. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, et al. Regional and
629 strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A.
630 2000;97: 11038–43. doi:10.1073/pnas.97.20.11038

631 4. Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, et al. The core
632 meiotic transcriptome in budding yeasts. Nat Genet. 2000;26: 415–423. doi:10.1038/82539

633 5. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions
634 of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet.

635 2001;29: 389–95. doi:10.1038/ng766

636 6. Oleksiak MF, Churchill GA, Crawford DL. Variation in gene expression within and among
637 natural populations. Nat Genet. 2002;32: 261–266. doi:10.1038/ng983

638 7. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. Genetics of gene
639 expression surveyed in maize, mouse and man. Nature. 2003;422: 297–302.
640 doi:10.1038/nature01434

641 8. Kirst M, Myburg AA, De León JPG, Kirst ME, Scott J, Sederoff R. Coordinated genetic
642 regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA
643 microarray data in an interspecific backcross of eucalyptus. Plant Physiol. 2004;135: 2368–78.
644 doi:10.1104/pp.103.037960

645 9. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic
646 analysis of genome-wide variation in human gene expression. Nature. 2004;430: 743–747.
647 doi:10.1038/nature02797

648 10. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping
649 determinants of human gene expression by regional and genome-wide association. Nature.
650 2005;437: 1365–9. doi:10.1038/nature04244

651 11. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, et al. Integrated
652 transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat
653 Genet. 2005;37: 243–253. doi:10.1038/ng1522

654 12. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide
655 associations of gene expression variation in humans. PLoS Genet. 2005;1: 0695–0704.

656 doi:10.1371/journal.pgen.0010078

657 13. DeCook R, Lall S, Nettleton D, Howell SH. Genetic Regulation of Gene Expression During
658 Shoot Development in *Arabidopsis*. Genetics. 2006;172: 1155–1164. Available:
659 http://www.genetics.org/cgi/content/abstract/172/2/1155

660 14. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, et al. A genome-wide
661 association study of global gene expression. Nat Genet. 2007;39: 1202–1207.
662 doi:10.1038/ng2109

663 15. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise
664 of eQTL studies. Trends Genet. 2008;24: 408–415. doi:10.1016/j.tig.2008.06.001

665 16. Kim J, Gibson G. Insights from GWAS into the quantitative genetics of transcription in
666 humans. Genet Res (Camb). 2010;92: 361–9. doi:10.1017/S001667231000056X

667 17. Powell JE, Henders AK, McRae AF, Kim J, Hemani G, Martin NG, et al. Congruence of
668 Additive and Non-Additive Effects on Gene Expression Estimated from Pedigree and SNP
669 Data. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003502

670 18. Wang RL, Stec A, Hey J, Lukens L, Doebley J. The limits of selection during maize
671 domestication. Nature. 1999;398: 236–239. doi:10.1038/18435

672 19. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell.
673 2000;101: 577–580. doi:10.1016/S0092-8674(00)80868-5

674 20. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in
675 budding yeast. Science (80-). 2002;296: 752–5. doi:10.1126/science.1069516

676 21. Ayroles JF, Carbone MA, Stone E a, Jordan KW, Lyman RF, Magwire MM, et al. Systems
677 genetics of complex traits in *Drosophila melanogaster*. Nat Genet. 2009;41: 299–307.
678 doi:10.1038/ng.332

679 22. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and
680 prospects. Nat Rev Genet. 2009;10: 565–77. doi:10.1038/nrg2612

681 23. Liao B-Y, Weng M-P, Zhang J. Contrasting genetic paths to morphological and physiological
682 evolution. Proc Natl Acad Sci U S A. 2010;107: 7353–8. doi:10.1073/pnas.0910339107

683 24. Hines HM, Papa R, Ruiz M, Papanicolaou A, Wang C, Nijhout HF, et al. Transcriptome
684 analysis reveals novel patterning and pigmentation genes underlying *Heliconius* butterfly wing
685 pattern variation. BMC Genomics. 2012;13: 288. doi:10.1186/1471-2164-13-288

686 25. Richards CL, Rosas U, Banta J, Bhambhra N, Purugganan MD. Genome-Wide Patterns of
687 *Arabidopsis* Gene Expression in Nature. PLoS Genet. 2012;8: e1002662.
688 doi:10.1371/journal.pgen.1002662

689 26. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet.
690 2001;17: 388–91. doi:10.1016/S0168-9525(01)02310-1

691 27. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. Nat
692 Rev Genet. 2002;3: 43–52. doi:10.1038/nrg703

693 28. Josephs EB, Lee YW, Stinchcombe JR, Wright SI. Association mapping reveals the role of
694 purifying selection in the maintenance of genomic variation in gene expression. Proc Natl
695 Acad Sci U S A. 2015;112: 15390–5. doi:10.1073/pnas.1503027112

696 29. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and

697 molecular biology. Genome Res. 2015;25: 1427–1431. doi:10.1101/gr.190983.115.

698 30. Flint J, and Mackay TFC. Genetic architecture of quantitative traits in flies, mice and humans.
699 Genome Res. 2009;19: 723–733. doi:10.1101/gr.086660.108

700 31. Hindorff L a, Sethupathy P, Junkins H a, Ramos EM, Mehta JP, Collins FS, et al. Potential
701 etiologic and functional implications of genome-wide association loci for human diseases and
702 traits. Proc Natl Acad Sci U S A. 2009;106: 9362–7. doi:10.1073/pnas.0903103106

703 32. Ku CS, Loy EY, Pawitan Y, Chia KS. The pursuit of genome-wide association studies: where
704 are we now? J Hum Genet. 2010;55: 195–206. doi:10.1038/jhg.2010.19

705 33. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:
706 206–16. doi:10.1038/nrg2063

707 34. Kliebenstein DJ, West MAL, van Leeuwen H, Kim K, Doerge RW, Michelmore RW, et al.
708 Genomic survey of gene expression diversity in *Arabidopsis thaliana*. Genetics. 2006;172: 1179–
709 89. doi:10.1534/genetics.105.049353

710 35. Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, et al.
711 Regulatory network construction in *Arabidopsis* by using genome-wide gene expression
712 quantitative trait loci. Proc Natl Acad Sci U S A. 2007;104: 1708–13.
713 doi:10.1073/pnas.0610429104

714 36. van Leeuwen H, Kliebenstein DJ, West MAL, Kim K, van Poecke R, Katagiri F, et al. Natural
715 variation among *Arabidopsis thaliana* accessions for transcriptome response to exogenous
716 salicylic acid. Plant Cell. 2007;19: 2099–110. doi:10.1105/tpc.107.050641

717 37. West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, et al.

718 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in
719 *Arabidopsis*. Genetics. 2007;175: 1441–50. doi:10.1534/genetics.106.064972

720 38. Zhang X, Cal AJ, Borevitz JO. Genetic architecture of regulatory variation in *Arabidopsis*
721 *thaliana*. Genome Res. 2011;21: 725–733. doi:10.1101/gr.115337.110

722 39. Lowry DB, Logan TL, Santuari L, Hardtke CS, Richards JH, DeRose-Wilson LJ, et al.
723 Expression quantitative trait locus mapping across water availability environments reveals
724 contrasting associations with genomic features in *Arabidopsis*. Plant Cell. 2013;25: 3266–79.
725 doi:10.1105/tpc.113.115352

726 40. Swanson-Wagner R a, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, et al. Paternal dominance
727 of trans-eQTL influences gene expression patterns in maize hybrids. Science. 2009;326: 1118–
728 1120. doi:10.1126/science.1178294

729 41. Holloway B, Luck S, Beatty M, Rafalski J-A, Li B. Genome-wide expression quantitative trait
730 loci (eQTL) analysis in maize. BMC Genomics. 2011;12: 336. doi:10.1186/1471-2164-12-336

731 42. Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, et al. RNA sequencing reveals the
732 complex regulatory network in the maize kernel. Nat Commun. 2013;4: 2832.
733 doi:10.1038/ncomms3832

734 43. Wang J, Yu H, Xie W, Xing Y, Yu S, Xu C, et al. A global analysis of QTLs for expression
735 variations in rice shoots at the early seedling stage. Plant J. 2010;63: 1063–1074.
736 doi:10.1111/j.1365-313X.2010.04303.x

737 44. Wang J, Yu H, Weng X, Xie W, Xu C, Li X, et al. An expression quantitative trait loci-guided
738 co-expression analysis for constructing regulatory network using a rice recombinant inbred

739 line population. J Exp Bot. 2014;65: 1069–1079. doi:10.1093/jxb/ert464

740 45. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR. Genetic Architecture of Transcript-
741 Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. Genetics. 2005;169: 2295–
742 2303. Available: http://www.genetics.org/cgi/content/abstract/169/4/2295

743 46. Drost DR, Benedict CI, Berg A, Novaes E, Novaes CRDB, Yu Q, et al. Diversification in the
744 genetic architecture of gene expression and transcriptional networks in organ differentiation
745 of *Populus*. Proc Natl Acad Sci U S A. 2010;107: 8492–7. doi:10.1073/pnas.0914709107

746 47. Kullan AR, van Dyk MM, Hefer C a, Jones N, Kanzler A, Myburg A a. Genetic dissection of
747 growth, wood basic density and gene expression in interspecific backcrosses of *Eucalyptus*
748 *grandis* and *E. urophylla*. BMC Genet. 2012;13: 60. doi:10.1186/1471-2156-13-60

749 48. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression
750 traits in yeast. Proc Natl Acad Sci U S A. 2005;102: 1572–7. doi:10.1073/pnas.0408709102

751 49. Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, et al. Segregating
752 Variation in the Transcriptome: Cis Regulation and Additivity of Effects. Genetics. 2006;173:
753 1347–1355. doi:10.1534/genetics.105.051474

754 50. Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. Rapid evolution of male-biased gene
755 expression in *Drosophila*. Proc Natl Acad Sci U S A. 2003;100: 9894–9.
756 doi:10.1073/pnas.1630690100

757 51. Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M. Gene expression quantitative
758 trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide
759 transcriptional regulation. Plant J. 2008;53: 90–101. doi:10.1111/j.1365-313X.2007.03315.x

760 52. Whitehead A, Crawford DL. Variation within and among species in gene expression: raw
761 material for evolution. Mol Ecol. 2006;15: 1197–1211. doi:10.1111/j.1365-294X.2006.02868.x

762 53. Leder EH, McCairns RJS, Leinonen T, Cano JM, Viitaniemi HM, Nikinmaa M, et al. The
763 evolution and adaptive potential of transcriptional variation in sticklebacks–signatures of
764 selection and widespread heritability. Mol Biol Evol. 2015;32: 674–89.
765 doi:10.1093/molbev/msu328

766 54. Mäkinen H, Papakostas S, Vøllestad LA, Leder EH, Primmer CR. Plastic and evolutionary
767 gene expression responses are correlated in European grayling (*Thymallus thymallus*)
768 subpopulations adapted to different thermal environments. J Hered. 2016;107: 82–89.
769 doi:10.1093/jhered/esv069

770 55. Kohn MH, Shapiro J, Wu C-I. Decoupled differentiation of gene expression and coding
771 sequence among *Drosophila* populations. Genes Genet Syst. 2008;83: 265–273.
772 doi:10.1266/ggs.83.265

773 56. Holloway AK, Lawniczak MKN, Mezey JG, Begun DJ, Jones CD. Adaptive gene expression
774 divergence inferred from population genomics. PLoS Genet. 2007;3: 2007–2013.
775 doi:10.1371/journal.pgen.0030187

776 57. Leinonen T, McCairns RJS, O'Hara RB, Merilä J. Q_{ST}–F_{ST} comparisons: evolutionary and
777 ecological insights from genomic heterogeneity. Nat Rev Genet. 2013;14: 179–90.
778 doi:10.1038/nrg3395

779 58. Nourmohammad A, Rambeau J, Held T, Berg J, Lassig M. Pervasive adaptation of gene
780 expression in *Drosophila*. arXiv. 2015; Available: http://arxiv.org/abs/1502.06406

781 59. Jansson S, Douglas CJ. *Populus*: A Model System for Plant Biology. Annu Rev Plant Biol.
782 2007;58: 435–458. doi:10.1146/annurev.arplant.58.032806.103956

783 60. Wang J, Street NR, Scofield DG, Ingvarsson PK. Variation in linked selection and
784 recombination drive genomic divergence during allopatric speciation of European and
785 American aspens. Mol Biol Evol. 2016; msw051. doi:10.1093/molbev/msw051

786 61. Wang J, Street NR, Scofield DG, Ingvarsson PK. Natural Selection and Recombination Rate
787 Variation Shape Nucleotide Polymorphism Across the Genomes of Three Related *Populus*
788 Species. Genetics. 2016;202: 1185–1200. doi:10.1534/genetics.115.183152

789 62. Sundell D, Mannapperuma C, Netotea S, Delhomme N, Lin Y-C, Sjödin A, et al. The Plant
790 Genome Integrative Explorer Resource: PlantGenIE.org. New Phytol. 2015;208: 1149–1156.
791 doi:10.1111/nph.13557

792 63.
793 ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/VCF/SwAsp_94samples.filte
794 r.maf005.beagle.maf.recode.vcf.gz.

795 64. Luquez V, Hall D, Albrectsen BR, Karlsson J, Ingvarsson P, Jansson S. Natural phenological
796 variation in aspen (*Populus tremula*): the SwAsp collection. Tree Genet Genomes. 2008;4: 279–
797 292. doi:10.1007/s11295-007-0108-y

798 65. Ingvarsson PK. Nucleotide polymorphism and linkage disequilibrium within and among
799 natural populations of European aspen (*Populus tremula* L., *Salicaceae*). Genetics. 2005;169: 945–
800 53. doi:10.1534/genetics.104.034959

801 66. Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK. Adaptive population

802 differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremula*,
803 L.): a comparison of neutral markers, candidate genes and phenotypic traits. Evolution.
804 2007;61: 2849–60. doi:10.1111/j.1558-5646.2007.00230.x

805 67. Zan Y, Shen X, Forsberg SKG, Carlborg Ö. Genetic regulation of transcriptional variation in
806 wild-collected *Arabidopsis thaliana* accessions. bioRxiv. 2016; Available:
807 http://biorxiv.org/content/early/2016/01/18/037077.abstract

808 68. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of
809 population epigenomic diversity. Nature. 2013;495: 193–8. doi:10.1038/nature11968

810 69. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA
811 methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife.
812 2015;4: e05255. doi:10.7554/eLife.05255

813 70. Robinson KM, Delhomme N, Mähler N, Schiffthaler B, Onskog J, Albrectsen BR, et al.
814 *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. BMC Plant Biol.
815 2014;14: 276. doi:10.1186/s12870-014-0276-5

816 71. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S. The *Populus* Genome Integrative
817 Explorer (PopGenIE): a new resource for exploring the *Populus* genome. New Phytol.
818 2009;182: 1013–1025. doi:10.1111/j.1469-8137.2009.02807.x

819 72. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA
820 polymorphism. Genetics. 1989;123: 585–595. doi:PMC1203831

821 73. Robinson KM, Ingvarsson PK, Jansson S, Albrectsen BR. Genetic variation in functional
822 traits influences arthropod community composition in aspen (*Populus tremula* L.). PLoS One.

823 2012;7: e37679. doi:10.1371/journal.pone.0037679

824 74. Wright FA, Sullivan PF, Brooks AI, Zou F, Sun W, Xia K, et al. Heritability and genomics of
825 gene expression in peripheral blood. Nat Genet. 2014;46: 430–7. doi:10.1038/ng.2951

826 75. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization.
827 Nat Rev Genet. 2004;5: 101–13. doi:10.1038/nrg1272

828 76. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing
829 the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.
830 Genome Res. 2013; doi:10.1101/gr.155192.113

831 77. MacNeil LT, Walhout AJM. Gene regulatory networks and the role of robustness and
832 stochasticity in the control of gene expression. Genome Res. 2011;21: 645–657.
833 doi:10.1101/gr.097378.109

834 78. Whitacre JM, Bender A. Networked buffering: a basic mechanism for distributed robustness
835 in complex adaptive systems. Theor Biol Med Model. 2010;7: 20. doi:10.1186/1742-4682-7-20

836 79. Whitacre JM. Biological robustness: Paradigms, mechanisms, systems principles. Front Genet.
837 2012;3: 1–15. doi:10.3389/fgene.2012.00067

838 80. Papakostas S, Vøllestad LA, Bruneaux M, Aykanat T, Vanoverbeke J, Ning M, et al. Gene
839 pleiotropy constrains gene expression changes in fish adapted to different thermal conditions.
840 Nat Commun. 2014;5: 4071. doi:10.1038/ncomms5071

841 81. Popadin KY, Gutierrez-Arcelus M, Lappalainen T, Buil A, Steinberg J, Nikolaev SI, et al.
842 Gene age predicts the strength of purifying selection acting on gene expression variation in
843 humans. Am J Hum Genet. 2014;95: 660–74. doi:10.1016/j.ajhg.2014.11.003

844 82. Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution of
845 gene interaction networks in the human and mouse genomes. Genome Biol. 2015;16: 202.
846 doi:10.1186/s13059-015-0772-4

847 83. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine
848 trees. Plant Mol Biol Report. 1993;11: 113–116. doi:10.1007/BF02670468

849 84. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European
850 Nucleotide Archive. Nucleic Acids Res. 2011;39: D28-31. doi:10.1093/nar/gkq967

851 85. Delhomme N, Mähler N, Schiffthaler B, Sundell D, Mannepperuma C, Hvidsten TR, et al.
852 Guidelines for RNA-Seq data analysis. Epigenesys. 2014; Available:
853 http://www.epigenesys.eu/en/protocols/bio-informatics/1283-guidelines-for-rna-seq-data-
854 analysis

855 86. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
856 Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

857 87. Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs
858 in metatranscriptomic data. Bioinformatics. 2012;28: 3211–3217.
859 doi:10.1093/bioinformatics/bts611

860 88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
861 universal RNA-seq aligner. Bioinformatics. 2012;29: 15–21.
862 doi:10.1093/bioinformatics/bts635

863 89. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput
864 sequencing data. Bioinformatics. 2014;31: 166–169. doi:10.1093/bioinformatics/btu638

865 90. Andrews S. FastQC: A quality control tool for high throughput sequence data [Internet]. 2016
866 [cited 27 Sep 2016]. Available: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

867 91. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-
868 seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

869 92. Dohm MR. Repeatability estimates do not always set an upper limit to heritability. Funct Ecol.
870 2002;16: 273–280. doi:10.1046/j.1365-2435.2002.00621.x

871 93. Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, et al. Marker-based
872 estimation of heritability in immortal populations. Genetics. 2015;199: 379–398.
873 doi:10.1534/genetics.114.167916

874 94. Spitze K. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation.
875 Genetics. 1993;135: 367–74. Available: http://www.genetics.org/content/135/2/367.short

876 95. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J
877 Stat Softw. 2015;67: 1–48. doi:10.18637/jss.v067.i01

878 96. Manier MK, Palumbi SR. Intraspecific divergence in sperm morphology of the green sea
879 urchin, *Strongylocentrotus droebachiensis*: implications for selection in broadcast spawners. BMC
880 Evol Biol. 2008;8: 283. doi:10.1186/1471-2148-8-283

881 97. Hyun MK, Ye C, Eskin E. Accurate discovery of expression quantitative trait loci under
882 confounding from spurious and genuine regulatory hotspots. Genetics. 2008;180: 1909–1925.
883 doi:10.1534/genetics.108.094201

884 98. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding
885 mechanisms underlying human gene expression variation with RNA sequencing. Nature.

886 2010;464: 768–72. doi:10.1038/nature08872

887 99. Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, et al. Normalizing
888 RNA-sequencing data by modeling hidden covariates with prior knowledge. PLoS One.
889 2013;8: e68141. doi:10.1371/journal.pone.0068141

890 100. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations.
891 Bioinformatics. 2012;28: 1353–1358. doi:10.1093/bioinformatics/bts163

892 101. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool
893 set for whole-genome association and population-based linkage analyses. Am J Hum Genet.
894 2007;81: 559–575. doi:10.1086/519795

895 102. Storey J. qvalue: Q-value estimation for false discovery rate control [Internet]. 2015. Available:
896 http://github.com/jdstorey/qvalue

897 103. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis.
898 BMC Bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559

899 104. ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/Annotation/DoubleMCL-tab-
900 annotation.txt.gz.

901 105. Korneliussen T, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing
902 Data. BMC Bioinformatics. 2014;15: 356. doi:10.1186/s12859-014-0356-4

903 106. Zhang C, Wang J, Long M, Fan C. GKaKs: The pipeline for genome-level Ka/Ks calculation.
904 Bioinformatics. 2013;29: 645–646. doi:10.1093/bioinformatics/btt009

905 107. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:

906 1586–1591. doi:10.1093/molbev/msm088

907

908 # Figures

909 **Fig 1. Population location and gene expression overview.** (A) Map of the original locations of
910 the SwAsp populations. The red arrow points to the location of the common garden used in this
911 study. (B) Distribution of gene expression heritability. (C) Distribution of gene expression Q_{ST}. (D)
912 Sample clustering based on all samples, including biological replicates. The heatmap represents the
913 sample correlation matrix based on the 500 genes with the highest expression variance. Darker
914 colour indicates higher correlation. The coloured bar represents the populations the samples belong
915 to. The small clusters on the diagonal correspond to biological replicates of each genotype.

916 **Fig 2. eQTL overview.** (A) Expression variance explained *(R^2)* for local and distant eQTLs. Box
917 plots show the maximum variance explained by a single eQTL for each gene and the total variance
918 explained by all eQTLs for each gene. The widths of the boxes are proportional to the number of
919 genes represented. The pairwise significiance of a Mann-Whitney test is indicated by asterisks: ***$p <$
920 0.001, **$p < 0.01$. (B) Genomic context of local and distant eSNPs. When an eSNP overlapped with
921 features on both strands, both of them were counted. For both local and distant eQTLs the features
922 are based on the gene that is closest to the eSNP, and furthermore, for local eQTLs, the features are
923 divided into whether the eSNP is located in or near the same gene that it is associated to or not.
924 Flanking regions represent 2 kb upstream and downstream from the gene. (C) Scatter plot showing
925 the positions of all significant eQTLs in this study. No evidence of eQTL hotspots can be observed.
926 Numbers indicate chromosome. (D) Broad sense heritability distributions for eGenes and non-

927 eGenes. (E) Manhattan plots for local eQTLs (upper) and distant eQTLs (lower). Each point
928 represents an eQTL.

929 **Fig 3. Pairwise gene expression correlation distribution and co-expression characteristics of**
930 **eGenes.** (A) Co-expression correlation distribution of all pairs of genes for the SwAsp data and the
931 exAtlas data. (B) Distribution of co-expression connectivity for eGenes and non-eGenes. (C)
932 Network module 23 and module 22 with core genes coloured blue and eGenes are indicated by
933 square nodes. In module 23 (upper network), all core genes are eGenes, while for module 22, there
934 are very few eGenes in the core. Red edges indicate negative correlation in expression while blue
935 edges indicate positive correlation. (D) eGene connectivity plotted against the total variance
936 explained of all eQTL associated with each eGene. Colors indicate the effect size where more blue
937 represents a more positive effect and red is a more negative effect.

938 **Fig 4. Correlation within paralog pairs as a function of the number of eGenes in the paralog**
939 **pair.** The widths of the boxes are proportional to the number of genes in each set. The mean
940 correlations for paralog pairs with 0, 1, or 2 eGenes were 0.17, 0.10, and 0.06, respectively. The
941 pairwise significance of a Mann-Whitney test is indicated by asterisks: ***$p < 0.001$, **$p < 0.01$.

942 **Fig 5. Measures of diversity and divergence.** eGenes and non-eGenes were contrasted, as well as
943 core and peripheral genes from the gene expression network. (A) Nucleotide diversity, (B) Tajima's
944 D, (C) $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and (D) d_N/d_S. Significance between each pair of gene categories was evaluated
945 using Mann-Whitney tests and significance is indicated by asterisks above the boxplot: ***$p < 0.001$,
946 **$p < 0.01$, *$p < 0.05$, n.s. $p > 0.05$.

947 # Supporting information

948 **S1 File. GO enrichment analysis results.** Contains GO enrichment results for local and distant
949 eGenes, network module cores, and subsets of genes with extreme H^2 and Q_{ST} values.

950 **S2 File. Various gene statistics.** Contains statistics for all genes, as well as all genes included in the
951 generation of the co-expression network and the eQTL mapping.

952 **S1 Fig. Clustering of genotypes.** (Left) Heatmap of the sample correlation matrix based on the 500
953 most variably expressed genes. Darker colour indicates higher correlation. The coloured bar
954 represents the populations the genotypes belong to. (Right) The two first principal components from
955 a principal component analysis (PCA) based on all genes. Again, colours represent the genotype
956 population. The percentages in the axis labels indicate the amount of variance explained by each
957 component.

958 **S2 Fig. Correlations between gene expression principal components (PCs) and**
959 **environmental variables.** The values in each tile represent the Pearson correlation between the
960 gene expression PC (x-axis) and environmental variable (y-axis). Small asterisks represent a nominal
961 p-value < 0.05 while large asterisks represent Benjamini-Hochberg (BH) adjusted p-values < 0.05.
962 The only factor with significant correlations to expression PCs was "Hours to harvest", which is the
963 number of hours into the sampling period that the buds were harvested. It was significantly
964 associated with PC4 (BH-adjusted $p = 4.6\times10^{-6}$), PC7 (BH-adjusted $p = 0.030$) and PC2 (BH-
965 adjusted $p = 0.033$).

966 **S3 Fig. Relationship between gene expression ranks and the number of genotypes with**
967 **detectable expression of the genes.** The number of genotypes that a gene was expressed in was
968 determined by counting the number of genotypes with non-zero expression for each gene. The gene

969 expression ranks were calculated by ranking the mean gene expression values where the mean was
970 calculated only considering samples with non-zero expression.

971 **S4 Fig. Distribution of local eSNPs relative to the gene that they are associated to.** The zero-
972 position represents both the start and the end of the gene feature, i.e. intragenic features are not
973 shown.

974 **S5 Fig. Heritability distributions for four non-overlapping sets of genes:** genes with both local
975 and distant eQTLs, genes with only distant eQTLs, genes with only local eQTLs, and genes with no
976 significant eQTLs. The box widths are proportional to the number of genes in each set.

977 **S6 Fig. eSNPs per eGene and eGenes per eSNP.** (A) Bar plot of the number of associated
978 eGenes per eSNP. (B) Histogram of the number of associated eSNPs per eGene.

979 **S7 Fig. Sliding window approach to detect eQTL hotspots.** The upper panel shows the number
980 of eSNPs in genomic windows of 100 kb for each of the 19 chromosomes. The lower panel shows
981 the number of unique genes that are associated to each genomic window (nominal eQTL p-value <
982 1×10^{-6}).

983 **S8 Fig. Gene expression variance plotted against co-expression network connectivity.** eGenes
984 are indicated by red points.

985 **S9 Fig. Minor allele frequency of eSNPs.** (A) Comparison of minor allele frequency between all
986 SNPs, distant eSNPs and local eSNPs. (B) The relationship between minor allele frequency and
987 effect size (absolute value of beta) of eQTLs.

988 **S10 Fig. Measures of diversity and selection for 0-fold and 4-fold degenerate sites.** eGenes and
989 non-eGenes were contrasted, as well as genes in network module cores and those not. Panel (A) and

990 (B) show values of mean nucleotide diversity (θ_π) and Tajima's D for genes in each categories,
991 respectively. Significance of a Mann-Whitney test comparing each pair of gene categories is indicated
992 by asterisks above the boxplot: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, n.s. $p > 0.05$.

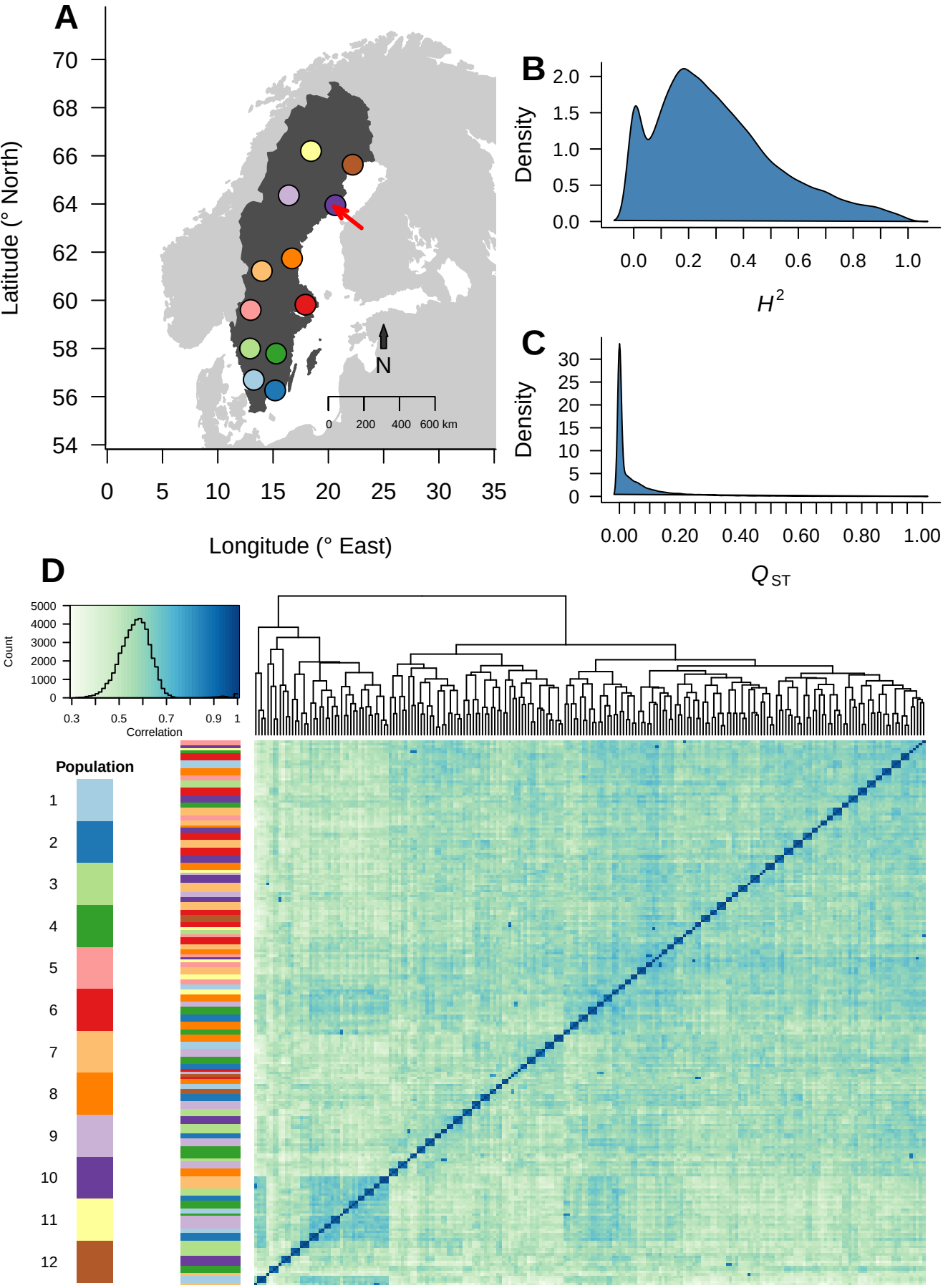
993 **S11 Fig. Heatmap representing the sample clustering based on gene expression values after**
994 **hidden confounder removal.** The 500 most variable genes in the original expression data were used
995 for calculating the sample correlations (i.e. the same genes as in **Fig 1**D). The colour bar represents
996 the population of the samples with the same colour scheme as in **Fig 1**D. The small clusters on the
997 diagonal represent genotype replicates.

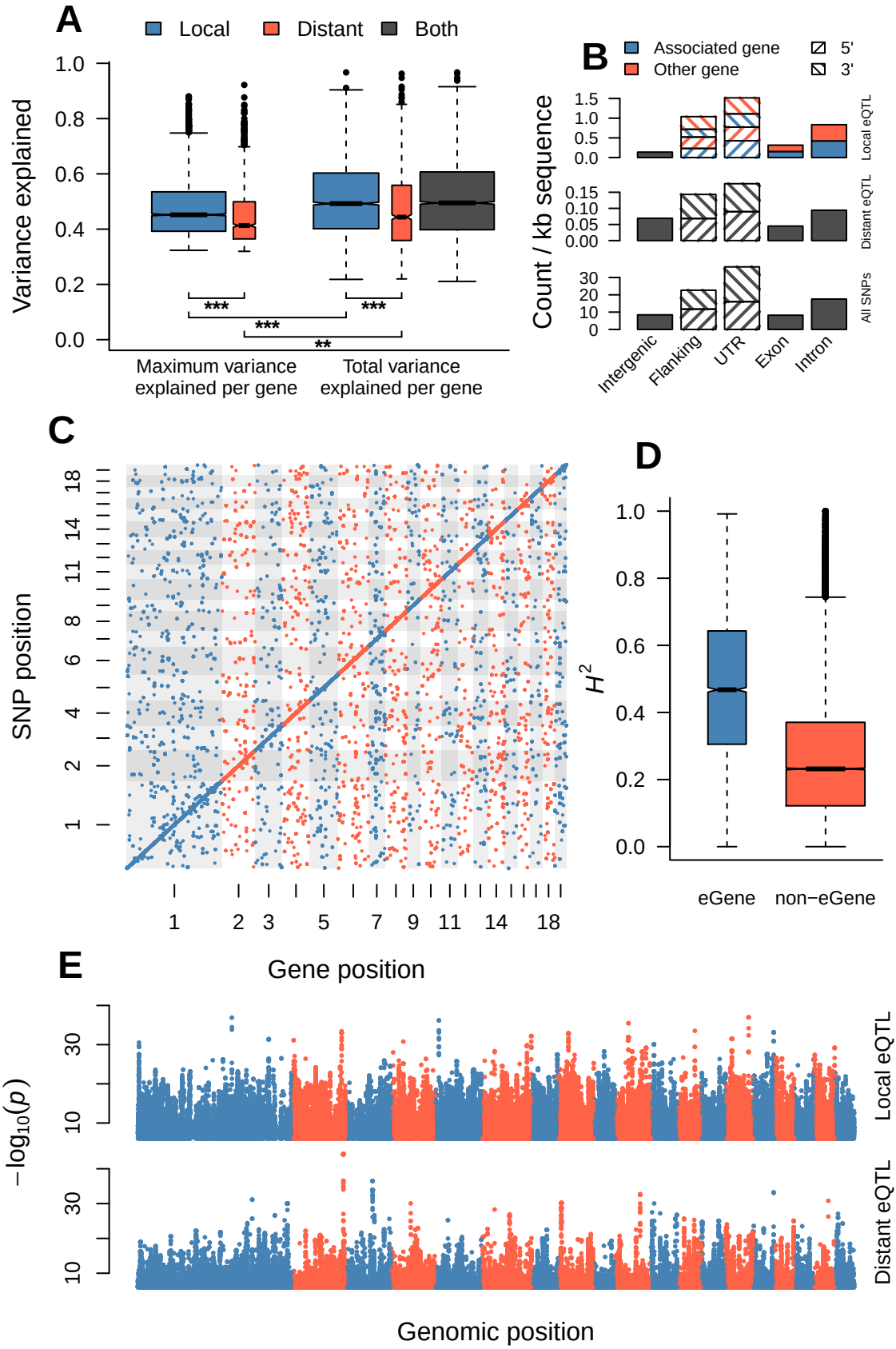
998 **S12 Fig. Representative photo of the sampled bud flush stage.**

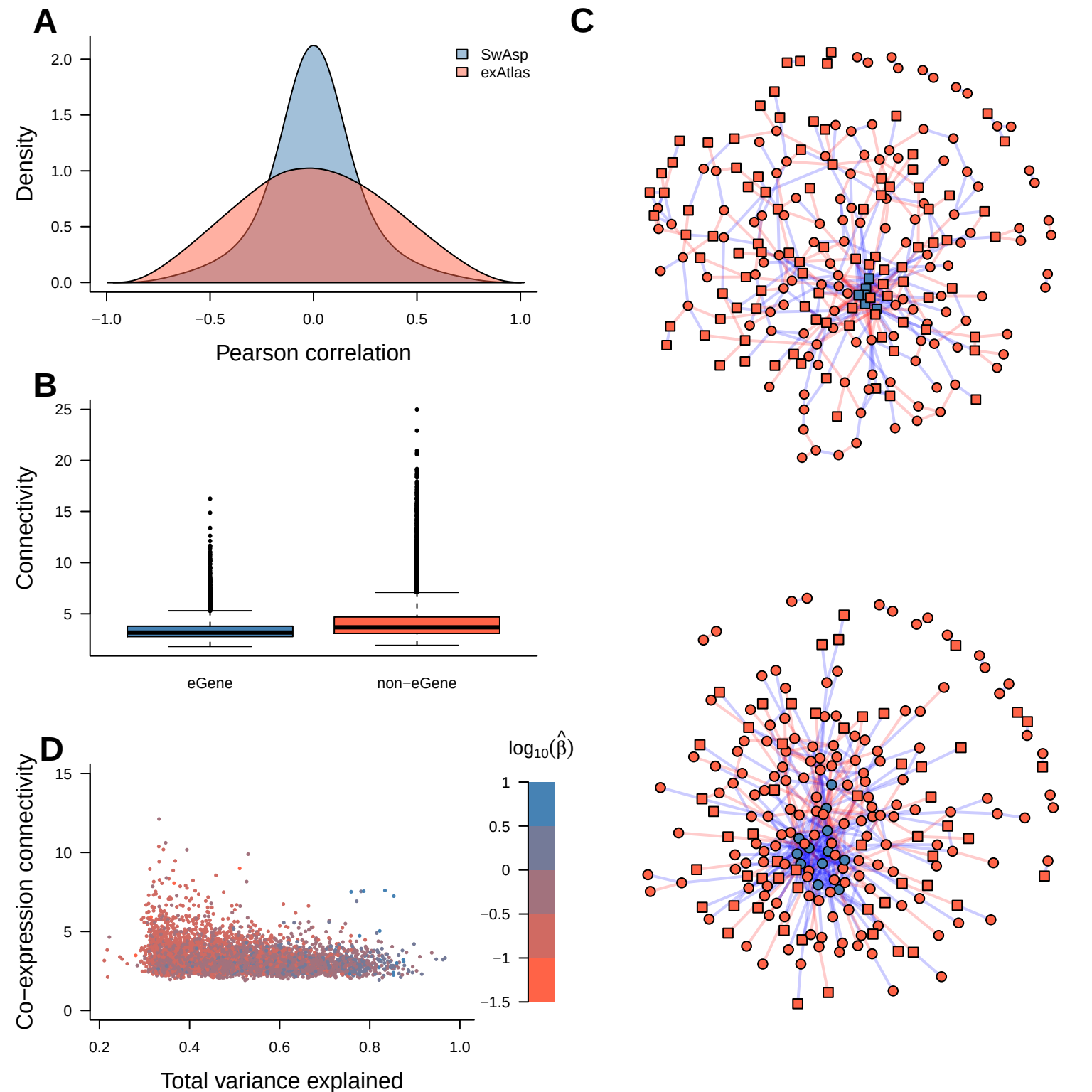
999 **S13 Fig. Sampling time and day of bud flush.** (A) Box plot distributions of the Julian day of
1000 sampling for the SwAsp sub-populations. (B) The relationship between Julian day of bud flush for
1001 the greenhouse sampled buds and Julian day of bud flush in the field for the same year (2012).

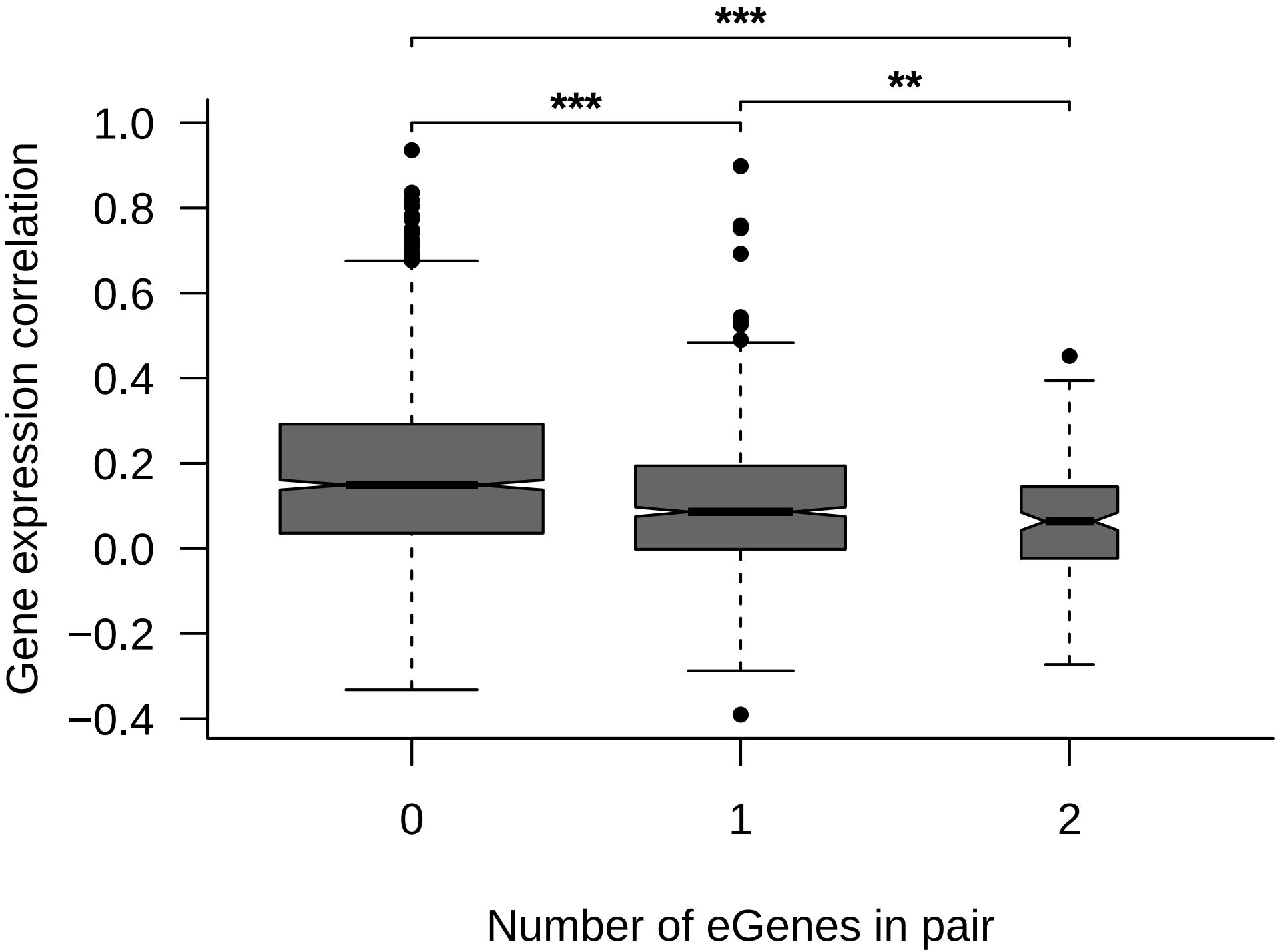
1002 **S14 Fig. The number of eQTLs detected for different levels of hidden confounder removal.**
1003 The y-axis shows the number of eQTLs detected at 5% FDR (prior to the empirical FDR
1004 calculation). The x-axis shows the number of principal components regressed out of the gene
1005 expression data.

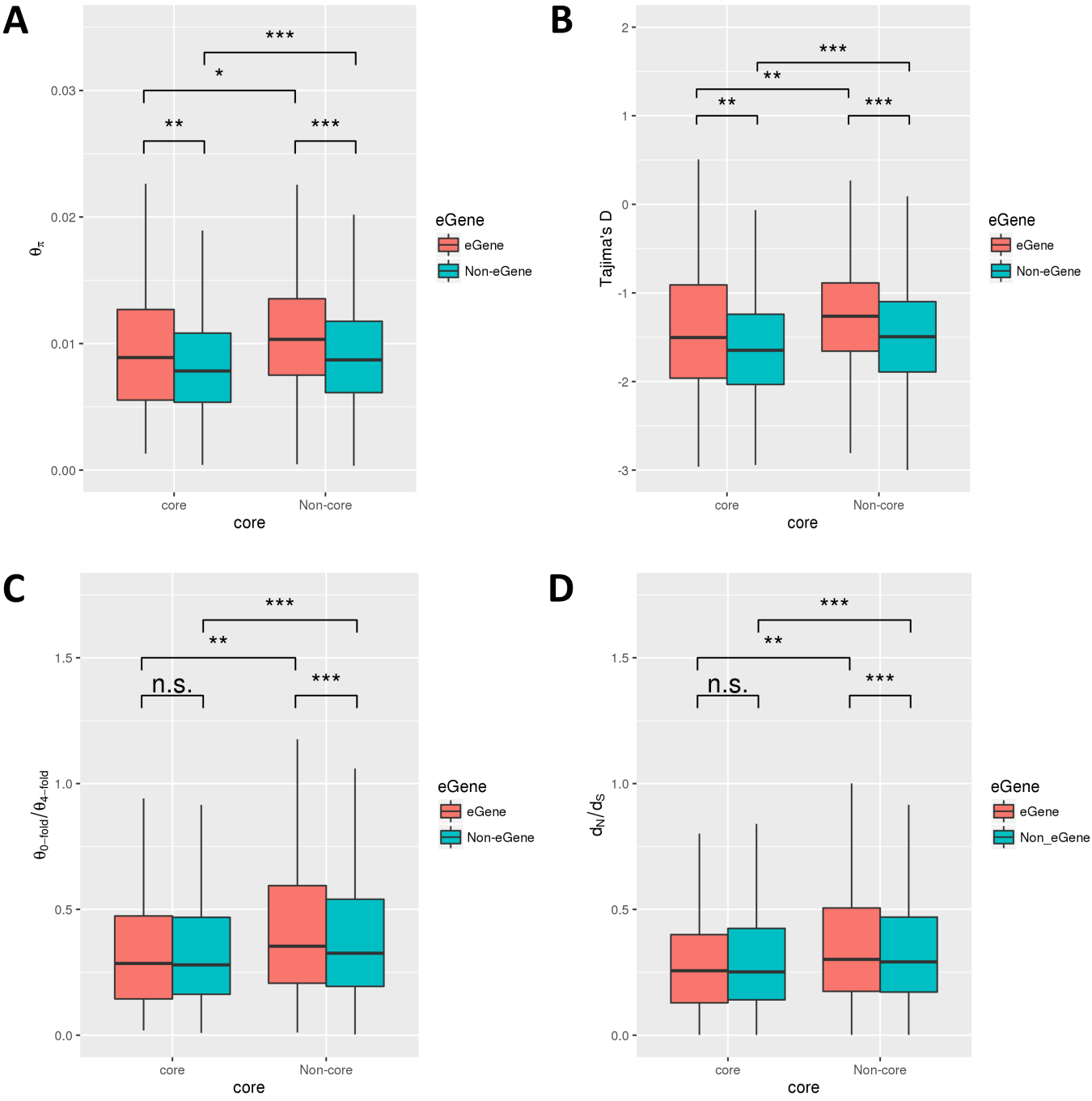
1006 **S15 Fig. Motivating example for major genotype filtering.** In this case the SNP is heterozygous
1007 in all samples but one, resulting in a minor allele frequency close to 0.5, but a major genotype
1008 frequency close to one. The resulting association turns out very significant, but is only supported by
1009 a single sample.

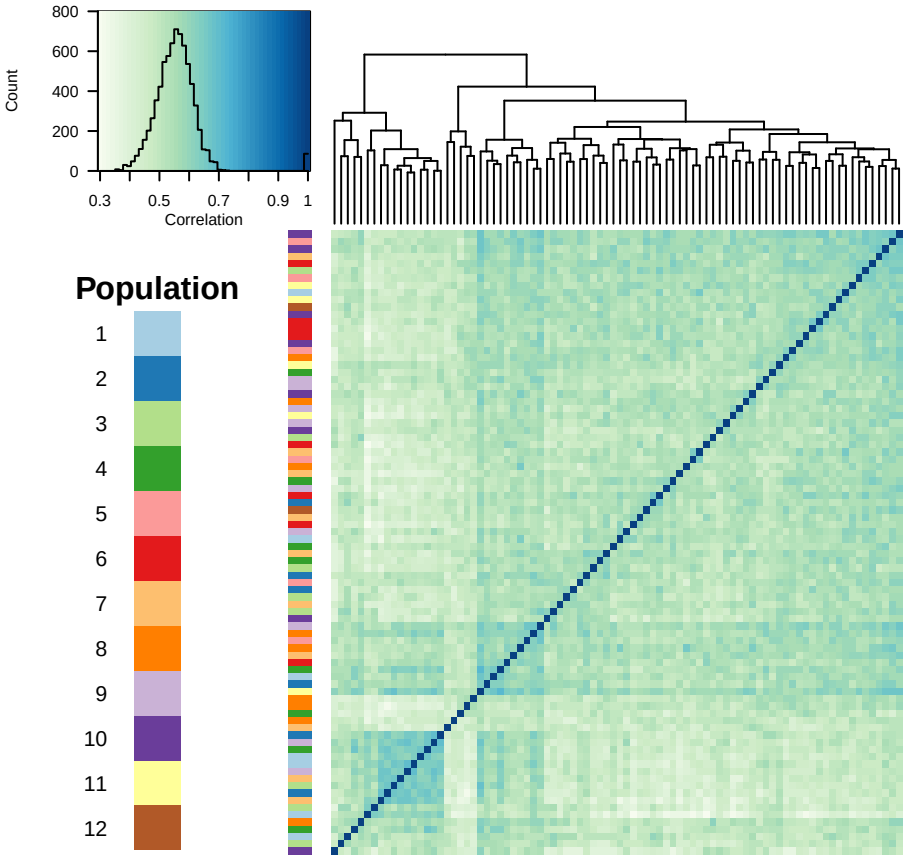

A
Latitude (° North)
Longitude (° East)
N
0 200 400 600 km
B
Density
H^2
C
Density
Q_{ST}
D
Count
Correlation
Population

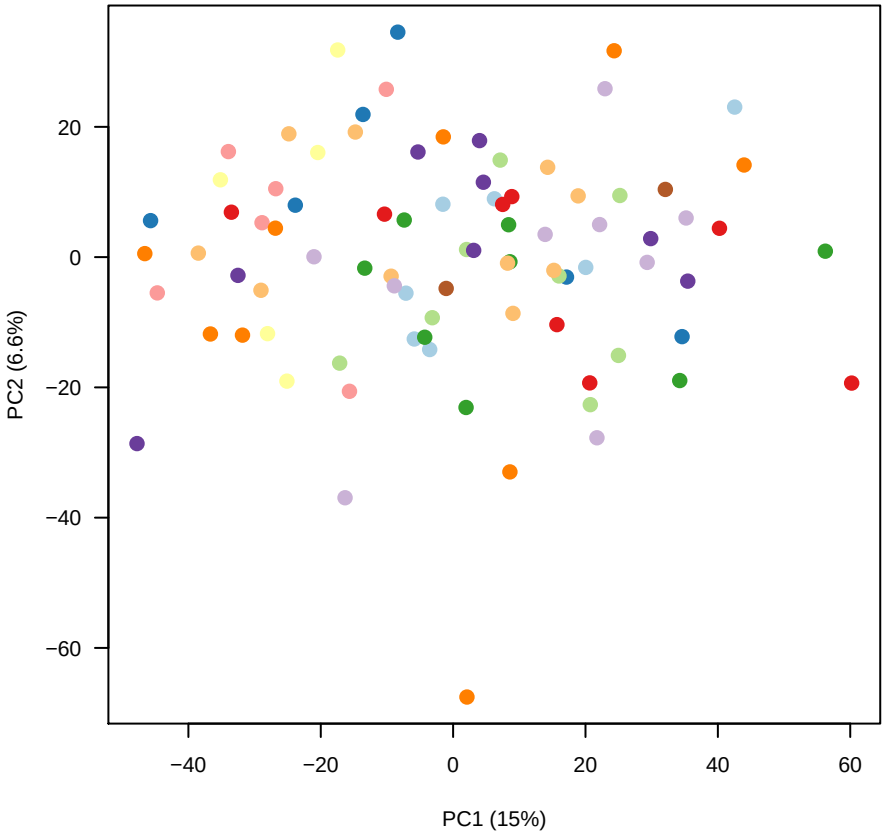

A
Local
Distant
Both
Variance explained
1.0
0.8
0.6
0.4
0.2
0.0
***
***
***
**
Maximum variance explained per gene
Total variance explained per gene
B
Associated gene
Other gene
5'
3'
Count / kb sequence
Local eQTL
Distant eQTL
All SNPs
Intergenic
Flanking
UTR
Exon
Intron
C
SNP position
Gene position
D
H^2
eGene
non−eGene
E
$-\log_{10}(p)$
Local eQTL
Distant eQTL
Genomic position

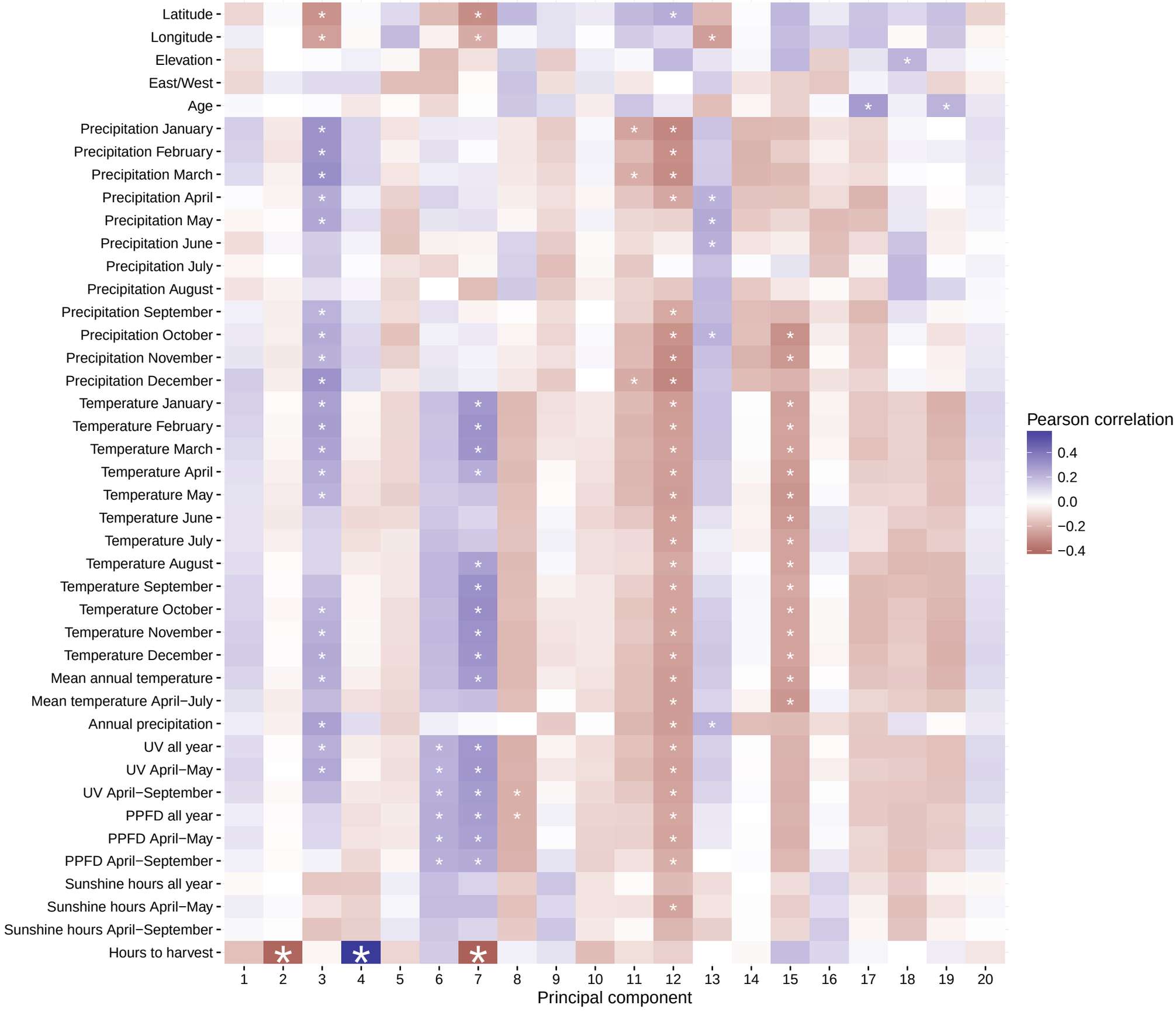

A
SwAsp
exAtlas
Density
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
Pearson correlation
B
Connectivity
25
20
15
10
5
eGene
non−eGene
C
D
Co−expression connectivity
15
10
5
0
0.2
0.4
0.6
0.8
1.0
Total variance explained
$\log_{10}(\hat{\beta})$
1
0.5
0
−0.5
−1
−1.5

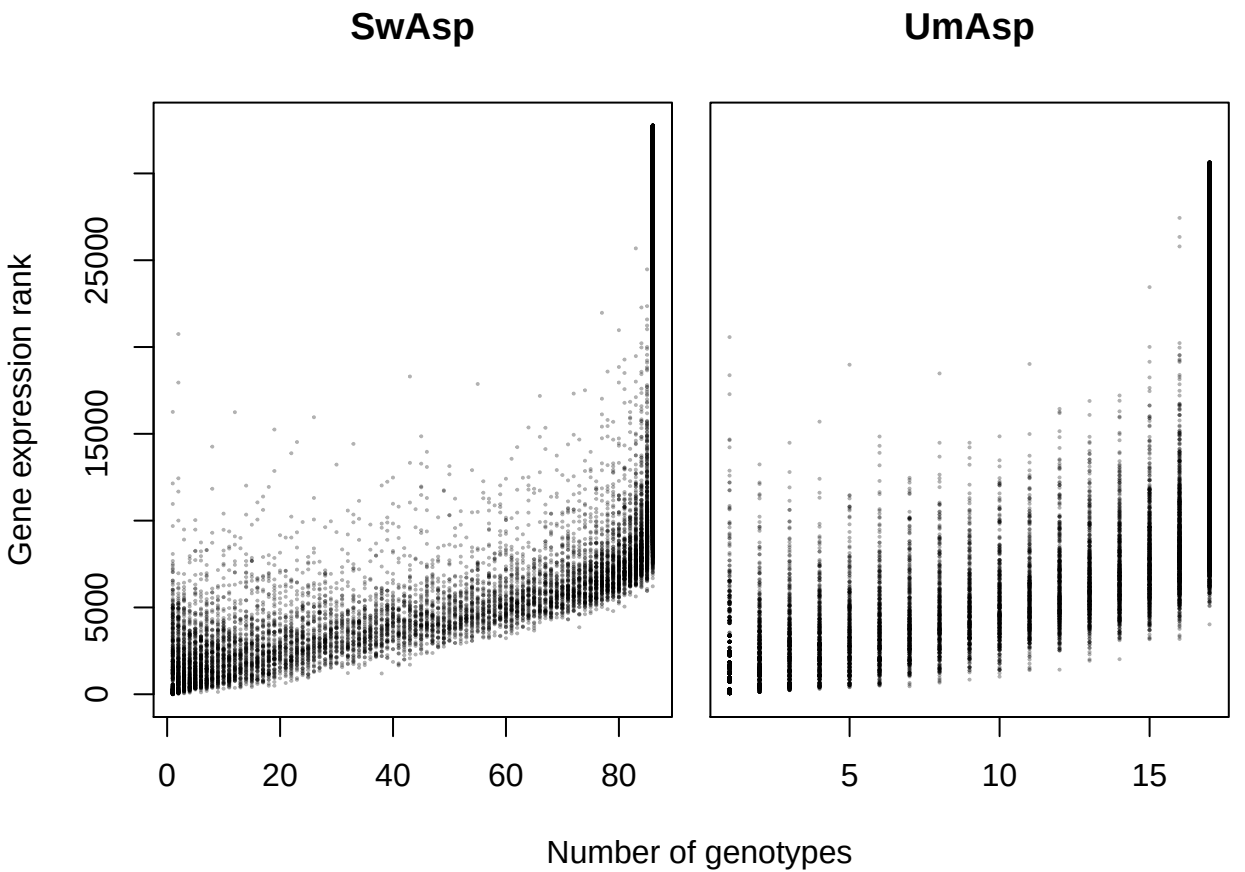

***
***
**
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Gene expression correlation
0
1
2
Number of eGenes in pair

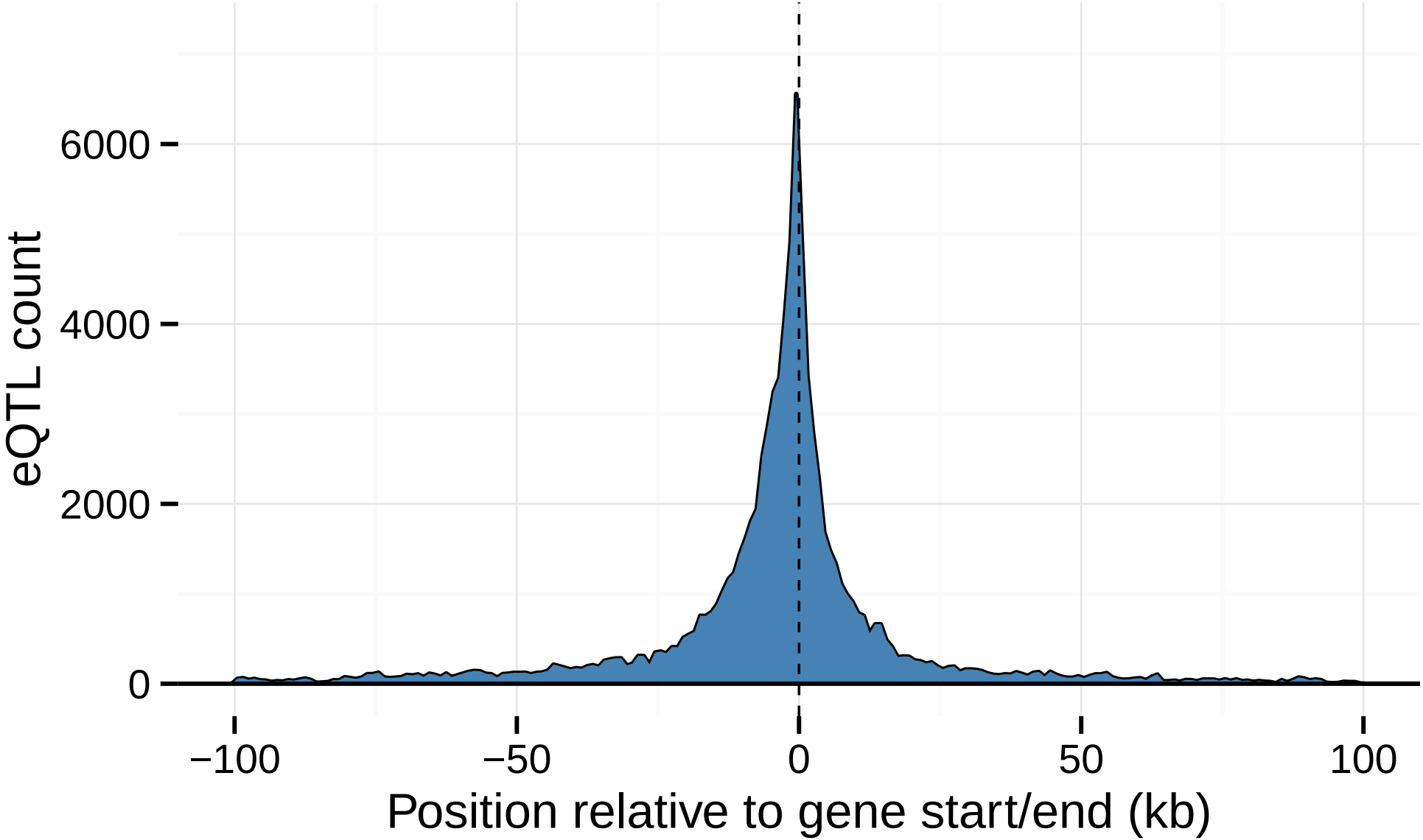

A
B
C
D
θπ
Tajima's D
θ0-fold/θ4-fold
dN/dS
core
Non-core
eGene
Non-eGene
Non_eGene
n.s.
*
**
***

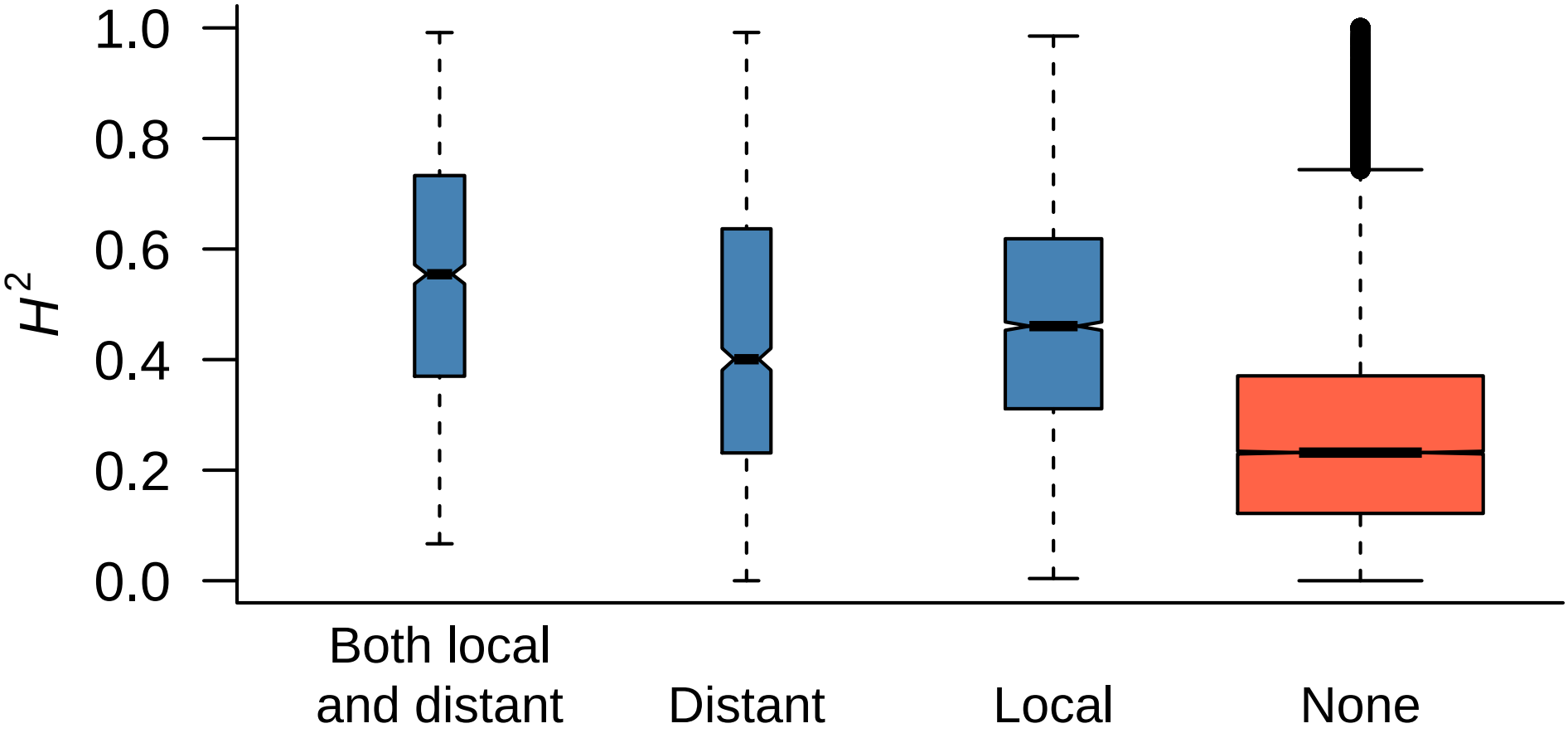

Count
800
600
400
200
0
0.3
0.5
0.7
0.9
1
Correlation
Population
1
2
3
4
5
6
7
8
9
10
11
12

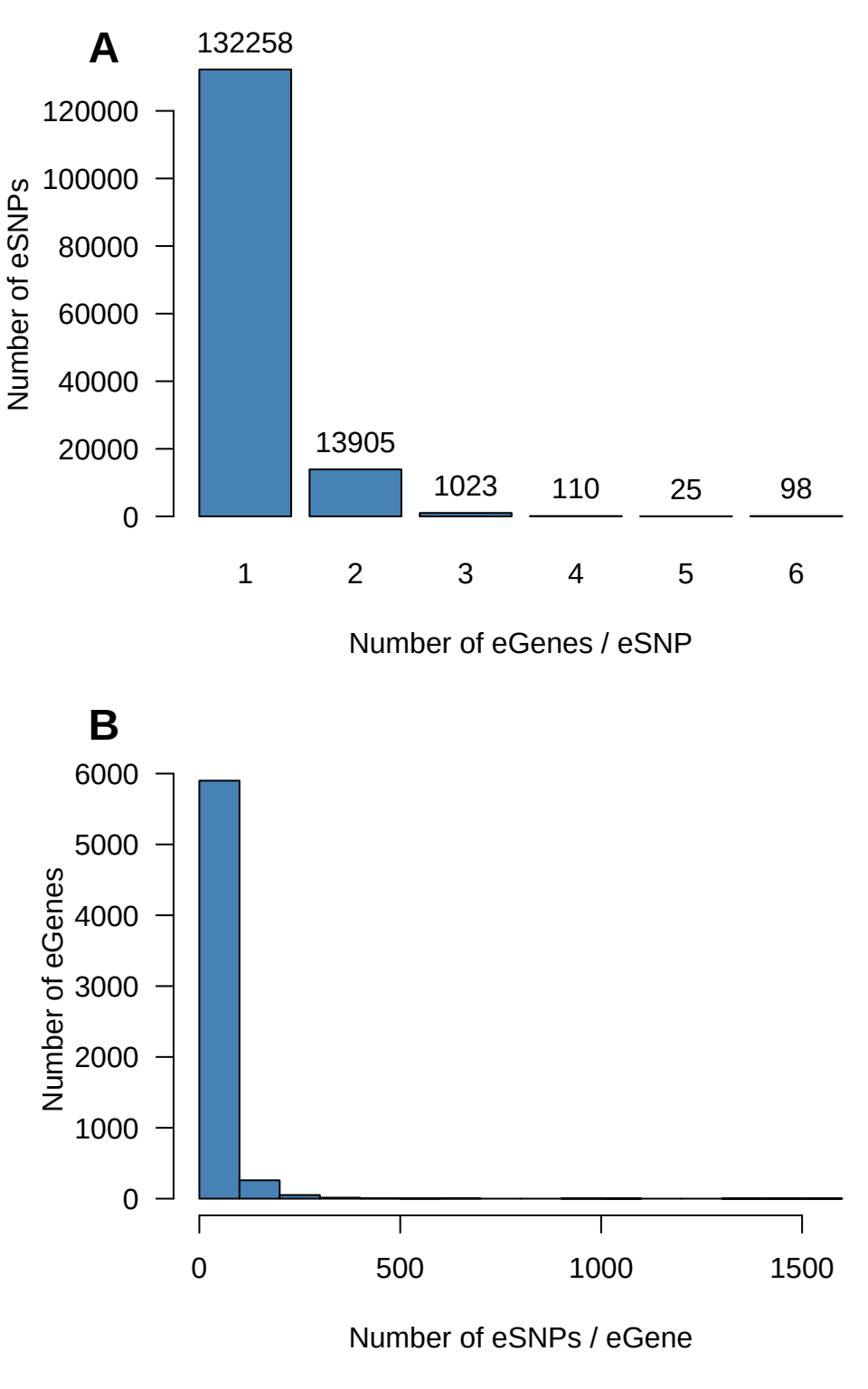

PC2 (6.6%)
20
0
−20
−40
−60
−40
−20
0
20
40
60
PC1 (15%)

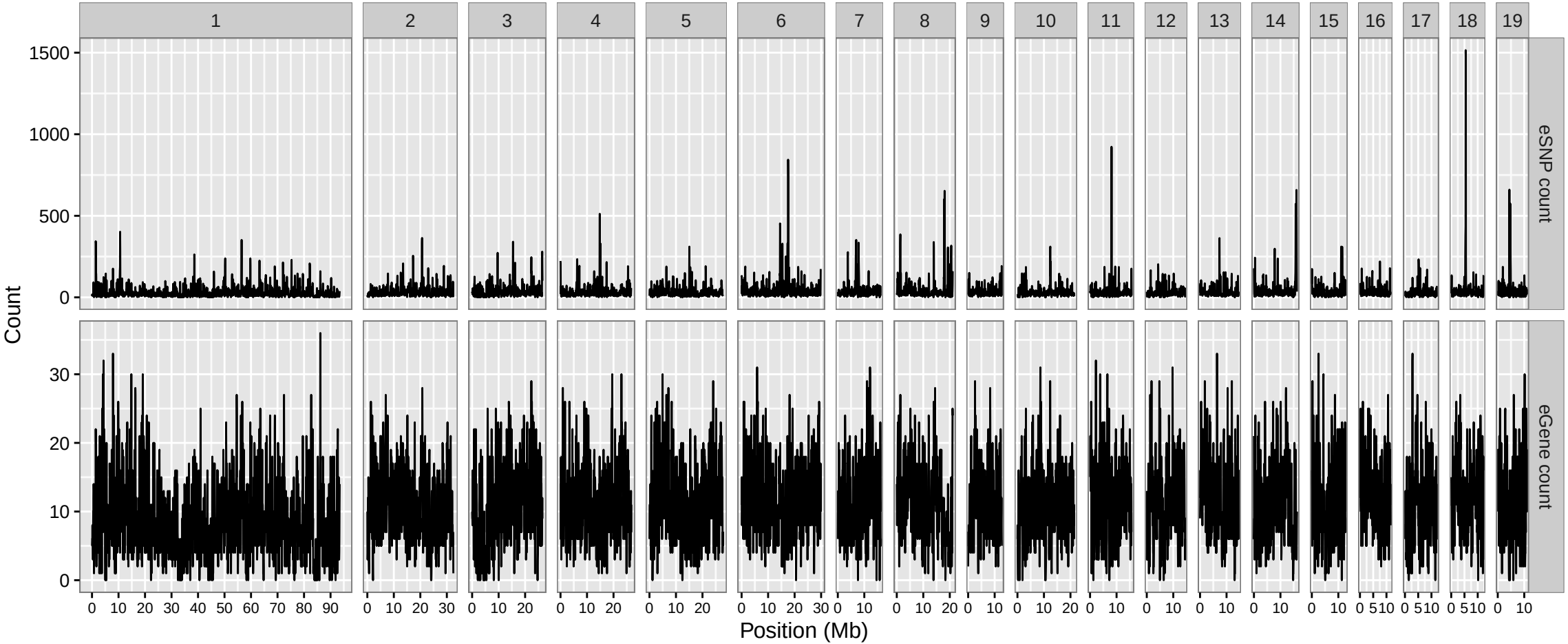

Latitude
Longitude
Elevation
East/West
Age
Precipitation January
Precipitation February
Precipitation March
Precipitation April
Precipitation May
Precipitation June
Precipitation July
Precipitation August
Precipitation September
Precipitation October
Precipitation November
Precipitation December
Temperature January
Temperature February
Temperature March
Temperature April
Temperature May
Temperature June
Temperature July
Temperature August
Temperature September
Temperature October
Temperature November
Temperature December
Mean annual temperature
Mean temperature April−July
Annual precipitation
UV all year
UV April−May
UV April−September
PPFD all year
PPFD April−May
PPFD April−September
Sunshine hours all year
Sunshine hours April−May
Sunshine hours April−September
Hours to harvest
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Principal component
Pearson correlation
0.4
0.2
0.0
−0.2
−0.4

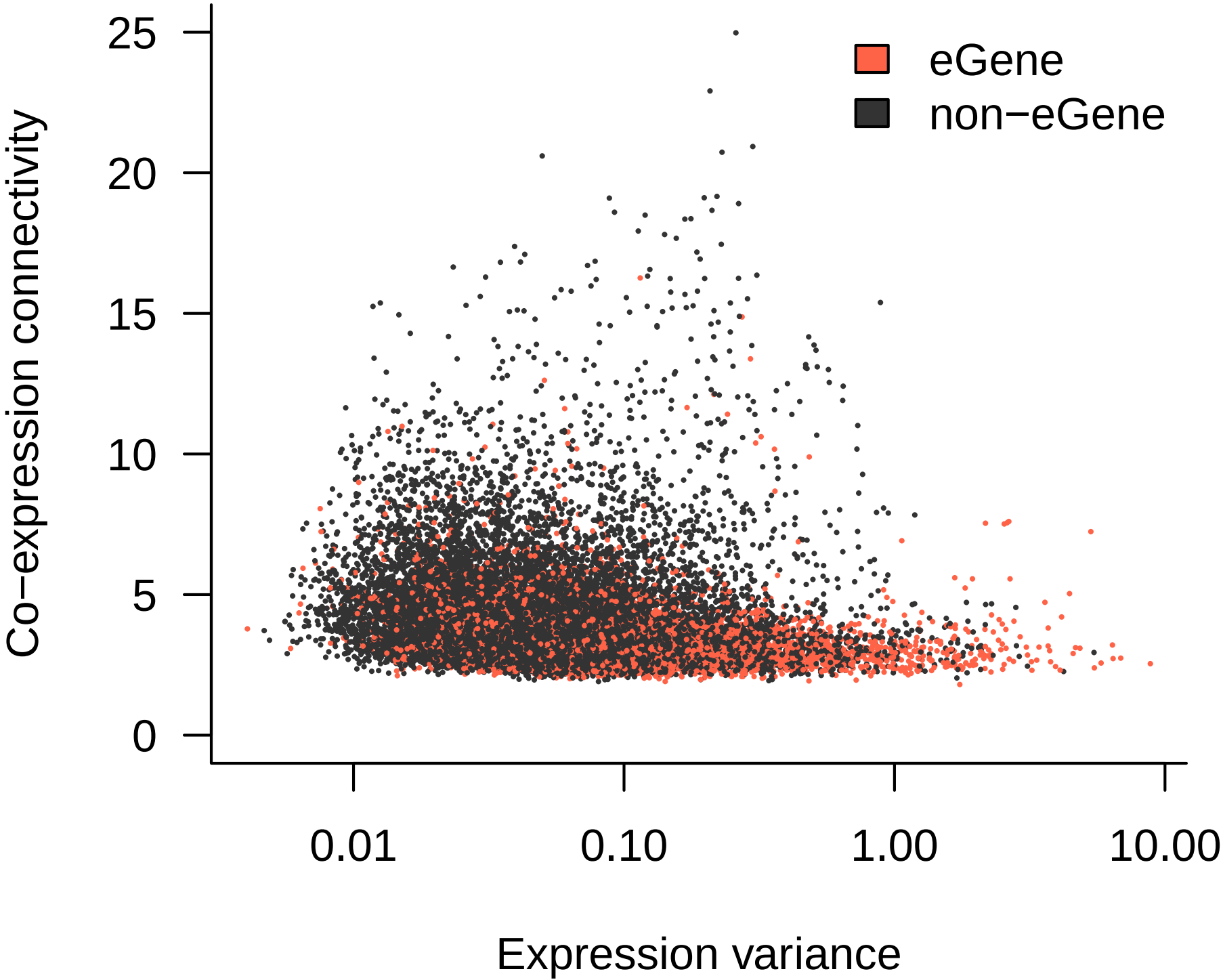

SwAsp
UmAsp
Gene expression rank
0
5000
15000
25000
0
20
40
60
80
5
10
15
Number of genotypes

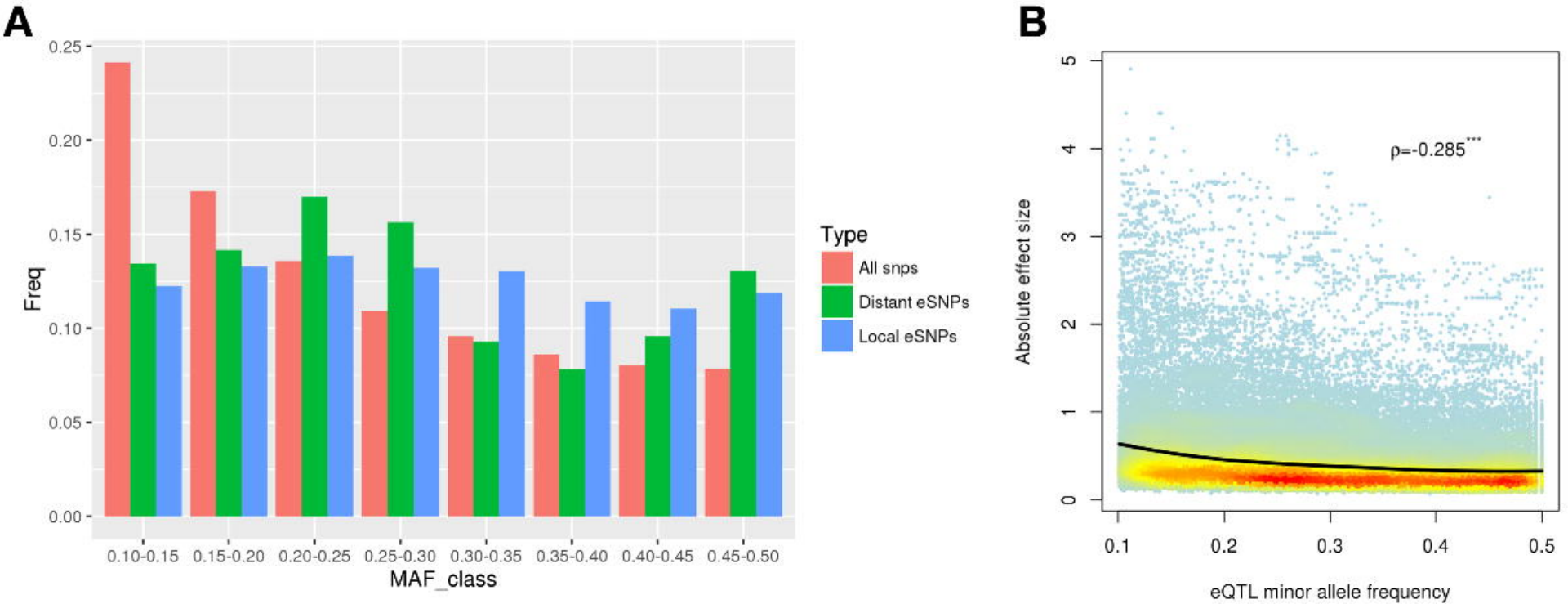

eQTL count
6000
4000
2000
0
−100
−50
0
50
100
Position relative to gene start/end (kb)

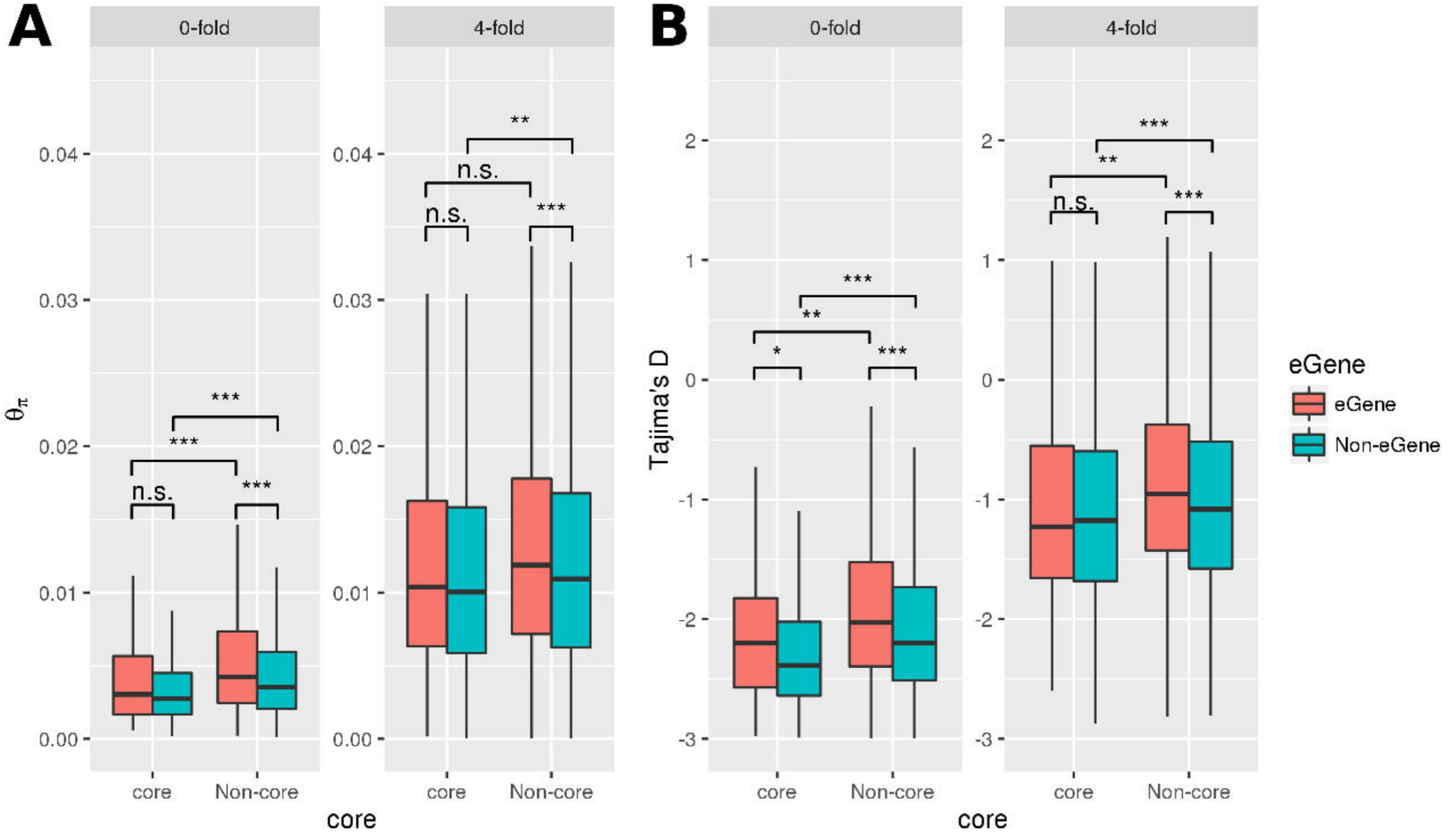

1.0
0.8
0.6
0.4
0.2
0.0
H^2
Both local and distant
Distant
Local
None

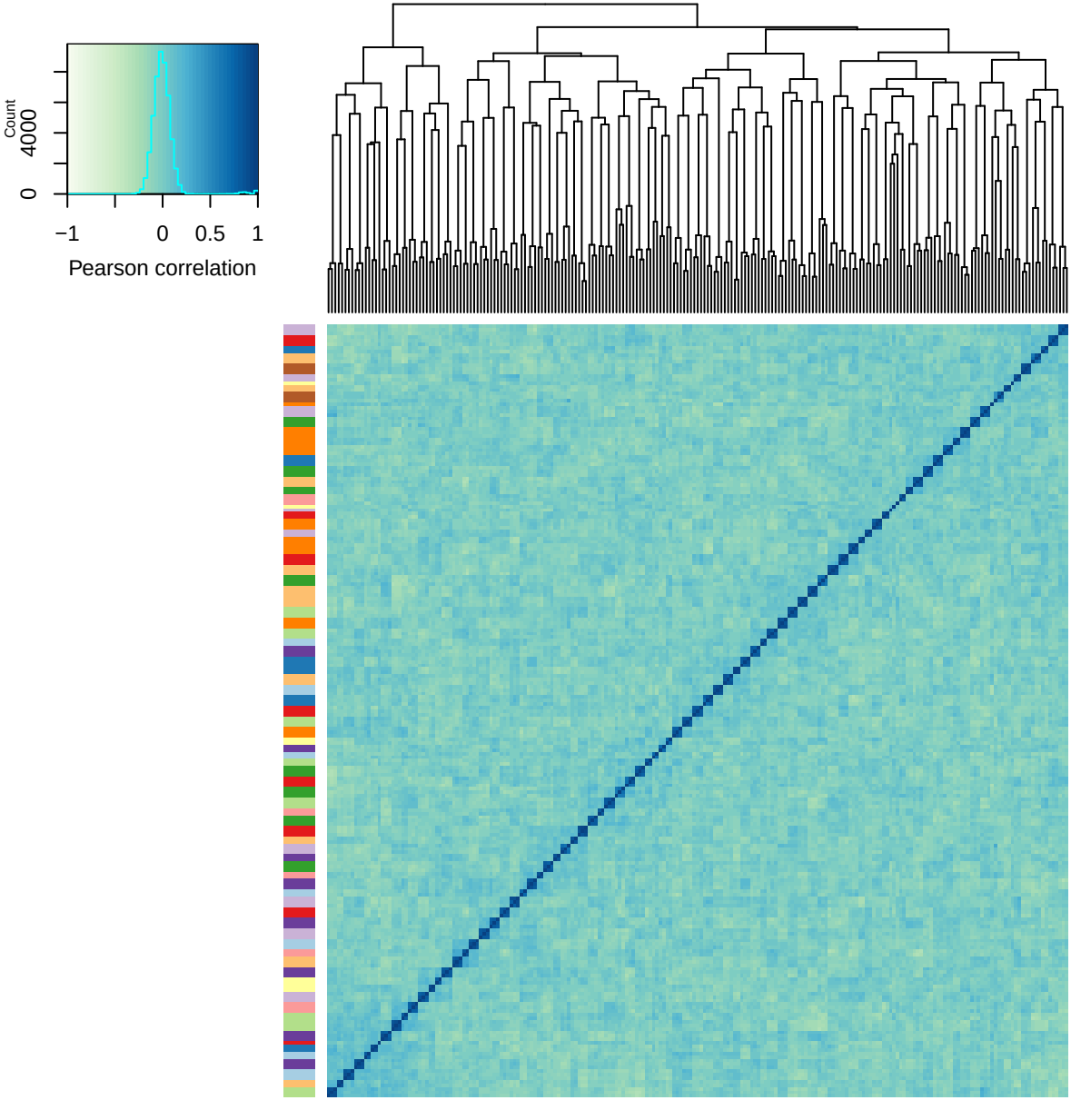

A
132258
13905
1023
110
25
98
120000
100000
80000
60000
40000
20000
0
Number of eSNPs
1
2
3
4
5
6
Number of eGenes / eSNP
B
6000
5000
4000
3000
2000
1000
0
Number of eGenes
0
500
1000
1500
Number of eSNPs / eGene


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
eSNP count
eGene count
Count
Position (Mb)

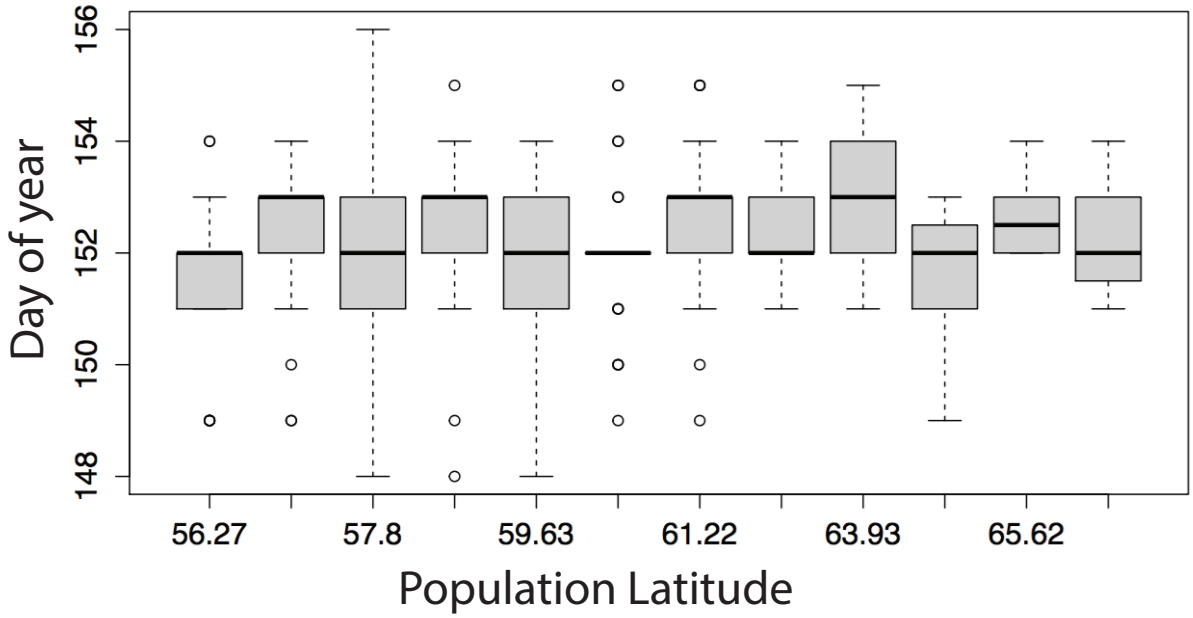

eGene
non−eGene
Co−expression connectivity
25
20
15
10
5
0
0.01
0.10
1.00
10.00
Expression variance


A
0.25
0.20
0.15
0.10
0.05
0.00
Freq
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
MAF_class
Type
All snps
Distant eSNPs
Local eSNPs
B
ρ=-0.285***
Absolute effect size
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
eQTL minor allele frequency

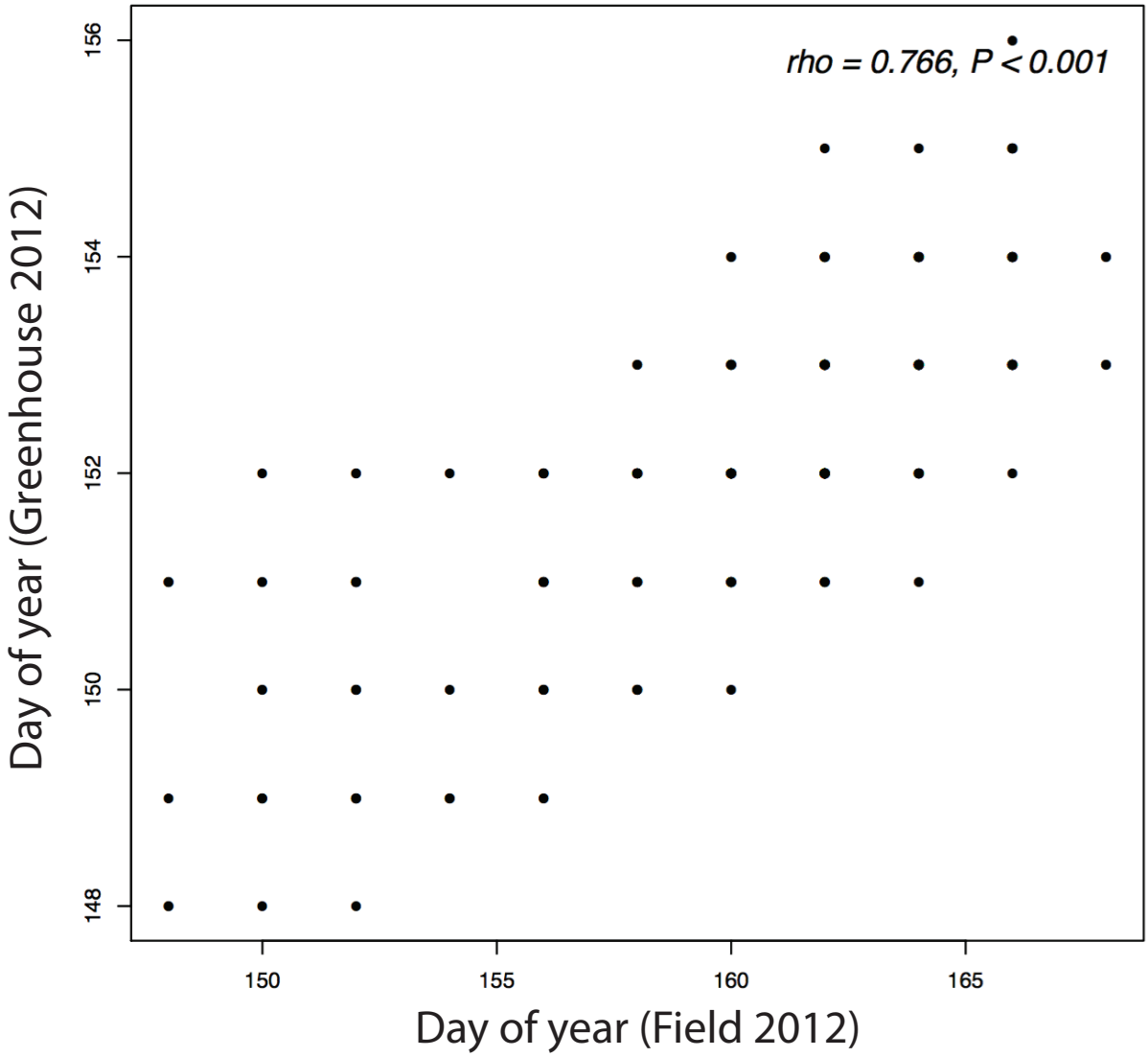

A
0-fold
4-fold
θ_π
0.04
0.03
0.02
0.01
0.00
***
***
n.s.
***
**
n.s.
n.s.
***
core
Non-core
core
B
0-fold
4-fold
Tajima's D
2
1
0
-1
-2
-3
***
**
*
***
***
**
n.s.
***
core
Non-core
core
eGene
eGene
Non-eGene

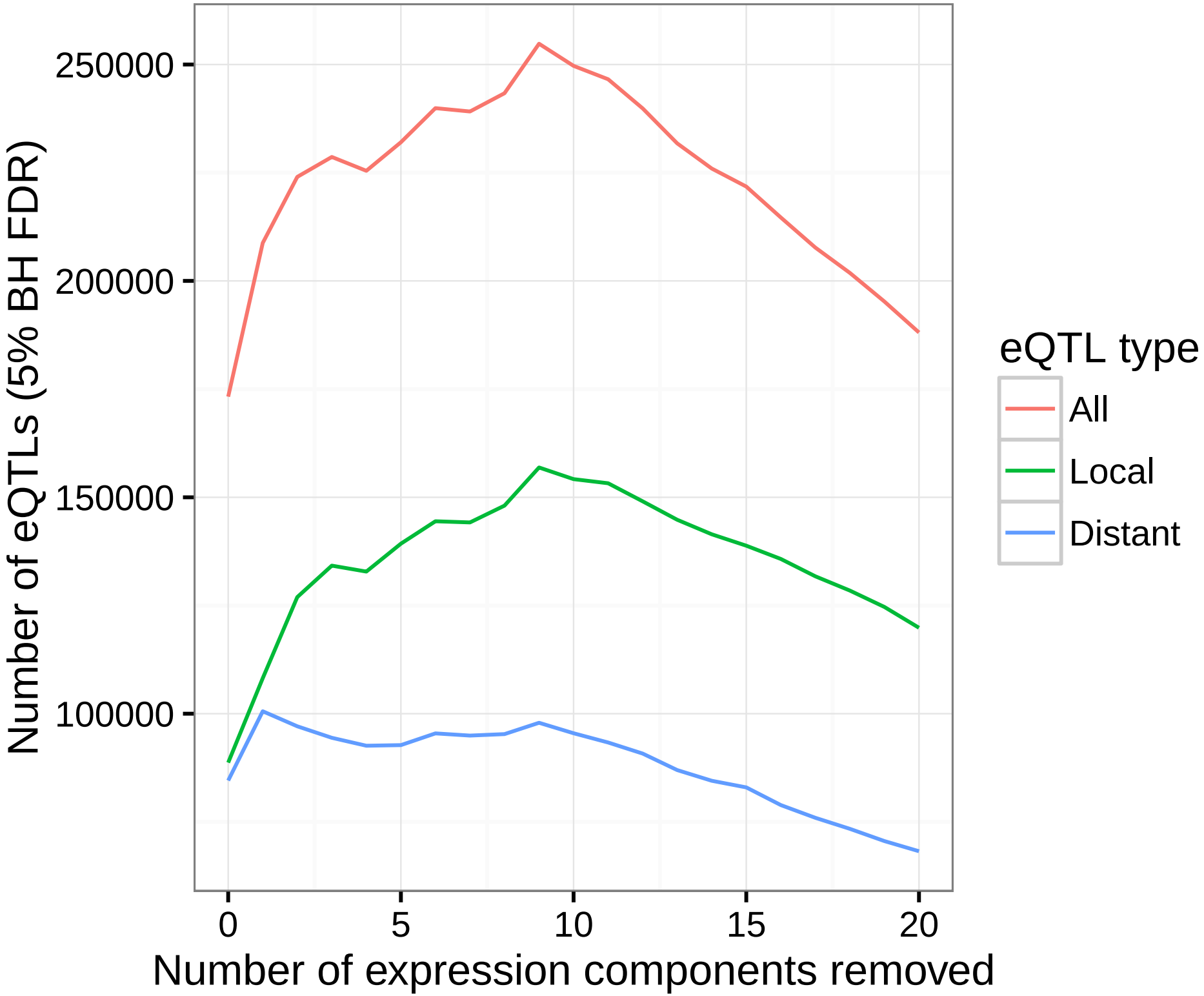

Count
4000
0
−1
0
0.5
1
Pearson correlation

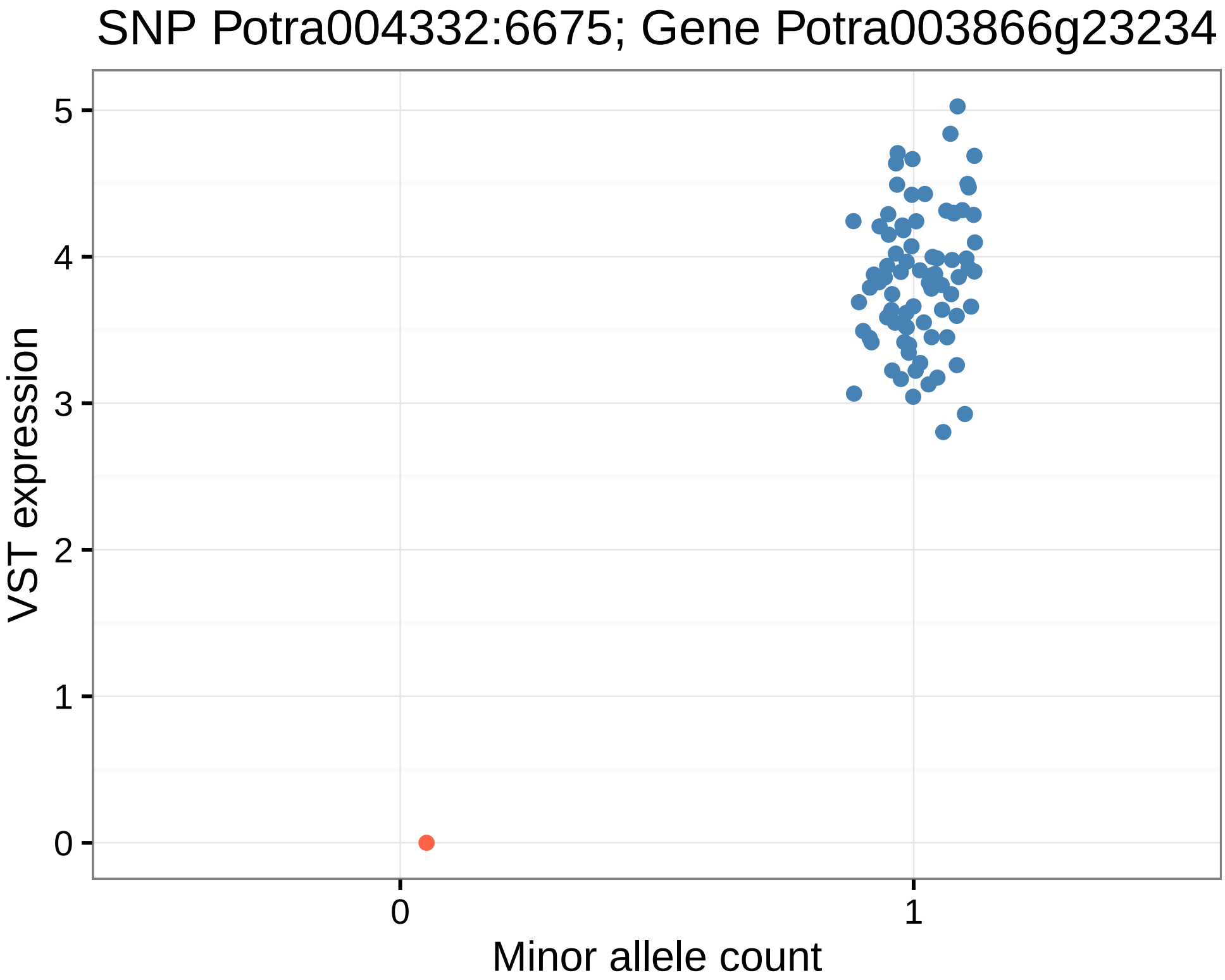
A


Day of year
148
150
152
154
156
56.27
57.8
59.63
61.22
63.93
65.62
Population Latitude

B


rho = 0.766, P < 0.001
Day of year (Greenhouse 2012)
148
150
152
154
156
150
155
160
165
Day of year (Field 2012)


250000
200000
150000
100000
Number of eQTLs (5% BH FDR)
0
5
10
15
20
Number of expression components removed
eQTL type
All
Local
Distant


SNP Potra004332:6675; Gene Potra003866g23234
VST expression
5
4
3
2
1
0
0
1
Minor allele count