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, *]

[*]Equal contribution

[†]To whom correspondence should be addressed.

[1]Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1430 Ås, Norway

[2]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, 901 87 Umeå, Sweden

[3]Umeå Plant Science Centre, Department of Ecology and Environmental Science, Umeå University, 901 87 Umeå, Sweden

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.

Introduction

A central aim of biology is to understand how emergent phenotypes are encoded in the genome and how genetic variation engenders phenotypic variation within populations. While the biological function of many genes can be ascertained through use of mutant screens, such approaches tend to identify genes essential to a biological process. There is no clear consensus as to whether variation in these genes underlies natural variation in phenotypic traits, how this relates to the evolutionary history of the traits or to the properties of the associated transcriptional networks. Additionally, many traits are complex, being controlled by the action and interaction of numerous genes, often rendering mutational approaches inadequate or infeasible. In such cases, turning to nature and

42 employing population genetics and systems biology approaches (systems genetics) to uncover genetic
43 and genomic co-variation between assayed traits (genomic markers, gene expression, protein
44 abundance, metabolites etc.) and 'classical' phenotypes (physiognomy, morphology, physiology)
45 represents an appealing alternative.

46 The availability of massively parallel sequencing technologies affords new possibilities for addressing
47 biological questions, for example enabling the generation of *de novo* genome assemblies and
48 population-wide resequencing data that can be used to perform genome-wide association studies
49 (GWAS), even in species with large genomes that harbour high levels of polymorphism or that
50 display rapid linkage disequilibrium (LD) decay [1]. The use of genome-wide resequencing data
51 allows the discovery of, in theory, all genetic polymorphisms within an individual. These genetic
52 markers, of which single nucleotide polymorphisms (SNPs) are currently the most commonly
53 considered, can then be used to perform association or linkage mapping to identify the subset of
54 polymorphisms engendering phenotypic variation among individuals. While the majority of SNPs are
55 expected to have no, or insignificant, consequence for a given trait, an implication of the infinitesimal
56 model [2] is the expectation that the control of variation for any given trait will be contributed to by
57 a large number of loci, each of small effect size. It is not clear how this holds across the scale of
58 molecular to complex, integrative phenotypes (*e.g.* from transcription to morphology) or whether this
59 pattern contrasts between adaptive and selectively neutral phenotypic trait variation.

60 Advances in sequencing technologies have concordantly revolutionised transcriptomics studies,
61 particularly in non-model organisms. Following seminal work [3,4], numerous early studies in a range
62 of species established that there is a significant heritable component underlying natural variation of
63 gene expression levels among individuals within populations [5–17] and that this variation underlies a
64 number of phenotypes [18–25]. Given these findings, it became apparent that gene expression values

65 could be considered in the same way as any other quantitative phenotype and be subjected to linkage
66 or association mapping to identify polymorphisms contributing to expression level variation among
67 individuals [26], as first reported in [20], with the identified loci termed expression Quantitative Trait
68 Locus (eQTL; [7]) or, less commonly, expression level polymorphisms (ELPs; [27]). eQTLs are
69 classified as either local or distant acting depending on the physical location of the associated
70 polymorphism in relation to the gene that the eQTL is mapped for: local eQTLs are usually defined
71 as being located within a specified physical distance of the gene location (typically up to 2 Mbp,
72 although this varies depending on species) while distant eQTLs represent polymorphisms that are
73 located beyond that threshold distance or on another chromosome. eQTLs can further be classified
74 as acting in *cis* or *trans*: *cis* eQTLs act in an allele-specific manner and are usually considered to be
75 local, although long-range *cis* interactions can occur, for example when a polymorphism is located in
76 an enhancer that is physically distant from the gene of interest; *trans* acting eQTLs affect both alleles
77 of a gene and are most commonly located distant to that gene. There continues to be strong interest
78 in eQTLs as they can identify mechanistic links between phenotype and genotype [28,29].
79 Importantly, the majority of polymorphisms that have been associated to phenotypes using GWAS
80 in a wide range of species are located outside of protein coding or transcribed regions [30–33],
81 suggesting that they influence expression rather than altering protein or transcript function.

82 A number of previous eQTL studies have been conducted using plant species including *Arabidopsis*
83 *thaliana* [34–39], *Zea mays* [40–42] and *Oryza sativa* [43,44], and in forest tree species [8,45–47]. These,
84 together with studies in other eukaryotic systems, have yielded generalities concerning the genetic
85 architecture of gene expression variation, including that a greater number of local eQTLs are
86 typically identified and that these individually explain a larger proportion of gene expression variance
87 than do distant eQTLs [35,48–51]. With the exception of human studies, the majority of previous
88 work was conducted using controlled, often inter-specific, crosses and it is not clear how generally

89 applicable their conclusions are for natural populations of unrelated individuals. Few studies have
90 considered whether observed, heritable variation is adaptive [52,53] or whether signatures of
91 population differentiation are observed at the transcriptome level [53–58], and there is a lack of
92 consensus as to whether natural variation in gene expression is selectively neutral [53].

93 Species in the *Populus* genus have been established as a powerful model system for forest tree
94 genomics due to their relatively small genome, rapid growth, propensity for clonal propagation and
95 ease of genetic transformation [59]. *P. tremula* (European aspen) has many features that render it a
96 particularly useful model for population genetics and speciation studies [60,61]. To facilitate use of *P.*
97 *tremula* as a model system we have produced a draft *de novo* genome assembly of the *P. tremula*
98 genome (available at [62]) and resequencing data [63] for all individuals comprising the Swedish
99 Aspen (SwAsp) collection [64].

100 Here, we utilise this resource in combination with population-wide RNA-Seq data, assaying gene
101 expression in winter buds at the point of bud break. We performed eQTL mapping and constructed
102 a co-expression network, with the results integrated to provide insight into the genetic architecture of
103 natural variation in gene expression levels. Our analyses revealed a scale free and modular co-
104 expression network, with highly connected genes being under-represented for eQTLs and with
105 eQTL effect size being negatively correlated with both gene connectivity and minor allele frequency.
106 Patterns of polymorphism and divergence within highly connected genes demonstrate that these
107 experience stronger selective constraint relative to genes in the network periphery. Our results
108 therefore highlight the importance of purifying selection for explaining observed standing variation
109 in gene expression in natural populations in addition to explaining how buffering of the
110 transcriptional network may be achieved. We propose that network modules can be explained by

111 eQTLs in one or a few highly connected module genes and that these are evolving under strong
112 functional constraint compared to genes in the network periphery.

113 Results

114 Population level gene expression similarity

115 We utilised the northern common garden (located at 63.9° N, near Umeå, Sweden) of the SwAsp
116 collection [64], which comprises 116 *P. tremula* genotypes sampled across the species distribution
117 range in Sweden (56.2° to 66.4° N, Fig 1A). The SwAsp collection was previously shown to contain
118 abundant genetic variation, with low linkage disequilibrium (LD) [61,65] and minimal population
119 structure [66]. Based on whole genome re-sequencing data that were aligned to a *de novo* assembly of
120 the *P. tremula* genome (available at PopGenIE.org; [62]), we called 4,509,654 SNPs after stringent
121 filtering, which we utilised to perform eQTL mapping.

122 To enable examination of the genetic architecture of natural variation in gene expression within the
123 SwAsp collection, we generated paired-end RNA-Seq expression data from winter buds at the point
124 of spring bud flush for 219 individuals (clonal replicates), representing 86 genotypes, and first
125 examined the distribution of broad-sense heritability (H^2) and population differentiation (Q_{ST}) (Fig
126 1B,C respectively) for all expressed, annotated genes. H^2 ranged from 0.0 to 1.0 with a mean (± s.d)
127 of 0.30 (0.22) and with 5,924 genes (17%) having $H^2 > 0.5$. There was a weak positive correlation
128 between H^2 and median expression (Pearson r = 0.09, $p < 2.2\times10^{-16}$), and a relatively strong positive
129 correlation to expression variance (Pearson r = 0.43, $p < 2.2\times10^{-16}$). Q_{ST} ranged from 0.0 to 1.0 with a
130 mean (± s.d) of 0.06 (0.12) and had a weak negative correlation with expression variance (Pearson r =
131 -0.02, $p < 4.5\times10^{-4}$) and a positive correlation with median expression (Pearson r = 0.18, p < 2.2×10-

132 [16]). To further examine whether population structure was apparent on the basis of expression
133 variation among genotypes, we performed hierarchical clustering of all individuals (Fig 1D) or
134 genotypes (S1 Fig). We observed some evidence of clustering among genotypes but this could not be
135 explained by population structure.

136 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)
137 and subjected these to Gene Ontology (GO) enrichment analysis to determine their biological
138 relevance (see S1 File for all results). Genes with high H^2 were enriched for categories including
139 protein phosphorylation (GO:0006468; $p = 3.7\times10^{-6}$), while high Q_{ST} genes were enriched in terms
140 including translation (GO:0006412; $p = 4.2\times10^{-20}$) and gene expression (GO:0010467; $p = 4.6\times10^{-12}$).
141 Likewise, we considered the genes with the lowest values, which revealed enrichment of terms
142 including cell wall modification (GO:0042545; $p = 2.6\times10^{-7}$) for the 2,289 genes with an H^2 of zero
143 and enrichment of terms including amino acid activation (GO:0043038; p = 0.0012) among the
144 11,895 genes with a Q_{ST} of zero.

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

153 We subsequently filtered expression values to remove unexpressed genes and uninformative
154 expression profiles with low variance. Of 35,154 annotated genes, 20,835 were expressed in all
155 samples, including biological replicates, and 23,183 were expressed in all genotypes when considering
156 genotype means. Filtering to remove uninformative expression retained 22,306 genes with the 12,848
157 removed genes representing those that were either not expressed in our bud samples (6,736 genes
158 with median expression of zero of which 2,385 had no detectable expression at all), or that were
159 weakly expressed (1,762 genes with variance < 0.05 and median expression < 2), together with genes
160 that had stable expression among genotypes (4,350 genes with expression variance < 0.05 and
161 median expression >= 2). The latter potentially represent genes with canalised gene expression.
162 Analysis of this set of stably expressed genes identified enrichment for GO categories including
163 protein transport (GO:0015031, $p = 6.8\times10^{-11}$) and protein localisation (GO:0008104, $p = 2.2\times10^{-10}$).
164 In contrast, the 500 genes with the highest variance across all samples were enriched for GO
165 categories related to protein phosphorylation (GO:0006468, $p < 10^{-6}$), chitin metabolic process
166 (GO:0006030, $p < 10^{-4}$), and cell wall macromolecule catabolic process (GO:0016998, $p < 10^{-4}$).
167 Comparing the variance of these 500 genes with mean F_{ST} calculated using SNPs within those genes
168 revealed no apparent relationship.

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

174 ## eQTL mapping

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

180 In total 164,290 eQTLs were identified at 5% empirical FDR: 131,103 local and 33,187 distant.
181 These eQTLs represented pairwise associations between 6,241 unique genes (eGenes) and 147,419
182 unique SNPs (eSNPs). The genomic context of eSNPs was determined by overlapping eSNP
183 positions with gene annotations. After normalising for feature length, the majority of local eSNPs
184 were located within untranslated regions (UTRs) and up- or down-stream (regulatory) regions of
185 genes, with distinctly lower representation within exons than introns (Fig 2B). The genomic context
186 distribution of local and distant eSNPs was largely similar, although there were distinctly more
187 eSNPs located within intergenic regions for distant eQTLs. Dividing the local eQTLs on the basis of
188 whether they were located within or near to the associated eGene or within or near any other gene
189 inside of our local distance threshold revealed that approximately half were located within or near to
190 the eGene itself. There was a clear tendency for a local eSNP to be located proximal to the
191 transcription start site (TSS) or the stop codon (S4 Fig). eGenes had significantly higher heritability
192 (difference in median of 0.24, Mann-Whitney $p < 2.2\times10^{-16}$) than non eGenes (Fig 2D), with this
193 trend being slightly higher for local than distant eQTLs (S5 Fig). There were no significant
194 differences in median expression between eGenes and non eGenes or between eGenes with local
195 and distant eQTLs. eGenes with at least one local eQTL were enriched for GO categories related to
196 tRNA metabolic process (GO:0006399, $p = 1.5\times10^{-5}$), ncRNA metabolic process (GO:0034660, p =

197 2.6×10^{-5}) and organonitrogen compound biosynthetic process (GO:1901566, $p = 2.2\times10^{-5}$), among
198 others, while eGenes with at least one distant eQTL were enriched for categories including protein
199 phosphorylation (GO:0006468, $p = 0.0064$; see S1 File for all results).

200 The vast majority of eSNPs were associated with a single eGene (132,258 eSNPs) with a maximum
201 of six eGenes associated with a single eSNP (S6A Fig). In contrast only 1,248 of the 6,241 eGenes
202 were associated with a single eSNP, while there were eGenes associated with up to 1,547 eSNPs (S6B
203 Fig). To partially account for linkage, we fitted linear models between the expression of each eGene
204 and all the significant eSNPs for that gene, both local and distant. The use of a linear model masks
205 eSNPs that contain identical/redundant information and thus effectively identifies haplotype blocks
206 present in all individuals (which we refer to as 'unique eSNPs'), while also producing a measure of
207 how well the combination of eSNPs explains the expression of the corresponding eGene (in terms of
208 percentage variance explained, %VE). Of the 4,993 eGenes associated with more than one eSNP,
209 4,703 were also associated to more than one unique eSNP. The adjusted %VE for the combination
210 of eQTLs was, in general, higher than for single eSNPs. Local eSNPs explained significantly more of
211 the variance than did distant eSNPs (local mean adjusted %VE = 51, distant mean adjusted %VE =
212 47, Mann-Whitney $p < 2.2\times10^{-16}$, Fig 2A) and also had higher statistical significance (Fig 2E). There
213 was a positive correlation between the maximum %VE of the eSNPs associated with an eGene and
214 gene expression H^2 (Pearson $r = 0.47$, $p < 2.2\times10^{-16}$).

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

220 ## Co-expression network

221 We used genotype mean gene expression values to calculate a co-expression network and
222 subsequently considered the network characteristics for genes with and without a mapped eQTL
223 (eGenes). In common with other biological networks, the network had a good scale-free fit (R^2 =
224 0.97). We compared our network with the *P. tremula* expression atlas (exAtlas; [62]), which represents
225 different tissues collected from a single genotype, and observed that the correlation distribution for
226 the exAtlas samples was much wider (mean correlation 0.01 ± s.d. 0.36) than that of our population
227 expression data (mean correlation 0.00 ± s.d. 0.12; Kolmogorov-Smirnov D = 0.14, $p < 2.2\times10^{-16}$;
228 Fig 3A).

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

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

239 To test whether eQTLs could explain expression variation of highly connected genes, which are
240 those that drive the network structure, we examined the relationship of eGenes to network
241 connectivity. In general, eGenes had lower connectivity and betweenness centrality than non-eGenes

242 (Mann-Whitney $p < 2.2\times10^{-16}$ for both, Fig 3B). Moreover, genes with a positive k_{diff} were
243 significantly under-represented for eGenes (hypergeometric test, $p = 1.7\times10^{-27}$). We defined the core
244 of each module as the 10% of genes in a module with the highest normalised k_{diff} while also having
245 an intra-modular connectivity >1. Using this definition, all 38 modules contained at least one core
246 gene, with the percentage of core genes ranging from 2-10% (S2 File). Among these module cores,
247 28 contained at least one eGene, with 29 module cores being significantly under-represented for
248 eGenes (hypergeometric test, 5% FDR). This further emphasised that eGenes were not central in the
249 network. Subsequently, we tested whether the periphery of the network was overrepresented for
250 eGenes. Sixty-four of 142 peripheral genes were eGenes, representing a significant enrichment
251 (hypergeometric test $p = 2.6\times10^{-5}$). On the other hand, network module cores were enriched for
252 both transcription factors (hypergeometric test, $p = 5.6\times10^{-7}$) and phylogenetically conserved genes
253 (hypergeometric test, $p = 8.6\times10^{-5}$) defined as genes with orthologs in *P. tremuloides*, *P. trichocarpa* and
254 *A. thaliana* (20,318 genes; [71]). Furthermore, *P. tremula*-specific genes, *i.e.* genes without orthologs in
255 *P. tremuloides*, *P. trichocarpa* or *A. thaliana* (1,614 genes), were slightly under-represented in network
256 module cores (hypergeometric test, $p = 0.013$) while being slightly over-represented among eGenes
257 (hypergeometric test, $p = 0.010$).

258 In addition to eGenes having generally lower connectivity, there was also a negative relationship
259 between eQTL effect size and co-expression connectivity, with eGenes having the highest
260 connectivity also having lower effect size (Pearson $r = -0.15$, $p < 10^{-10}$, Fig 3D). As expected, eGenes
261 also had higher expression heritability (H^2) than non-eGenes (difference in median heritability 0.24,
262 Mann-Whitney $p < 2.2\times10^{-16}$). In addition, H^2 correlated positively with eQTL effect size (Pearson r
263 $= 0.59$, $p < 2.2\times10^{-16}$) and negatively with connectivity (Pearson $r = -0.30$, $p < 2.2\times10^{-16}$).

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

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

272 In *P. tremula* 3,910 paralog pairs were detected [71], with 2,140 of these (4,185 unique genes) passing
273 our gene expression and variance filtering criteria. These paralogs were significantly under-
274 represented for eGenes, with 1,078 of the 4,185 genes having at least one associated eSNP
275 (hypergeometric test, $p = 0.0004$). Comparing the expression correlation of paralog pairs to that of
276 random gene pairs showed that paralogs exhibited conserved regulation (permutation test, $p <$
277 0.001). We compared the expression correlation distributions of paralog pairs containing 0, 1, and 2
278 eGenes (Fig 4) and found that a higher number of eGenes in a pair was associated with lower
279 expression correlation (linear model $\hat{\beta} = -10.52$, $p < 10^{-16}$). Excluding paralogs did not alter the
280 fact that eGenes had significantly lower network connectivity than non-eGenes.

281 ## Population genetics signatures of eQTLs

282 We examined allele frequency distribution of eSNPs compared to all SNPs, finding that that eSNPs
283 included proportionally fewer low-frequency variants and proportionally more intermediate
284 frequency variants relative to the genome-wide level (S9A Fig). We detected a significant negative
285 correlation between minor allele frequency and eQTL effect size (Spearman's rho = -0.285, p <

286 0.001; S9B Fig), indicating that variants of large effect were present only at relatively low frequencies
287 with few large effect eQTLs detected. While purifying selection acting on eSNPs is a likely
288 explanation for this pattern, limited statistical power to detect associations between low frequency
289 variants and gene expression levels could also contribute to this observed pattern. However, as
290 shown in [28], even if the pattern is confounded by statistical artefacts, it likely represents a true
291 biological phenomenon, where large effect eQTLs are constrained to segregate at low frequencies.

292 We further compared patterns of genetic variation between genes with and without eQTLs (eGenes
293 vs. non-eGenes) and found that eGenes exhibited significantly higher genetic variation than non-
294 eGenes (Fig 5A). Furthermore, genes in network module cores harboured significantly lower genetic
295 variation compared to genes in the periphery of the network (Fig 5A and S10A Fig). To test whether
296 different gene categories were experiencing different levels of selective constraint, we also assessed
297 the site frequency spectrum across genes using Tajima's D [72]. Even if Tajima's D values are overall
298 negative in *P. tremula*, likely due to a historical range expansion [61], we would expect genes in
299 module cores to have lower values of Tajima's D compared to genome-wide averages if they are
300 indeed evolving under stronger purifying selection than peripheral genes. Likewise, if eGenes are
301 experiencing relaxed selective constraints, they should exhibit higher Tajima's D values relative to the
302 genome-wide average. Indeed, both of these expectations were supported by our data (Fig 5B, S10B
303 Fig): In accordance with stronger selective constraints for genes in network cores, we found
304 significantly reduced ratios of non-synonymous to synonymous polymorphisms ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and
305 substitutions (d_N/d_S) for these genes compared to peripheral genes in the network (Fig 5C,D). For
306 peripheral genes, eGenes had significantly higher $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and d_N/d_S compared to non-eGenes,
307 again suggesting that genes with eQTLs in the network periphery might be under relaxed purifying
308 selection compared non-eGenes. In comparison, there was no significant difference of $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ or

309 d_N/d_S between eGenes and non-eGenes located in network cores, suggesting that, within the
310 network core, both these classes of genes are evolving under comparably strong selective constraints.

311 Discussion

312 Expression heritability and population differentiation

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

319 In agreement with results reported for non-phenology related phenotypic traits in the SwAsp
320 collection [73], we observed no evidence of population structure, clinal variation or sub-population
321 differentiation (Q_{ST}) in our gene expression data, which is not surprising given the lack of population
322 structure present at the genome level [66]. Although some evidence of clustering was apparent (Fig
323 1D), this disappeared after hidden confounder removal (S12 Fig). Reassuringly, biological replicates
324 of a genotype clustered together for both the original and the adjusted data, indicating that our data
325 were reliable and that reproducible biological differences (corresponding to genetic differences) in
326 gene expression among genotypes could be detected. We were not able to identify the underlying
327 factors responsible for the clustering, but were able to exclude a set of climatic, environmental and
328 experimental factors that represented the most obvious candidates. However, it is of course possible
329 that environmental factors, especially latitude, may correlate with the expression of specific genes.
330 We did identify enrichment of GO categories in the set of the 500 most variable genes, suggesting

331 that this represents a signature with biological relevance. At the other extreme, for genes with
332 canalised expression, there was also enrichment for GO categories, with these potentially
333 representing genes having housekeeping functions.

334 eQTLs

335 In line with other published work, we found more local than distant eQTLs, with local eQTLs
336 explaining significantly more of the variance in gene expression than distant eQTLs [35,48–51] (Fig
337 2A). Although each eSNP was typically associated with only a single gene, many genes were
338 associated with more than one unique eSNP, with the combined variance explained by these eSNPs
339 being higher than for the individual eSNPs, both for local and distant eQTLs. This suggests that the
340 observed expression variation resulted from the combined effects of several eSNPs, consistent with
341 the infinitesimal model. Looking specifically at only the subset of 1,100 eGenes with both local and
342 distant eQTLs, we observed an average increase in the proportion of variance explained by all eSNPs
343 compared to only local eSNPs (an increase of 3.4 percentage points after adjustment).

344 Local eQTLs were most frequently located in regulatory regions (Fig 2B), with UTRs having the
345 highest density of local eSNPs (~1.5 eSNP per kbp sequence) followed by flanking regions (2 kbp
346 up- and downstream, ~1 per kbp sequence) and introns (~0.75 per kbp sequence), which is
347 consistent with previous findings in natural populations of humans, *Drosophila* and *Capsella grandiflora*
348 [28]. The distributions of genomic context for local and distant eQTLs were broadly similar, however
349 distant eQTLs were somewhat more common in exonic regions compared to local ones, as would be
350 expected by indirectly acting eQTLs, and were more frequently located in intergenic regions,
351 potentially representing distantly acting regulatory elements such as enhancers. It has previously been
352 reported that distant eQTLs in humans are most frequently located in introns, followed by intergenic
353 regions [74]. Our results only partially agree with this, with distant eQTLs being far more commonly

354 associated with UTRs and flanking regions, followed by introns and intergenic regions. Finally, the
355 global distribution of genomic contexts for all investigated SNPs was similar to that of local and
356 distant eSNPs, suggesting that SNPs affecting expression do not have a characteristic genomic
357 signature and most likely representing the concentration of SNPs to regions of active recombination
358 [61].

359 ## Co-expression

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

370 Genes with the highest expression variance, and the highest heritability, were enriched for eGenes.
371 Perhaps counterintuitively given that genetic variation underlies heritable gene expression variation,
372 genes co-expressed with many other genes (*i.e.* genes with high connectivity) were under-represented
373 for eGenes, while genes with few co-expression partners (located in the network periphery) were
374 enriched for eGenes (Fig 3B). Furthermore, eGene connectivity was negatively correlated with
375 variance explained by the associated eSNPs (Fig 3D). It has similarly been reported that conserved
376 genes and hubs in human protein-protein interaction networks were less likely to be associated with a

377 detectable eQTL and that the effect size of eQTLs was negatively correlated with connectivity in the
378 protein-protein interaction network [76].

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

389 Regulatory networks are known to be redundant, scale-free and to have an inherent ability to buffer
390 against single mutations of large negative effect [77–79]. In our case, genes with the potential to
391 cause such large negative effects would be observed as hubs in the co-expression network, and we
392 would therefore expect hubs to be buffered against single polymorphisms of large effect size and, in
393 particular, the persistence of large negative effect mutations. As such, we could expect the expression
394 of highly connected genes to be subjected to both strong stabilizing and purifying selection, which
395 would result in them being controlled largely by eSNPs of small additive effects. However, due to
396 statistical limitations, many of these small-effect eSNPs would represent false negatives (*i.e.* would be
397 unobserved), likely explaining why the core of 10 of the 38 modules within our network did not
398 contain any eGenes. Additionally, several other types of complex regulatory mechanisms can
399 preclude our ability to detect eSNPs: Epistatic interactions between eSNPs, such as AND-logics,

400 would result in very little correlation between the individual AND-interacting eSNPs and the
401 expression of the controlled gene. Likewise, the expression of a gene will frequently not only be a
402 function of the associated eSNPs but also of the expression of the regulators acting through those
403 eSNPs, which are themselves affected by additional eSNPs. Furthermore, the regulator(s) that
404 initiated an observed co-expression module may no longer be expressed at the sampling point used
405 for RNA isolation, and thus the eSNPs that determined the expression variation of that regulator will
406 no longer be visible to us (or a 'ghost signature' may remain, but be too weak to reach significance),
407 which again could explain the lack of eGenes in some module cores.

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

418 The Salicaceae lineage underwent a recent whole-genome duplication shared by all member species
419 and that remains represented by a large number of paralogous gene pairs in the genomes of *Populus*
420 species. If many of these duplicated genes are functionally redundant or in the process of diverging,
421 one would expect them to be overrepresented for eGenes as sub- or neo-functionalisation requires
422 derived SNPs to drive expression differences. However, we saw an under-representation of eGenes

423 in paralog pairs, suggesting that SNPs which initially induced expression divergence have reached
424 fixation. Interestingly, we found progressively lower expression correlation between paralog pairs
425 containing zero, one or two eGenes (Fig 4) indicating that a subset of paralogs are still undergoing
426 sub- or neo-functionalisation, with their associated eSNPs driving expression divergence.

427 ## Signatures of selection

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

439 Consistent with the observation that network core genes are evolving under stronger functional
440 constraint than genes in the network periphery, we found that conserved genes (*i.e.* old genes) were
441 significantly over-represented in network cores while *P. tremula* specific genes (*i.e.* young genes) were
442 under-represented. This is in agreement with previous studies that have shown that old genes tend to
443 be more central in regulatory networks, with increased constraints on expression and fewer
444 associated eQTLs [81,82]. Finally, we detected a significant negative correlation between eQTL effect
445 size and eSNP allele frequency (S9B Fig), suggesting that prevalent purifying selection may be acting

446 on expression variation of the eQTLs detected in our study, which is consistent with recent results
447 from *Capsella grandiflora* [28].

448 Conclusion

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

462 Materials and Methods

463 Samples

464 We collected branches form the SwAsp collection common garden in north of Sweden on 27th May
465 2012, before natural bud break but as close to the point of natural spring bud break as possible.

466 Branches were placed in the greenhouse facility at the Umeå Plant Science Centre under conditions
467 selected to induce rapid bud break (24 h light, temperature of 20 °C and humidity 50-70 %). At a
468 defined point of emergence (S13 Fig), buds were harvested, flash frozen in liquid nitrogen and stored
469 at -80 °C until used for RNA isolation. Only terminal buds were sampled (i.e. no lateral buds were
470 included). The time from the day branches were placed in the greenhouse until bud flush sampling
471 ranged from one day to eight days (S14A Fig) and there was a high, positive correlation to bud flush
472 date recorded in the field for the same year (S14B Fig; $r=0.776$, $p < 2.2\times10^{-16}$). As has previously
473 been reported [73], there was no apparent Q_{ST} for bud flush, either in the field or for the greenhouse
474 material (Q_{ST} 0.13 and 0.07 respectively), however H^2 was high ($H^2 = 0.82$ and 0.71 respectively).

475 ## RNA isolation

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

RNA-Sequencing and analysis

RNA-Sequencing was performed as in [70]. Briefly, paired-end (2 × 100 bp) RNA-Seq data were generated using standard Illumina protocols and kits (TruSeq SBS KIT-HS v3, FC-401-3001; TruSeq PE Cluster Kit v3, PE-401-3001) and all sequencing was performed using the Illumina HiSeq 2000 platform at the Science for Life Laboratory, Stockholm, Sweden. Raw data is available at the European Nucleotide Archive (ENA, [84]) with accession number ERP014886.

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

Gene expression H^2 and Q_{ST}

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

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

where V_G is the genetic component of the variance calculated as the expression variance between genotypes for a particular gene (*i.e.* variance among genotype means) and V_P is the total phenotypic variance calculated as the sum of V_G and V_E, where V_E is the environment component of the variance calculated as the expression variance within genotypes for a particular gene (i.e. the mean

507 variance among clonal replicates). Point estimates of H^2 were obtained using the repeatability
508 function from the heritability v1.1 R package [93].

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

510 $$Q_{ST} = \frac{V_{between}}{V_{between} + 2V_{genetic}}$$

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

513 ```
expression ~ 1 + (1|population) + (1|clone)
```

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

517 ## Hidden confounder removal

518 Gene expression data was adjusted for hidden confounders before mapping eQTLs and constructing
519 the co-expression network. Hidden confounders in the gene expression data was accounted for by
520 regressing out the 9 first principal components (PCs) of the gene expression data [97–99]. The
521 number of components to remove was determined by running the eQTL mapping with 0 to 20 PCs
522 removed and selecting the number of components that yielded the largest number of significant
523 eQTLs (Benjamini-Hochberg $p < 0.05$) (S15 Fig). This approach is based on the assumption that the
524 number of identified eQTLs will increase if the removed PCs are removing unwanted, systematic
525 variation (i.e. noise) rather than informative biological variation [97–99].
```

526 ## eQTL mapping

527 eQTL mapping was performed by associating gene expression with biallelic SNPs using the R
528 package Matrix eQTL v2.1.1 [100]. The corresponding raw resequencing data for all SwAsp
529 genotypes is available at the NCBI SRA resource as BioProject PRJNA297202 (SRA: SRP065057)
530 and the resultant Variant Call Format (VCF) file used for association mapping is available from [63].
531 Before performing association mapping, genes were filtered on variance so that only genes with a
532 gene expression variance above 0.05 were included. SNPs were also filtered on minor allele
533 frequency (MAF) and major genotype frequency (MGF); any SNPs with MAF < 0.1 or MGF > 0.9
534 were excluded to avoid spurious associations. A motivating example for the MGF filtering can be
535 seen in S16 Fig. We also generated an LD-trimmed SNP set by removing one SNP from each pair of
536 SNPs with a between SNP correlation coefficient (r^2) >0.2 in blocks of 50 SNPs using PLINK v1.9
537 [101], yielding 217,489 independent SNPs that were retained for analyses of population structure.
538 The first genotype principal component based on the set of independent SNPs was used as a
539 covariate in the linear model used by Matrix eQTL to account for the weak signature of population
540 structure. Permutation testing was used to determine eQTL significance whereby genotype sample
541 labels were permuted 1000 times and the maximum absolute t-statistic from Matrix eQTL was
542 recorded for each gene across all SNPs for each permutation. Empirical p-values were calculated
543 with the empPvals function in the qvalue v2.4.2 R package [102], and q-values (empirical FDR) were
544 calculated with the qvalue function in the same package.

545 When determining the genomic context of eSNPs, there were some cases where introns overlapped
546 exons as a result of overlapping gene model being present on the same strand. These 41 eSNPs were
547 discarded from the counting. Another type of overlap that was discarded were cases where an eSNP
548 overlapped a gene feature, but no sub-feature inside that gene (e.g. UTR, exon or intron). These

549 1961 eSNPs were excluded from the counting. Since many of the features overlap (e.g. exon and
550 untranslated regions), the priority for counting was untranslated region, exon/intron,
551 upstream/downstream and intergenic.

552 ## Co-expression

553 The R package WGCNA v1.51 [103] was used for constructing a co-expression network. The input
554 gene expression values were per-gene genotype means. We chose to use the unsigned network type
555 for this study with the motivation that we did not want to discard negative relationships. By looking
556 at this from an eQTL perspective, an eSNP can be positively associated with one gene while
557 negatively associated with another. The relationship between these two genes would be missed if we
558 used the signed or the signed-hybrid approaches. Using the unsigned approach, we assure that genes
559 with strong negative correlation end up in the same network modules. A soft thresholding power of
560 5 was used to calculate adjacencies. The topological overlap matrix (TOM) was generated using the
561 TOMsimilarity function with the signed approach in order to take negative edges into account (see
562 [103] for details). In order to identify network modules, hierarchical clustering was applied to the
563 TOM dissimilarity matrix (1 – TOM) and the resulting dendrogram was divided into modules using
564 the cutreedynamic function. The connectivity of the network was then defined as the adjacency sum
565 for each node, i.e. the weights of the edges that are connected to this node. This concept was applied
566 to modules as well to obtain measures of intra- and inter-modular connectivity, i.e. the connectivity
567 based on edges connecting the gene with other genes inside the same module, and connectivity based
568 on edges connecting the gene with genes outside of the module.

569 To define the periphery of the network we applied a hard edge-threshold to the network where only
570 gene-pairs with an absolute Pearson correlation > 0.22 were linked, which corresponded to the top
571 1% most correlated gene pairs. Genes were then classed as peripheral if they linked to only one other

572 gene.

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

575 ## Population genetic analysis

576 The minor allele frequency (MAF) was compared between eSNPs (both local and distant) and all
577 SNPs throughout the genome. To test for the signature of selection on eQTLs, we also estimate the
578 correlation between minor allele frequency and eQTL effect size. Furthermore, we estimated and
579 compared nucleotide diversity (θ_π) and Tajima's D [72] between eGenes and non-eGenes, as well as
580 genes locating at central or peripheral in the gene network (core and peripheral genes). We only used
581 the reads with mapping quality above 30 and the bases with quality score higher than 20 for the
582 estimation of θ_π and Tajima's D by ANGSD [105]. Genes with less than 50 covered sites left from
583 previous quality filtering steps (Wang et al. In prep) were excluded. In the end, there were 131 and
584 1,563 eGenes and non-eGenes that were located in module cores, respectively. Similarly there were
585 5,759 eGenes and 12,893 non-eGenes located the periphery of the gene expression network. In
586 addition, θ_π and Tajima's D were also calculated for 0-fold and 4-fold degenerate sites within genes,
587 representing non-synonymous and synonymous sites, respectively. We further compared the
588 estimates of $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ between different categories of genes. We used the transcript with the
589 highest content of protein-coding sites to categorize synonymous and non-synonymous sites within
590 each gene. Finally, a total of 33,039 orthologous genes were determined by best BLASTp sequence
591 homology searches between *P. tremula* and *P. trichocarpa* [71,104]. We estimated the ratio of non-
592 synonymous substitution and synonymous substitution rate (d_N/d_S) using gKaKs pipeline [106] with
593 codeml method [107] for 133 and 1,599 orthologous eGenes and non-eGenes within module cores

594 of gene expression network, and for 5,758 and 12,828 orthologous eGenes and non-eGenes locating
595 at peripheral in the gene network. Significance for the above statistical measurements between each
596 pair of gene categories was evaluated using Mann-Whitney tests.

597 Acknowledgements

598 PKI has been supported by a grant from Vetenskapsrådet. JW was supported by a scholarship from
599 the Chinese Scholarship Council. NRS is supported by the Trees and Crops for the Future (TC4F)
600 project. The authors also would like to acknowledge support from the Science for Life Laboratory,
601 the National Genomics Infrastructure (NGI), and the Uppsala Multidisciplinary Center for
602 Advanced Computational Science (UPPMAX) for providing assistance in massive parallel sequencing
603 and computational infrastructure.

604 References

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

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

609 3. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, et al. Regional and
610 strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A.
611 2000;97: 11038–43. doi:10.1073/pnas.97.20.11038

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

614 5. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions
615 of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet.
616 2001;29: 389–95. doi:10.1038/ng766

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

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

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

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

629 10. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping
630 determinants of human gene expression by regional and genome-wide association. Nature.
631 2005;437: 1365–9. doi:10.1038/nature04244

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

635 12. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide
636 associations of gene expression variation in humans. PLoS Genet. 2005;1: 0695–0704.
637 doi:10.1371/journal.pgen.0010078

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

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

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

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

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

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

653 19. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell.
654 2000;101: 577–580. doi:10.1016/S0092-8674(00)80868-5

655 20. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in

656 budding yeast. Science (80-). 2002;296: 752–5. doi:10.1126/science.1069516

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

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

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

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

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

670 26. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet.
671 2001;17: 388–91. doi:10.1016/S0168-9525(01)02310-1

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

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

677 29. Lappalainen T. Functional genomics bridges the gap between quantitative genetics and
678 molecular biology. Genome Res. 2015;25: 1427–1431. doi:10.1101/gr.190983.115.

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

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

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

686 33. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:
687 206–16. doi:10.1038/nrg2063

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

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

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

698 37. West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, et al.
699 Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in
700 *Arabidopsis*. Genetics. 2007;175: 1441–50. doi:10.1534/genetics.106.064972

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

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

707 40. Swanson-Wagner R a, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, et al. Paternal dominance
708 of trans-eQTL influences gene expression patterns in maize hybrids. Science. 2009;326: 1118–
709 1120. doi:10.1126/science.1178294

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

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

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

718 44. Wang J, Yu H, Weng X, Xie W, Xu C, Li X, et al. An expression quantitative trait loci-guided

719 co-expression analysis for constructing regulatory network using a rice recombinant inbred
720 line population. J Exp Bot. 2014;65: 1069–1079. doi:10.1093/jxb/ert464

721 45. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR. Genetic Architecture of Transcript-
722 Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. Genetics. 2005;169: 2295–
723 2303. Available: http://www.genetics.org/cgi/content/abstract/169/4/2295

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

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

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

732 49. Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, et al. Segregating
733 Variation in the Transcriptome: Cis Regulation and Additivity of Effects. Genetics. 2006;173:
734 1347–1355. doi:10.1534/genetics.105.051474

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

738 51. Potokina E, Druka A, Luo Z, Wise R, Waugh R, Kearsey M. Gene expression quantitative
739 trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide

740 transcriptional regulation. Plant J. 2008;53: 90–101. doi:10.1111/j.1365-313X.2007.03315.x

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

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

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

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

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

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

760 58. Nourmohammad A, Rambeau J, Held T, Berg J, Lassig M. Pervasive adaptation of gene

761 expression in *Drosophila*. arXiv. 2015; Available: http://arxiv.org/abs/1502.06406

762 59. Jansson S, Douglas CJ. *Populus*: A Model System for Plant Biology. Annu Rev Plant Biol.
763 2007;58: 435–458. doi:10.1146/annurev.arplant.58.032806.103956

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

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

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

773 63.
774 ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/VCF/SwAsp_94samples.filte
775 r.maf005.beagle.maf.recode.vcf.gz.

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

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

782 66. Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK. Adaptive population
783 differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremula*,
784 L.): a comparison of neutral markers, candidate genes and phenotypic traits. Evolution.
785 2007;61: 2849–60. doi:10.1111/j.1558-5646.2007.00230.x

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

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

791 69. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA
792 methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife.
793 2015;4: e05255. doi:10.7554/eLife.05255

794 70. Robinson KM, Delhomme N, Mähler N, Schiffthaler B, Onskog J, Albrectsen BR, et al.
795 *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. BMC Plant Biol.
796 2014;14: 276. doi:10.1186/s12870-014-0276-5

797 71. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S. The *Populus* Genome Integrative
798 Explorer (PopGenIE): a new resource for exploring the *Populus* genome. New Phytol.
799 2009;182: 1013–1025. doi:10.1111/j.1469-8137.2009.02807.x

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

802 73. Robinson KM, Ingvarsson PK, Jansson S, Albrectsen BR. Genetic variation in functional

803 traits influences arthropod community composition in aspen (*Populus tremula* L.). PLoS One.
804 2012;7: e37679. doi:10.1371/journal.pone.0037679

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

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

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

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

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

817 79. Whitacre JM. Biological robustness: Paradigms, mechanisms, systems principles. Front Genet.
818 2012;3: 1–15. doi:10.3389/fgene.2012.00067

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

822 81. Popadin KY, Gutierrez-Arcelus M, Lappalainen T, Buil A, Steinberg J, Nikolaev SI, et al.
823 Gene age predicts the strength of purifying selection acting on gene expression variation in

824 humans. Am J Hum Genet. 2014;95: 660–74. doi:10.1016/j.ajhg.2014.11.003

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

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

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

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

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

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

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

844 89. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput

845 sequencing data. Bioinformatics. 2014;31: 166–169. doi:10.1093/bioinformatics/btu638

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

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

850 92. Dohm MR. Repeatability estimates do not always set an upper limit to heritability. Funct Ecol.
851 2002;16: 273–280. doi:10.1046/j.1365-2435.2002.00621.x

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

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

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

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

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

865 98. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding
866 mechanisms underlying human gene expression variation with RNA sequencing. Nature.
867 2010;464: 768–72. doi:10.1038/nature08872

868 99. Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, et al. Normalizing
869 RNA-sequencing data by modeling hidden covariates with prior knowledge. PLoS One.
870 2013;8: e68141. doi:10.1371/journal.pone.0068141

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

873 101. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool
874 set for whole-genome association and population-based linkage analyses. Am J Hum Genet.
875 2007;81: 559–575. doi:10.1086/519795

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

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

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

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

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

886 107. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:
887 1586–1591. doi:10.1093/molbev/msm088

888

889 Figures

890 Fig 1. (A) Map of the original locations of the SwAsp populations. The red arrow points to the
891 location of the common garden used in this study. (B) Distribution of gene expression heritability.
892 (C) Distribution of gene expression Q_{ST}. (D) Sample clustering based on all samples, including
893 biological replicates. The heatmap represents the sample correlation matrix based on the 500 genes
894 with the highest expression variance. Darker colour indicates higher correlation. The coloured bar
895 represents the populations the samples belong to. The small clusters on the diagonal correspond to
896 biological replicates of each genotype.

897 Fig 2. (A) Expression variance explained *(R^2)* for local and distant eQTLs. Box plots show the
898 maximum variance explained by a single eQTL for each gene and the total variance explained by all
899 eQTLs for each gene. The widths of the boxes are proportional to the number of genes represented.
900 The pairwise significiance of a Mann-Whitney test is indicated by asterisks: ***$p < 0.001$, **$p < 0.01$.
901 (B) Genomic context of local and distant eSNPs. When an eSNP overlapped with features on both
902 strands, both of them were counted. For both local and distant eQTLs the features are based on the
903 gene that is closest to the eSNP, and furthermore, for local eQTLs, the features are divided into
904 whether the eSNP is located in or near the same gene that it is associated to or not. Flanking regions
905 represent 2 kb upstream and downstream from the gene. (C) Scatter plot showing the positions of all
906 significant eQTLs in this study. No evidence of eQTL hotspots can be observed. Numbers indicate

907 chromosome. (D) Broad sense heritability distributions for eGenes and non-eGenes. (E) Manhattan
908 plots for local eQTLs (upper) and distant eQTLs (lower). Each point represents an eQTL.

909 Fig 3. (A) Co-expression correlation distribution of all pairs of genes for the SwAsp data and the
910 exAtlas data. (B) Distribution of co-expression connectivity for eGenes and non-eGenes. (C)
911 Network module 23 and module 22 with core genes coloured blue and eGenes are indicated by
912 square nodes. In module 23 (upper network), all core genes are eGenes, while for module 22, there
913 are very few eGenes in the core. Red edges indicate negative correlation in expression while blue
914 edges indicate positive correlation. (D) eGene connectivity plotted against the total variance
915 explained of all eQTL associated with each eGene. Colors indicate the effect size where more blue
916 represents a more positive effect and red is a more negative effect.

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

921 Fig 5. Measures of diversity and divergence for eGenes and non-eGenes as well as for core and
922 peripheral genes from the gene expression network. (A) Nucleotide diversity, (B) Tajima's D, (C) θ_{0-}
923 $_{fold}/\theta_{4-fold}$ and (D) d_N/d_S. Significance between each pair of gene categories was evaluated using
924 Mann-Whitney tests and significance is indicated by asterisks above the boxplot: ***$p < 0.001$, **$p <$
925 0.01, *$p < 0.05$, n.s. $p > 0.05$.

926 # Supporting information

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

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

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

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

944 S3 Fig. Relationship between gene expression ranks and the number of genotypes with detectable
945 expression of the genes. The number of genotypes that a gene was expressed in was determined by
946 counting the number of genotypes with non-zero expression for each gene. The gene expression
947 ranks were calculated by ranking the mean gene expression values where the mean was calculated
948 only considering samples with non-zero expression.

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

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

954 S6 Fig. (A) Bar plot of the number of associated eGenes per eSNP. (B) Histogram of the number of
955 associated eSNPs per eGene.

956 S7 Fig. Sliding window approach to detect eQTL hotspots. The upper panel shows the number of
957 eSNPs in genomic windows of 100 kb for each of the 19 chromosomes. The lower panel shows the
958 number of unique genes that are associated to each genomic window (nominal eQTL p-value <
959 1×10^{-6}).

960 S8 Fig. Gene expression variance plotted against co-expression network connectivity. eGenes are
961 indicated by red points.

962 S9 Fig. (A) Comparison of minor allele frequency between all SNPs, distant eSNPs and local eSNPs.
963 (B) The relationship between minor allele frequency and effect size (absolute value of beta) of
964 eQTLs.

965 S10 Fig. Measures of diversity and selection for 0-fold and 4-fold degenerate sites in eGenes and
966 non-eGenes, as well as genes in network module cores and those not. Panel (A) and (B) show values
967 of mean nucleotide diversity (θ_π) and Tajima's D for genes in each categories, respectively.
968 Significance of a Mann-Whitney test comparing each pair of gene categories is indicated by asterisks
969 above the boxplot: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, n.s. $p > 0.05$.

970 S11 Fig. Comparison of signals of positive selection (F_{ST} and $|nS_L|$) between SNPs (blue dots and
971 bar) involved in regulating gene expression levels and other SNPs across the genome (gray dots and
972 bar). Strong signals of positive selection are revealed by $|nS_L|>2$ as indicated by horizontal dashed
973 line. Significance of a Mann-Whitney test comparing eSNPs and other SNPs is indicated by asterisks
974 below the boxplot: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, n.s. $p > 0.05$.

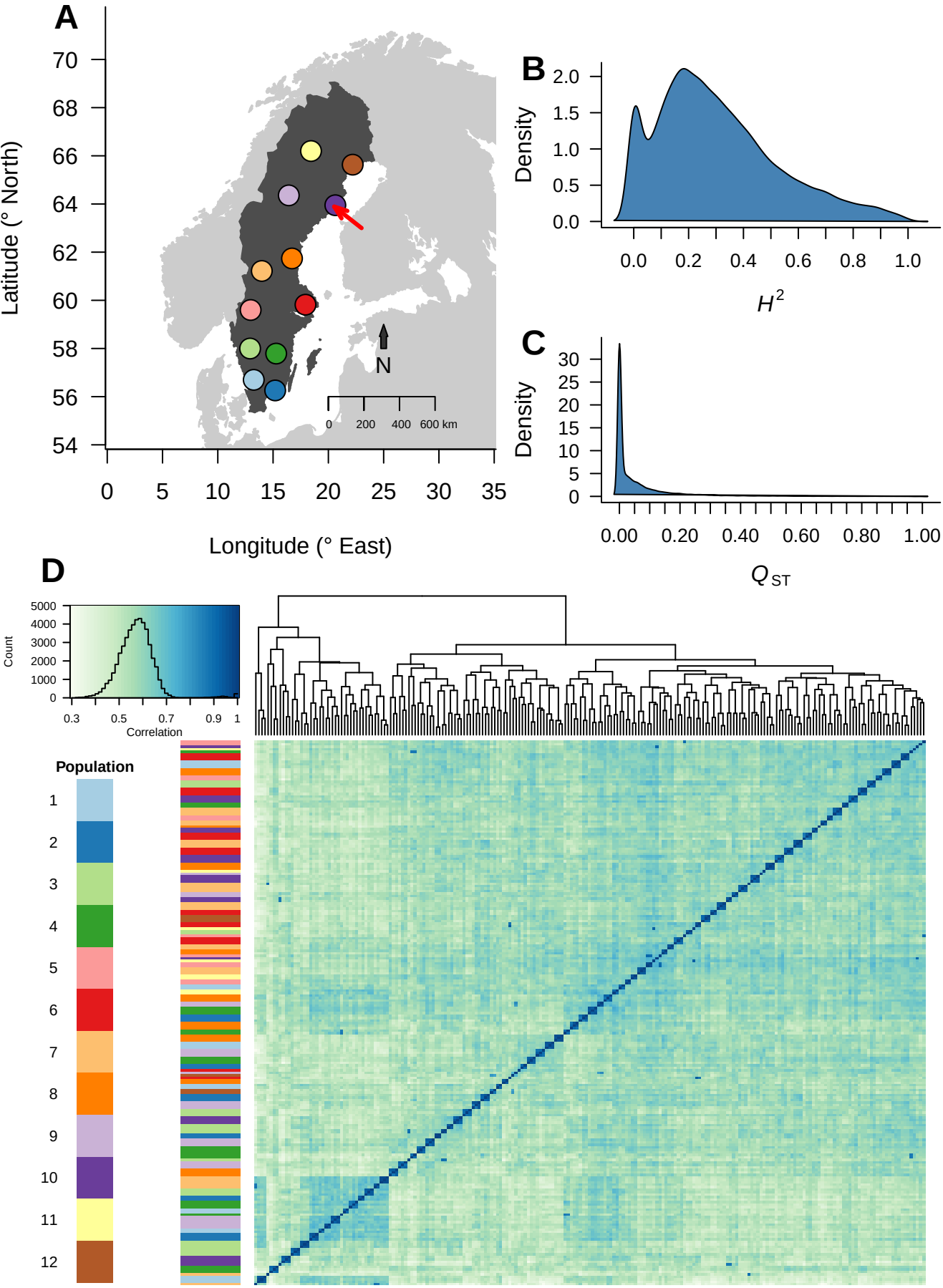
975 S12 Fig. Heatmap representing the sample clustering based on gene expression values after hidden
976 confounder removal. The 500 most variable genes in the original expression data were used for
977 calculating the sample correlations (i.e. the same genes as in Fig 1D). The colour bar represents the
978 population of the samples with the same colour scheme as in Fig 1D. The small clusters on the
979 diagonal represent genotype replicates.

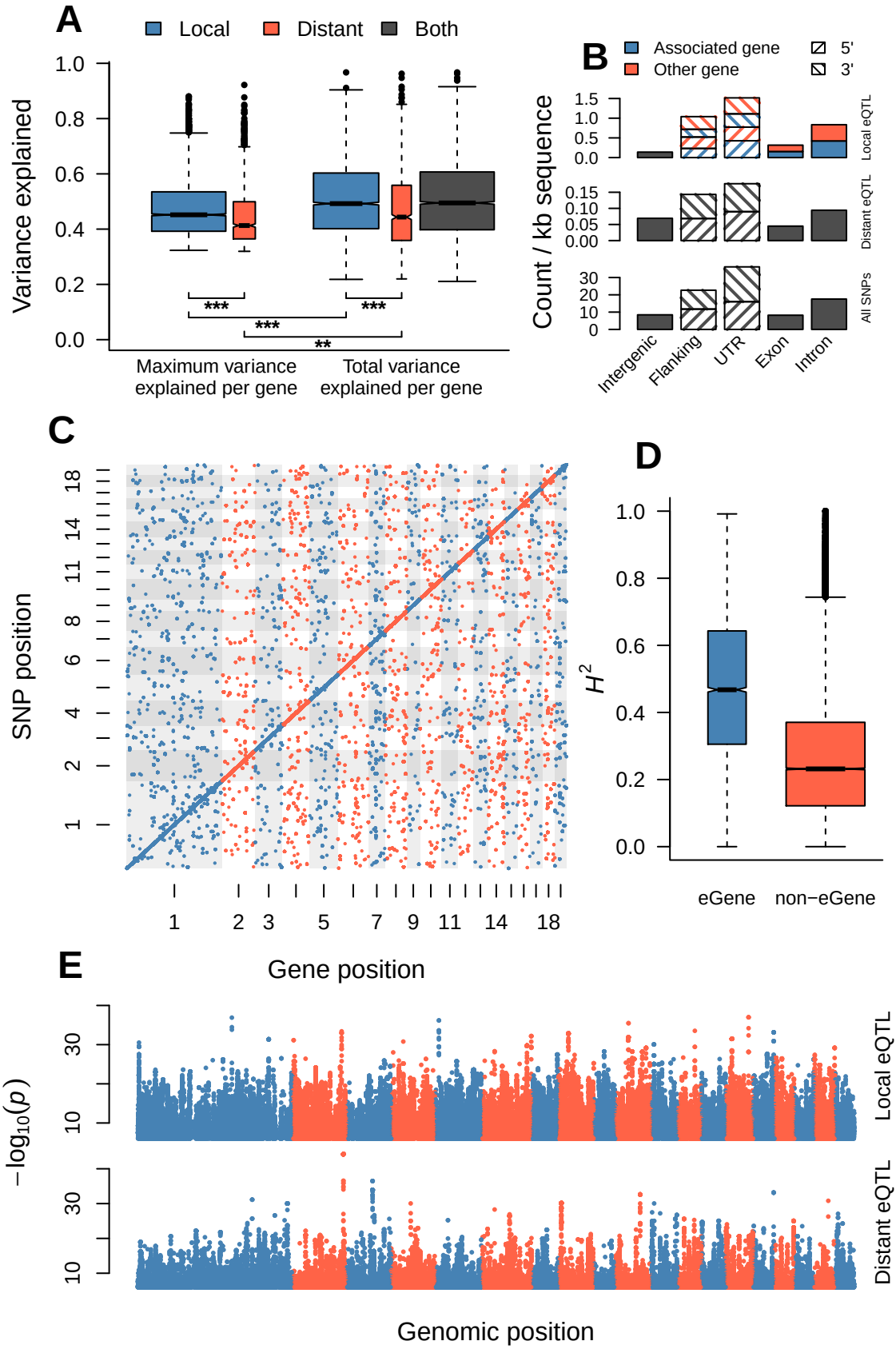
980 S13 Fig. Representative photo of the sampled bud flush stage.

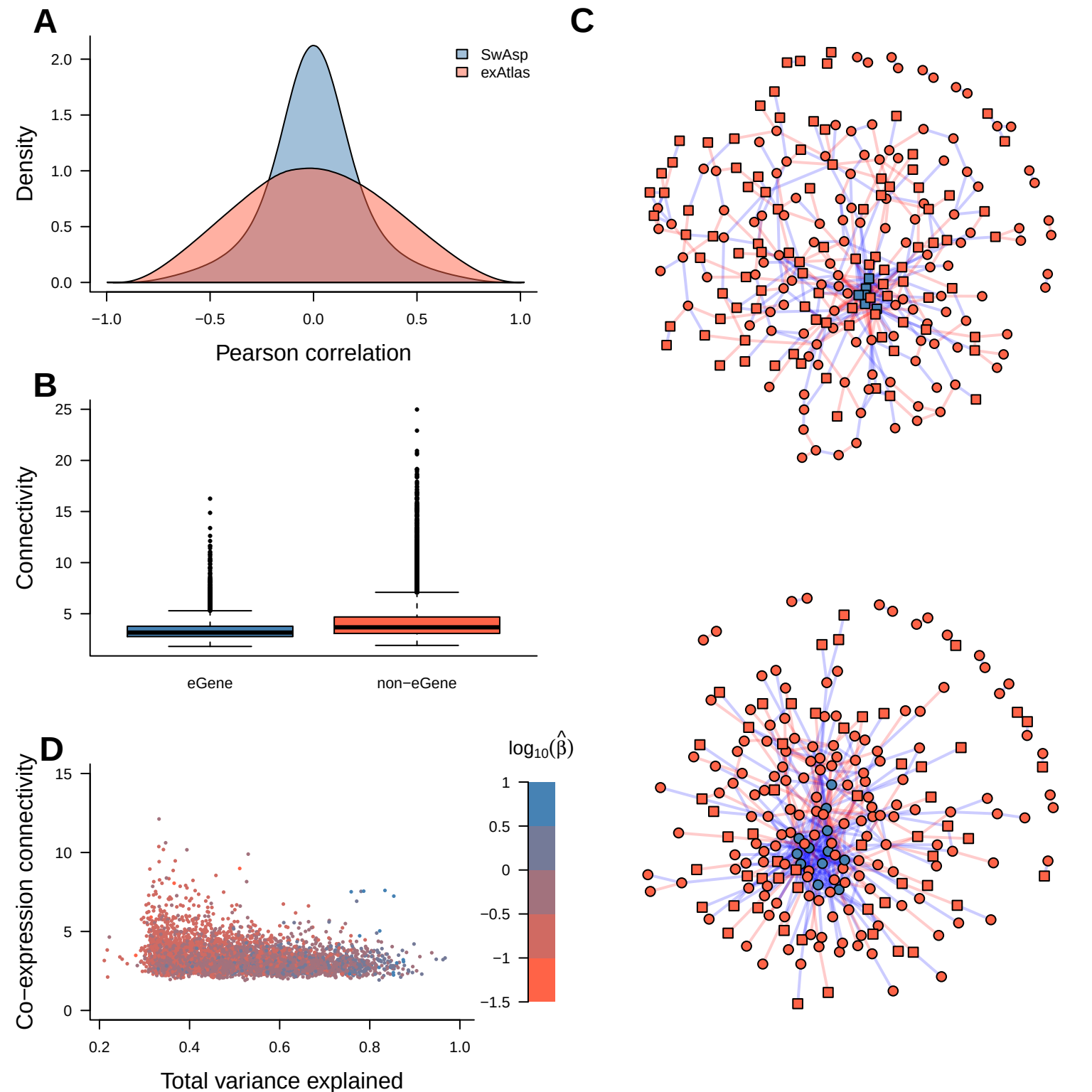
981 S14 Fig. (A) Box plot distributions of the Julian day of sampling for the SwAsp sub-populations. (B)
982 The relationship between Julian day of bud flush for the greenhouse sampled buds and Julian day of
983 bud flush in the field for the same year (2012).

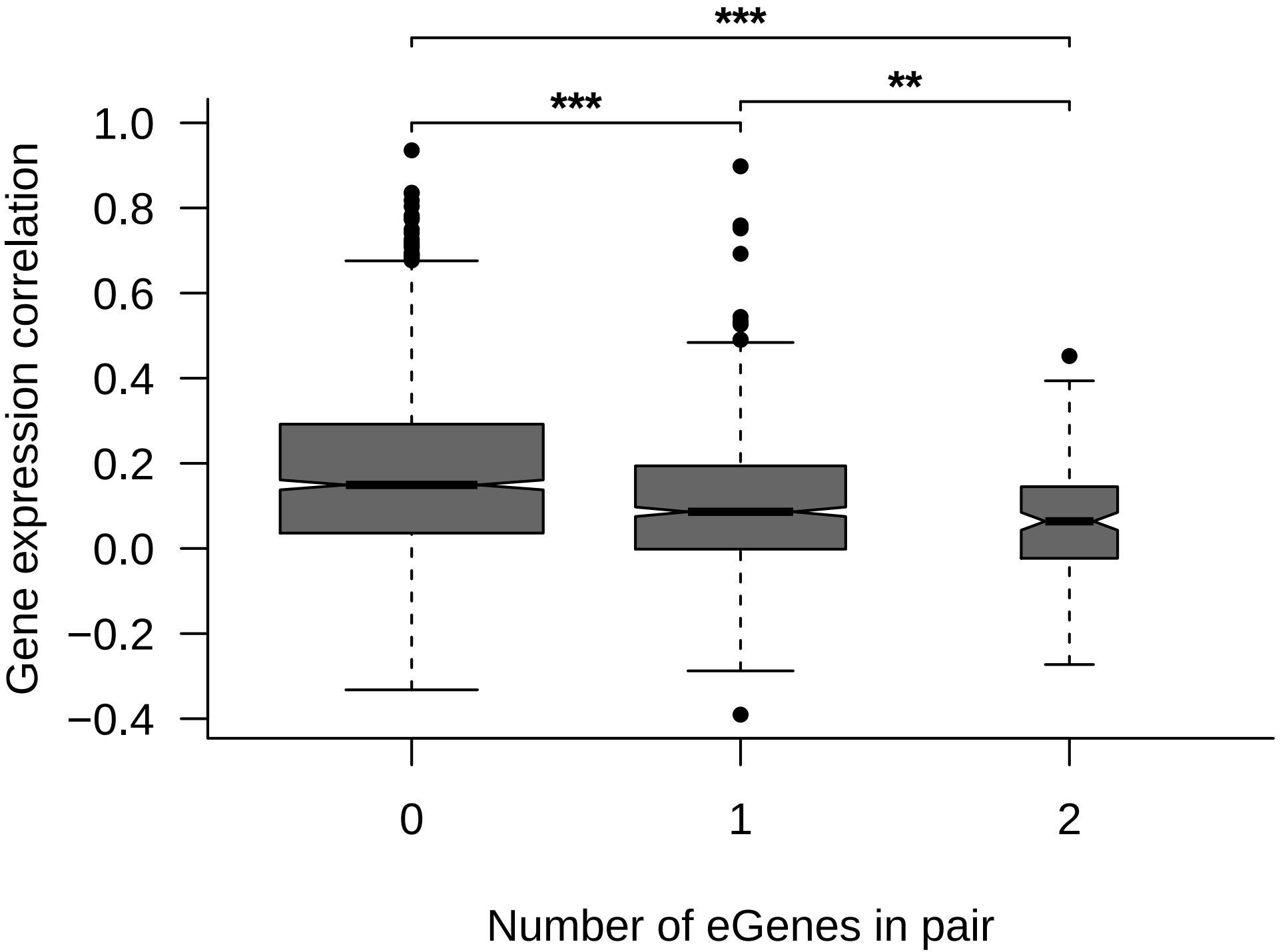
984 S15 Fig. The number of eQTLs detected as a function of the number of principal components
985 regressed out of the gene expression data.

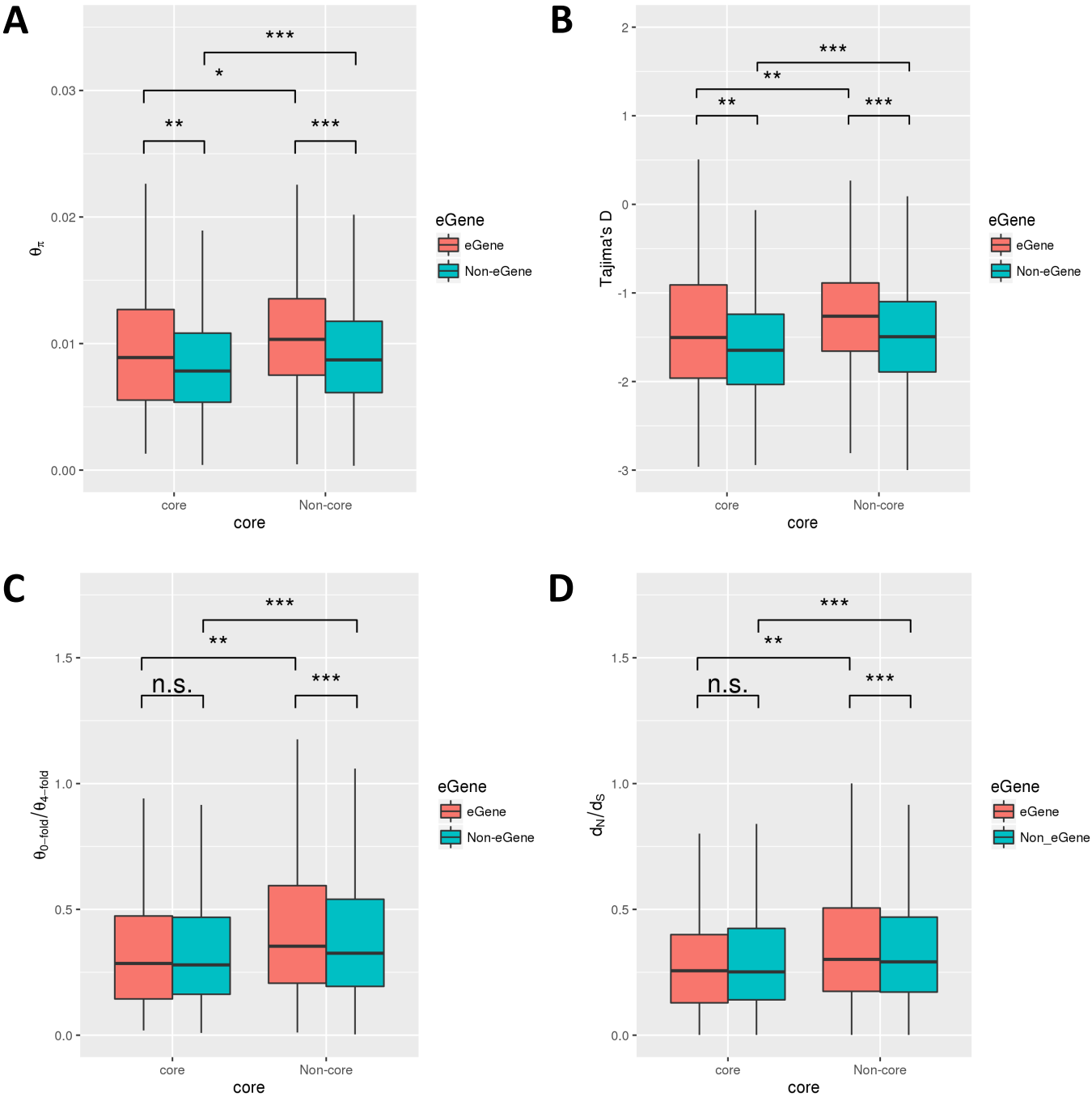
986 S16 Fig. Motivating example for major genotype filtering. In this case the SNP is heterozygous in all
987 samples but one, resulting in a minor allele frequency close to 0.5, but a major genotype frequency
988 close to one. The resulting association turns out very significant, but is only supported by a single
989 sample.

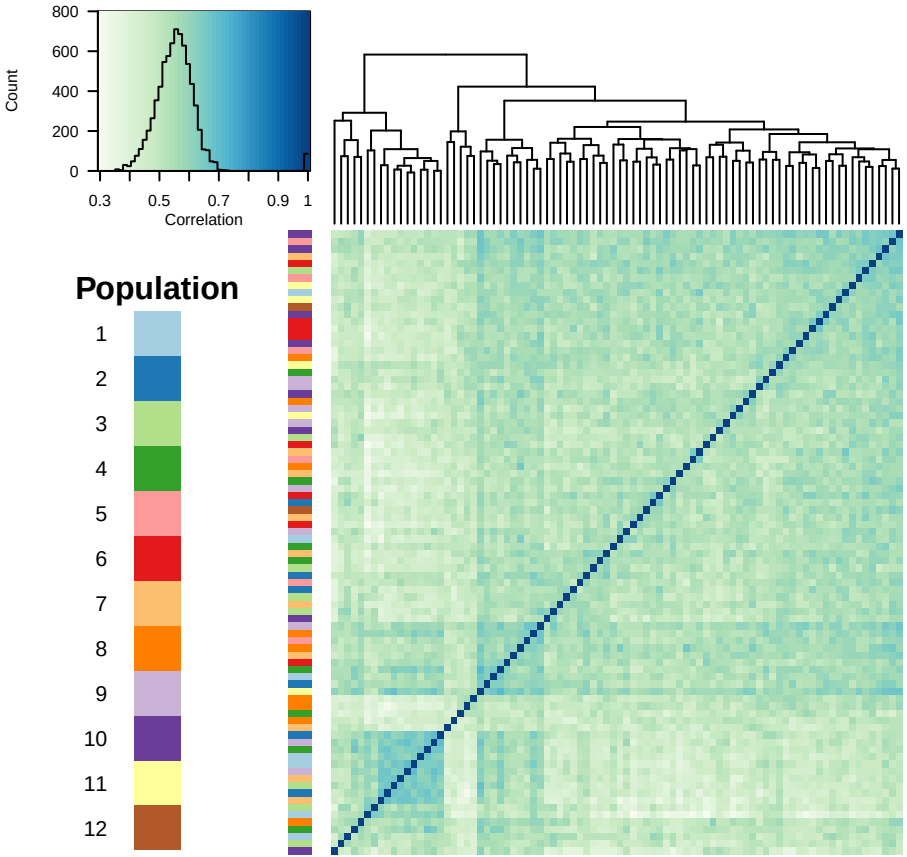

A
Latitude (° North)
Longitude (° East)
N
0 200 400 600 km
B
Density
H^2
C
Density
Q_{ST}
D
Count
Correlation
Population

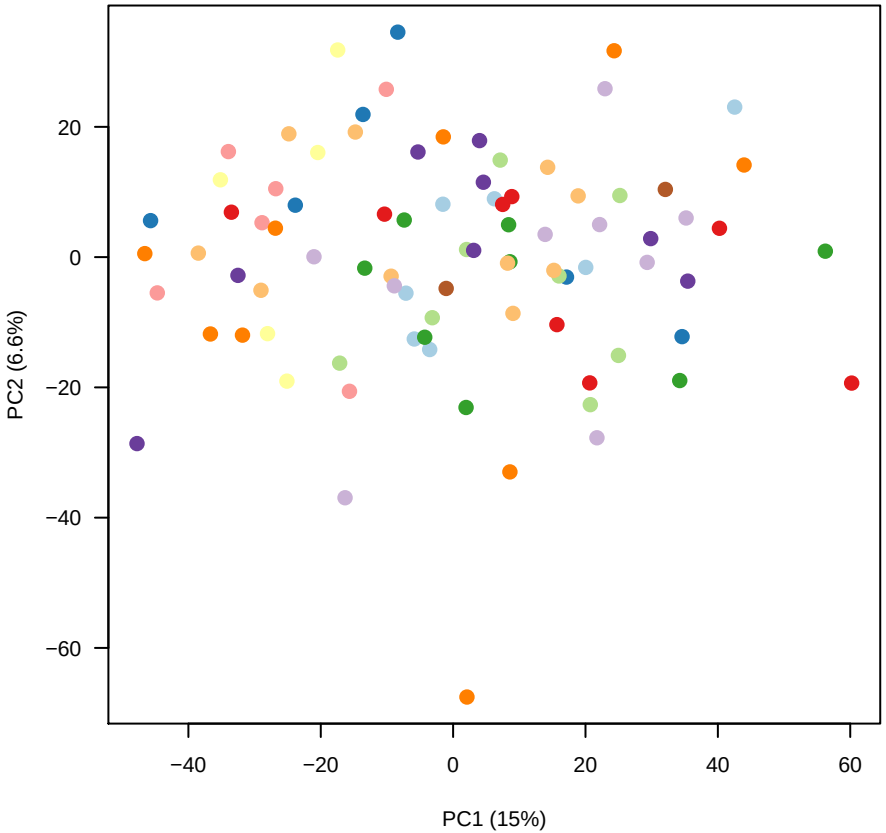

A
Local
Distant
Both
Variance explained
1.0
0.8
0.6
0.4
0.2
0.0
***
***
***
**
Maximum variance explained per gene
Total variance explained per gene
B
Associated gene
Other gene
5'
3'
Count / kb sequence
Local eQTL
Distant eQTL
All SNPs
Intergenic
Flanking
UTR
Exon
Intron
C
SNP position
Gene position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

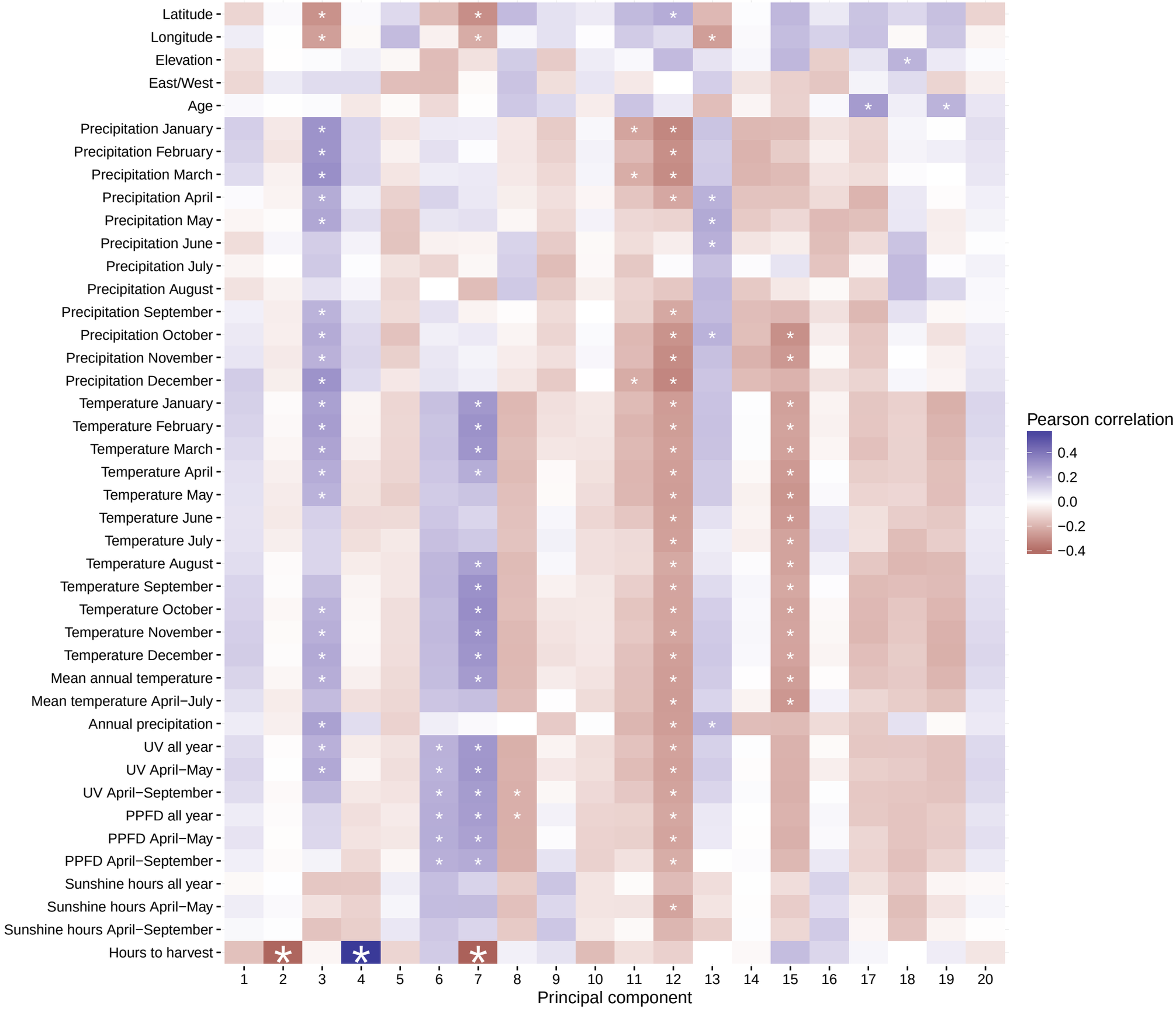

A
SwAsp
exAtlas
Density
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
Pearson correlation
B
Connectivity
25
20
15
10
5
eGene
non−eGene
C
D
Co−expression connectivity
15
10
5
0
0.2
0.4
0.6
0.8
1.0
Total variance explained
$\log_{10}(\hat{\beta})$
1
0.5
0
−0.5
−1
−1.5

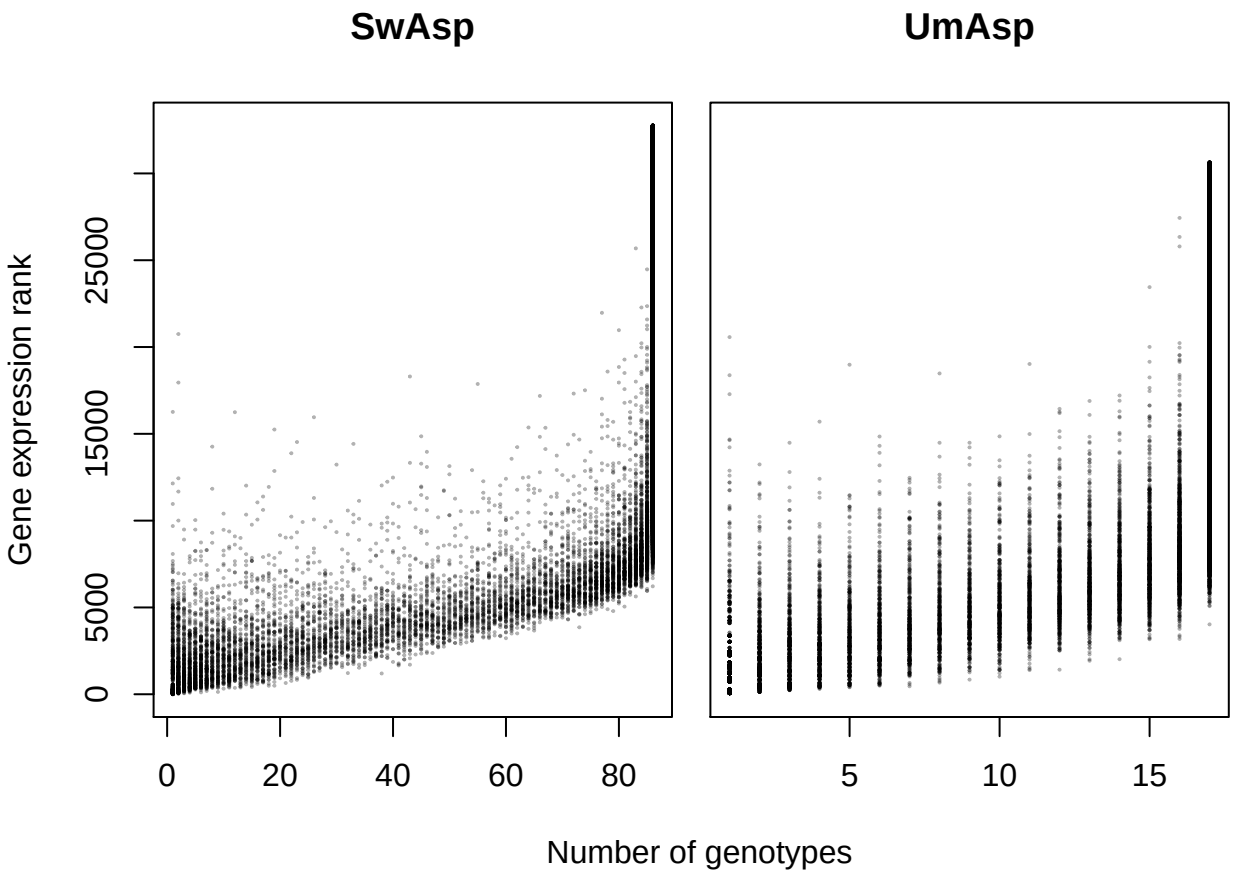

***
***
**
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Gene expression correlation
0
1
2
Number of eGenes in pair

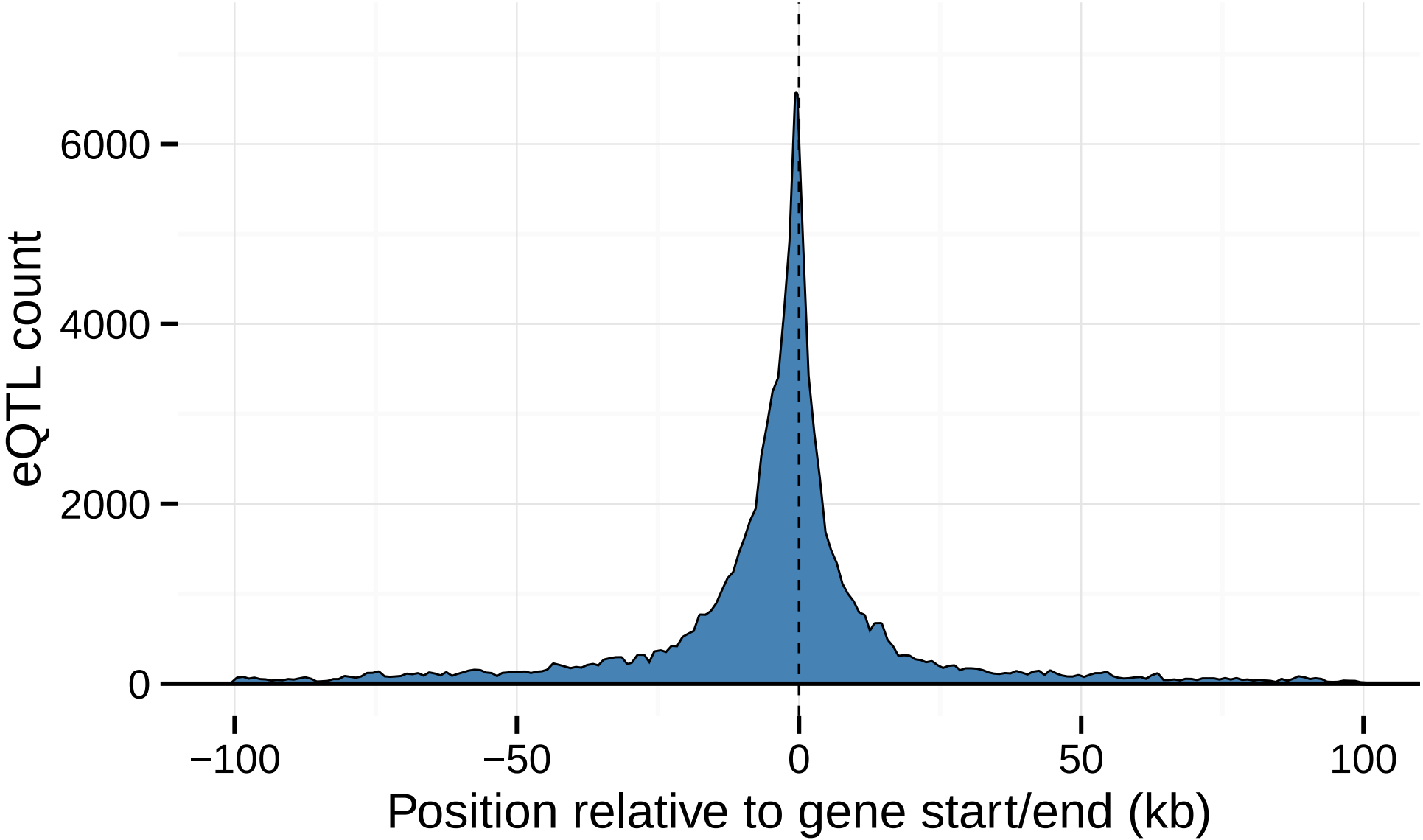

A
B
C
D
θπ
Tajima's D
θ0-fold/θ4-fold
dN/dS
core
Non-core
eGene
Non-eGene
Non_eGene
n.s.
*
**
***

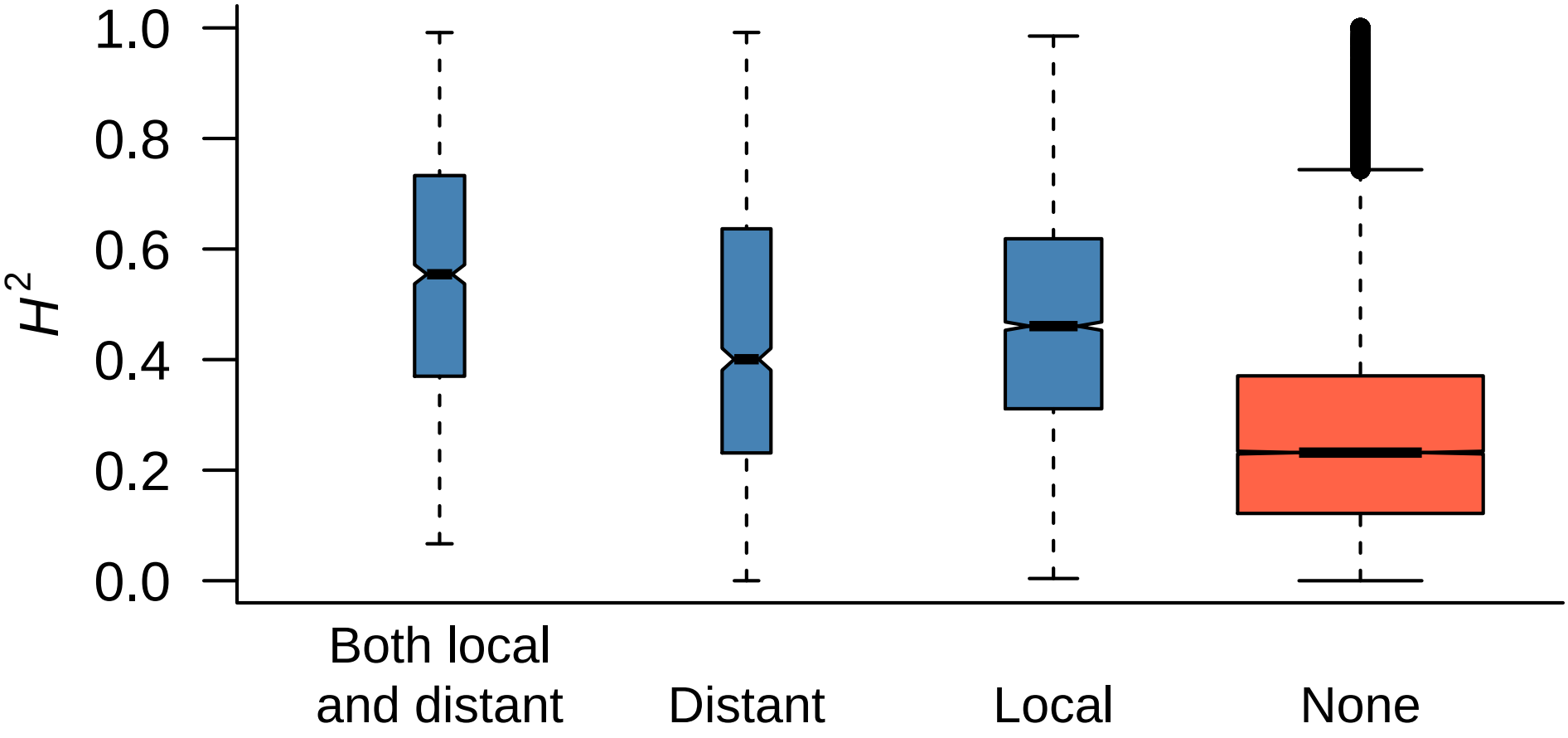

Count
800
600
400
200
0
0.3
0.5
0.7
0.9
1
Correlation
Population
1
2
3
4
5
6
7
8
9
10
11
12

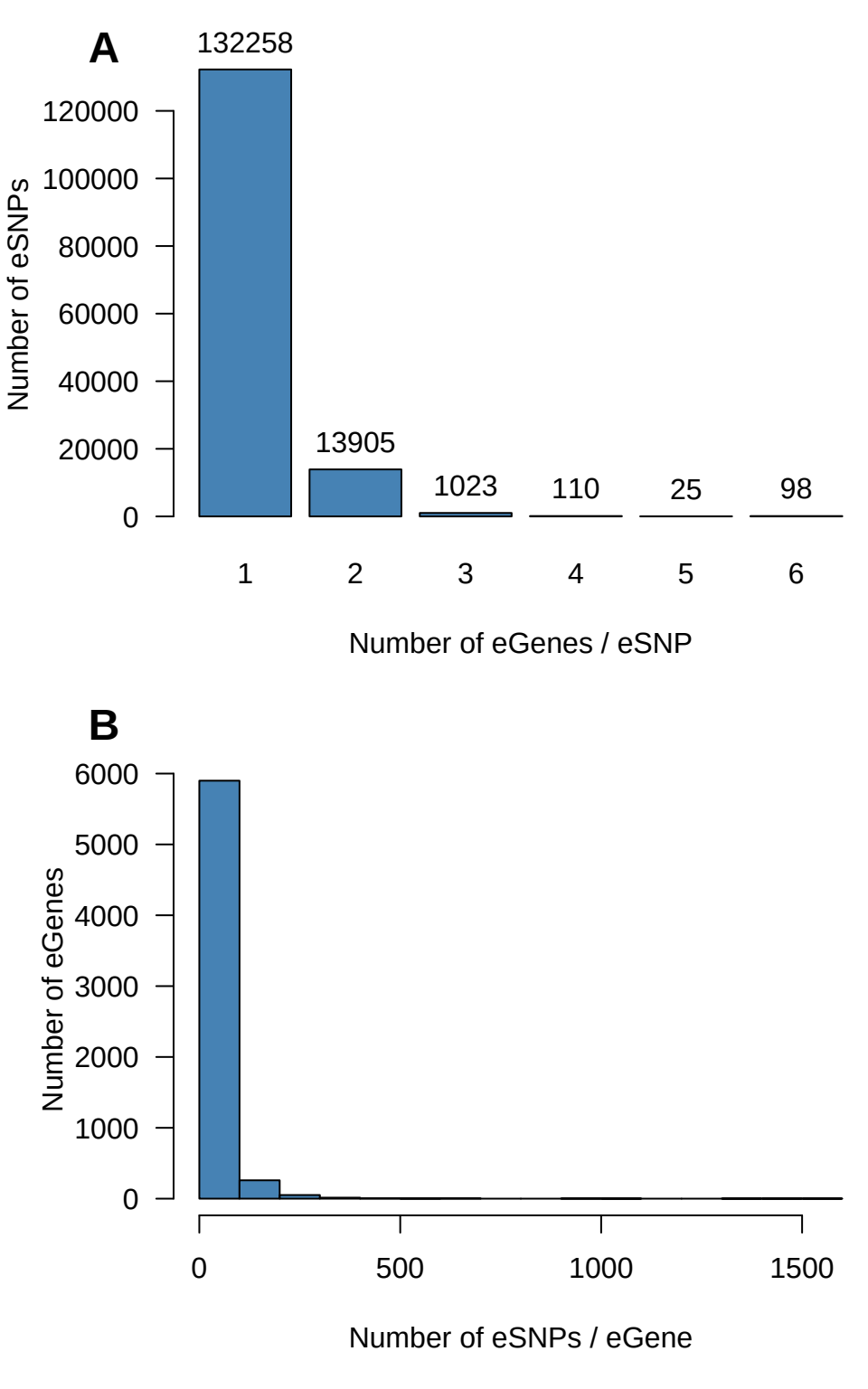

PC2 (6.6%)
20
0
−20
−40
−60
−40
−20
0
20
40
60
PC1 (15%)

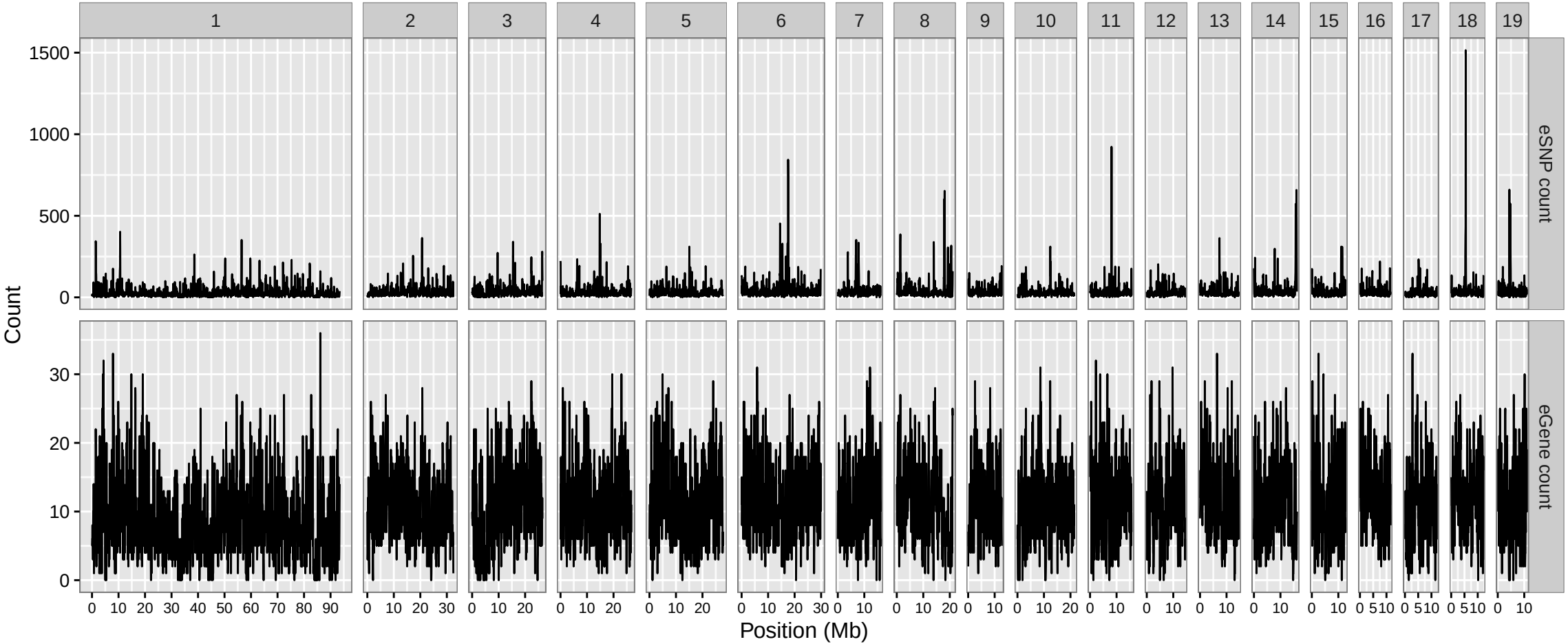

Latitude
Longitude
Elevation
East/West
Age
Precipitation January
Precipitation February
Precipitation March
Precipitation April
Precipitation May
Precipitation June
Precipitation July
Precipitation August
Precipitation September
Precipitation October
Precipitation November
Precipitation December
Temperature 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

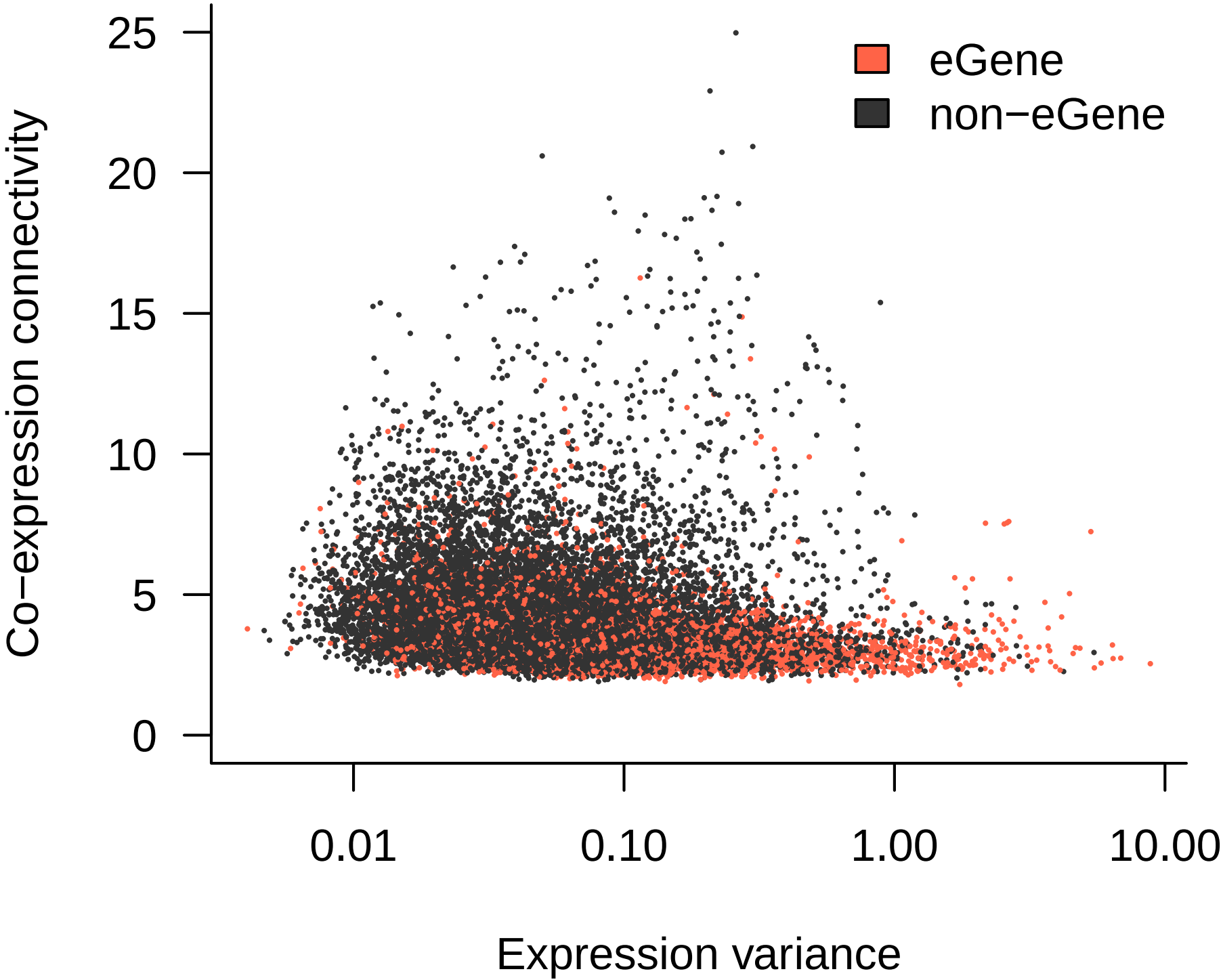

SwAsp
UmAsp
Gene expression rank
0
5000
15000
25000
0
20
40
60
80
5
10
15
Number of genotypes

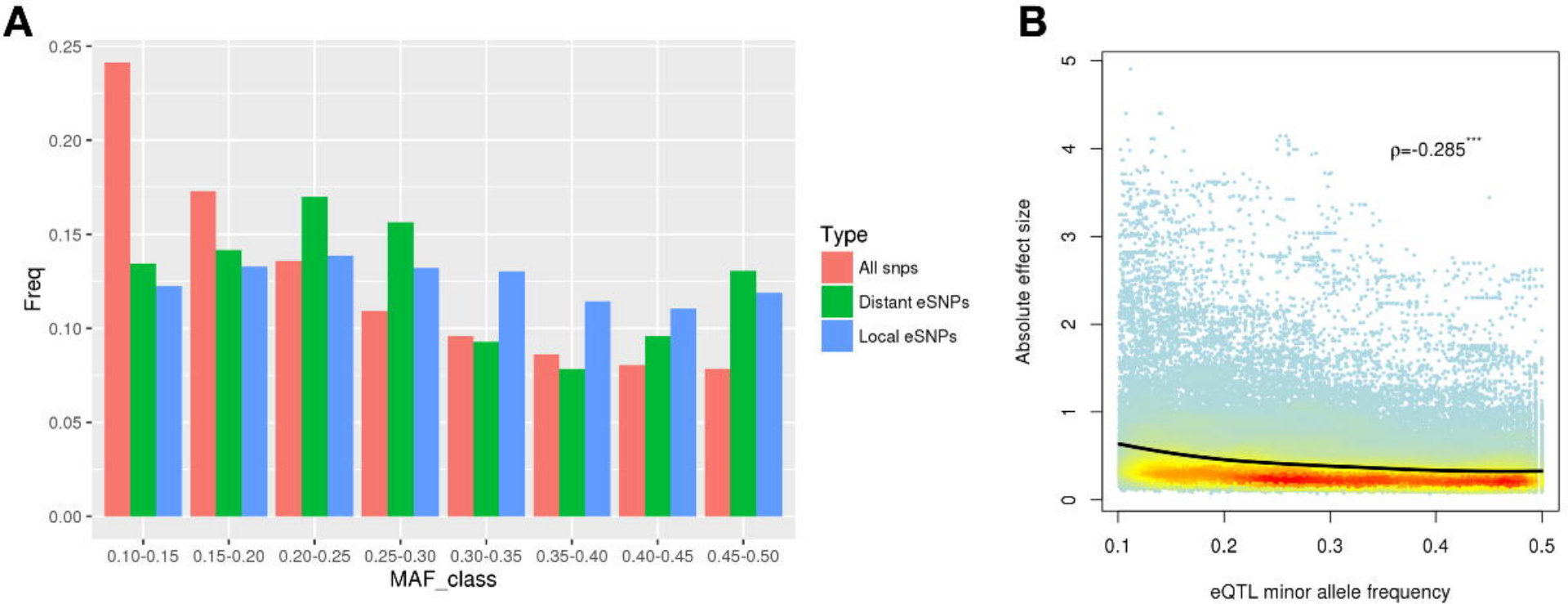

eQTL count
6000
4000
2000
0
−100
−50
0
50
100
Position relative to gene start/end (kb)

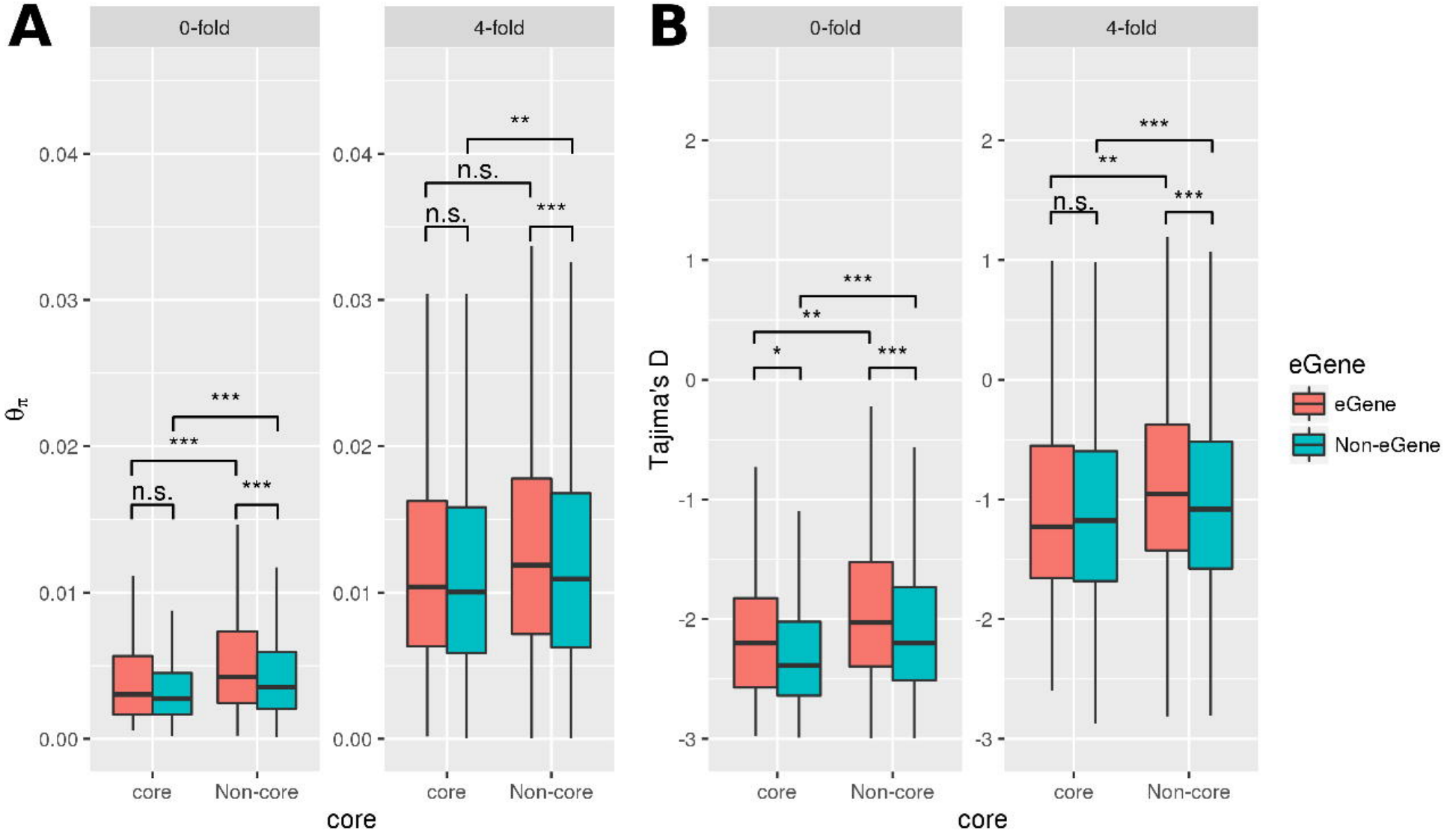

1.0
0.8
0.6
0.4
0.2
0.0
H^2
Both local and distant
Distant
Local
None

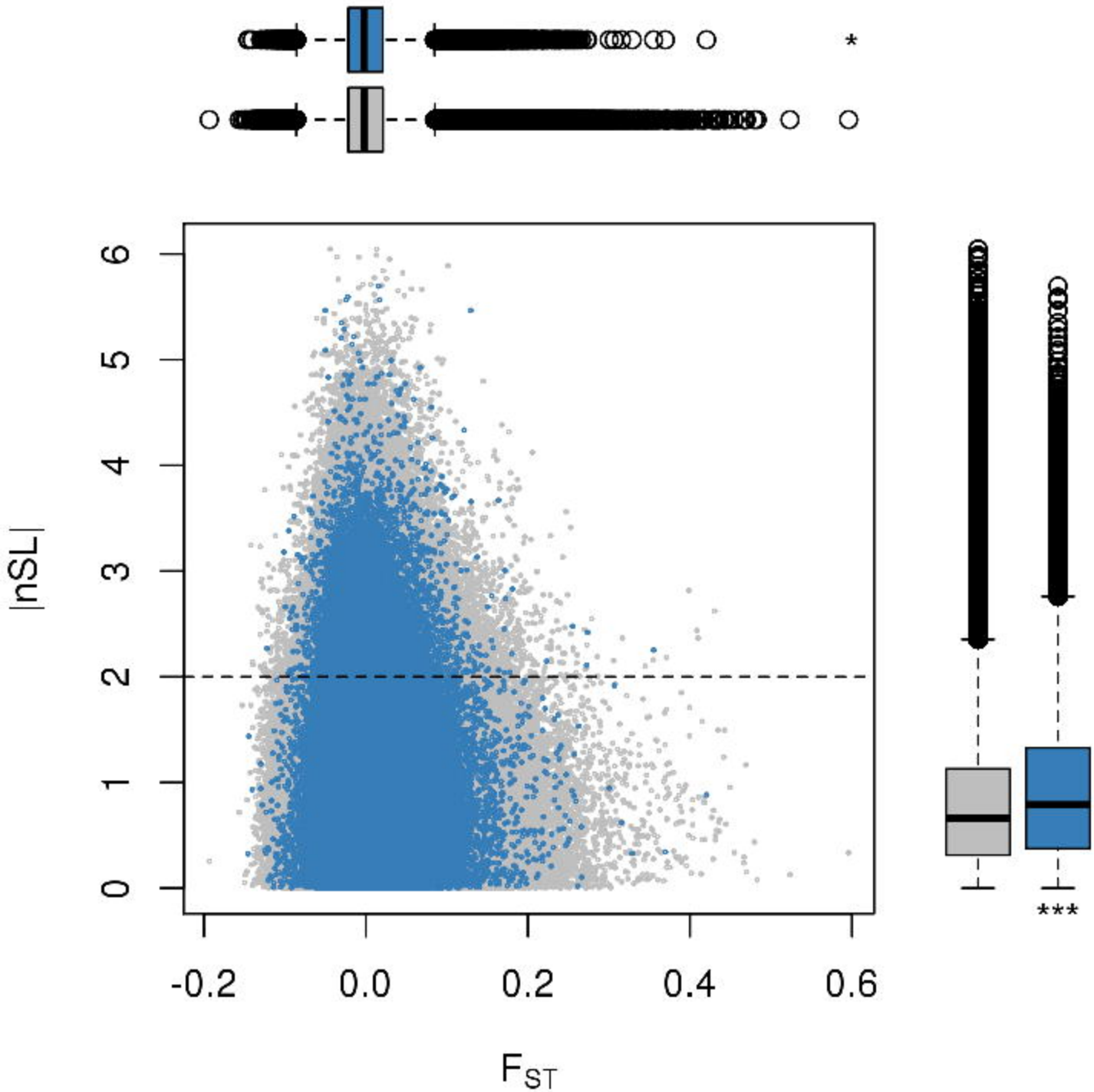

A
132258
13905
1023
110
25
98
120000
100000
80000
60000
40000
20000
0
Number of eSNPs
1
2
3
4
5
6
Number of eGenes / eSNP
B
6000
5000
4000
3000
2000
1000
0
Number of eGenes
0
500
1000
1500
Number of eSNPs / eGene

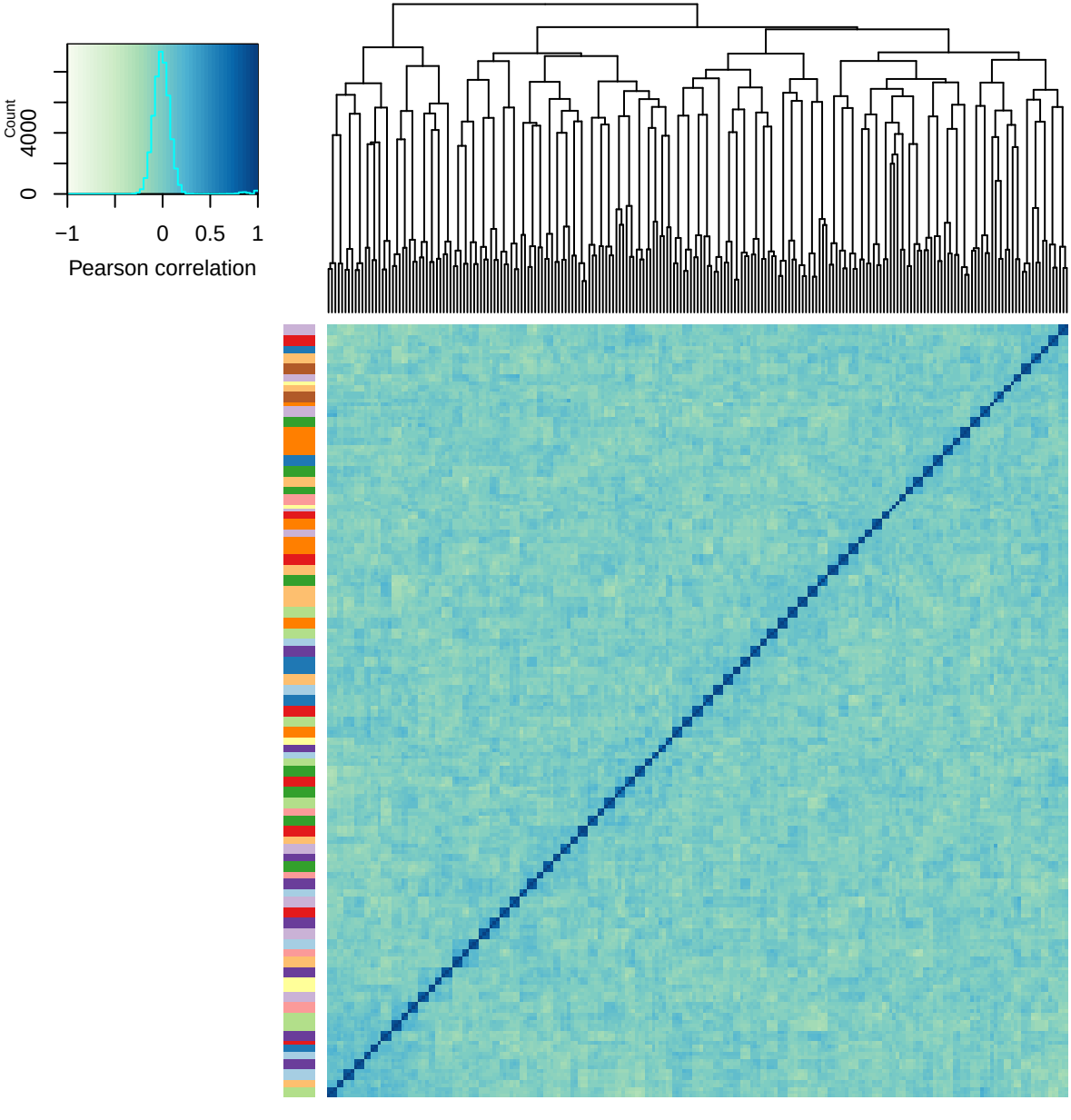

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
eSNP count
eGene count
Count
Position (Mb)


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

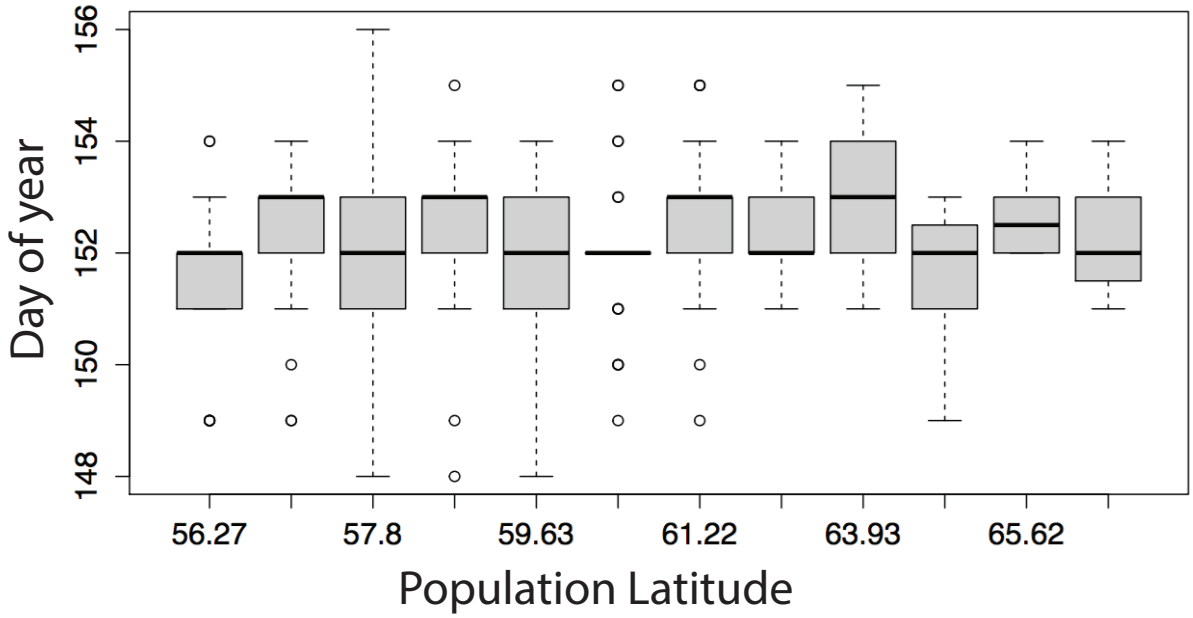

A
0.25
0.20
0.15
0.10
0.05
0.00
Freq
0.10-0.15
0.15-0.20
0.20-0.25
0.25-0.30
0.30-0.35
0.35-0.40
0.40-0.45
0.45-0.50
MAF_class
Type
All snps
Distant eSNPs
Local eSNPs
B
ρ=-0.285***
Absolute effect size
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
eQTL minor allele frequency


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

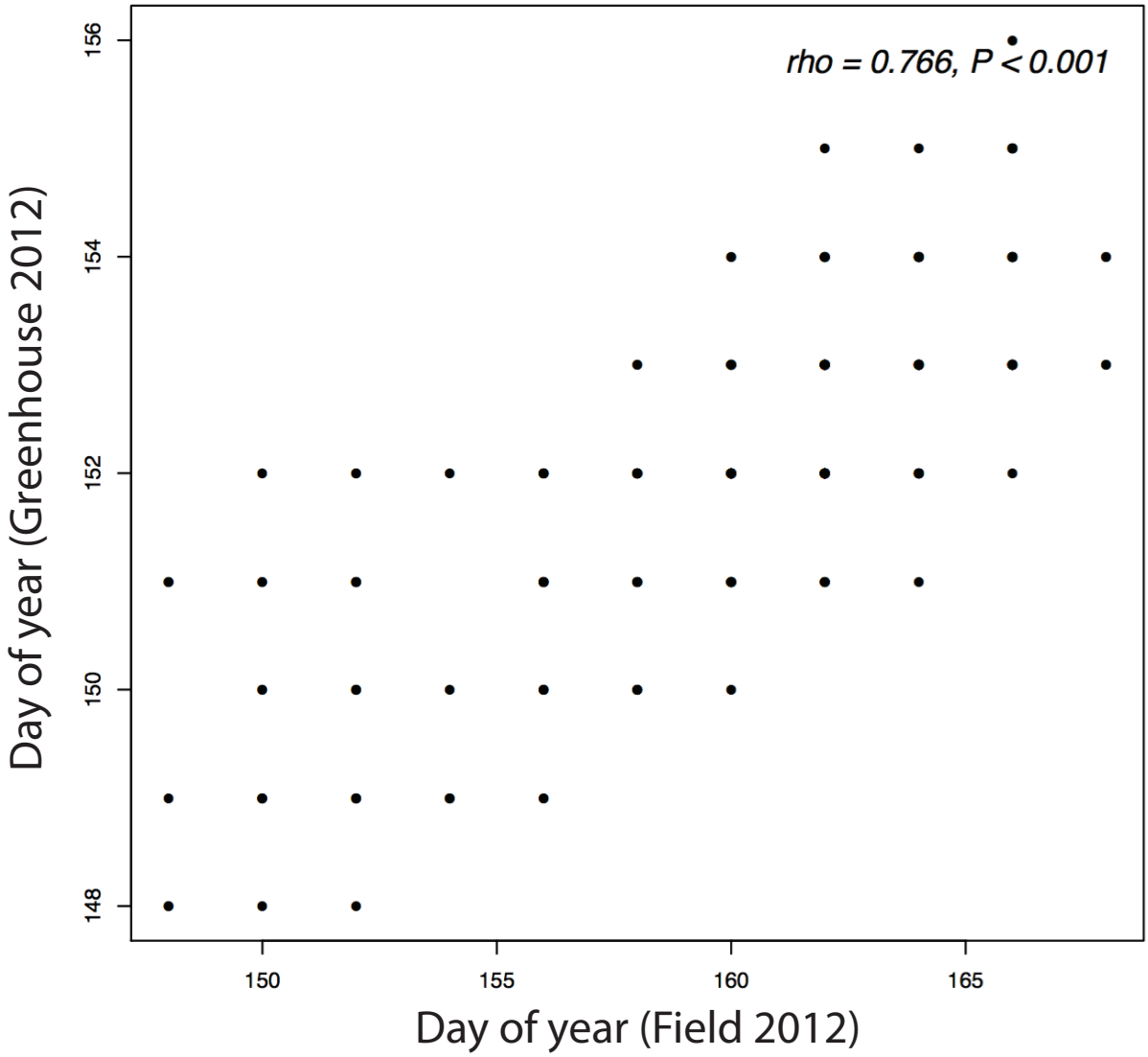

*
|nSL|
0
1
2
3
4
5
6
-0.2
0.0
0.2
0.4
0.6
F_{ST}
***

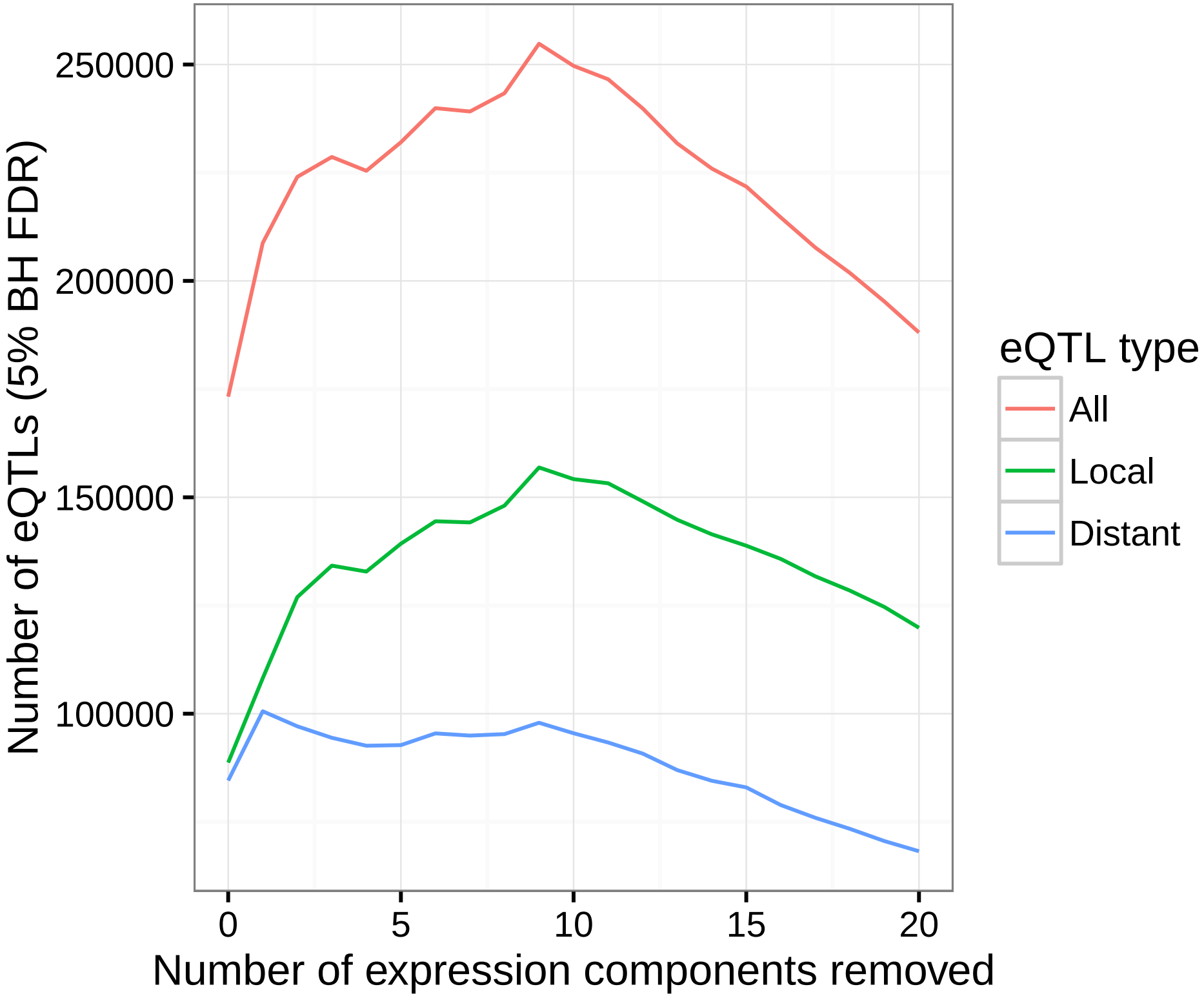

Count
4000
0
−1
0
0.5
1
Pearson correlation

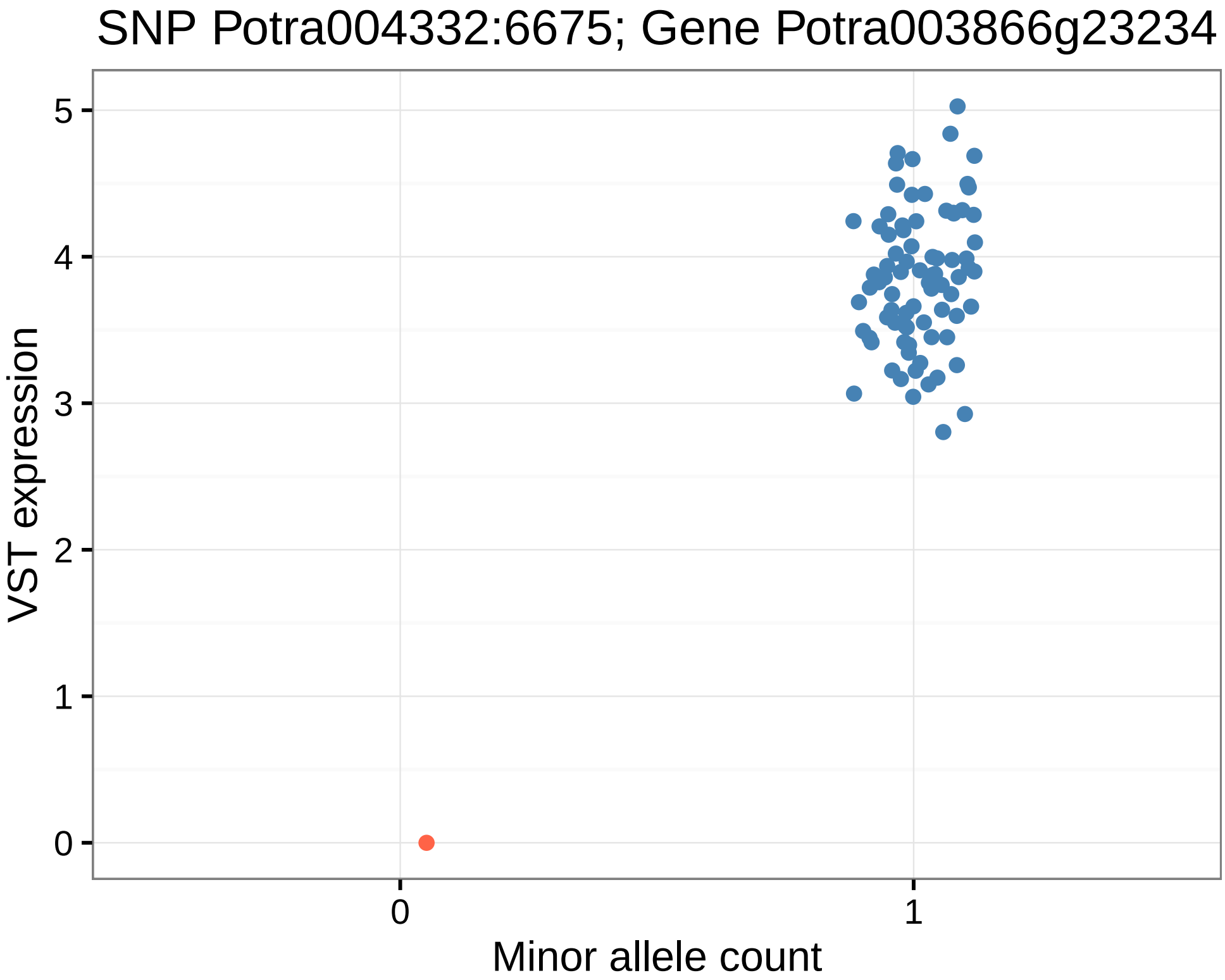
A


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