Gene co-expression network connectivity is an important determinant of selective constraint

Niklas Mähler[1,2], Jing Wang[3,4], Barbara K Terebieniec[2], Pär K Ingvarsson[3,5], Nathaniel R Street[2¶*] and Torgeir R Hvidsten[1, 2¶]

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

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

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

[4] Centre for Integrative Genetics, Faculty of Biosciences, Norwegian University of Life Sciences, Ås, Norway

[5]Department of Plant Biology, Swedish University of Agricultural Sciences, Uppsala, Sweden

[*] Corresponding author

E-mail: nathaniel.street@umu.se

[¶] These authors contributed equally to this work

Abstract

While several studies have investigated general properties of the genetic architecture of natural variation in gene expression, few of these have considered natural, outbreeding populations. In parallel, systems biology has established that a general feature of biological networks is that they are scale-free, rendering them buffered against random mutations. To date, few studies have attempted

to examine the relationship between the selective processes acting to maintain natural variation of gene expression and the associated co-expression network structure. Here we utilised RNA-Sequencing to assay gene expression in winter buds undergoing bud flush in a natural population of *Populus tremula*, and outbreeding forest tree species. 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, with local eQTLs having significantly higher effect sizes. 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 the relationship between network topology, the genetic architecture of gene expression and signatures of selection. Within the co-expression network, eGenes were underrepresented in network module cores (hubs) and overrepresented in the periphery of the network, with a negative correlation between eQTL effect size and network connectivity. We additionally found that module core genes have experienced stronger selective constraint on coding and non-coding sequence, with connectivity associated with signatures of selection. Our integrated genetics and genomics results suggest that purifying selection is the primary mechanism underlying the genetic architecture of natural variation in gene expression assayed in flushing leaf buds of *P. tremula* and that connectivity within the co-expression network is linked to the strength of purifying selection.

Author summary

Numerous studies have shown that many genomic polymorphisms contributing to phenotypic variation are located outside of protein coding regions, suggesting that they act by modulating gene expression. Furthermore, phenotypes are seldom explained by individual genes, but rather emerge from networks of interacting genes. The effect of regulatory variants and the interaction of genes can

be described by co-expression networks, which are known to contain a small number of highly connected nodes and many more lowly connected nodes, making them robust to random mutation. While previous studies have examined the genetic architecture of gene expression variation, few were performed in natural populations with fewer still integrating the co-expression network.

We undertook a study using a natural population of European aspen (*Populus tremula*), showing that highly connected genes within the co-expression network had lower levels of polymorphism, had polymorphisms segregating at lower frequencies and with lower than average effect sizes, suggesting purifying selection acts on central components of the network. Furthermore, the most highly connected genes within co-expression network hubs were underrepresented for identified expression quantitative trait loci, suggesting that purifying selection on individual SNPs is driven by stabilising selecting on gene expression. In contrast, genes in the periphery of the network displayed signatures of relaxed selective constraint. Highly connected genes are therefore buffered against large expression modulation, providing a mechanistic link between selective pressures and network toplogy, which act in cohort to maintain the robustness at the population level of the co-expression network derived from flushing buds in *P. tremula*.

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 much emphasis was, and is, placed on studying the genetics of those emergent phenotypes, less attention has been paid to the genetics of the various steps along the central dogma of molecular biology (in essence, the progression of genome to RNA to protein) that underlie the emergence of a phenotype of interest. The availability of massively parallel sequencing technologies affords new possibilities for addressing

64 biological questions, for example enabling the generation of *de novo* genome assemblies and
65 population-wide resequencing data that can be used to perform genome-wide association studies
66 (GWAS), even in species with large genomes that harbour high levels of polymorphism or that
67 display rapid linkage disequilibrium (LD) decay [1]. The use of genome-wide resequencing data
68 allows the discovery of, in theory, all genetic polymorphisms within an individual. These genetic
69 markers, of which single nucleotide polymorphisms (SNPs) are currently the most commonly
70 considered, can then be used to perform association or linkage mapping to identify the subset of
71 polymorphisms engendering phenotypic variation among individuals.

72 Advances in sequencing technologies have concordantly revolutionised transcriptomics studies,
73 particularly in non-model organisms. Following seminal work [2,3], numerous early studies in a range
74 of species established that there is a significant heritable component underlying natural variation of
75 gene expression levels among individuals within populations [4–16] and that this variation underlies a
76 number of phenotypes [17–24]. Given these findings, it became apparent that gene expression values
77 could be considered in the same way as any other quantitative phenotype and be subjected to linkage
78 or association mapping to identify polymorphisms contributing to expression level variation among
79 individuals [25], as first reported in [19], with the identified loci termed expression Quantitative Trait
80 Locus (eQTL; [6]) or, less commonly, expression level polymorphisms (ELPs; [26]). eQTLs are
81 classified as either local or distant acting depending on the physical location of the associated
82 polymorphism in relation to the gene that the eQTL is mapped for: local eQTLs are usually defined
83 as being located within a specified physical distance of the gene location on the same chromosome,
84 while distant eQTLs represent polymorphisms that are located beyond that threshold distance or on
85 another chromosome. eQTLs can be further classified as acting in *cis* or *trans*: *cis* eQTLs act in an
86 allele-specific manner and are usually considered to be local, although long-range *cis* interactions can
87 occur, for example when a polymorphism is located in an enhancer that is physically distant from the

88 gene of interest; *trans* acting eQTLs affect both alleles of a gene and are most commonly located
89 distant to that gene. There continues to be strong interest in eQTLs as they can identify mechanistic
90 links between phenotype and genotype [27,28]. Importantly, the majority of polymorphisms that
91 have been associated to phenotypes using GWAS in a wide range of species are located outside of
92 protein coding or transcribed regions [29–32], suggesting that they influence expression rather than
93 altering protein or transcript function.

94 A number of previous eQTL studies have been conducted using plant species including *Arabidopsis*
95 *thaliana* [33–38], *Zea mays* [39–41] and *Oryza sativa* [42,43], and in forest tree species [7,44–46]. These,
96 together with studies in other eukaryotic systems, have yielded generalities concerning the genetic
97 architecture of gene expression variation, including that a greater number of local eQTLs are
98 typically identified and that these individually explain a larger proportion of gene expression variance
99 than do distant eQTLs [34,47–50]. Much of the previous work was conducted using controlled, and
100 for tree species in particular inter-specific, crosses, comparisons of accessions or was performed in
101 non-natural systems and it is not clear how generally applicable their conclusions are for natural
102 populations of unrelated individuals. Few studies have considered whether observed, heritable
103 variation is adaptive [51,52] or whether signatures of population differentiation are observed at the
104 transcriptome level [52–57], and it is not yet clear the extent to which selection acting on gene
105 expression underlies adaptive phenotypic trait variation [52].

106 Systems biology has greatly improved our understanding of the shared regulation of genes, revealing
107 the topological properties of transcriptional co-expression networks. A salient feature of the
108 topology of co-expression networks is that they are scale-free, having few highly connected nodes
109 (genes) and many nodes with few connections. This property imparts an inherent ability to buffer
110 against single mutations of large negative effect as random mutation of an expression pattern (*i.e.* an

111 eQTL) or coding sequence will more often affect a network node of low connectivity [58–60].
112 Although there have been a number of eQTL studies performed, the context of eQTLs within the
113 co-expression network and how this relates to patterns of selection have not been a focus.

114 Species in the *Populus* genus have been established as a powerful model system for forest tree
115 genomics due to their relatively small genome, rapid growth, propensity for clonal propagation and
116 ease of genetic transformation [61]. *P. tremula* (European aspen) has many features that render it a
117 particularly useful model for population genetics and speciation studies [62,63], studies of which are
118 facilitated by availability of a draft *de novo* genome assembly [64] and population resequencing data
119 [65]. In this study, we aimed to determine the evolutionary forces that maintain the genetic variation
120 of gene expression within the context of the corresponding co-expression network using a natural
121 collection of *P. tremula*. Specifically, we wished to test whether co-expressed sets of genes are
122 enriched for specific biological functions, whether network topology influences gene expression and
123 sequence evolution of the constituent genes, and how selection interacts with network topology to
124 affect the patterns of genetic variation within populations. To address these questions we generated
125 population-wide RNA-Seq data in, assaying gene expression in winter buds at the point of bud break.
126 We performed eQTL mapping and constructed a co-expression network, which was scale free and
127 modular, with highly connected genes in the module cores being under-represented for eQTLs and
128 with eQTL effect size being negatively correlated with gene connectivity. Patterns of polymorphism
129 and divergence within genes in module cores imply that they are likely experiencing stronger selective
130 constraint relative to genes in the network periphery. Our results suggest that purifying selection
131 plays an important role in buffering the transcriptional network against large perturbations and that
132 natural variation in gene expression is more prevalent in genes of low network connectivity as a
133 result of relaxed selective constraint.

Results

We utilised the northern common garden (located at 63.9° N, near Umeå, Sweden) of the Swedish Aspen (SwAsp) collection [66], which comprises 116 *P. tremula* genotypes sampled from twelve geographic locations spanning the species distribution range in Sweden (56.2° to 66.4° N, Fig 1A). The SwAsp collection was previously shown to contain abundant genetic variation, with low linkage disequilibrium (LD) [63,67] and minimal population structure [68]. Based on whole genome re-sequencing data aligned to a *de novo* assembly of the *P. tremula* genome (available at PopGenIE.org; [64]), we called 4,509,654 SNPs after stringent filtering, which we utilised to perform eQTL mapping. To examine the genetic architecture of natural variation in gene expression within the SwAsp collection, we generated paired-end RNA-Seq expression data from winter buds at the point of spring bud break/flush for 219 individuals (clonal replicates), representing 86 genotypes.

Population level gene expression similarity

We first examined the distribution of broad-sense heritability (H^2) and population differentiation (Q_{ST}) (Fig 1B,C respectively) for all expressed, annotated genes. H^2 ranged from 0.0 to 1.0 with a mean (± s.d) of 0.30 (0.22) and with 5,924 genes (17%) having $H^2 > 0.5$. Permutation testing showed that 21,219 genes had significantly higher heritability than expected by chance ($p < 0.01$). There was a weak positive correlation between H^2 and median expression level (Pearson r = 0.09; df = 32,767; $p < 2.2\times10^{-16}$), and a relatively strong positive correlation to expression variance (Pearson r = 0.43; df = 32,767; $p < 2.2\times10^{-16}$). Q_{ST} ranged from 0.0 to 1.0 with a mean (± s.d) of 0.06 (0.12) and had a weak negative correlation with expression variance (Pearson r = -0.02; df = 29,670; $p < 4.5\times10^{-4}$) and a positive correlation with median expression (Pearson r = 0.18; df = 29,670; $p < 2.2\times10^{-16}$). These findings are similar to those reported for a number of species [5–17,53], suggesting that the

156 expression of a large proportion of genes is under substantial genetic control and that the expression
157 of highly expressed genes is under generally tighter genetic control than genes with lower expression.

158 While a number of previous studies have identified evidence for heritability of gene expression
159 [52,69,70], the relationship between expression variation and population structure has been explored
160 less. Our previous work has established that there is minimal population structure at the genetic level
161 in the SwAsp collection [68]. To examine whether population structure was apparent on the basis of
162 expression variation among genotypes, we performed hierarchical clustering of all individuals (Fig
163 1D) or genotypes (S1 Fig). While some evidence of clustering among genotypes was apparent
164 (regions of blue in Fig 1D, S1 Fig), genotypes (or individuals) did not cluster according to population
165 of origin (as indicated by the y axis color bars), suggesting that the observed clustering does not
166 result from population structure. We tested whether the clustering could be used to predict the
167 population of origin for genotypes by cutting the dendrogram to produce 12 clusters that were used
168 as the response in a multinomial logistic regression. The mean accuracy for a 10-fold cross validation
169 was extremely low (0.09). We additionally performed permutation tests, which showed that the mean
170 Q_{ST} for all genes was significantly lower than expected by chance ($p < 0.001$).

171 To identify whether genes with the highest H^2 and Q_{ST} were enriched for characteristic biological
172 functional signatures we selected the 500 genes with the highest H^2 (0.88-1.0, 0.93 ± 0.03) and 500
173 genes with the highest Q_{ST} (0.54-1.0, 0.71 ± 0.13) and subjected these to GO enrichment analysis
174 (see S1 File for all results). Genes with high H^2 were enriched for categories including protein
175 phosphorylation (GO:0006468; $p = 3.7\times10^{-6}$), while high Q_{ST} genes were enriched in terms including
176 translation (GO:0006412; $p = 4.2\times10^{-20}$) and gene expression (GO:0010467; $p = 4.6\times10^{-12}$). Likewise,
177 we considered the genes with the lowest values, which revealed enrichment of terms including cell
178 wall modification (GO:0042545; $p = 2.6\times10^{-7}$) for the 2,289 genes with an H^2 of zero and

179 enrichment of terms including amino acid activation (GO:0043038; p = 0.0012) among the 11,895
180 genes with a Q_{ST} of zero.

181 We performed a regression analysis to ascertain whether a set of geographic (latitude, longitude,
182 elevation), climatic (temperature, precipitation) or other (time since sample collection) factors
183 significantly explained the global patterns of gene expression similarity among genotypes (S2 Fig), as
184 identified by performing a PCA of the expression data. None of the gene expression principal
185 components (PCs) were significantly explained by these environmental factors, with the only
186 significant results found between PCs 2, 5 and 7 and the number of hours from collecting branches
187 from the field until bud samples were collected in the greenhouse, which explained 6.6%, 3.2%, and
188 total 2.1% expression variance, respectively.

189 We subsequently filtered expression values to remove unexpressed genes and uninformative
190 expression profiles with low variance, as these are uninformative for association mapping or for co-
191 expression analyses. Of 35,154 annotated genes, 20,835 were expressed in all samples, including
192 biological replicates, and 23,183 were expressed in all genotypes when considering genotype means.
193 Filtering to remove uninformative expression retained 22,306 genes, with the 12,848 removed genes
194 representing those that were either not expressed in our bud samples (6,736 genes with median
195 expression of zero of which 2,385 had no detectable expression at all), or that were weakly expressed
196 (1,762 genes with variance < 0.05 and median expression < 2), together with genes that had stable
197 expression among genotypes (4,350 genes with expression variance < 0.05 and median expression
198 >= 2). The latter potentially represent genes with canalised gene expression. Analysis of this set of
199 stably expressed genes identified enrichment for GO categories including protein transport
200 (GO:0015031, $p = 6.8\times10^{-11}$) and protein localisation (GO:0008104, $p = 2.2\times10^{-10}$). In contrast, the
201 500 genes with the highest variance across all samples were enriched for GO categories related to

protein phosphorylation (GO:0006468, $p < 10^{-6}$), chitin metabolic process (GO:0006030, $p < 10^{-4}$), and cell wall macromolecule catabolic process (GO:0016998, $p < 10^{-4}$). Comparing the variance of these 500 genes with mean F_{ST} calculated using SNPs within those genes revealed no apparent relationship, suggesting that these patterns were not the result of population structure.

A recent reanalysis [71] of two existing datasets assaying gene expression among natural accessions of *A. thaliana* [72,73] observed that thousands of genes displayed clear present/absent expression among accessions. In contrast, when filtering our data using a similar approach, we did not find any genes displaying this pattern of expression variation (S3 Fig), an observation that we also confirmed in an independent *P. tremula* dataset [74], albeit containing substantially fewer genotypes.

eQTL mapping

To explore the genetic architecture of gene expression variation among genotypes, we performed eQTL mapping, defining an eQTL as a significant association between a SNP (termed an eSNP) and the expression of a gene (termed an eGene). Furthermore, we classified an eQTL as *local* if the eSNP was located on the same chromosome and not more than 100 kbp from the associated eGene, and as *distant* otherwise. Our threshold distance for local/distant classification was empirically determined based on the distribution of distances between eSNPs and their associated genes and the assumption that most detectable eQTLs located within one chromosome were local (S4 Fig). We did not consider whether eQTLs acted in *cis* or *trans*. In common with other studies [75–77] we removed hidden confounders in the expression data prior to mapping eQTLs by removing variance attributable to the first nine PCs of the expression data, removal of which maximised the number of eQTLs identified (S5 Fig). After removing these hidden confounders, we repeated the gene expression clustering analysis and observed that the previous sample clustering was no longer apparent (S6 Fig).

225 In total we identified 164,290 eQTLs at a 5% empirical FDR: 131,103 local and 33,187 distant. These
226 eQTLs represented pairwise associations between 6,241 unique genes (eGenes; 28% of genes
227 considered) and 147,419 unique SNPs (eSNPs), with a mean of 21.0 local and 5.3 distant eSNPs per
228 eGene, respectively. 4,091 genes had only local eQTLs, 1,050 had only distant eQTLS while 1,100
229 had both. Local eSNPs explained significantly more of the variance than distant eSNPs (local mean
230 adjusted %VE = 51, distant mean adjusted %VE = 47, Mann-Whitney $p < 2.2\times10^{-16}$, Fig 2A) and
231 also had higher statistical significance (Mann-Whitney $p = 6.9\times10^{-12}$, Fig 2E). As expected there was a
232 clear tendency for a local eSNP to be located proximal to the transcription start site (TSS) or the stop
233 codon (S10 Fig). eGenes had 229 significantly higher heritability than non eGenes (median
234 heritability difference was 0.16, permutation 230 test $p < 0.0001$) (Fig 2B), with this trend being
235 slightly higher for local than distant eQTLs (S25A Fig). There was also an expected, positive
236 correlation between the maximum %VE of the eSNPs associated with an eGene and gene
237 expression H^2 (Pearson $r = 0.47$, df = 6,232, $p < 2.2\times10^{-16}$). These patterns are broadly similar to
238 those reported in a number of previous studies [34,47–50], although the ratio of local-to-distant
239 eQTLs differs among studies and is highly influenced by sample size. Before hidden confounder
240 removal, eGenes had marginally higher mean expression than non-eGenes (mean expression 3.5 and
241 3.3, respectively; permutation test p-value < 0.0001). There were no significant differences after
242 hidden confounder removal, regardless of whether the eQTL was local or distant. eGenes with at
243 least one local eQTL were enriched for GO categories related to tRNA metabolic process
244 (GO:0006399, $p = 1.5\times10^{-5}$), ncRNA metabolic process (GO:0034660, $p = 2.6\times10^{-5}$) and
245 organonitrogen compound biosynthetic process (GO:1901566, $p = 2.2\times10^{-5}$), among others, while
246 eGenes with at least one distant eQTL were enriched for categories including protein
247 phosphorylation (GO:0006468, $p = 0.0064$; see S1 File for all results).

248 In contrast to a number of previous studies [19,34,36,47,50], we did not find evidence for any
249 distantly acting hotspots (Fig 2C, S7 Fig), which represent loci where large numbers of *trans*-acting
250 variants are co-located. Although the removal of hidden confounders has been shown to improve
251 the signal:noise ratio for eQTL mapping [75–77], it is possible that the process may remove the
252 signature of large effect distantly-acting hotspots. We performed eQTL mapping before hidden
253 confounder removal and observed one hotspot representing 12 eSNPs assocaiated with the
254 expression of 278 genes that was not present after removing hidden confounders (S8 Fig). The 12
255 eSNPs are located in close physical proximity and are therefore likely linked. In our data, the vast
256 majority of eSNPs were associated with a single eGene (132,258 eSNPs) with a maximum of six
257 eGenes associated with a single eSNP (S9A Fig). In contrast, only 1,248 of the 6,241 eGenes were
258 associated with a single eSNP, with eGenes associated with up to 1,547 eSNPs (S9B Fig). In cases
259 where eSNPs associated with the expression variation of a single eGene are physically close together,
260 these eSNPs may be identified due to linkage rather than all being causative. To account for this we
261 fitted linear models between the expression of each eGene and all the significant eSNPs for that
262 gene, both local and distant. The use of a linear model masks eSNPs that contain
263 identical/redundant information and thus effectively identifies haplotype blocks present in all
264 individuals (which we refer to as 'unique eSNPs'), while also producing a measure of how well the
265 combination of eSNPs explains the expression of the corresponding eGene (in terms of percentage
266 variance explained, %VE). Of the 4,993 eGenes associated with more than one eSNP, 4,703 were
267 also associated to more than one unique eSNP, of which 4,210 genes were associated with at least
268 one local eSNP and 1,203 were associated with at least one distant eSNP. The adjusted %VE for the
269 combination of eQTLs was, in general, higher (mean %VE 51.1) than for single eSNPs (mean %VE
270 44.3).

271 We next considered the genomic context of eSNPs, which was determined by intersecting eSNP
272 positions with gene annotations. After normalising for feature length, the majority of local eSNPs
273 were located within untranslated regions (UTRs) and up- or down-stream (regulatory) regions of
274 genes, with distinctly lower representation within exons than introns (Fig 2D). The genomic context
275 distribution of local and distant eSNPs was largely similar, although there were distinctly more
276 eSNPs located within intergenic regions for distant eQTLs. These distributions patterns are
277 consistent with previous findings in natural populations of humans [76,78], *Drosophila* [79] and
278 *Capsella grandiflora* [27]. A local eSNP/eQTL can be located within the region of the associated eGene
279 (5'/3' 2 kbp flanking, 5'/3' UTR, exon, intron), within the region of a gene other than the associated
280 eGene or within an intergenic region. We found that approximately half of the local eSNPs were
281 located within the region of the eGene itself and half within another gene, with relatively few local
282 eSNPs located in intergenic regions. When the eSNP was located within the gene region of the
283 associated eGene there was a clear tendency for that eSNP to be located proximal to the
284 transcription start site (TSS) or the stop codon (S10A Fig). This patterns was not present in the cases
285 where the eSNP was located in the adjacent or another gene (S10B,D Fig), even after accounting for
286 strand (S10C,D Fig). In these cases there was also a lower tendency for the eSNP to be located
287 within the gene body, with a generally higher presence of eSNPs in the flanking gene regions. Given
288 this pattern, we therefore examined the expression correlation of the eGene to the gene in which the
289 eSNPs was located, contrasting this to pairs of non-eGenes and pairs where the eSNP and located
290 within the eGene and the adjacent gene (S11 Fig). For those cases where the eSNP was not located
291 within the eGene, we observed a higher expression correlation between the eGene and the gene in
292 which the eSNP was located, potentially indicating that the eSNP induces a local and more general
293 influence on expression. As UTRs are currently not well annotated in *P. tremula* it should be noted
294 that many SNPs currently classified as being located in flanking regions may actually reside within

UTRs. The global distribution of genomic contexts for all investigated SNPs (regardless of whether they were eSNPs or not) was similar to that of both local and distant eSNPs, suggesting no notable ascertainment bias for eSNPs.

Co-expression network

Systems biology studies typically consider datasets assaying gene expression throughout development, among tissue types or in response to abiotic, biotic or genetic perturbations. A characteristic and salient feature of the resultant co-expression networks is that they are scale-free [80]. To determine whether the co-expression network representing expression variation among individuals within our natural population displayed the same properties, we used genotype mean gene expression values, after removal of hidden confounders, to calculate a co-expression network. In common with other biological networks, the network was scale-free (R^2 = 0.97), suggesting that the genetic polymorphisms underlying the observed expression variance induce similar co-expression structures to those observed in previous systems biology studies. We compared the correlation and variance properties of our dataset to that of the *P. tremula* expression atlas (exAtlas; [64]), which represents different tissues collected from a single genotype. The correlation distribution for the exAtlas samples was much wider (mean correlation 0.01 ± s.d. 0.36) than that of our population expression data (mean correlation 0.00 ± s.d. 0.12; Kolmogorov-Smirnov D = 0.14, $p < 2.2\times10^{-16}$). The expression variance for the SwAsp expression data was also significantly lower than in the exAtlas data (Wilcoxon signed rank test, V = 18274602, $p < 2.2\times10^{-16}$; Fig 3A).

Clustering analysis of the co-expression network identified 38 co-expression modules (S12 Fig) containing a total of 20,686 genes (min=86 genes, max=1591 genes). These were enriched for a number of Gene Ontology (GO) categories including translation (modules 9, 10, and 14),

photosynthesis (module 22) and oxidation-reduction process (module 29; for all results see S1 File). Despite the narrow distribution of correlation values, the modules were reasonably well defined, as indicated by the normalised connectivity difference (k_{diff}), *i.e.* the difference between intra- and inter-modular connectivity. All modules exhibited a positive mean k_{diff}, with only 157 genes (0.7%) having a negative k_{diff}. This was in stark contrast to genes assigned to the 'junk' module (*i.e.* all genes not assigned to any well-defined module), where there were 480 genes with negative k_{diff} (29%).

eGenes are under-represented in network module cores

We examined the relationship of eGenes and eQTLs to network connectivity, determining significance of these results using permutations tests shuffling gene assignments while maintaining the network structure (*i.e.* node gene IDs were shuffled while edges remained constant). In general, eGenes had lower connectivity and betweenness centrality than non-eGenes, but the effect was minimal (mean difference of 0.79 and 8.9×10^{-5}, respectively; permutation test, $p < 0.0001$ for both, Fig 3B). Moreover, genes with a positive k_{diff} were significantly under-represented for eGenes (permutation test, $p < 0.0001$). We defined the core of each module as the 10% of genes in a module with the highest normalised k_{diff} while also having an intra-modular connectivity >1. Using this definition, all 38 modules contained at least one core gene, with the percentage of core genes ranging from 2-10% (S2 File, S12 Fig). Before removal of hidden confounders, core genes had both higher mean expression and variance than non-core genes (difference of 1.1 and 0.16, respectively; permutation test, $p < 0.0001$). However, the co-expression network was inferred from the expression data after the removal of hidden confounds, which removed these difference (no difference in mean expression: permutation test $p = 0.39$, and marginal difference in variance: mean variance of 0.12 and 0.11, permutation test $p = 0.0184$, S13 Fig). Additionally, there was a weak negative relationship between network connectivity and gene expression variance in the network as a whole (Pearson r = -

340 0.08; df = 22,304; $p < 2.2\times10^{-16}$; S14 Fig) and eGenes of the highest connectivity had lower effect
341 size (based on the maximum effect size across all eQTLs associated to each eGene, Pearson r = -
342 0.15; df = 6,239; $p < 2.2\times10^{-16}$, S15). Among the module cores, 28 contained at least one eGene, with
343 25 module cores being significantly under-represented for eGenes (permutation test, $p < 0.05$; Fig
344 3D). This further emphasised that, in general, eGenes were not central in the network. There was at
345 least one eSNP located within the region of a core gene (5'/3' flanking, 5'/3' UTR, exon, intron) for
346 32 of the 38 modules, with module cores also being under-represented for eSNPs (permutation test,
347 $p < 0.0001$). For a number of metrics, module 23 had notable differences to the general pattern: It
348 contained few core genes, all of which were eGenes; the core genes had high variance compared to
349 non-core genes; the module in general was enriched for eGenes.

350 We tested whether the periphery of the network was over-represented for eGenes. Sixty-four of 145
351 peripheral genes were eGenes, representing a significant enrichment (permutation test, $p < 0.0001$).
352 On the other hand, network module cores were enriched for both transcription factors (permutation
353 test, $p < 0.0001$), which had higher connectivity than non-transcription factors (permutation test, $p <$
354 0.0001), and phylogenetically conserved genes (permutation test, $p < 0.0001$), defined as genes with
355 orthologs in *P. tremuloides*, *P. trichocarpa* and *A. thaliana* (20,318 genes; [81]). Furthermore, *P. tremula*-
356 specific genes, *i.e.* genes without orthologs in *P. tremuloides*, *P. trichocarpa* or *A. thaliana* (1,614 genes),
357 were slightly under-represented in network module cores (permutation test, $p = 0.009$) while being
358 slightly over-represented among eGenes (permutation test, $p = 0.0076$).

359 In addition to eGenes having generally lower connectivity, there was also a negative relationship
360 between eQTL effect size and co-expression connectivity for both local (Pearson r = -0.15; df =
361 5,189; $p < 2.2\times10^{-16}$) and distant eQTL (Pearson r = -0.12; df = 2,148; $p < 6.2\times10^{-8}$). The H^2 of
362 eGenes within the core was lower than for eGenes outside the core (mean difference of 0.10,

permutation test $p < 0.0001$, Fig 3C) and H^2 correlated negatively with connectivity in the network as a whole (Pearson $r = -0.30$; $df = 22{,}304$; $p < 2.2\times10^{-16}$).

We examined the distribution of the mode of action (Fig 3D) and genomic context (Fig 3E) of eSNPs within the network. There were distinctly more distantly acting eSNPs within the core than the non-core (Fig 3D) and, of these, there were more distantly acting eSNPs located within exons of core eGenes compared to core non-eGenes, which had higher representation of distantly acting eSNPs located in UTRs (Fig 3E). The genomic context distribution of local acting eSNPs was similar in all cases (core/non-core and eGene/non-eGene; Fig 3E), however it is clearly apparent that non-core eGenes contained by far the greatest density of eSNPs.

Paralogs with diverged expression are more likely to be eGenes

The Salicaceae lineage underwent a relatively recent (58 million years ago) whole-genome duplication (WGD) shared by all member species and that remains represented by a large number of paralogous gene pairs in the genomes of *Populus* species [82]. If many of these duplicated genes are functionally redundant or in the process of diverging, one would expect them to be overrepresented for eGenes as sub- or neo-functionalisation requires derived SNPs to drive expression or coding divergence. To test for evidence of this we considered paralogous pairs of genes derived from the WGD event. In *P. tremula* 3,910 paralog pairs were detected [81], with 2,140 of these (4,185 unique genes) passing our gene expression and variance filtering criteria. These paralogs were significantly under-represented for eGenes, with 1,078 of the 4,185 genes having at least one associated eSNP (hypergeometric test, $p = 0.0004$). Comparing the expression correlation of paralog pairs to that of random gene pairs showed that paralogs exhibited conserved regulation (permutation test, $p < 0.001$). We compared the expression correlation distributions of paralog pairs containing 0, 1, and 2 eGenes (Fig 4) and found that a higher number of eGenes in a pair was associated with lower expression correlation (linear

model $\hat{\beta} = -0.06$, $p < 2.2\times10^{-16}$). Excluding paralogs did not alter the fact that eGenes had significantly lower network connectivity than non-eGenes.

The rate of sequence evolution is associated with network topology

To assess whether genes with and without eQTLs (eGenes vs. non-eGenes) are experiencing different levels of selective constraint, we assessed nucleotide diversity (θ_π) and the site frequency spectrum of segregating mutations using Tajima's D in different genomic contexts. We found that, irrespective of whether eQTLs were local or distant, eGenes exhibited significantly higher genetic variation than non-eGenes (Fig 5A, S1 Table). Furthermore, although Tajima's D values are overall negative in *P. tremula,* likely reflecting a historical range expansion [63], eGenes exhibited significantly higher Tajima's D values compared to non-eGenes (Fig 5C, S1 Table), again regardless of whether the associated eQTL was local or distant. As such, non-eGenes appear to be experiencing stronger selective constraint than eGenes. These patterns were consistent for the different genetic contexts considered (S16 Fig), likely reflecting the effects of linked selection [63]. To test this hypothesis we looked for differences in selection efficacy on protein sequence evolution between eGenes and non-eGenes by examining the ratios of intraspecific nonsynonymous to synonymous polymorphisms ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and interspecific nonsynonymous to synonymous substitutions (d_N/d_S) across these genes (using *P. trichocarpa* as an outgroup). Both $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and d_N/d_S estimates were significantly lower in non-eGenes compared to eGenes (Fig 5E,G, S1 Table), likely reflecting stronger purifying selection acting on non-eGenes. As has been reported previously for *Capsella grandiflora* [27] and a wild population of baboons [83], we observed a negative correlation between minor allele frequency and eQTL effect size (S17 Fig). We examined this relationship in permuted data, which revealed an excess of low MAF SNPs in the permuted data compared to the original data (S18 Fig). This is in contrast to the results in *Capsella* [27], where the reverse was observed. However, in both [27] and in

409 our results there was a consistent negative relationship between MAF and effect size. The
410 comparative differences in MAF distributions for real and permuted data between the two studies
411 may largely result from statistical differences in how the test statistic and p-values were calculated,
412 but may also reflect differences in the population genetics of the two systems. One concern is that
413 this relationship may result from a higher false positive rate at lower MAF due to the concomitant
414 decrease in sample size. To address this concern, we performed a subsampling analysis, similarly to
415 [27], to remove the effect of MAF and examined the correlation of effect sizes estimated in the
416 original and sub-sampled datasets. The high correlation observed (S19 Fig) suggests that the
417 observed negative relationship between effect size and MAF is not artefactual.

418 To further examine the relationship between network topology and sequence evolution, we
419 contrasted genes located in module cores and those not (core vs. non-core). Independent of whether
420 a gene was an eGene or not, core genes had significantly lower levels of genetic diversity and
421 Tajima's D values compared to non-core genes (Fig 5B,D). In addition, core-genes also had
422 significantly reduced ratios of non-synonymous to synonymous polymorphisms ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and
423 substitutions (d_N/d_S) (Fig 5F,H). Again, these patterns were consistent across different genomic
424 contexts (S16 Fig). Taken together, these results suggest that genes in network module cores
425 experience reduced rates of molecular evolution due to stronger purifying selection, i.e. selective
426 removal of deleterious mutations, and are therefore evolving under stronger selective constraint
427 compared to non-core genes. Stronger purifying selection on mutations within core genes is likely
428 driven by stronger stabilising selection of gene expression noise or modulation acting to maintain the
429 optimal level of expression in core, compared to peripheral, genes [84].

430 As the sequence evolution of a given gene is known to correlate with different factors [85], such as
431 gene expression level or variance, the evolutionary age of a gene [86,87] and, as we show, the

presence or absence of an eQTL and the topology of the co-expression network (S20 Fig), we performed analyses to ascertain their relative roles in determining patterns of sequence evolution. Owing to the collinearity of various characteristics of expression (S20 Fig), we performed principal component analysis (PCA) on representative gene expression measures (before hidden confounder removal) to examine the extent to which these measures were interdependent. This analysis revealed that PC1, which explained 37.03% of the variation in these five measures, was mainly dominated by the connectivity of genes in the co-expression network and whether they are located within network module cores or not (Fig 6A). Gene expression variance and eGene status showed a strong influence on PC2 (Fig 6A) while expression level contributed largely to PC3 (Fig 6A).

In order to test whether the correlations between sequence evolution and various characteristics of expression are independent of one another, we calculated correlations (Spearman's rank) between sequence evolution rates and PCs. The rates of sequence evolution over both short (represented by within-species diversity: θ_π, Tajima's D and $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and long (represented by between-species divergence, d_N/d_S) timescales showed significantly negative correlations with the connectivity and core status of genes in the co-expression network (PC1) (Fig 6B-E). This indicates that the connectivity of genes within the network, both globally and within the local context of expression modules (core status), are key factors associated with the rates of sequence evolution. PC2, which largely reflected gene expression variance showed significant positive correlations with patterns of genetic diversity within species (θ_π and Tajima's D), and had significant, although relatively weak, negative correlation with the rate of protein sequence evolution (Fig 6B-E). In addition, in accordance with other studies [88], gene expression level (largely represented by PC3) showed significantly negative correlations with the rate of protein sequence evolution (Fig 6B-E). We then performed partial Spearman correlation, aiming to estimate the relationship between two variables

455 while controlling for other variables, between sequence evolution measures and the five measures of
456 gene expression used in the above PCA (Table 1). We found the correlation between sequence
457 evolution and connectivity of genes within the co-expression network persists, although slightly
458 lower, after accounting for the effects of gene expression level and other factors (Table 1).

459

460 **Table 1**. Summary of the correlation coefficients (Spearman's rank correlation coefficient) between
461 measures of sequence evolution and gene expression-related measures.

	θ_π		Tajima's D		$\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$		d_N/d_S	
	Pairwise	Partial[a]	Pairwise	Partial[a]	Pairwise	Partial[a]	Pairwise	Partial[a]
Connectivity	-0.2197***	-0.1440***	-0.1929***	-0.1164***	-0.1352***	-0.0417***	-0.1529***	-0.0520***
Core vs. Non-core	-0.0848***	-0.0045ns	-0.0938***	-0.0153*	-0.0652***	-0.0146*	-0.0588***	0.0009ns
Expression level	-0.1312***	-0.0729***	-0.0817***	-0.0370***	-0.2213***	-0.1914***	-0.2792***	-0.2441***
Expression variance	0.1238***	0.1263***	0.0580***	0.0587***	0.0356***	0.0320**	0.0302**	0.0241**
eGene vs. Non-eGene	0.1677***	0.1159***	0.1869***	0.1419***	0.0540***	0.0457***	0.0309**	0.0265**

462 [a] Partial correlations between two specific variables after controlling for all other gene-expression
463 related measures.

464 ns=not significant (P>0.05)

465 *P<0.05

466 **P<0.001

467 ***P<1e-5

468 # Discussion

469 Our primary aim was to determine the evolutionary forces maintaining genetic variance associated
470 with gene expression variation and to determine their relationship to the associated gene co-
471 expression network topology within a wild-collected, outbreeding population of the forest tree
472 *Populus tremula*. We first established that, in common to results in other species [4–16,52], there is
473 prevalent heritability in gene expression levels, with 17% of the genes (5,924) having $H^2 > 0.5$. To
474 examine the genetic architecture of heritable gene expression variation within the population, we
475 performed eQTL mapping, from which we identified more local than distant eQTLs, with local
476 eQTLs explaining significantly more of the variance in gene expression than distant eQTLs (Fig 2A),
477 which is similarly in agreement with previous studies across a range of species [34,47–50]. Although
478 each eSNP was typically associated with only a single gene, many genes were associated with more
479 than one unique eSNP, indicating that numerous loci influence the expression of genes with an
480 associated eQTL. For a large proportion of eGenes, the identified set of eSNPs explained a relatively
481 high proportion of the heritable expression variation (median 1.0 and s.d. 2.3; S21 Fig). While it
482 appeared that a single eSNP often accounted for a large proportion of the explained heritable
483 variance, due to linkage it is not possible to determine the relative contribution of each eSNP within
484 a haplotype block, with the apparent contribution of individual eSNPs largely resulting from the
485 order in which they are entered into the statistical model. While eGenes had higher H^2 than non-
486 eGenes, as is expected, there were, nonetheless, 2,780 non-eGenes with $H^2 > 0.5$, potentially
487 reflecting that the expression variance of these genes is controlled by many eSNPs of small effect
488 size that we lacked the power to detect.

489 To gain insight into how eQTLs may be influencing expression of the associated eGene, we
490 examined their genomic context. Local eQTLs were most frequently located in regulatory regions
491 (Fig 2D), with UTRs having the highest density of local eSNPs (~1.5 eSNP per kbp) followed by
492 flanking regions (2 kbp up- and downstream, ~1 per kbp) and introns (~0.75 per kbp), which is

493 consistent with previous findings in natural populations of humans [76,78], *Drosophila* [79] and
494 *Capsella grandiflora* [27], likely indicates that these loci cause regulatory changes in the transcriptional
495 dynamics of the gene.

496 To address our primary question, we used the gene expression values to construct a co-expression
497 network. Compared to networks more typical of systems biology studies, for example the *P. tremula*
498 exAtlas network [64], where samples originated from different tissues of a single genotype, the
499 pairwise expression correlations underlying our co-expression network were low. Despite this, the
500 network displayed typical characteristics, being scale-free with hubs and distinct modules [80]. To
501 determine whether network topology is related to the evolutionary history of its component genes,
502 we examined correlations between network connectivity and rates of sequence evolution. Scale-free
503 gene co-expression networks are defined as robust, or buffered, to the effects of random mutations.
504 As there are few highly connected genes (which are important determinants of the observed co-
505 expression structure), a random mutation would be unlikely to affect such a gene. In comparison, as
506 the majority of genes have low connectivity, a random mutation is more likely to affect such a low
507 connectivity gene, which have few connections to other genes in the network and the effect of this
508 mutation will therefore be minimal (i.e. the overall network is buffered from, or robust to, the
509 mutational effect) [89]. In network modules, where many genes share the same expression pattern, a
510 single eSNP modulating the expression of a central regulator would be sufficient to induce similar
511 expression variation (i.e. co-expression) in the set of connected genes within that network module.
512 The buffered characteristic of the network would therefore hold true even if all genes within the
513 network are exposed to the same evolutionary history – i.e. that all are equally likely to accumulate
514 mutations. However, natural selection may interact in combination with network topology, for
515 example to prevent the accumulation of mutations within specific genes [90]. If the distribution of
516 mutations is not random across the network, such an interaction can be inferred. In our data, genes

517 in module cores (the set of the most highly connected genes within a module) were under-
518 represented for eGenes and, respectively, those in the network periphery were enriched for eGenes
519 (Fig 3B), suggesting that polymorphisms associated with natural variation in expression are more
520 likely to affect genes of low connectivity. More generally, eGene connectivity was negatively
521 correlated with variance explained by the associated eSNPs (S15 Fig). It has similarly been reported
522 that hubs in human protein-protein interaction networks are less likely to be associated with a
523 detectable eQTL and that the effect size of eQTLs is negatively correlated with connectivity in the
524 protein-protein interaction network [91]. Furthermore, connectivity and core status contributed
525 primarily to PCs that were independent to expression level (Fig 6A), with connectivity being
526 correlated negatively with the rates of sequence evolution at both short (Fig 6B-C) and long (Fig 6D-
527 E) timescales. In comparison, expression level exhibited low correlation with sequence evolution at
528 short timescales, but higher negative correlations over long timescales (Fig 6B-E). Genes with high
529 co-expression connectivity, mostly those within network module cores, have significantly reduced
530 genetic diversity and exhibit reduced rates of protein evolution compared to more peripheral genes
531 (Fig 5). This suggests that core genes, which represent genes of high potential effect, have evolved
532 under stronger evolutionary constraint than genes in the periphery of the network. As genes in
533 module cores generally have higher levels of pleiotropy and lower levels of dispensability [58], they
534 are consequently more constrained against changes in both gene expression and protein sequence
535 compared to genes in the periphery of the network [92]. Furthermore, we show that eGenes are
536 overrepresented among non-core genes of the network and that they have experienced lower levels
537 of purifying selection relative to non-eGenes, corroborating a relaxation of both expression and
538 coding constraints for these genes. Relaxed selection of peripheral eGenes is thus expected to result
539 in an accumulation of weakly deleterious mutations that will segregate as intra-specific
540 polymorphisms. Consistent with these observations, we found that conserved genes (*i.e.* old genes)

541 were significantly over-represented in network cores while *P. tremula* specific genes (*i.e.* young genes)
542 were under-represented. This is in agreement with previous studies that have shown that
543 evolutionarily ancient genes tend to be more central in regulatory networks, with increased
544 constraints on expression and fewer associated eQTLs [87,86]. Taken together, these results indicate
545 how the co-expression network can be buffered against large perturbations via constraint of core
546 genes while enabling flexibility and adaptation by tolerating an accumulation of mutations within the
547 network periphery.

548 The Salicaceae lineage underwent a relatively recent (58 million years ago) whole-genome duplication
549 shared by all member species and that remains represented by a large number of paralogous gene
550 pairs in the genomes of *Populus* species [82]. If many of these duplicated genes are functionally
551 redundant or in the process of diverging, one would expect them to be overrepresented for eGenes
552 as sub- or neo-functionalisation requires derived SNPs to drive expression or coding divergence.
553 However, we saw an under-representation of eGenes in paralog pairs, suggesting that the SNPs that
554 initially induced expression divergence have reached fixation. Interestingly, we found progressively
555 lower expression correlation between paralog pairs containing zero, one or two eGenes (Fig 4),
556 indicating that a subset of paralogs are still undergoing sub- or neo-functionalisation, with their
557 associated eSNPs driving expression divergence.

558 Conclusion

559 We identified substantial heritability of gene expression within a natural, out-breeding population, of
560 *P. tremula*. Polymorphisms associated with expression variance were most frequently located within
561 regulatory regions, suggesting that they act by inducing expression variance. The gene co-expression
562 network displayed typical characteristics, being scale-free i.e. containing a small number of highly

563 connected genes. The most highly connected genes within module cores were underrepresented by
564 eQTLs, with a negative correlation between eQTL effect size and network connectivity. In contrast,
565 the network periphery was enriched for eQTLs, suggesting higher selective constraint on expression
566 variance within the network core (stabilising selection) and relaxed constraint within the periphery.
567 Integration of the eQTL and population genetics analyses with characteristics of the associated gene
568 co-expression network highlight that the context of a gene within the co-expression network appears
569 to be an important determinant of the evolutionary dynamics of transcribed loci. Our results point
570 towards stronger selection acting on network core genes compared to genes in the periphery of the
571 co-expression network, with a negative correlation between rates of sequence evolution and gene
572 connectivity. Taken together, this suggests that highly connected genes within the core of the co-
573 expression network derived from flushing buds of *P. tremula* have experienced stronger purifying
574 selection than those in the network periphery, the action of which is associated with higher
575 stabilising selection of gene expression variance for highly connected genes.

576 Materials and Methods

577 Samples

578 We collected branches form the SwAsp collection common garden in north of Sweden on 27th May
579 2012, before natural bud break but as close to the point of natural spring bud break as possible.
580 Branches were placed in the greenhouse facility at the Umeå Plant Science Centre under conditions
581 selected to induce rapid bud break (24 h light, temperature of 20 °C and humidity 50-70 %). At a
582 defined point of emergence (S22 Fig) buds were harvested, flash frozen in liquid nitrogen and stored
583 at -80 ° C until used for RNA isolation. Only terminal buds were sampled (i.e. no lateral buds were
584 included). The time from the day branches were placed in the greenhouse until bud flush sampling

ranged from one day to eight days (S23A Fig) and there was a high, positive correlation to bud flush date recorded in the field for the same year (S23B Fig ; r=0.776, $p < 2.2\times10^{-16}$). As has previously been reported [93], there was very low Q_{ST} for bud flush, either in the field or for the greenhouse material (Q_{ST} 0.13 and 0.07 respectively), however H^2 was high (H^2 = 0.82 and 0.71 respectively).

RNA isolation

One to two buds per clonal replicate were ground using one 3 mm stainless steel bead (Qiagen, Redwood city, USA) in Corning® 96 well PP 1.2 ml cluster tubes (Sigma-Aldrich, St. Louis, USA) using a Mixer Mill MM400 (Retsch, Haan, Germany) at 20 Hz for 2 x 15 sec. Total RNA was extracted from all samples according to [94] with the omission of the L spermidine. Buffer volumes were adjusted according to starting material (70 - 130 mg). RNA isolation was performed using one extraction with CTAB buffer followed by one chloroform : isoamyl alcohol IAA (24:1) extraction. All other steps were performed as in [94]. DNA contamination was removed using DNA-free™ DNA removal Kit (Life Technologies, Carlsbad, USA). RNA purity was measured using a NanoDrop 2000 (Thermo Scientific, Wilmington, USA) and RNA integrity was assessed using the Plant RNA Nano Kit for the Bioanalyzer (Agilent Technologies, Santa Clara, USA).

RNA-Sequencing and analysis

RNA-Sequencing was performed as in [74]. 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, [95]) with accession number ERP014886, and the normalised

606 and filtered gene expression data matrix is available at the PlantGenIE FTP resource
607 (ftp://plantgenie.org/Publications/Maehler2016/).

608 RNA-Seq FASTQ-files were pre-processed and aligned to v1.0 of the *P. tremula* reference genome
609 [81] as in [96]. In short, reads were quality and adapter trimmed using Trimmomatic v0.32 [97],
610 rRNA matching reads were filtered using SortMeRNA v1.9 [98], reads were aligned to the v1.0 *P.*
611 *tremula* reference genome using STAR 2.4.0f1 [99] and read counts were obtained using htseq-count
612 from HTSeq [100]. FastQC [101] was used to track read quality throughout the process. Normalised
613 gene expression values were obtained by applying a variance stabilising transformation (VST) to the
614 raw counts from HTSeq, as implemented in the DESeq2 R package [102].

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

616 We calculated repeatability as an assumed upper bound estimate of broad sense heritability of gene
617 expression (see [103] for discussion) from the variance estimates in our data according to the
618 equation

619 $$H^2 = \frac{V_G}{V_P}$$

620 where V_G is the genetic component of the variance calculated as the expression variance between
621 genotypes for a particular gene (*i.e.* variance among individual means) and V_P is the total phenotypic
622 variance calculated as the sum of V_G and V_E, where V_E is the environment component of the
623 variance calculated as the expression variance within genotypes for a particular gene (i.e. the mean
624 variance among clonal replicates). Point estimates of H^2 were obtained using the repeatability
625 function from the heritability v1.1 R package [104]. The significance of the broad sense heritability
626 was estimated by using an empirical null model where heritabilities were based on random genotype

assignments. For each gene, 1,000 permutations were performed and empirical p-values were calculated using the empPvals function in the qvalue 2.6.0 R package.

Population differentiation (Q_{ST}; [105]) was calculated as

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

where $V_{between}$ is the variance among populations and V_{within} is the residual genetic variance among genotypes within populations as computed using the lmer function from the lme4 v1.1.12 R package [106] using the formula

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

where *expression* is the expression of a gene, *population* is a factor representing the population of each sample, and *clone* is a factor representing genotype replicates. As we use repeatability as an upper bound estimate of H^2, our Q_{ST} estimates are conservative [107]. Similar to the broad sense heritability, the significance of the Q_{ST} was estimated using permutation testing where genotype labels were shuffled among the populations 1,000 times. Due to the characteristics of the data and the method, many of the permutations resulted in undefined QST estimates (an average of 511 missing values per gene). Thus, the empirical p-values for the genes were based on different number of null values.

Hidden confounder removal

Genotype mean gene expression data was adjusted for hidden confounders before mapping eQTLs and constructing the co-expression network. Hidden confounders in the gene expression data was accounted for by regressing out the 9 first principal components (PCs) of the gene expression data [75–77]. The number of components to remove was determined by running the eQTL mapping with

647 0 to 20 PCs removed and selecting the number of components that yielded the largest number of
648 significant eQTLs (Benjamini-Hochberg $p < 0.05$) (S5 Fig). This approach was based on the
649 assumption that the number of identified eQTLs would increase if the removed PCs were removing
650 unwanted, systematic variation (i.e. noise) rather than informative biological variation [75–
651 77,108,109].

652 ## eQTL mapping

653 eQTL mapping was performed by associating genotype mean gene expression levels with biallelic
654 SNPs using the R package Matrix eQTL v2.1.1 [110]. The corresponding raw resequencing data for
655 all SwAsp genotypes is available at the NCBI SRA resource as BioProject PRJNA297202 (SRA:
656 SRP065057) and the resultant Variant Call Format (VCF) file used for association mapping is
657 available from [65]. Before performing association mapping, genes were filtered on variance so that
658 only genes with a gene expression variance above 0.05 were included. SNPs were also filtered on
659 minor allele frequency (MAF) and major genotype frequency (MGF); any SNPs with MAF < 0.1 or
660 MGF > 0.9 were excluded to avoid spurious associations. A motivating example for the MGF
661 filtering can be seen in S24 Fig. We also generated an LD-trimmed SNP set by removing one SNP
662 from each pair of SNPs with a between SNP correlation coefficient (r^2) >0.2 in blocks of 50 SNPs
663 using PLINK v1.9 [111], yielding 217,489 independent SNPs that were retained for analyses of
664 population structure. The first genotype principal component based on the set of independent SNPs
665 was used as a covariate in the linear model used by Matrix eQTL to account for the weak signature
666 of population structure. Permutation testing was used to determine eQTL significance whereby
667 genotype sample labels were permuted 1,000 times and the maximum absolute t-statistic from Matrix
668 eQTL was recorded for each expressed gene across all SNPs for each permutation, resulting in 1,000
669 random t-statistics being collected for each gene. Empirical p-values were calculated for each eQTL

670 in the observed data using the permuted t-statistics for the observed eGene with the empPvals
671 function in the qvalue v2.4.2 R package [112], and q-values (empirical FDR) were calculated with the
672 qvalue function in the same package.

673 When determining the genomic context of eSNPs, there were some cases where introns overlapped
674 exons as a result of overlapping gene model being present on the same strand. These 41 eSNPs were
675 discarded from the counting. Another type of overlap that was discarded were cases where an eSNP
676 overlapped a gene feature, but no sub-feature inside that gene (e.g. UTR, exon or intron). These
677 1961 eSNPs were excluded from the counting. Since many of the features overlap (e.g. exon and
678 untranslated regions), the priority for counting was untranslated region, exon/intron,
679 upstream/downstream and intergenic.

680 Permutation tests involving eGenes and non-eGenes were performed by shuffling eGene
681 assignments among the 22,306 genes that were considered for eQTL mapping. This was repeated
682 10,000 times and empirical p-values were calculated using the empPvals function in the qvalue v2.6.0
683 R package. The minor allele frequency (MAF) was compared between eSNPs (both local and distant)
684 and all SNPs throughout the genome. To test for the signature of selection on eQTLs, we also
685 estimated the correlation between minor allele frequency and eQTL effect size.

686 Permutation analysis of the minor allele frequency distribution was performed by generating 20 sets
687 of 150,000 random SNPs from the set of MAF and MGF-filtered SNPs (sampled with replacement).
688 eQTL mapping was performed for each of these sets, and in addition, the sample labels of the
689 genotype data was shuffled 50 times for each SNP-set and consequently used for eQTL mapping,
690 resulting in a total of 1,000 permuted eQTL sets.

691 For the subsampling analysis, the significant eSNPs from the original eQTL mapping were
692 subsampled in order to fix the minor allele frequency. It was decided to fix the minor allele frequency
693 at 25% and use 40 samples as this retained 136,942 of the original 147,419 eSNPs (93%). The actual
694 subsampling was done by dropping samples recursively until the minor allele frequency fell within ±2
695 percentage points of the desired frequency. This was done using our custom program vcfsubsample
696 (https://github.com/maehler/vcfsubsample).

697 Co-expression

698 The R package WGCNA v1.51 [113] was used for constructing a co-expression network. The input
699 gene expression values were per-gene genotype means corrected for hidden confounders (the same
700 data used for eQTL mapping). We chose to use the unsigned network type for this study with the
701 motivation that we did not want to discard negative relationships. By looking at this from an eQTL
702 perspective, an eSNP can be positively associated with one gene while negatively associated with
703 another. The relationship between these two genes would be missed if we used the signed or the
704 signed-hybrid approaches. Using the unsigned approach, we assure that genes with strong negative
705 correlation end up in the same network modules. A soft thresholding power of 5 was used to
706 calculate adjacencies. The topological overlap matrix (TOM) was generated using the TOMsimilarity
707 function with the signed approach in order to take negative edges into account (see [113] for details).
708 In order to identify network modules, hierarchical clustering was applied to the TOM dissimilarity
709 matrix (1 – TOM) and the resulting dendrogram was divided into modules using the cutreedynamic
710 function. The connectivity of the network was then defined as the adjacency sum for each node, i.e.
711 the weights of the edges that are connected to this node. This concept was applied to modules as
712 well to obtain measures of intra- and inter-modular connectivity, i.e. the connectivity based on edges

713 connecting the gene with other genes inside the same module, and connectivity based on edges
714 connecting the gene with genes outside of the module.

715 To define the periphery of the network we applied a hard edge-threshold to the network where only
716 gene-pairs with an absolute Pearson correlation > 0.4 were linked, which corresponded to the top
717 0.1% most correlated gene pairs. Genes were then classed as peripheral if they linked to only one
718 other gene.

719 Conserved genes and *P. tremula*-specific genes were inferred from gene family analysis based on
720 protein sequence similarity available on the PopGenIE FTP server [114]. In short, two rounds of
721 TribeMCL were run using BLASTP with an E-value threshold of 10^{-5} where the first round was all-
722 vs-all while the second was all-vs-all inside each gene family.

723 Transcription factors were identified by BLASTing the the *P. tremula* protein sequences against the
724 protein sequences of transcription factors in *P. trichocarpa* with an E-value threshold of 10^{-10} and
725 taking the best hit for each *P. trichocarapa* transcription factor. The *P. trichocarpa* transcription factor
726 annotations were taken from Plant Transcription Factor Database v3.0 [115].

727 All permutation tests based on the network were performed by computing the measure of interest on
728 the original network, and then performing the same computation on 10,000 networks were the gene
729 labels were shuffled. Empirical p-values were calculated using the empPvals function in the qvalue
730 v2.6.0 R package [112].

731 Functional enrichment was tested using the R package topGO 2.24.0. As background the 22,306
732 genes included in the network construction were used.

Population genetic analysis

To estimate nucleotide diversity (θ_π) and Tajima's D [116] among genes, we only used the reads with mapping quality above 30 and the bases with quality score higher than 20 for the estimation of θ_π and Tajima's D by ANGSD [117]. Genes with less than 50 covered sites left from previous quality filtering steps (Wang et al. In prep) were excluded. In addition, θ_π and Tajima's D were also calculated for seven site categories of each gene: 1kbp upstream of genes, 1kbp downstream of genes, 3' UTR, 5' UTR, intron, 0-fold non-synonymous and 4-fold synonymous sites within genes. We further compared the estimates of $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ between different categories of genes. We used the transcript with the highest content of protein-coding sites to categorize the seven genomic features within each gene. Finally, we estimated the ratio of non-synonymous substitution and synonymous substitution rate (d_N/d_S) using gKaKs pipeline [118] with codeml method [119] for a total of 33,039 orthologous genes that were determined by best BLASTp sequence homology searches between *P. tremula* and *P. trichocarpa* [81,114]. Significance for the above statistical measurements between each pair of gene categories was evaluated using Mann-Whitney tests.

To test for the main effects of various gene expression factors (including connectivity within co-expression network, core vs. noncore, presence or absence of an eQTL, gene expression level and variance) on the rate of sequence evolution, we first performed principal component analysis (PCA) of the five gene expression-related variables to examine the extent to which they were interdependent. To further handle the problem of collinearity between them, the PCs were calculated. We then examined the Spearman's rank correlation coefficient of each PC with the parameters of sequence evolution. The data were scaled before doing PCA, and the computation of PCs was implemented in the "prcomp" function of the statistical software package R 3.2.0.

Finally, to account for the autocorrelation between gene expression measurements, we estimated partial Spearman correlations between sequence evolution parameters and each of gene expression related variables while controlling for other variables. The partial correlation were performed with R package ppcor.

Analysis script availability

Relevant processed data and R transcripts are available on the PlantGenIE FTP resource (ftp://plantgenie.org/Publications/Maehler2016/).

Acknowledgements

We thank Dr Kathryn Robinson for her help sampling material, for providing environmental data and for informative discussion. The authors also would like to acknowledge support from the Science for Life Laboratory, the National Genomics Infrastructure (NGI), and the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for providing assistance in massive parallel sequencing and computational infrastructure. The authors wish to thank the three reviewers, who offered invaluable feedback and suggestions during the review process. Their input substantially improved the presentation and interpretation of the study and we are grateful for the time they dedicated to this.

References

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

2. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, et al. Regional and

775 strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A.
776 2000;97: 11038–43. doi:10.1073/pnas.97.20.11038

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

779 4. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions
780 of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet.
781 2001;29: 389–95. doi:10.1038/ng766

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

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

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

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

794 9. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping
795 determinants of human gene expression by regional and genome-wide association. Nature.

796 2005;437: 1365–9. doi:10.1038/nature04244

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

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

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

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

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

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

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

816 17. Wang RL, Stec A, Hey J, Lukens L, Doebley J. The limits of selection during maize

817 domestication. Nature. 1999;398: 236–239. doi:10.1038/18435

818 18. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell.
819 2000;101: 577–580. doi:10.1016/S0092-8674(00)80868-5

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

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

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

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

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

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

835 25. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet.
836 2001;17: 388–91. doi:10.1016/S0168-9525(01)02310-1

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

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

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

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

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

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

851 32. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:
852 206–16. doi:10.1038/nrg2063

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

856 34. Keurentjes JJB, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, et al.
857 Regulatory network construction in *Arabidopsis* by using genome-wide gene expression

858 quantitative trait loci. Proc Natl Acad Sci U S A. 2007;104: 1708–13.
859 doi:10.1073/pnas.0610429104

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

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

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

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

872 39. Swanson-Wagner R a, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, et al. Paternal dominance
873 of trans-eQTL influences gene expression patterns in maize hybrids. Science. 2009;326: 1118–
874 1120. doi:10.1126/science.1178294

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

877 41. Fu J, Cheng Y, Linghu J, Yang X, Kang L, Zhang Z, et al. RNA sequencing reveals the
878 complex regulatory network in the maize kernel. Nat Commun. 2013;4: 2832.

879 doi:10.1038/ncomms3832

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

883 43. Wang J, Yu H, Weng X, Xie W, Xu C, Li X, et al. An expression quantitative trait loci-guided
884 co-expression analysis for constructing regulatory network using a rice recombinant inbred
885 line population. J Exp Bot. 2014;65: 1069–1079. doi:10.1093/jxb/ert464

886 44. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR. Genetic Architecture of Transcript-
887 Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. Genetics. 2005;169: 2295–
888 2303. Available: http://www.genetics.org/cgi/content/abstract/169/4/2295

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

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

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

897 48. Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, et al. Segregating
898 Variation in the Transcriptome: Cis Regulation and Additivity of Effects. Genetics. 2006;173:
899 1347–1355. doi:10.1534/genetics.105.051474

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

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

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

908 52. Leder EH, McCairns RJS, Leinonen T, Cano JM, Viitaniemi HM, Nikinmaa M, et al. The
909 Evolution and Adaptive Potential of Transcriptional Variation in Sticklebacks--Signatures of
910 Selection and Widespread Heritability. Mol Biol Evol. 2015;32: 674–689.
911 doi:10.1093/molbev/msu328

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

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

919 55. Holloway AK, Lawniczak MKN, Mezey JG, Begun DJ, Jones CD. Adaptive gene expression
920 divergence inferred from population genomics. PLoS Genet. 2007;3: 2007–2013.

921 doi:10.1371/journal.pgen.0030187

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

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

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

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

932 60. Whitacre JM. Biological robustness: Paradigms, mechanisms, systems principles. Front Genet.
933 2012;3: 1–15. doi:10.3389/fgene.2012.00067

934 61. Jansson S, Douglas CJ. *Populus*: A Model System for Plant Biology. Annu Rev Plant Biol.
935 2007;58: 435–458. doi:10.1146/annurev.arplant.58.032806.103956

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

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

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

945 65.
946 ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/VCF/SwAsp_94samples.filte
947 r.maf005.beagle.maf.recode.vcf.gz.

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

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

954 68. Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK. Adaptive population
955 differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremula*,
956 L.): a comparison of neutral markers, candidate genes and phenotypic traits. Evolution.
957 2007;61: 2849–60. doi:10.1111/j.1558-5646.2007.00230.x

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

960 70. Yang S, Liu Y, Jiang N, Chen J, Leach L, Luo Z, et al. Genome-wide eQTLs and heritability
961 for gene expression traits in unrelated individuals. BMC Genomics. 2014;15: 13.
962 doi:10.1186/1471-2164-15-13

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

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

968 73. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA
969 methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife.
970 2015;4: e05255. doi:10.7554/eLife.05255

971 74. Robinson KM, Delhomme N, Mähler N, Schiffthaler B, Onskog J, Albrectsen BR, et al.
972 *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. BMC Plant Biol.
973 2014;14: 276. doi:10.1186/s12870-014-0276-5

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

977 76. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding
978 mechanisms underlying human gene expression variation with RNA sequencing. Nature.
979 2010;464: 768–72. doi:10.1038/nature08872

980 77. Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, et al. Normalizing
981 RNA-sequencing data by modeling hidden covariates with prior knowledge. PLoS One.
982 2013;8: e68141. doi:10.1371/journal.pone.0068141

983 78. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing

984 the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.
985 Genome Res. 2014;24: 14–24. doi:10.1101/gr.155192.113

986 79. Massouras A, Waszak SM, Albarca-Aguilera M, Hens K, Holcombe W, Ayroles JF, et al.
987 Genomic variation and its impact on gene expression in Drosophila melanogaster. PLoS
988 Genet. 2012;8: e1003055. doi:10.1371/journal.pgen.1003055

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

991 81. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S. The *Populus* Genome Integrative
992 Explorer (PopGenIE): a new resource for exploring the *Populus* genome. New Phytol.
993 2009;182: 1013–1025. doi:10.1111/j.1469-8137.2009.02807.x

994 82. Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, et al. The willow genome and divergent
995 evolution from poplar after the common genome duplication. Cell Res. Shanghai Institutes
996 for Biological Sciences, Chinese Academy of Sciences; 2014;24: 1274–1277.
997 doi:10.1038/cr.2014.83

998 83. Tung J, Zhou X, Alberts SC, Stephens M, Gilad Y. The genetic architecture of gene
999 expression levels in wild baboons. Elife. 2015;4: 1–22. doi:10.7554/eLife.04729

1000 84. Wagner A. Genotype networks shed light on evolutionary constraints. Trends Ecol Evol.
1001 Elsevier Ltd; 2011;26: 577–584. doi:10.1016/j.tree.2011.07.001

1002 85. Alvarez-Ponce D, Aguadé M, Rozas J. Network-level molecular evolutionary analysis of the
1003 insulin / TOR signal transduction pathway across 12 Drosophila genomes. Genome Res.
1004 2009;19: 234–42. doi:10.1101/gr.084038.108.234

1005 86. Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution of
1006 gene interaction networks in the human and mouse genomes. Genome Biol. 2015;16: 202.
1007 doi:10.1186/s13059-015-0772-4

1008 87. Popadin KY, Gutierrez-Arcelus M, Lappalainen T, Buil A, Steinberg J, Nikolaev SI, et al.
1009 Gene age predicts the strength of purifying selection acting on gene expression variation in
1010 humans. Am J Hum Genet. 2014;95: 660–74. doi:10.1016/j.ajhg.2014.11.003

1011 88. Zhang J, Yang J-R. Determinants of the rate of protein sequence evolution. Nat Rev Genet.
1012 Nature Publishing Group; 2015;16: 409–420. doi:10.1038/nrg3950

1013 89. Albert R, Barabasi AL. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74:
1014 47–97. doi:10.1088/1478-3967/1/3/006

1015 90. Choi J, Shooshtari P, Samocha KE, Daly MJ, Cotsapas C. Network Analysis of Genome-Wide
1016 Selective Constraint Reveals a Gene Network Active in Early Fetal Brain Intolerant of
1017 Mutation. PLoS Genet. 2016;12: 1–16. doi:10.1371/journal.pgen.1006121

1018 91. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing
1019 the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.
1020 Genome Res. 2013; doi:10.1101/gr.155192.113

1021 92. Papakostas S, Vøllestad LA, Bruneaux M, Aykanat T, Vanoverbeke J, Ning M, et al. Gene
1022 pleiotropy constrains gene expression changes in fish adapted to different thermal conditions.
1023 Nat Commun. 2014;5: 4071. doi:10.1038/ncomms5071

1024 93. Robinson KM, Ingvarsson PK, Jansson S, Albrectsen BR. Genetic variation in functional
1025 traits influences arthropod community composition in aspen (*Populus tremula* L.). PLoS One.

1026 2012;7: e37679. doi:10.1371/journal.pone.0037679

1027 94. Chang S, Puryear J, Cairney J. A simple and efficient method for isolating RNA from pine
1028 trees. Plant Mol Biol Report. 1993;11: 113–116. doi:10.1007/BF02670468

1029 95. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European
1030 Nucleotide Archive. Nucleic Acids Res. 2011;39: D28-31. doi:10.1093/nar/gkq967

1031 96. Delhomme N, Mähler N, Schiffthaler B, Sundell D, Mannepperuma C, Hvidsten TR, et al.
1032 Guidelines for RNA-Seq data analysis. Epigenesys. 2014; Available:
1033 http://www.epigenesys.eu/en/protocols/bio-informatics/1283-guidelines-for-rna-seq-data-
1034 analysis

1035 97. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.
1036 Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

1037 98. Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs
1038 in metatranscriptomic data. Bioinformatics. 2012;28: 3211–3217.
1039 doi:10.1093/bioinformatics/bts611

1040 99. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
1041 universal RNA-seq aligner. Bioinformatics. 2012;29: 15–21.
1042 doi:10.1093/bioinformatics/bts635

1043 100. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput
1044 sequencing data. Bioinformatics. 2014;31: 166–169. doi:10.1093/bioinformatics/btu638

1045 101. Andrews S. FastQC: A quality control tool for high throughput sequence data [Internet]. 2016
1046 [cited 27 Sep 2016]. Available: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

1047 102. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-
1048 seq data with DESeq2. Genome Biol. 2014;15: 550. doi:10.1186/s13059-014-0550-8

1049 103. Dohm MR. Repeatability estimates do not always set an upper limit to heritability. Funct Ecol.
1050 2002;16: 273–280. doi:10.1046/j.1365-2435.2002.00621.x

1051 104. Kruijer W, Boer MP, Malosetti M, Flood PJ, Engel B, Kooke R, et al. Marker-based
1052 estimation of heritability in immortal populations. Genetics. 2015;199: 379–398.
1053 doi:10.1534/genetics.114.167916

1054 105. Spitze K. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation.
1055 Genetics. 1993;135: 367–74. Available: http://www.genetics.org/content/135/2/367.short

1056 106. Bates D, Mächler M, Bolker B, Walker S. Fitting Linear Mixed-Effects Models Using lme4. J
1057 Stat Softw. 2015;67: 1–48. doi:10.18637/jss.v067.i01

1058 107. Manier MK, Palumbi SR. Intraspecific divergence in sperm morphology of the green sea
1059 urchin, *Strongylocentrotus droebachiensis*: implications for selection in broadcast spawners. BMC
1060 Evol Biol. 2008;8: 283. doi:10.1186/1471-2148-8-283

1061 108. Listgarten J, Kadie C, Schadt EE, Heckerman D. Correction for hidden confounders in the
1062 genetic analysis of gene expression. Proc Natl Acad Sci U S A. 2010;107: 16465–70.
1063 doi:10.1073/pnas.1002425107

1064 109. Parts L, Stegle O, Winn J, Durbin R. Joint genetic analysis of gene expression data with
1065 inferred cellular phenotypes. PLoS Genet. 2011;7: 1–10. doi:10.1371/journal.pgen.1001276

1066 110. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations.
1067 Bioinformatics. 2012;28: 1353–1358. doi:10.1093/bioinformatics/bts163

1068 111. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool
1069 set for whole-genome association and population-based linkage analyses. Am J Hum Genet.
1070 2007;81: 559–575. doi:10.1086/519795

1071 112. Storey J. qvalue: Q-value estimation for false discovery rate control [Internet]. 2015. Available:
1072 http://github.com/jdstorey/qvalue

1073 113. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis.
1074 BMC Bioinformatics. 2008;9: 559. doi:10.1186/1471-2105-9-559

1075 114. ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/Annotation/DoubleMCL-tab-
1076 annotation.txt.gz.

1077 115. Jin J, Zhang H, Kong L, Gao G, Luo J. PlantTFDB 3.0: A portal for the functional and
1078 evolutionary study of plant transcription factors. Nucleic Acids Res. 2014;42: 1182–1187.
1079 doi:10.1093/nar/gkt1016

1080 116. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA
1081 polymorphism. Genetics. 1989;123: 585–595. doi:PMC1203831

1082 117. Korneliussen T, Albrechtsen A, Nielsen R. ANGSD: Analysis of Next Generation Sequencing
1083 Data. BMC Bioinformatics. 2014;15: 356. doi:10.1186/s12859-014-0356-4

1084 118. Zhang C, Wang J, Long M, Fan C. GKaKs: The pipeline for genome-level Ka/Ks calculation.
1085 Bioinformatics. 2013;29: 645–646. doi:10.1093/bioinformatics/btt009

1086 119. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:
1087 1586–1591. doi:10.1093/molbev/msm088

1088

1089 # Figures

1090 **Fig 1. Population location and gene expression overview.** (A) Map of the original locations of
1091 the SwAsp populations. The red arrow points to the location of the common garden used in this
1092 study. (B) Distribution of gene expression heritability for all genes and for the subset of genes after
1093 filtering to remove uninformative expression. (C) Distribution of gene expression Q_{ST} for all genes
1094 and for the subset of genes after filtering to remove uninformative expression. (D) Sample clustering
1095 based on all samples, including biological replicates. The heatmap represents the sample correlation
1096 matrix based on the 500 genes with the highest expression variance using gene expression data prior
1097 to hidden confounder removal. Darker colour indicates higher correlation. The coloured bar
1098 represents the populations the samples belong to. The small clusters on the diagonal correspond to
1099 biological replicates of each genotype.

1100 **Fig 2. eQTL overview.** (A) Expression variance explained *(R^2)* for local and distant eQTLs. Box
1101 plots show the maximum variance explained by a single eQTL for each gene and the total variance
1102 explained by all eQTLs for each gene. The widths of the boxes are proportional to the number of
1103 genes represented. The pairwise significiance of a Mann-Whitney test is indicated by asterisks: ***$p <$
1104 0.001, **$p < 0.01$. (B) Broad sense heritability distributions for eGenes and non-eGenes. (C) Scatter
1105 plot showing the positions of all significant eQTLs in this study. No evidence of eQTL hotspots can
1106 be observed. Numbers indicate chromosome. (D) Genomic context of local and distant eSNPs.
1107 Context categories are normalized for feature length. When an eSNP overlapped with features on
1108 both strands, both of them were counted. For both local and distant eQTLs the features are based
1109 on the gene that is closest to the eSNP, and furthermore, for local eQTLs, the features are divided
1110 into whether the eSNP is located in or near the same gene that it is associated to ("associated gene")
1111 or not ("other gene"). Flanking regions represent 2 kbp upstream and downstream from the gene.

1112 (E) Manhattan plots for local eQTLs (upper) and distant eQTLs (lower). Each point represents an
1113 eQTL.

1114 **Fig 3. Co-expression characteristics of eGenes and eSNPs.** (A) Gene expression variance
1115 distribution for all genes in the SwAsp data (before removal of hidden confounders) and the exAtlas
1116 data. (B) Distribution of co-expression connectivity for eGenes and non-eGenes. (C) Distributions
1117 of the proportion of total variance explained and heritability for eGenes and non-eGenes divided
1118 into core and non-core genes. (D) Genes having distantly, locally, or both distantly and locally acting
1119 eSNPs located within 2 kbp (or inside) the gene divided into core and non-core genes. (E) Genomic
1120 context of distantly, locally, or both distantly and locally acting eSNPs located within 2 kbp of an
1121 eGene divided into core and non-core eGenes. The eSNP counts are normalised for total feature
1122 length.

1123 **Fig 4. Correlation within paralog pairs as a function of the number of eGenes in the paralog**
1124 **pair.** The widths of the boxes are proportional to the number of genes in each set. The mean
1125 correlations for paralog pairs with 0, 1, or 2 eGenes were 0.17, 0.10, and 0.06, respectively. The
1126 pairwise significance of a Mann-Whitney test is indicated by asterisks: ***$p < 0.001$, **$p < 0.01$.

1127 **Fig 5. Measures of sequence diversity and divergence.** Nucleotide diversity (A,B), Tajima's D
1128 (C, D), $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ (E,F) and d_N/d_S (G,H) are compared between eGenes (with local eQTLs or with
1129 only distant eQTLs) and non-eGenes, as well as core and non-core genes from the gene expression
1130 network. Significance between each pair of gene categories was evaluated using Mann-Whitney tests
1131 and significance is indicated by asterisks above the boxplot: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, n.s.
1132 $p > 0.05$.

Fig 6. Associations between metrics of gene expression and sequence evolution in *Populus tremula*. (A) Percentage variance explained by five principal components comprising co-expression network connectivity, genes in network cores or not (Core vs. non-core genes), gene expression levels, gene expression variance and genes with eQTLs or not (eGenes vs. non-eGenes). Colour shadings depict the proportion contribution of each gene expression measure to each principal component. (B-E) Spearman's rank correlations between PCs and four metrics of sequence evolution: nucleotide diversity (θ_π), Tajima's D, $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and d_N/d_S. Stacking of the barplots shows the relative contribution of each gene expression measure to each PC. Plus or minus indicates the direction of correlation for individual variable on the corresponding PCs. Asterisks indicate significant correlations, ***P < 1e-5, **P < 0.001,*P<0.05, ns=not significant (P >0.05).

Supporting information

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

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

S1 Table. Statistic summary (median and central 95% range) for four measures of sequence evolution: θ_π, Tajima's D, $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and dN/dS in groups corresponding to Fig 5, and the summary of Wilcoxon rank sum test (location shift and 95% confidence interval) between two groups of estimates shown in Fig 5.

S1 Fig. Clustering of genotypes. (Left) Heatmap of the sample correlation matrix based on the 500 most variably expressed genes. Darker colour indicates higher correlation. The coloured bar

1154 represents the populations the genotypes belong to. (Right) The two first principal components from
1155 a principal component analysis (PCA) based on all genes. Again, colours represent the genotype
1156 population. The percentages in the axis labels indicate the amount of variance explained by each
1157 component.

1158 **S2 Fig. Correlations between gene expression principal components (PCs) and**
1159 **environmental variables.** The values in each tile represent the Pearson correlation between the
1160 gene expression PC (x-axis) and environmental variable (y-axis). Small asterisks represent a nominal
1161 p-value < 0.05 while large asterisks represent Benjamini-Hochberg (BH) adjusted p-values < 0.05.
1162 The only factor with significant correlations to expression PCs was "Hours to harvest", which is the
1163 number of hours into the sampling period that the buds were harvested. It was significantly
1164 associated with PC4 (BH-adjusted $p = 4.6\times10^{-6}$), PC7 (BH-adjusted $p = 0.030$) and PC2 (BH-
1165 adjusted $p = 0.033$).

1166 **S3 Fig. Relationship between gene expression ranks and the number of genotypes with**
1167 **detectable expression of the genes.** The number of genotypes that a gene was expressed in was
1168 determined by counting the number of genotypes with non-zero expression for each gene. The gene
1169 expression ranks were calculated by ranking the mean gene expression values where the mean was
1170 calculated only considering samples with non-zero expression.

1171 **S4 Fig. Physical distance between eSNPs and their associated genes for each of the 19**
1172 **chromosomes.**

1173 **S5 Fig. The number of eQTLs detected for different levels of hidden confounder removal.**
1174 The y-axis shows the number of eQTLs detected at 5% FDR (prior to the empirical FDR
1175 calculation). The x-axis shows the number of principal components regressed out of the gene

1176 expression data. Distant eQTLs are defined as eQTLs where the gene and the SNP are located on
1177 different scaffolds.

1178 **S6 Fig. Heatmap representing the sample clustering based on gene expression values after**
1179 **hidden confounder removal.** The 500 most variable genes in the original expression data were used
1180 for calculating the sample correlations (i.e. the same genes as in Fig 1D). The colour bar represents
1181 the population of the samples with the same colour scheme as in Fig 1D. The small clusters on the
1182 diagonal represent genotype replicates.

1183 **S7 Fig. Sliding window approach to detect eQTL hotspots.** The upper panel shows the number
1184 of eSNPs in genomic windows of 100 kb for each of the 19 chromosomes. The lower panel shows
1185 the number of unique genes that are associated to each genomic window (nominal eQTL p-value <
1186 1×10^{-6}).

1187 **S8 Fig. eQTL scatter plot using gene expression data before hidden confounder removal.** A
1188 clear hotspot can be seen at the end of chromosome 4.

1189 **S9 Fig. eSNPs per eGene and eGenes per eSNP.** (A) Bar plot of the number of associated
1190 eGenes per eSNP. (B) Histogram of the number of associated eSNPs per eGene. (C) The mean co-
1191 expression for genes associated with the same eSNP. The numbers on the x-axis represents the
1192 number of eGenes that a single eSNP is associated to, and the numbers above the boxes shows the
1193 total number of eSNPs that are represented, and in parentheses are the number of unique genes that
1194 they are associated with.

1195 **S10 Fig. Local eSNP distribution.** Local eSNPs were counted in cases where the eSNP was
1196 located in the gene it was associated with (A) as well as when it was located in a direct neighbour of
1197 the eGene (B, C) or in a gene further away from the eGene (D, E). For cases when the eSNP was

1198 located in neighbouring genes (B-E), this was plotted both in the direction of the gene (B, D) and
1199 relative to the eGene (C, E). Up and downstream regions were divided into 1 kbp bins while the
1200 intergenic region was normalised to 20 bins. In other words, up and downstream counts are not
1201 directly comparable to intragenic counts.

1202 **S11 Fig. Gene expression correlation distributions for neighbouring genes.** Neighbouring
1203 genes in the genome were divided into five different categories: non-eGene – non-eGene, eGene –
1204 gene, eGene with eSNP proximal to it – gene, eGene – first neighbour harbouring the eSNP, and
1205 eGene – neighbour further away harbouring the eSNP. The numbers in parenthesis in the legend
1206 represent the number of pairs in each category.

1207 **S12 Fig. The number of genes and the number of core genes in each network module.** Each
1208 bar is divided into eGene and non-eGene assignments. Module 0 contains genes not assigned to any
1209 module.

1210 **S13 Fig. Gene expression variance in module cores and non-cores for each of the 38**
1211 **modules.**

1212 **S14 Fig. Gene expression variance plotted against co-expression network connectivity.**
1213 eGenes are indicated by red points.

1214 **S15 Fig. Relationship between total variance explained by eQTL and co-expression**
1215 **connectivity.** The variance explained is the adjusted variance explained based on all eQTL
1216 associated with each gene.

1217 **S16 Fig. Comparisons of nucleotide diversity and Tajima's D for different genomic features.**
1218 Upstream and downstream are 1 kbp away from the gene start and end, respectively. 0-fold and 4-
1219 fold refers to 0-fold and 4-fold degenerate sites, respectively.

1220 **S17 Fig. Minor allele frequency of eSNPs.** (A) Comparison of minor allele frequency between all
1221 SNPs, distant eSNPs and local eSNPs. (B) The relationship between minor allele frequency and
1222 effect size (absolute value of beta) of eQTLs.

1223 **S18 Fig. Relationship between MAF and effect size in permuted eQTLs.** eQTL permutations
1224 were done by selecting 20 random subsets of 150,000 SNPs each. Each of these sets were associated
1225 with gene expression levels, and in addition, 50 permutations were performed for each set by
1226 shuffling the genotype labels of the SNP data, resulting in a total of 1,000 permuted eQTL sets. (A)
1227 The number of SNPs for the nominal p-value threshold 10^-6 (the same initial threshold that was
1228 used for the original associations). (B) The Spearman correlation between minor allele frequency and
1229 absolute effect size for the 20 sets and their corresponding permutations. (C) Allele frequency
1230 spectra for the 20 subsets and their permutations.

1231 **S19 Fig. Comparing the effect sizes between the original and the subsampled data.** The
1232 Pearson correlation is 0.98 (df = 157,020).

1233 **S20 Fig. Association between metrics of gene expression and sequence evolution.** Scatter
1234 plots (lower off-diagnal, the red-to-yellow-to-blue gradient indicates decreased density of observed
1235 events at a give location in the graph) and correlations with probability values (upper off-diagonal,
1236 spearman's rank correlations) for measures of gene expression and sequence evolution: kTotal, the
1237 total (global) connectivity in the network; is_core_gene, gene is part of module core or not;
1238 orig_expr_mean, mean expression before hidden confounder removal; orig_expr_var, expression
1239 variance before hidden confounder removal; is_egene, gene with eQTLs or not; tP, pairwise
1240 nucleotide diversity; tajD, Tajima's D; tP_0_4, the ratio of pairwise nucleotide diversity at zero-fold
1241 non-synonymous and four-fold synonymous sites; kaks, dN/dS.

1242 **S21 Fig. Variance explained by eSNPs and heritability.** This figure assumes that eSNPs only
1243 account for heritable variation, explaining why some of the eSNPs explain more than 100% of the
1244 heritability. The x-axis has been truncated at 300%.

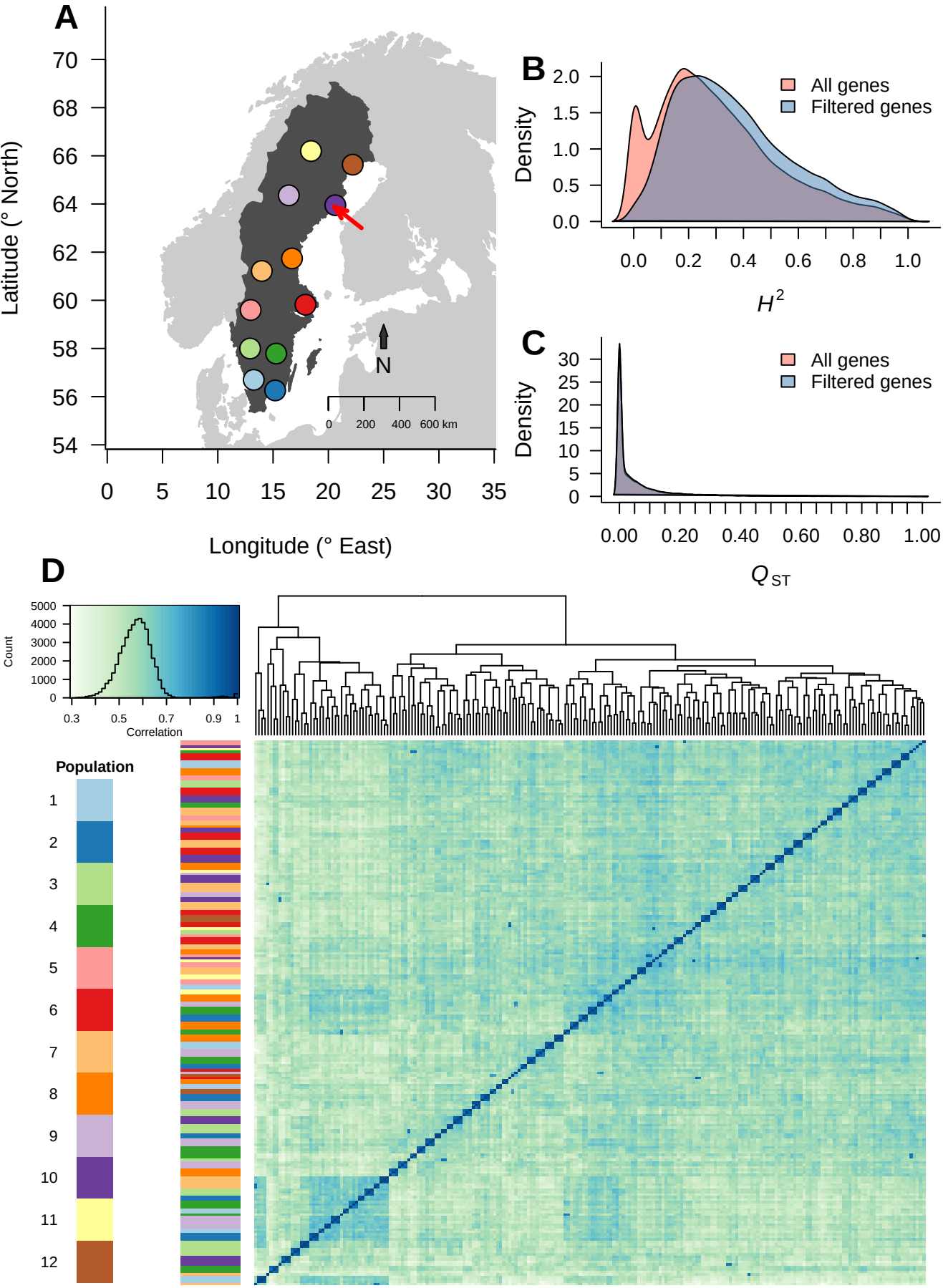
1245 **S22 Fig. Representative photo of the sampled bud flush stage.**

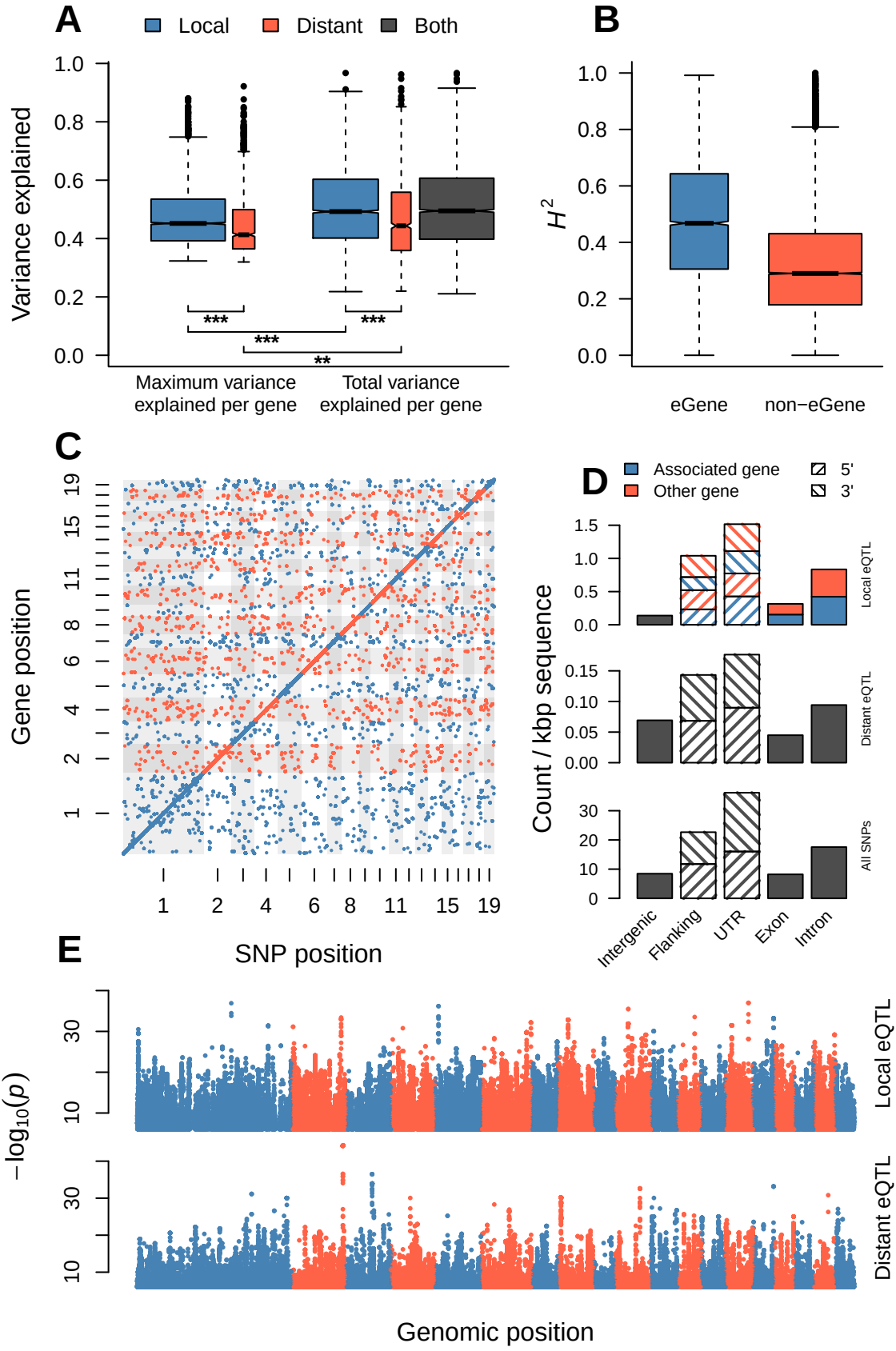
1246 **S23 Fig. Sampling time and day of bud flush.** (A) Box plot distributions of the Julian day of
1247 sampling for the SwAsp sub-populations. (B) The relationship between Julian day of bud flush for
1248 the greenhouse sampled buds and Julian day of bud flush in the field for the same year (2012).

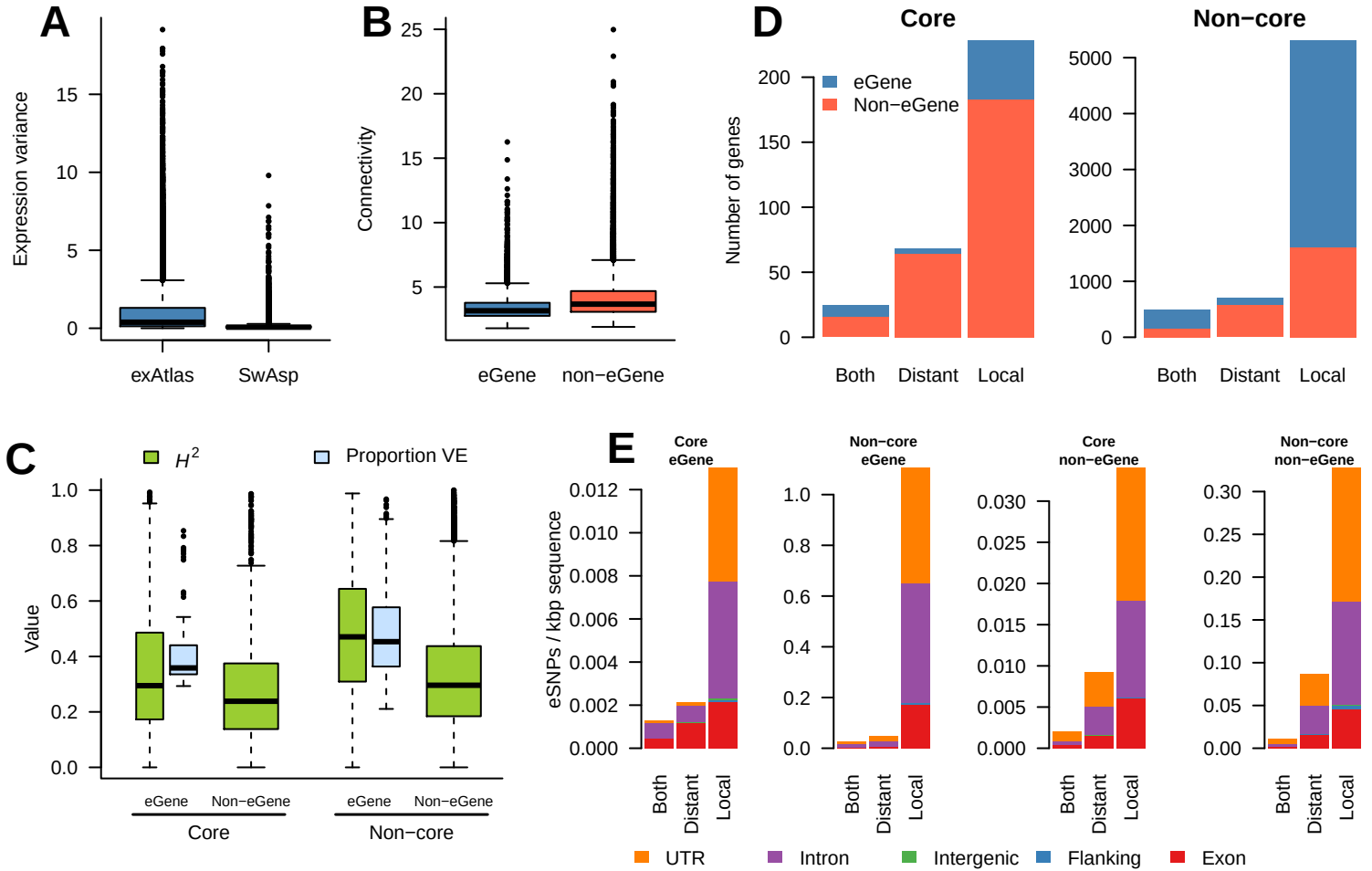
1249 **S24 Fig. Motivating example for major genotype filtering.** In this case the SNP is heterozygous
1250 in all samples but one, resulting in a minor allele frequency close to 0.5, but a major genotype
1251 frequency close to one. The resulting association turns out very significant, but is only supported by
1252 a single sample.

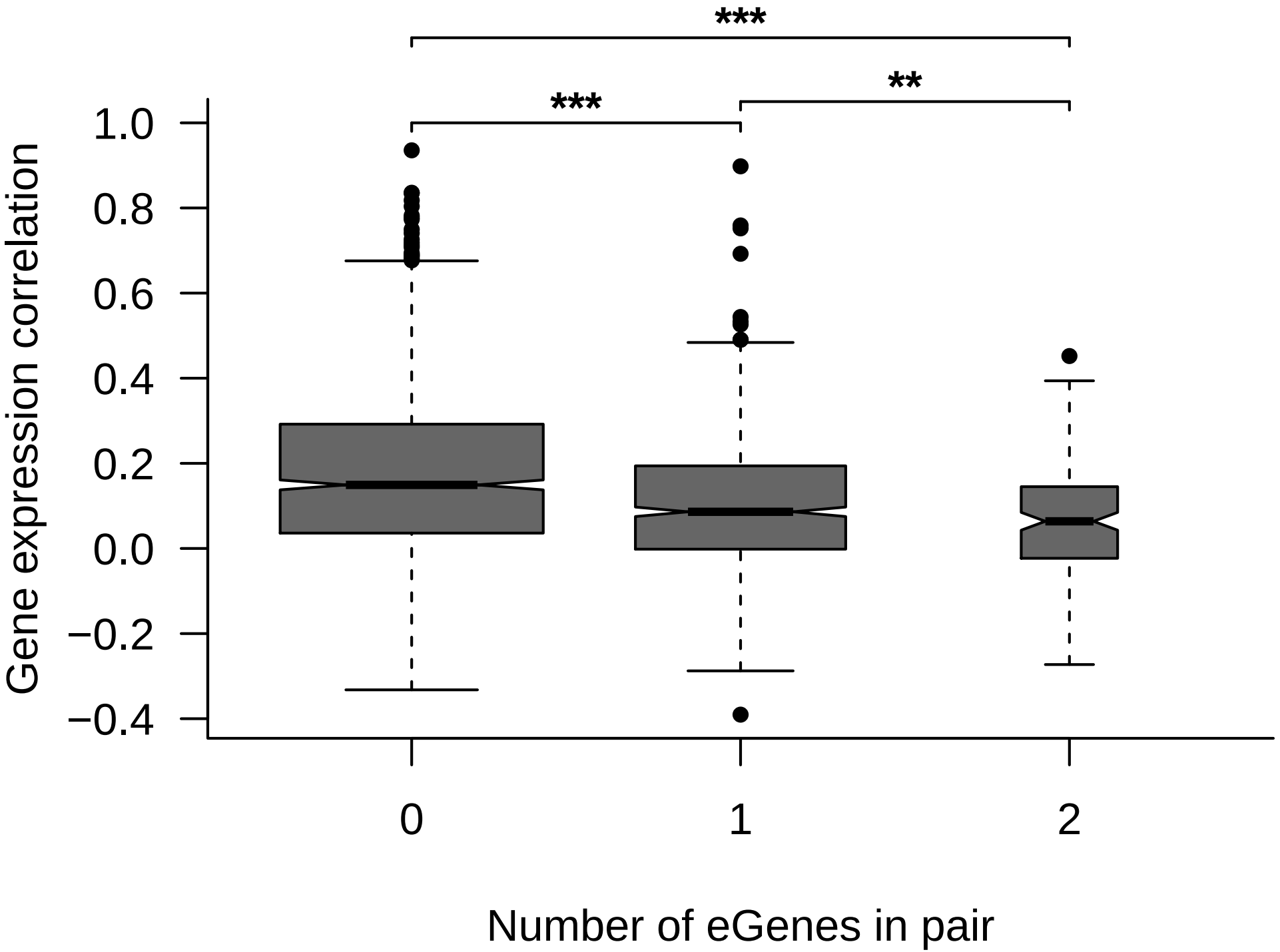
1253 **S25 Fig. Heritability distributions for four non-overlapping sets of genes:** genes with both local
1254 and distant eQTLs, genes with only distant eQTLs, genes with only local eQTLs, and genes with no
1255 significant eQTLs. The box widths are proportional to the number of genes in each set.

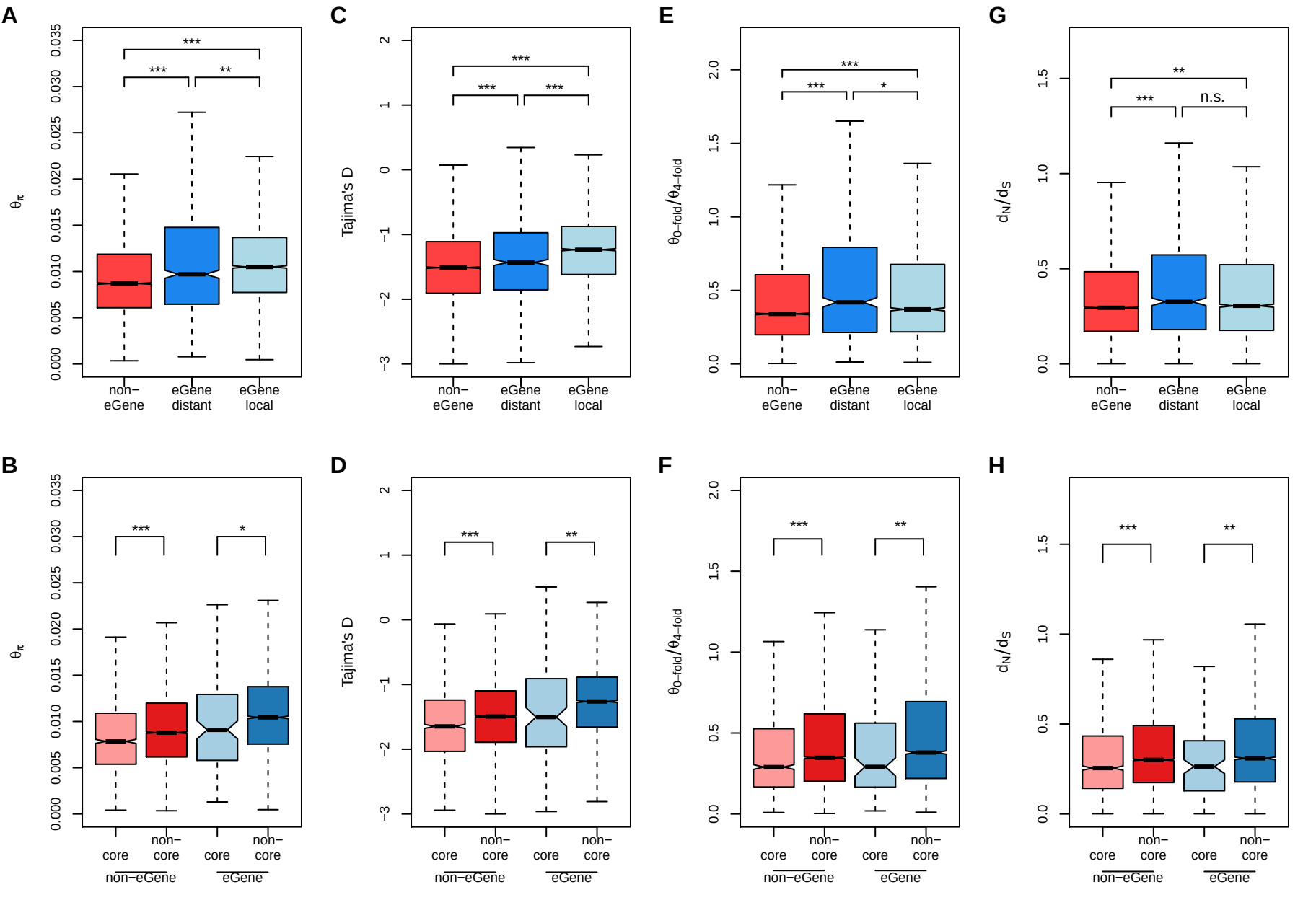
1256

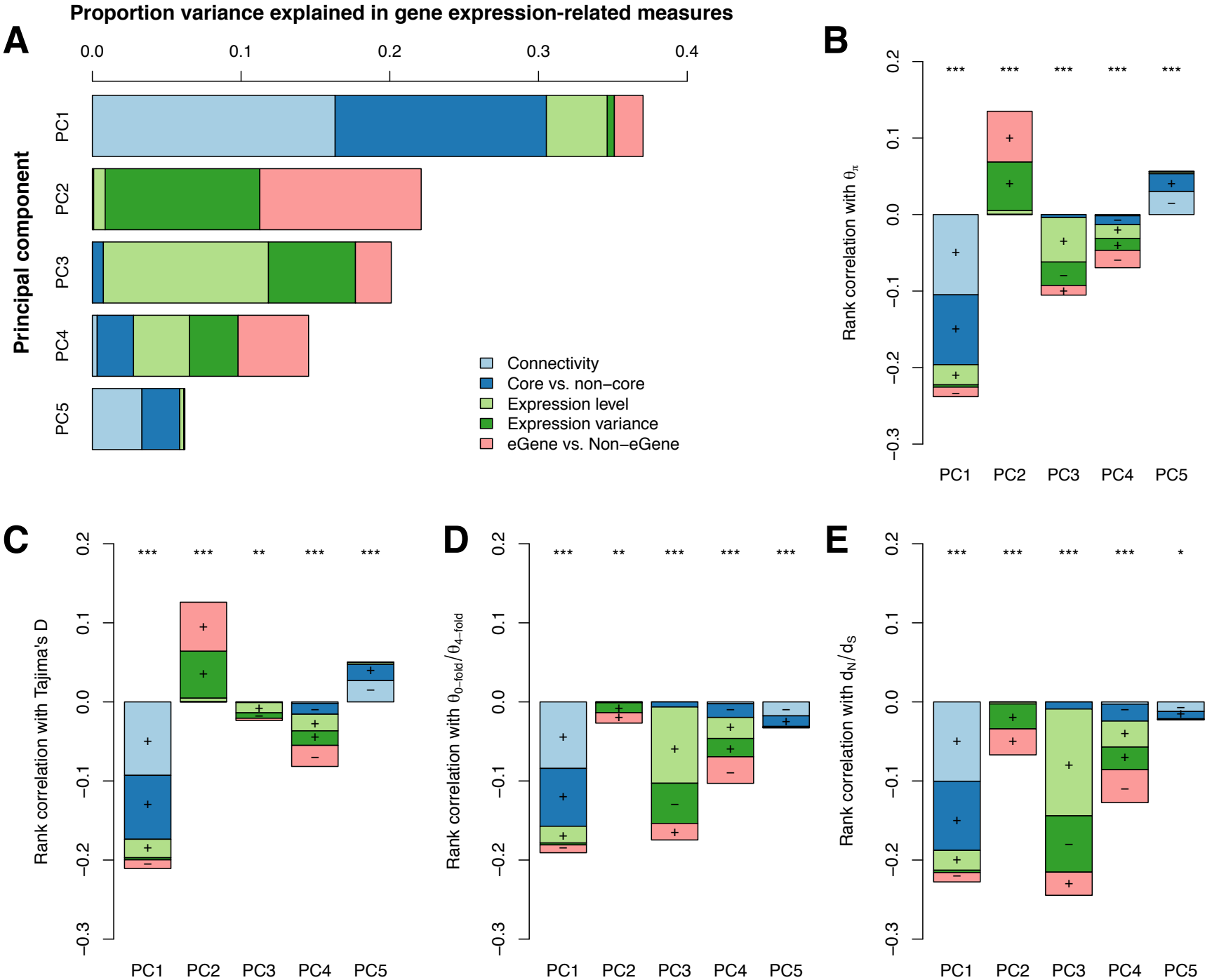

A
Latitude (° North)
Longitude (° East)
N
0 200 400 600 km
B
Density
All genes
Filtered genes
H^2
C
Density
All genes
Filtered genes
Q_{ST}
D
Count
Correlation
Population

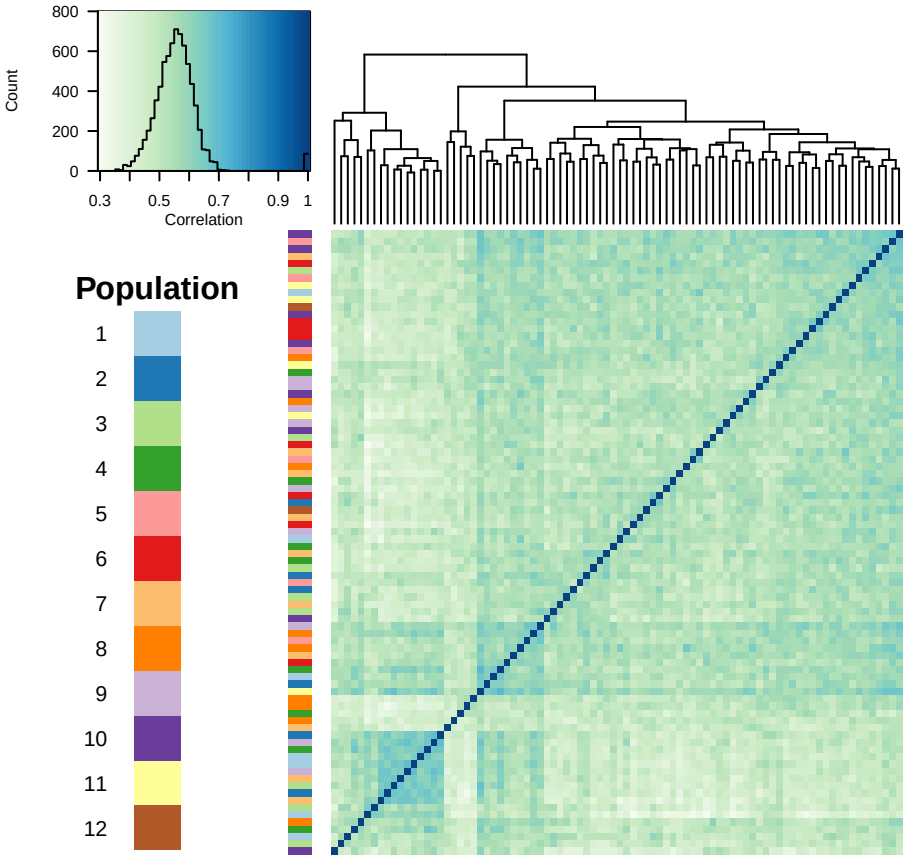

A
Local
Distant
Both
Variance explained
1.0
0.8
0.6
0.4
0.2
0.0
***
***
***
**
Maximum variance explained per gene
Total variance explained per gene
B
H^2
1.0
0.8
0.6
0.4
0.2
0.0
eGene
non−eGene
C
Gene position
19
15
11
8
6
4
2
1
1
2
4
6
8
11
15
19
SNP position
D
Associated gene
Other gene
5'
3'
Count / kbp sequence
1.5
1.0
0.5
0.0
Local eQTL
0.15
0.10
0.05
0.00
Distant eQTL
30
20
10
0
All SNPs
Intergenic
Flanking
UTR
Exon
Intron
E
$-\log_{10}(p)$
30
10
Local eQTL
30
10
Distant eQTL
Genomic position

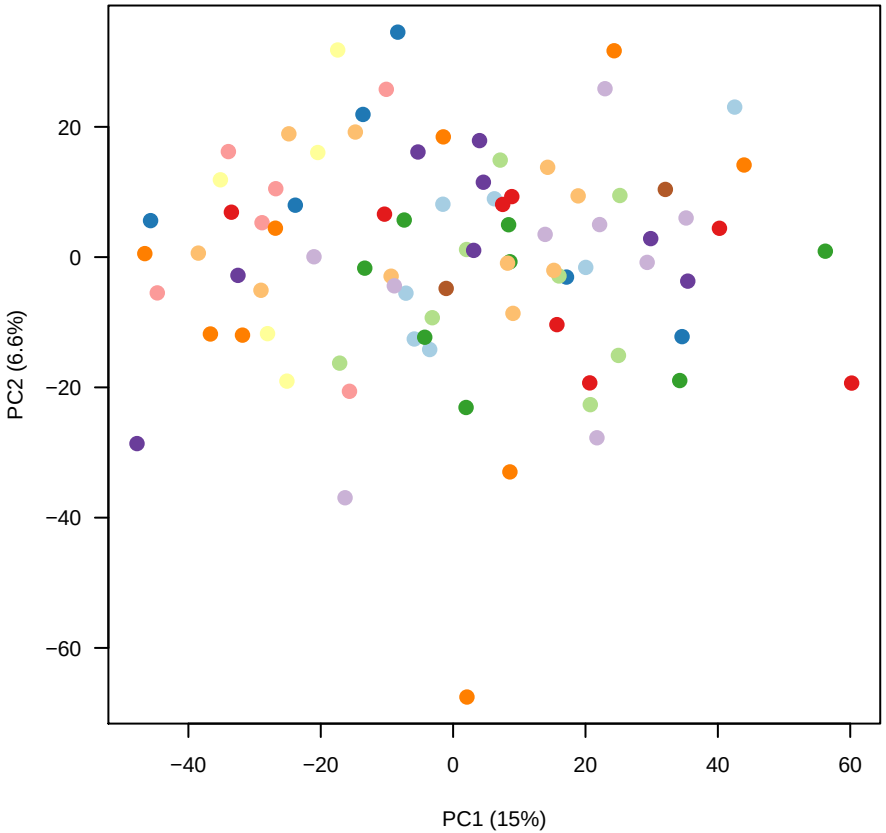

A
Expression variance
0
5
10
15
exAtlas
SwAsp
B
Connectivity
5
10
15
20
25
eGene
non−eGene
D
Core
Non−core
Number of genes
eGene
Non−eGene
0
50
100
150
200
0
1000
2000
3000
4000
5000
Both
Distant
Local
C
H^2
Proportion VE
Value
0.0
0.2
0.4
0.6
0.8
1.0
eGene
Non−eGene
Core
Non−core
E
Core eGene
Non−core eGene
Core non−eGene
Non−core non−eGene
eSNPs / kbp sequence
0.000
0.002
0.004
0.006
0.008
0.010
0.012
0.0
0.2
0.4
0.6
0.8
1.0
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Both
Distant
Local
UTR
Intron
Intergenic
Flanking
Exon

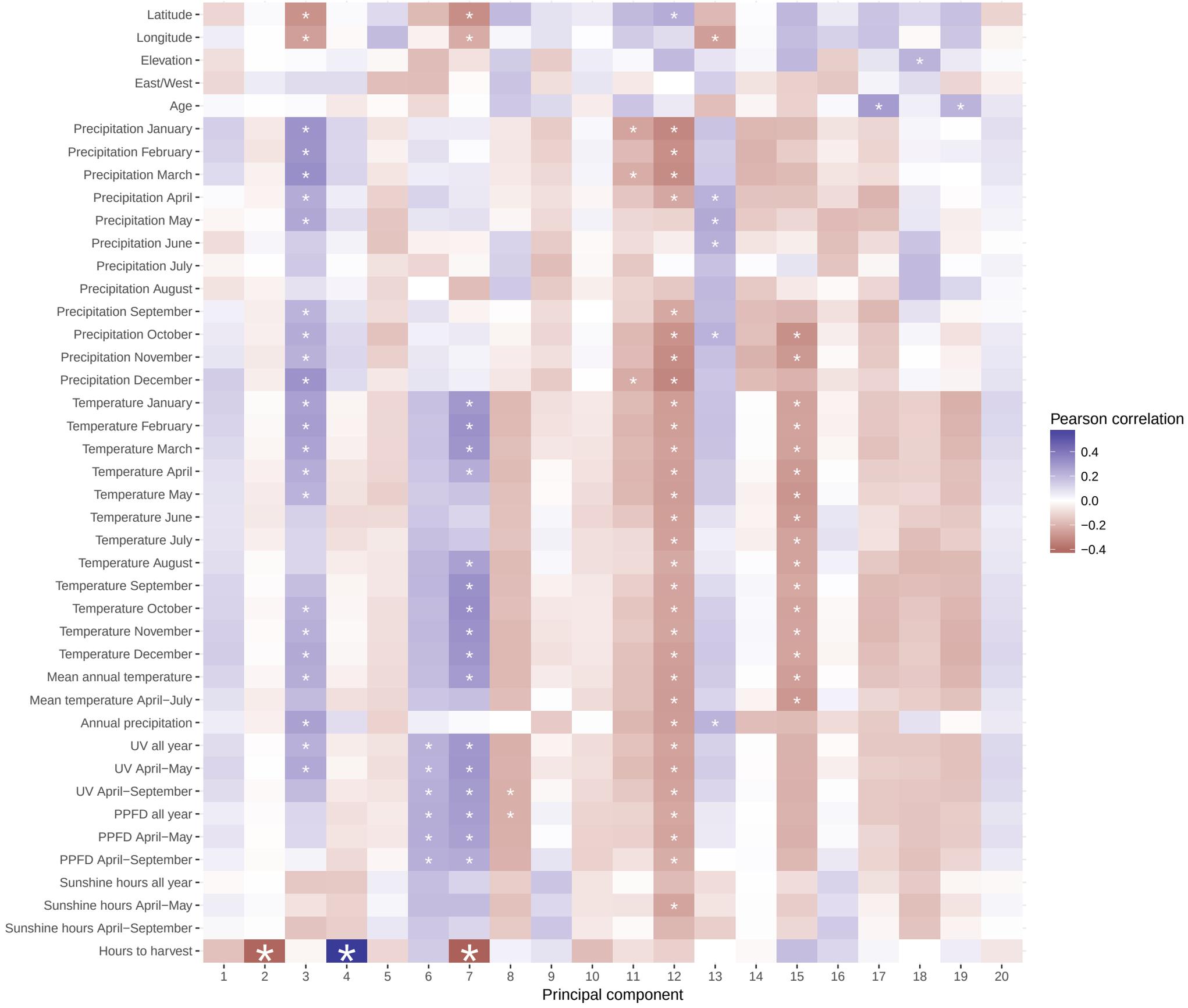

***
***
**
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Gene expression correlation
0
1
2
Number of eGenes in pair

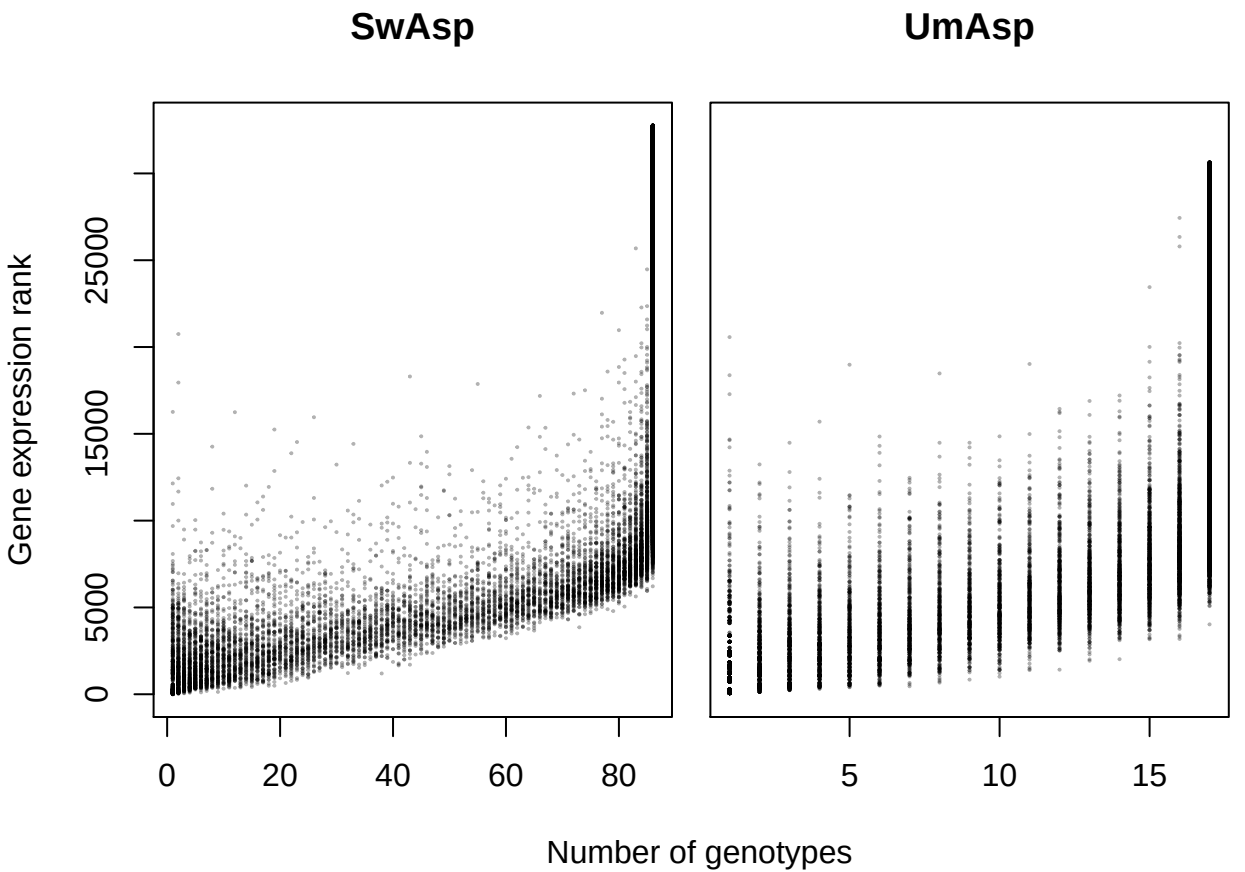

A
B
C
D
E
F
G
H
θπ
Tajima's D
θ0-fold/θ4-fold
dN/dS
non-eGene
eGene distant
eGene local
core
non-core
non-eGene
eGene
***
**
*
n.s.

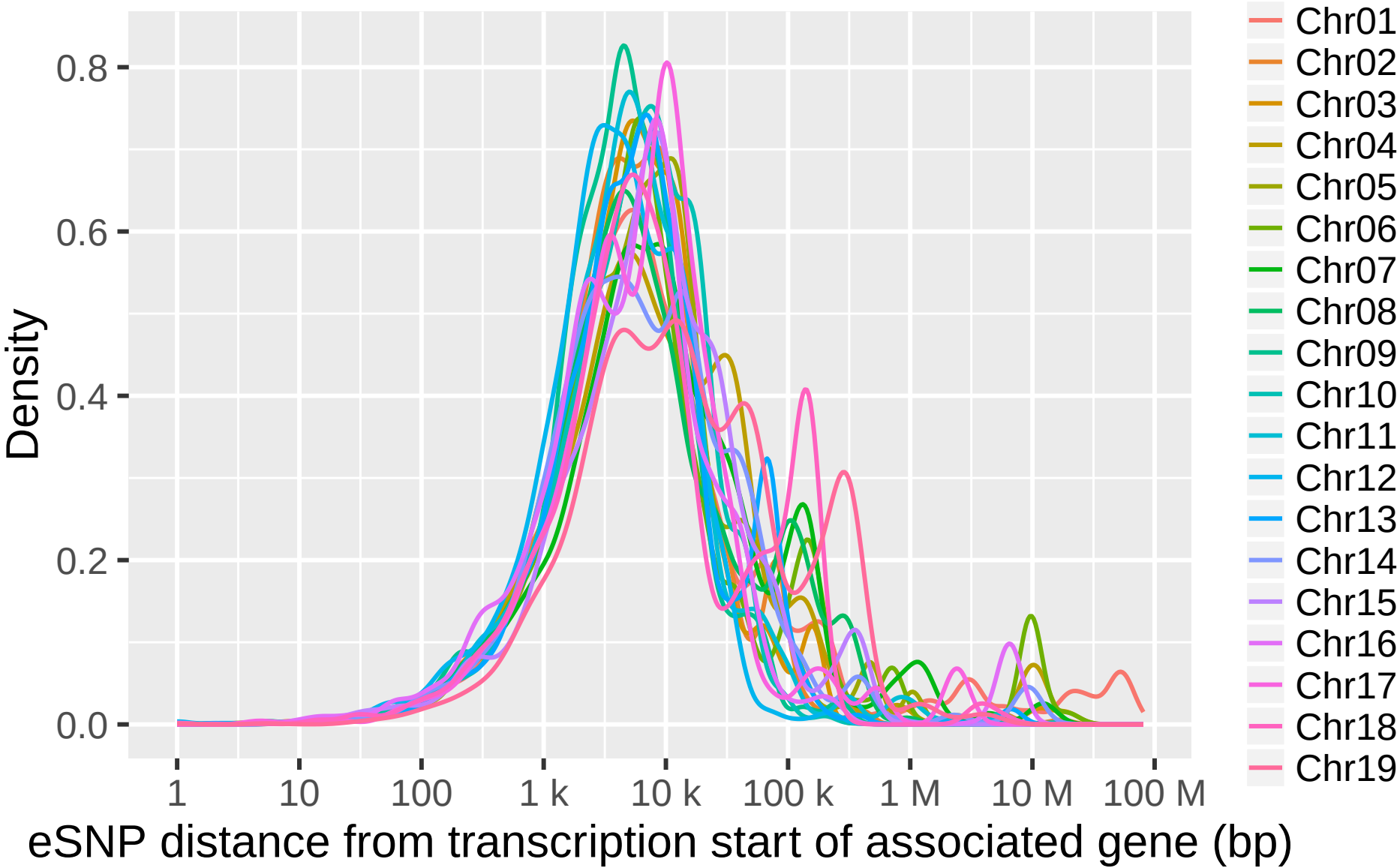

A
Proportion variance explained in gene expression-related measures
0.0
0.1
0.2
0.3
0.4
Principal component
PC1
PC2
PC3
PC4
PC5
Connectivity
Core vs. non−core
Expression level
Expression variance
eGene vs. Non−eGene
B
Rank correlation with θ_π
C
Rank correlation with Tajima's D
D
Rank correlation with $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$
E
Rank correlation with d_N/d_S
0.2
0.1
0.0
−0.1
−0.2
−0.3
PC1
PC2
PC3
PC4
PC5

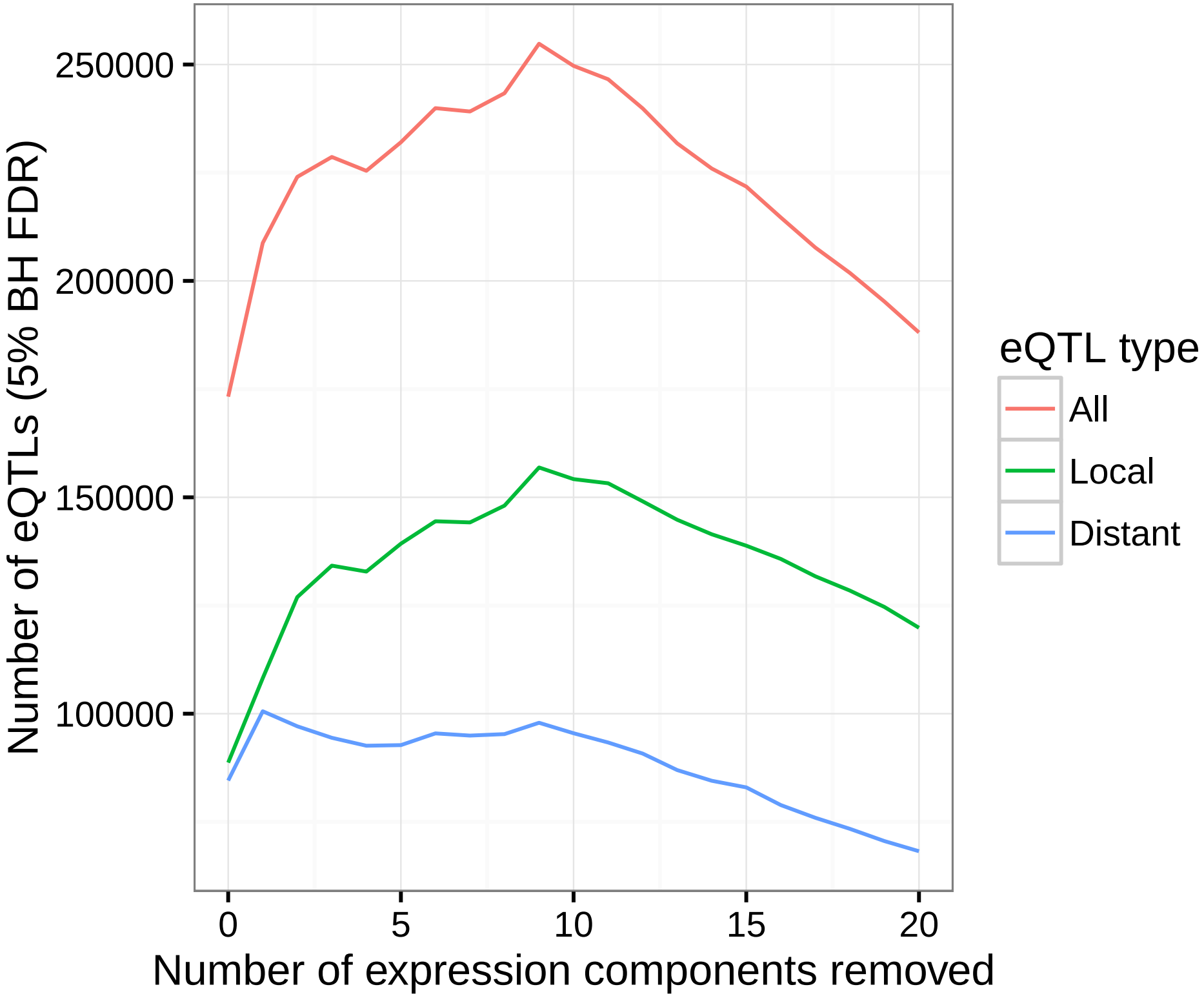

Count
800
600
400
200
0
0.3
0.5
0.7
0.9
1
Correlation
Population
1
2
3
4
5
6
7
8
9
10
11
12

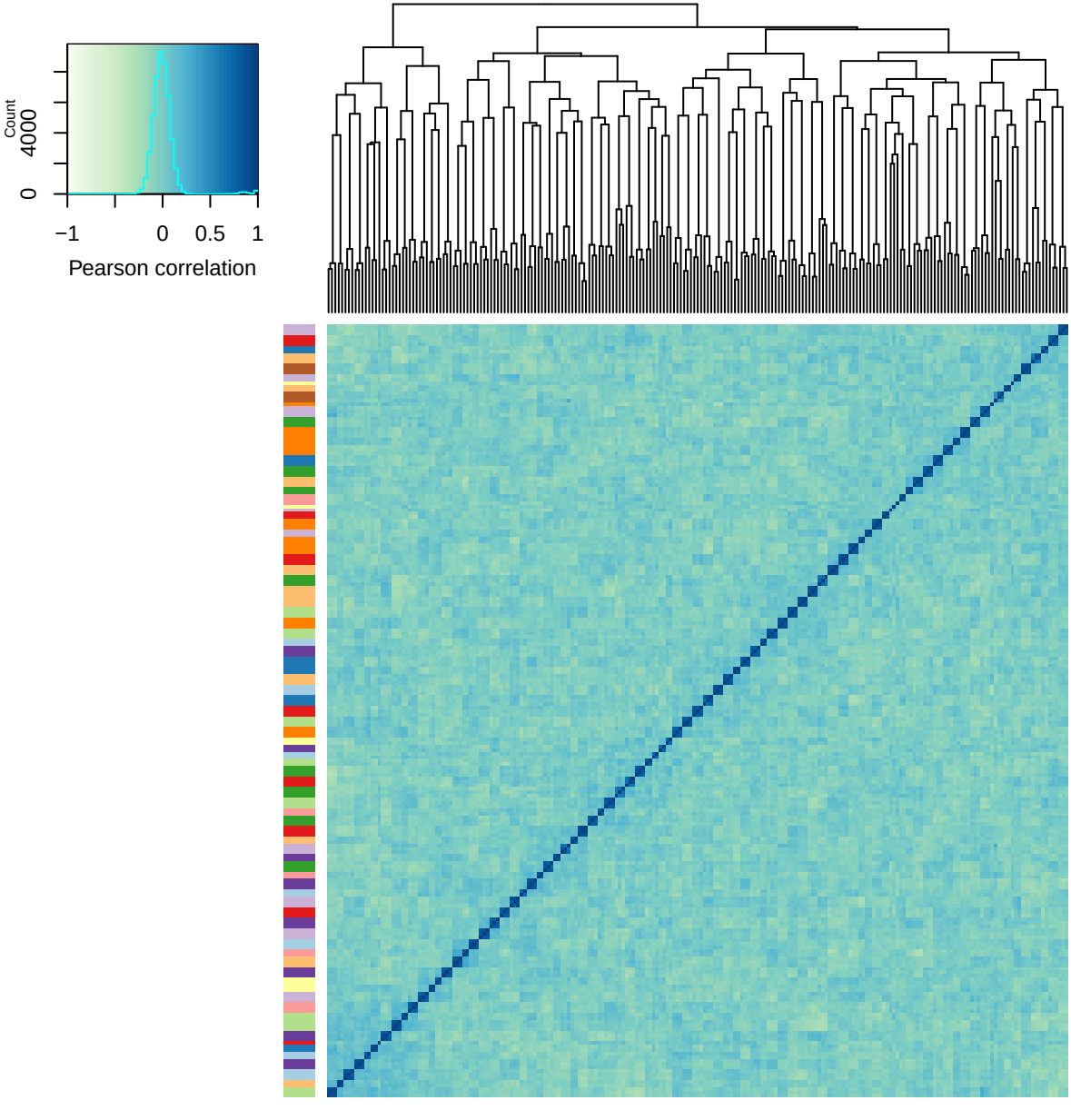

PC2 (6.6%)
20
0
−20
−40
−60
−40
−20
0
20
40
60
PC1 (15%)

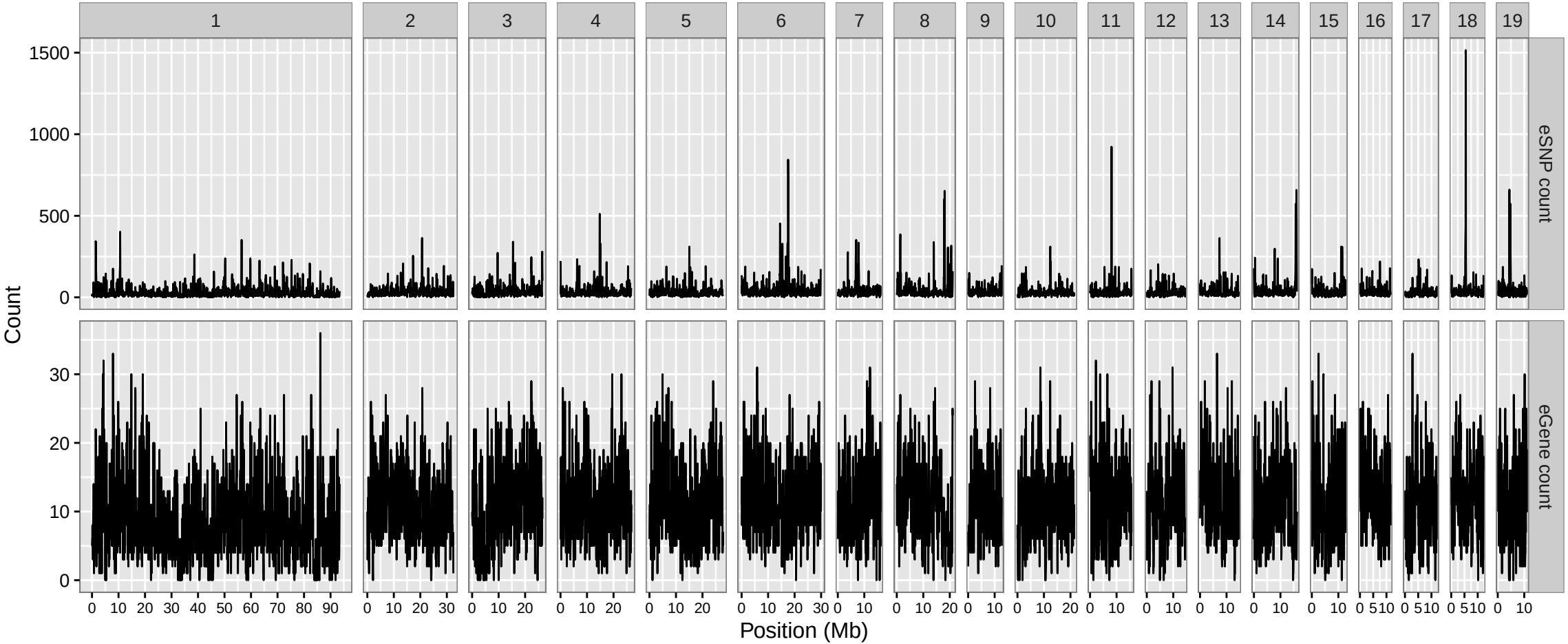

Latitude
Longitude
Elevation
East/West
Age
Precipitation January
Precipitation February
Precipitation March
Precipitation April
Precipitation May
Precipitation June
Precipitation July
Precipitation August
Precipitation September
Precipitation October
Precipitation November
Precipitation December
Temperature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

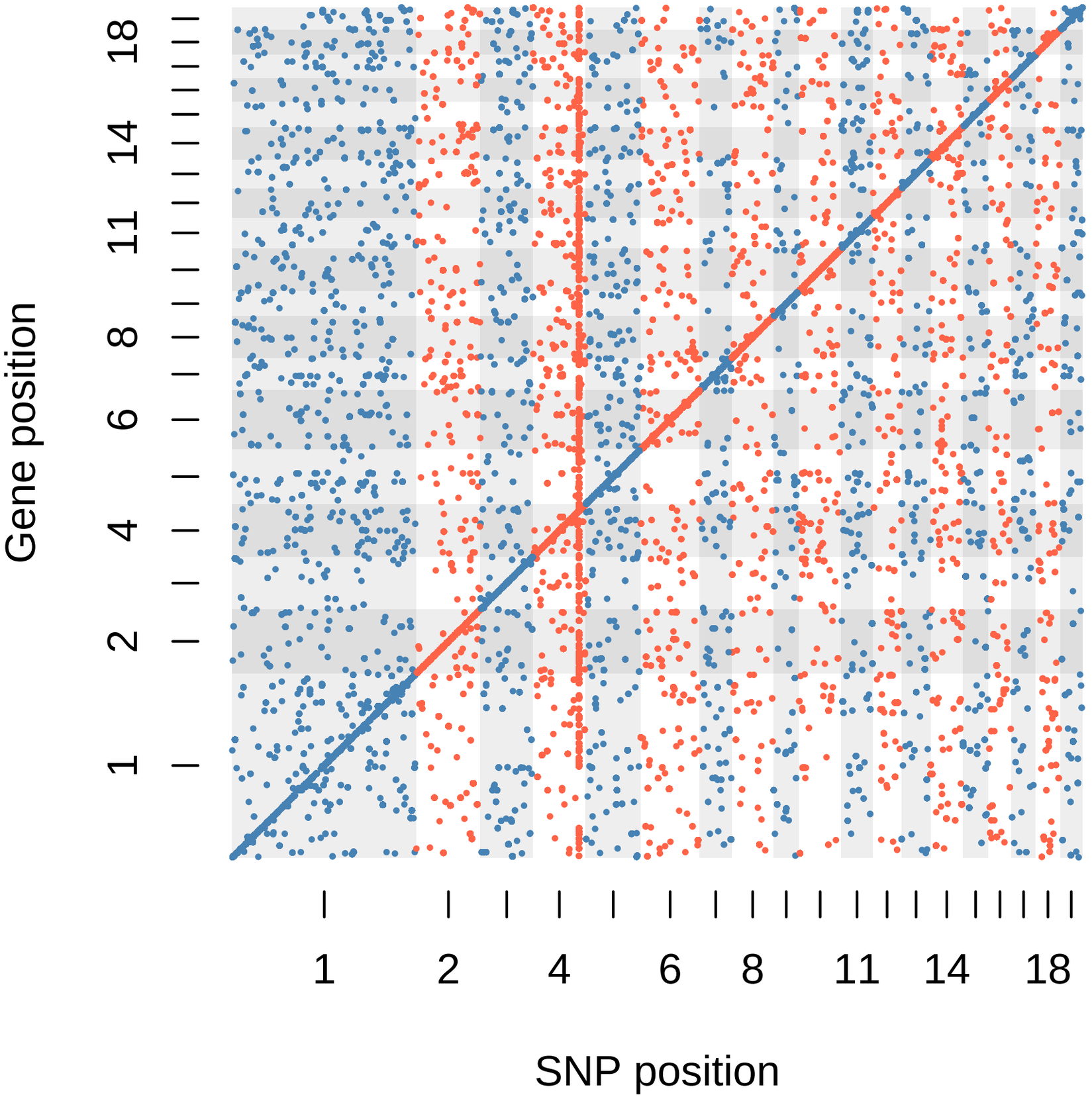

SwAsp
UmAsp
Gene expression rank
0
5000
15000
25000
0
20
40
60
80
5
10
15
Number of genotypes

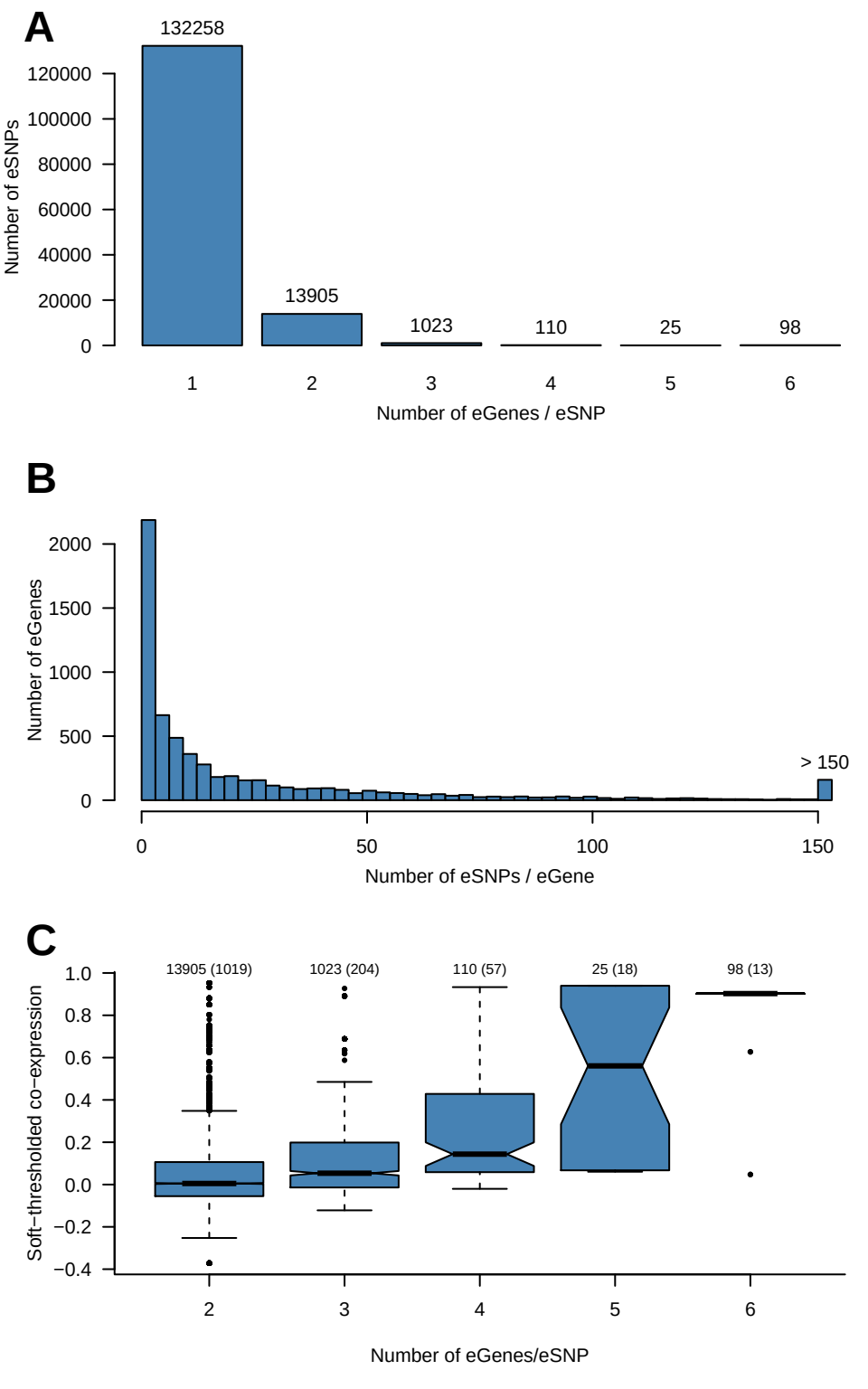

Density
0.8
0.6
0.4
0.2
0.0
1
10
100
1 k
10 k
100 k
1 M
10 M
100 M
eSNP distance from transcription start of associated gene (bp)
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19

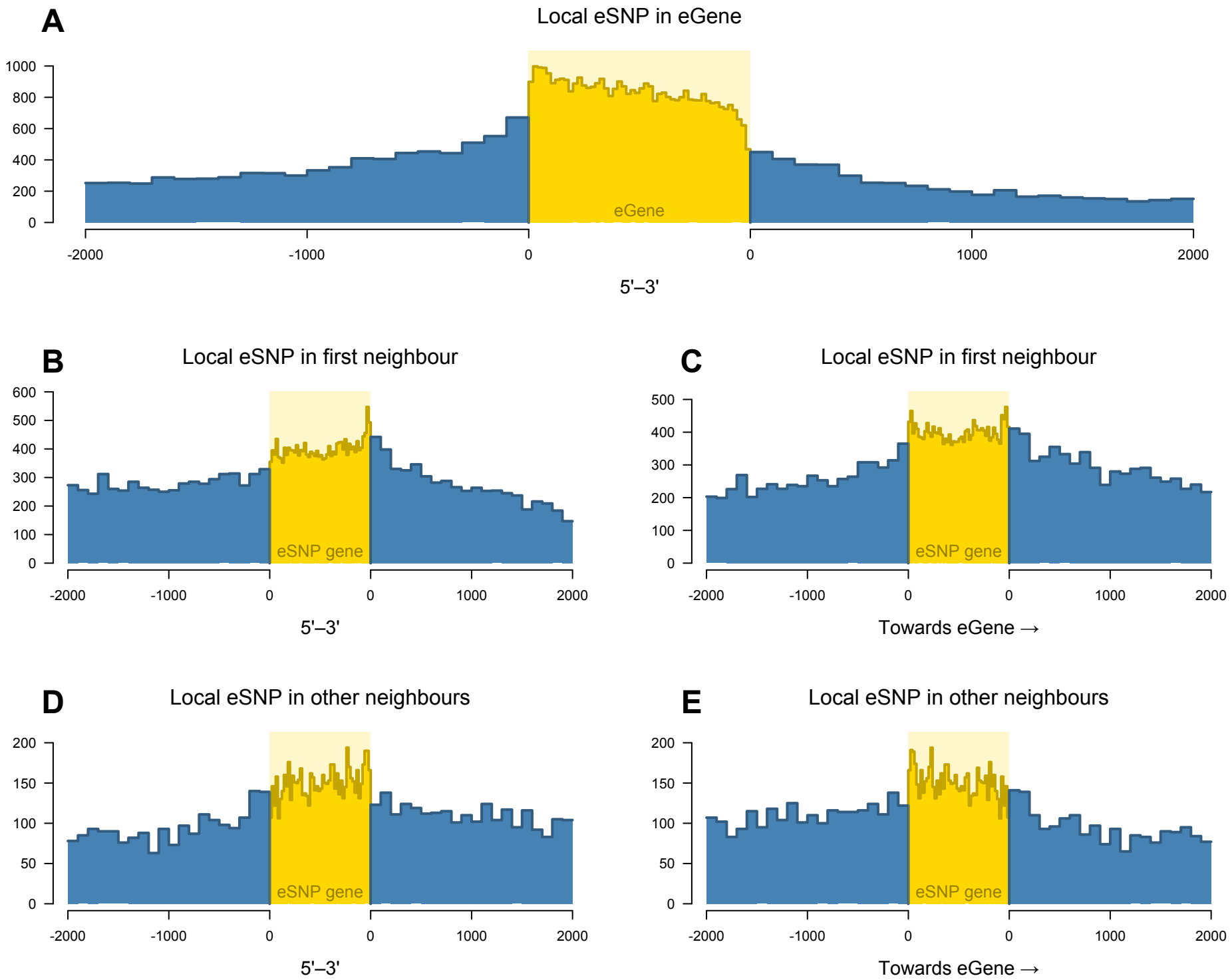

250000
200000
150000
100000
Number of eQTLs (5% BH FDR)
0
5
10
15
20
Number of expression components removed
eQTL type
All
Local
Distant

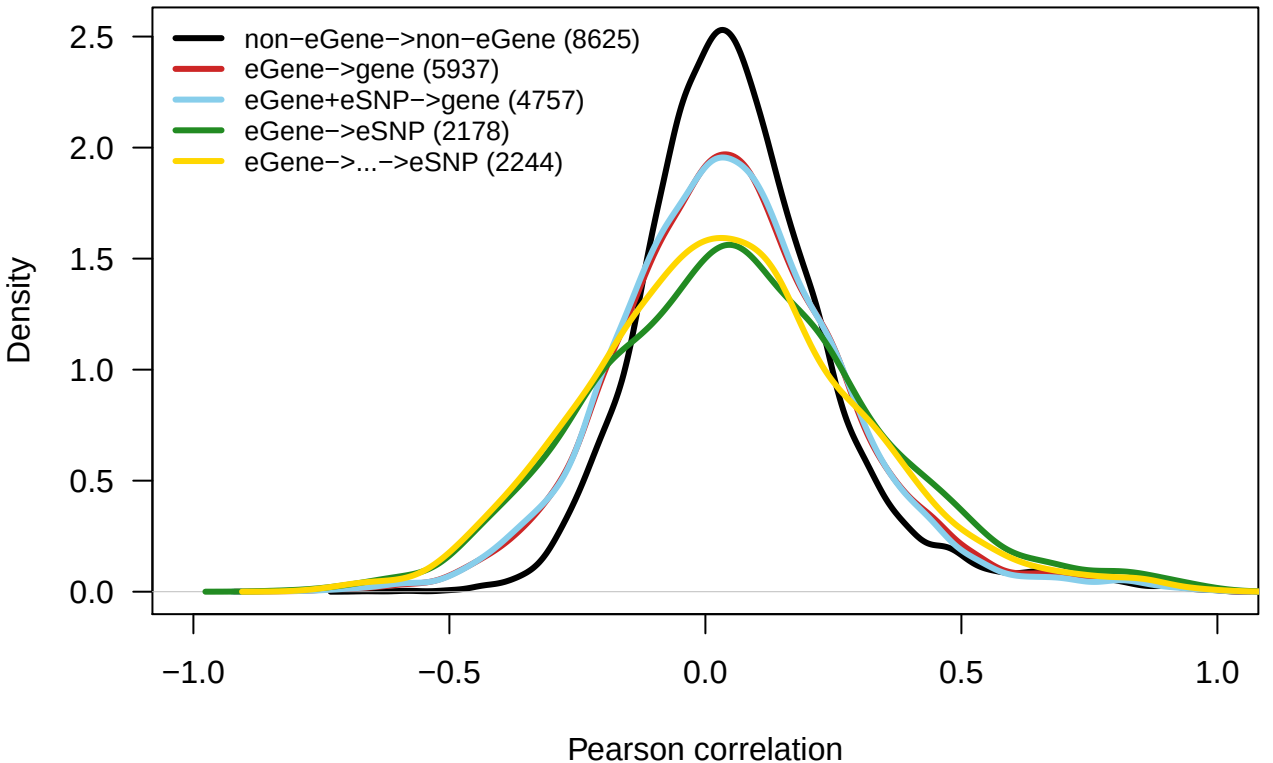

Count
4000
0
−1
0
0.5
1
Pearson correlation

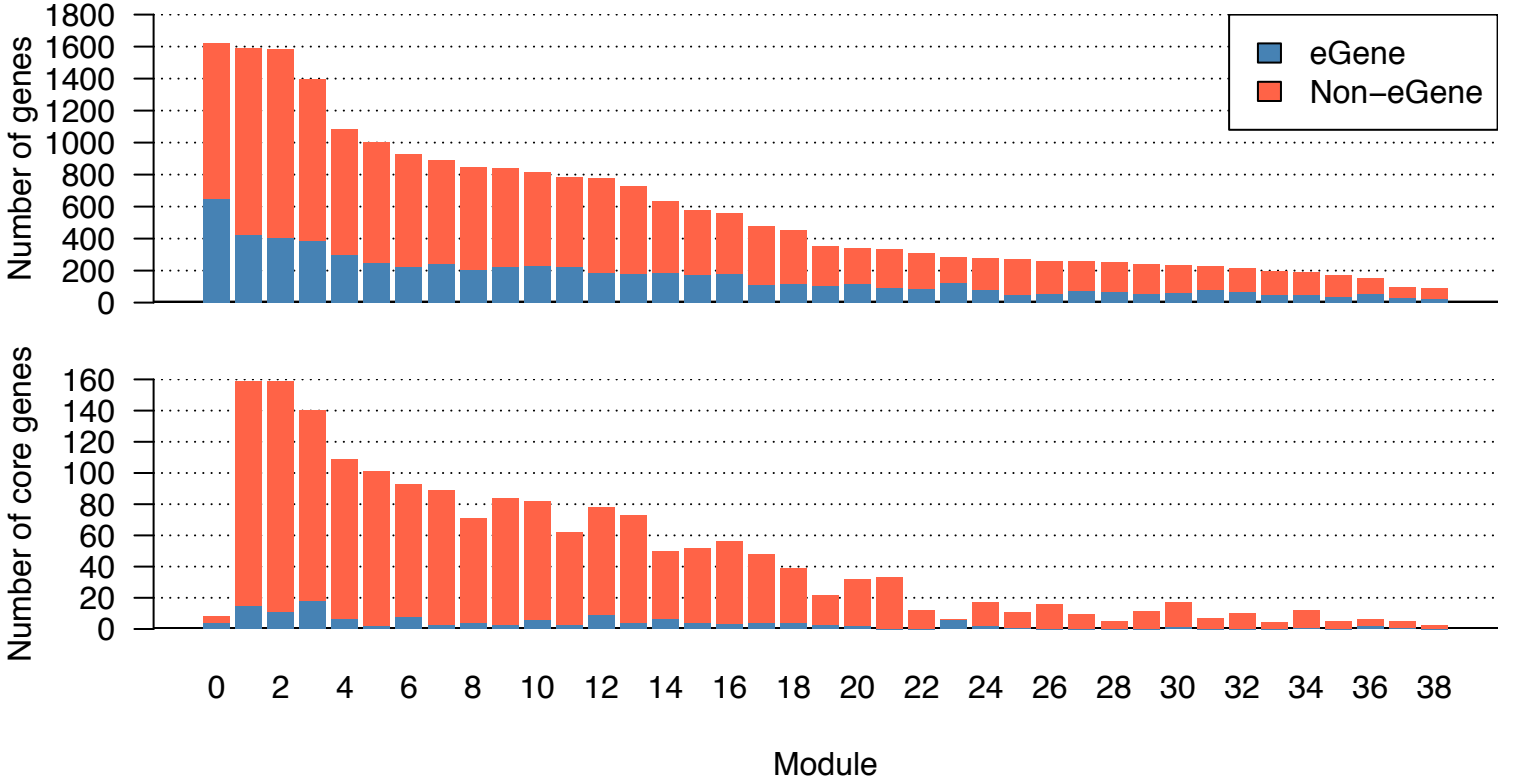

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
eSNP count
eGene count
Count
Position (Mb)

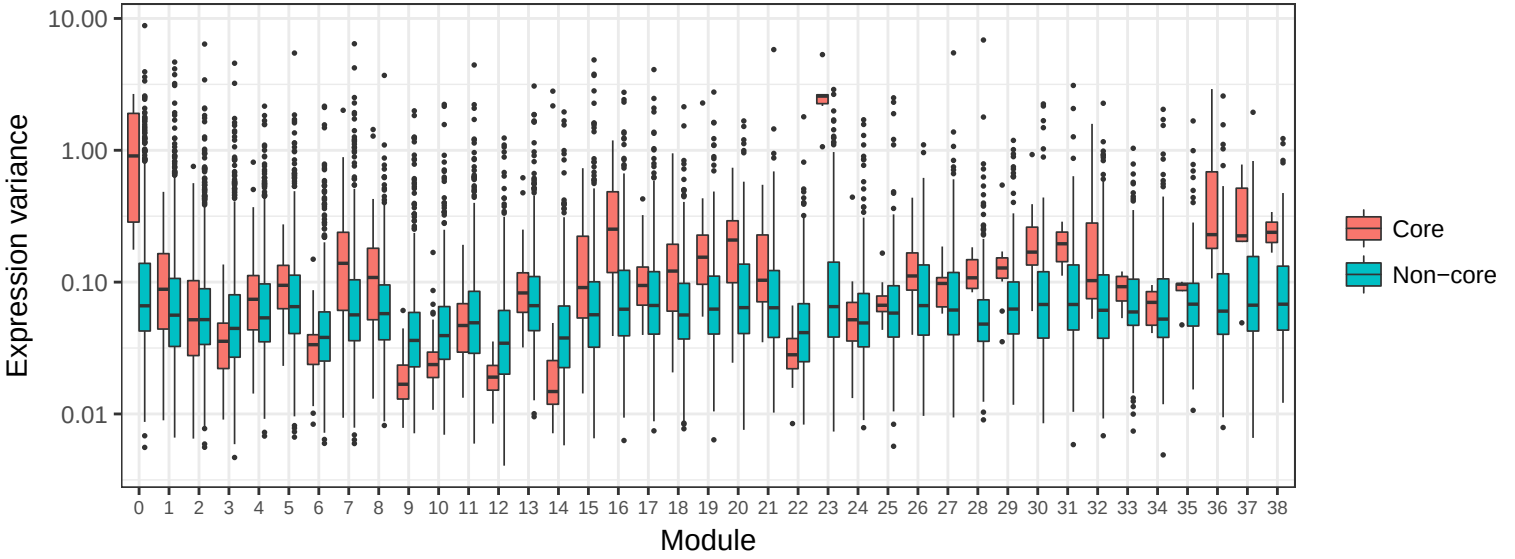

Gene position
1
2
4
6
8
11
14
18
SNP position
1
2
4
6
8
11
14
18

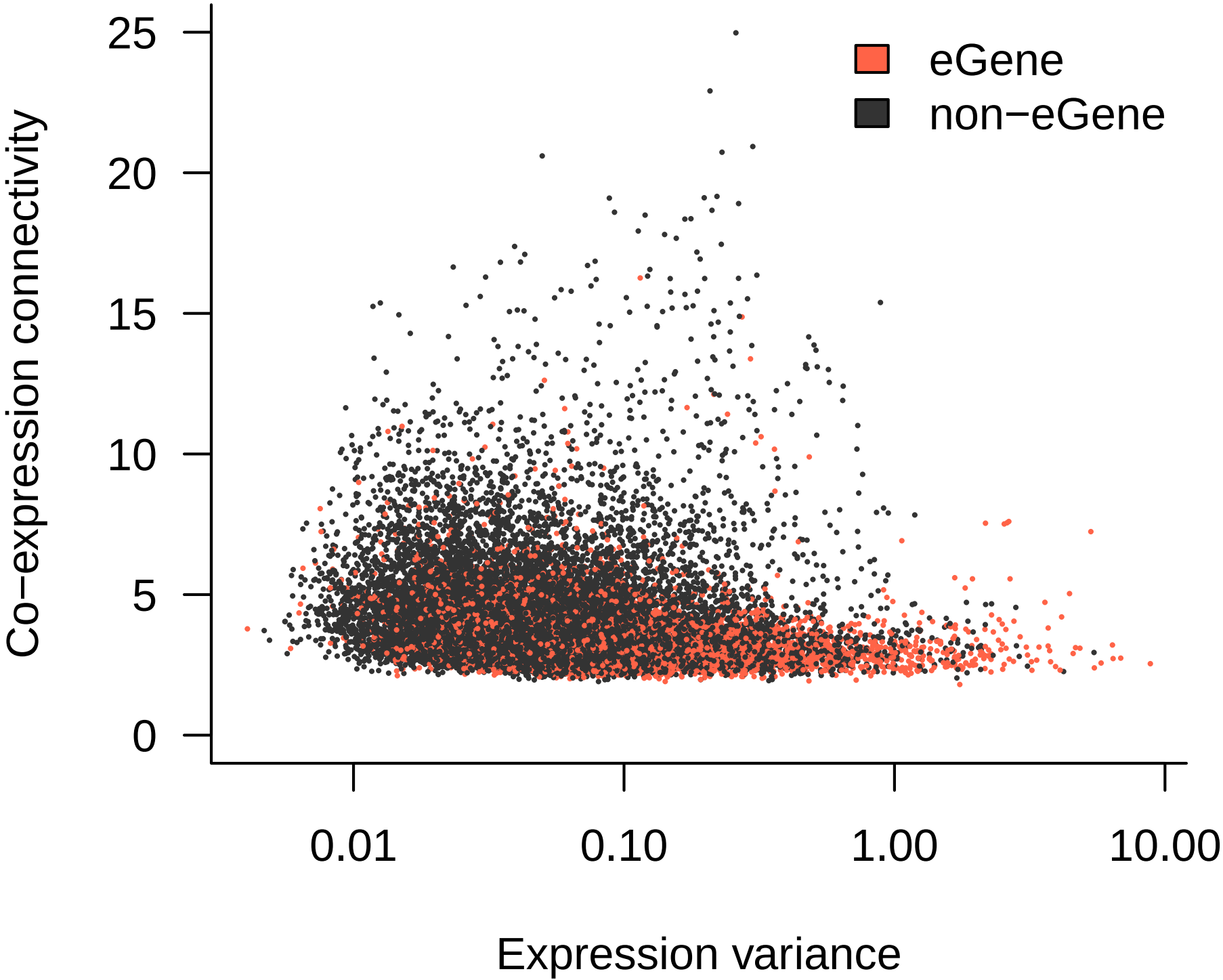

A
132258
13905
1023
110
25
98
Number of eSNPs
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
Number of eGenes / eSNP
B
Number of eGenes
2000
1500
1000
500
0
> 150
0
50
100
150
Number of eSNPs / eGene
C
13905 (1019)
1023 (204)
110 (57)
25 (18)
98 (13)
Soft−thresholded co−expression
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
2
3
4
5
6
Number of eGenes/eSNP

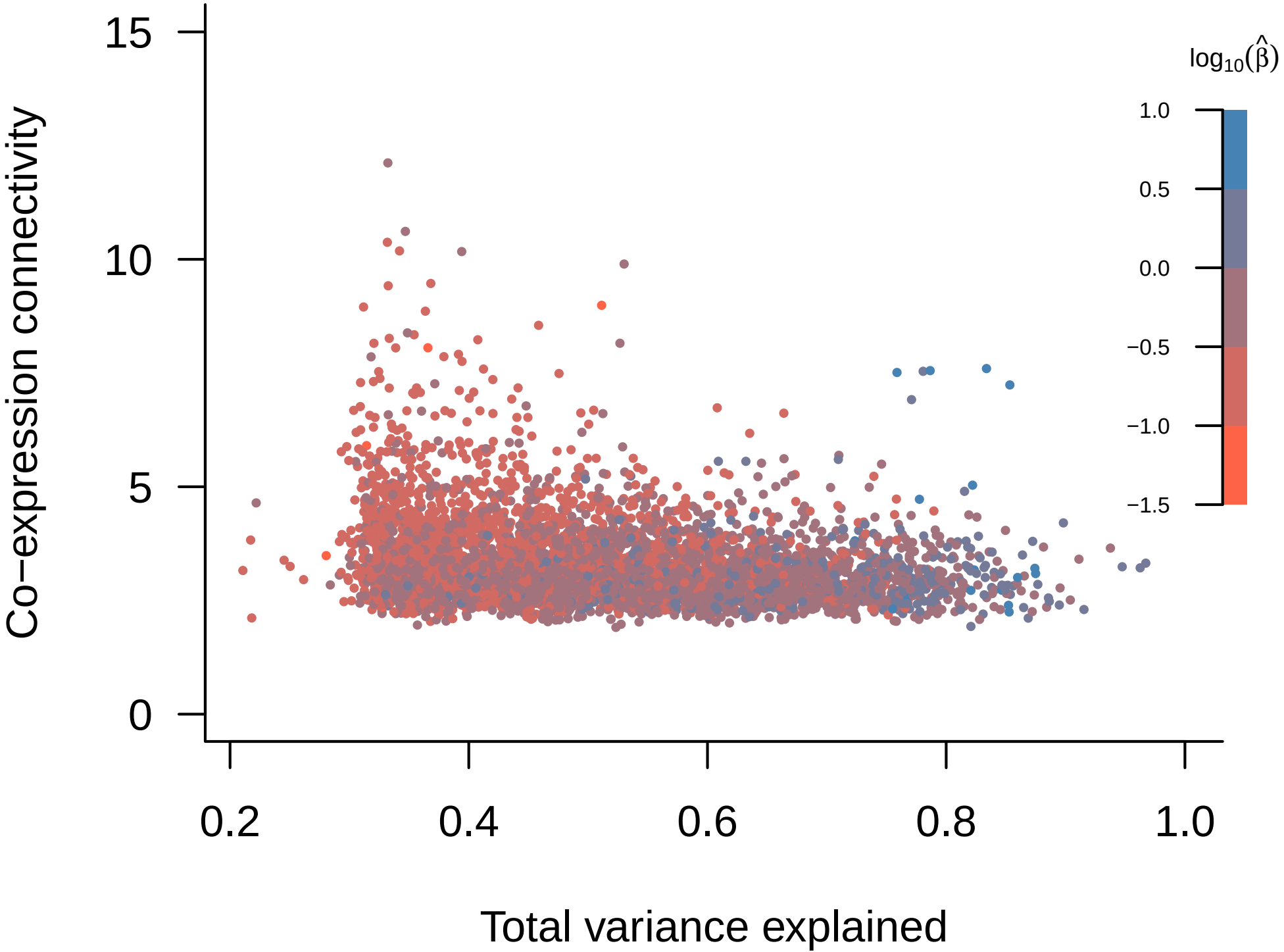

A
Local eSNP in eGene
1000
800
600
400
200
0
eGene
-2000
-1000
0
0
1000
2000
5'–3'
B
Local eSNP in first neighbour
600
500
400
300
200
100
0
eSNP gene
-2000
-1000
0
0
1000
2000
5'–3'
C
Local eSNP in first neighbour
500
400
300
200
100
0
eSNP gene
-2000
-1000
0
0
1000
2000
Towards eGene →
D
Local eSNP in other neighbours
200
150
100
50
0
eSNP gene
-2000
-1000
0
0
1000
2000
5'–3'
E
Local eSNP in other neighbours
200
150
100
50
0
eSNP gene
-2000
-1000
0
0
1000
2000
Towards eGene →

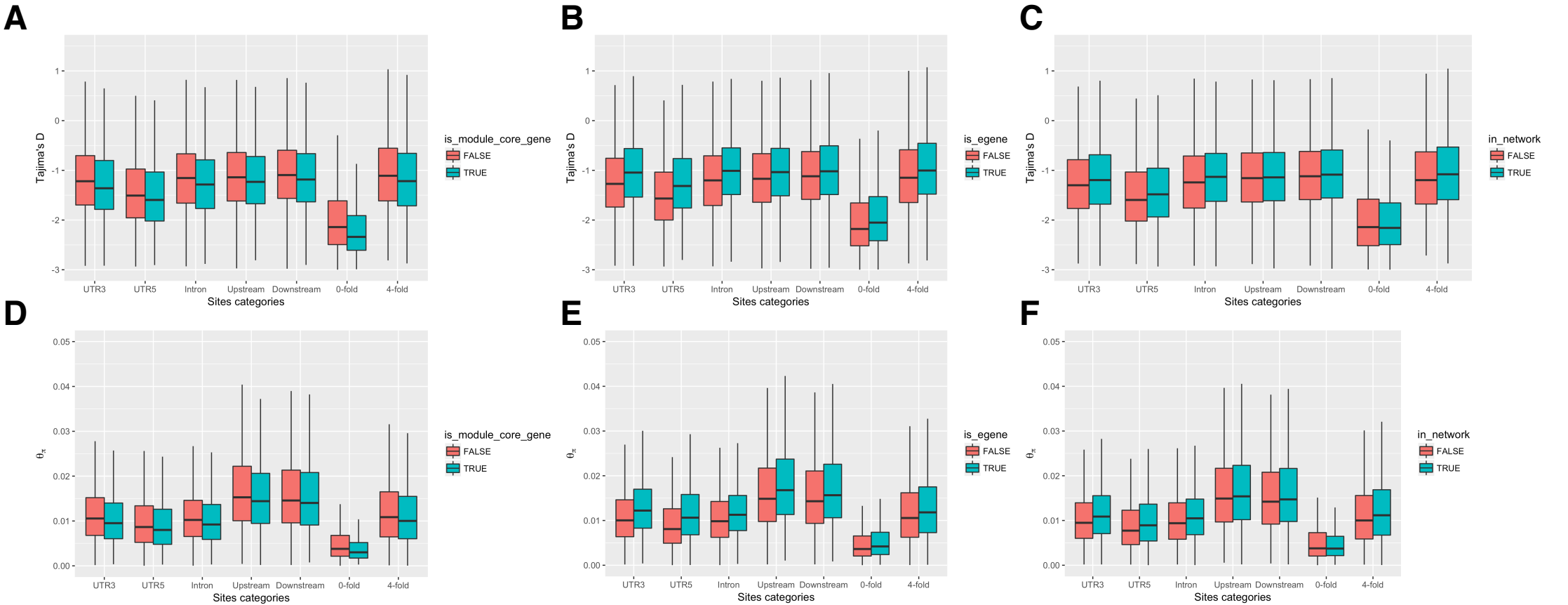

non−eGene−>non−eGene (8625)
eGene−>gene (5937)
eGene+eSNP−>gene (4757)
eGene−>eSNP (2178)
eGene−>...−>eSNP (2244)
Density
2.5
2.0
1.5
1.0
0.5
0.0
−1.0
−0.5
0.0
0.5
1.0
Pearson correlation

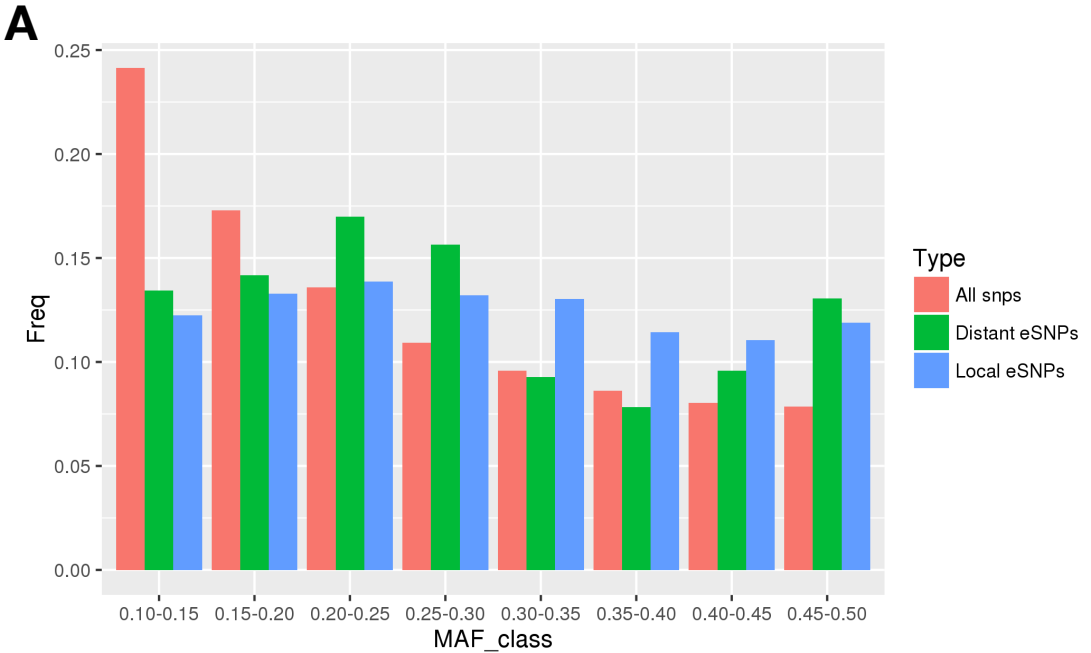

eGene
Non−eGene
Number of genes
1800
1600
1400
1200
1000
800
600
400
200
0
Number of core genes
160
140
120
100
80
60
40
20
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38
Module

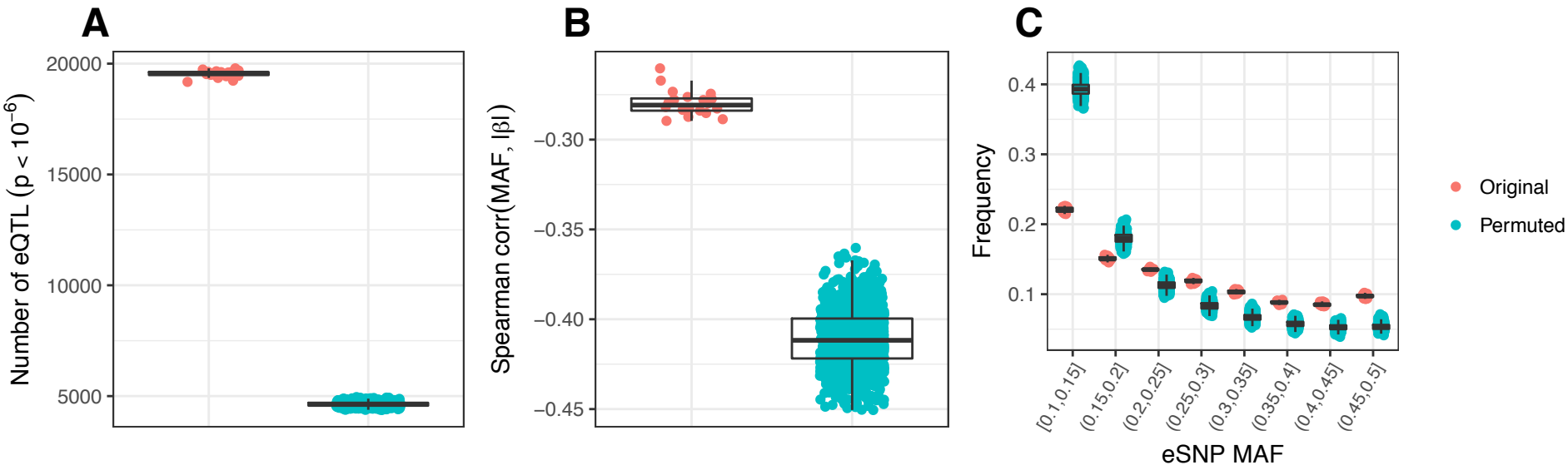

Expression variance
10.00
1.00
0.10
0.01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Module
Core
Non−core

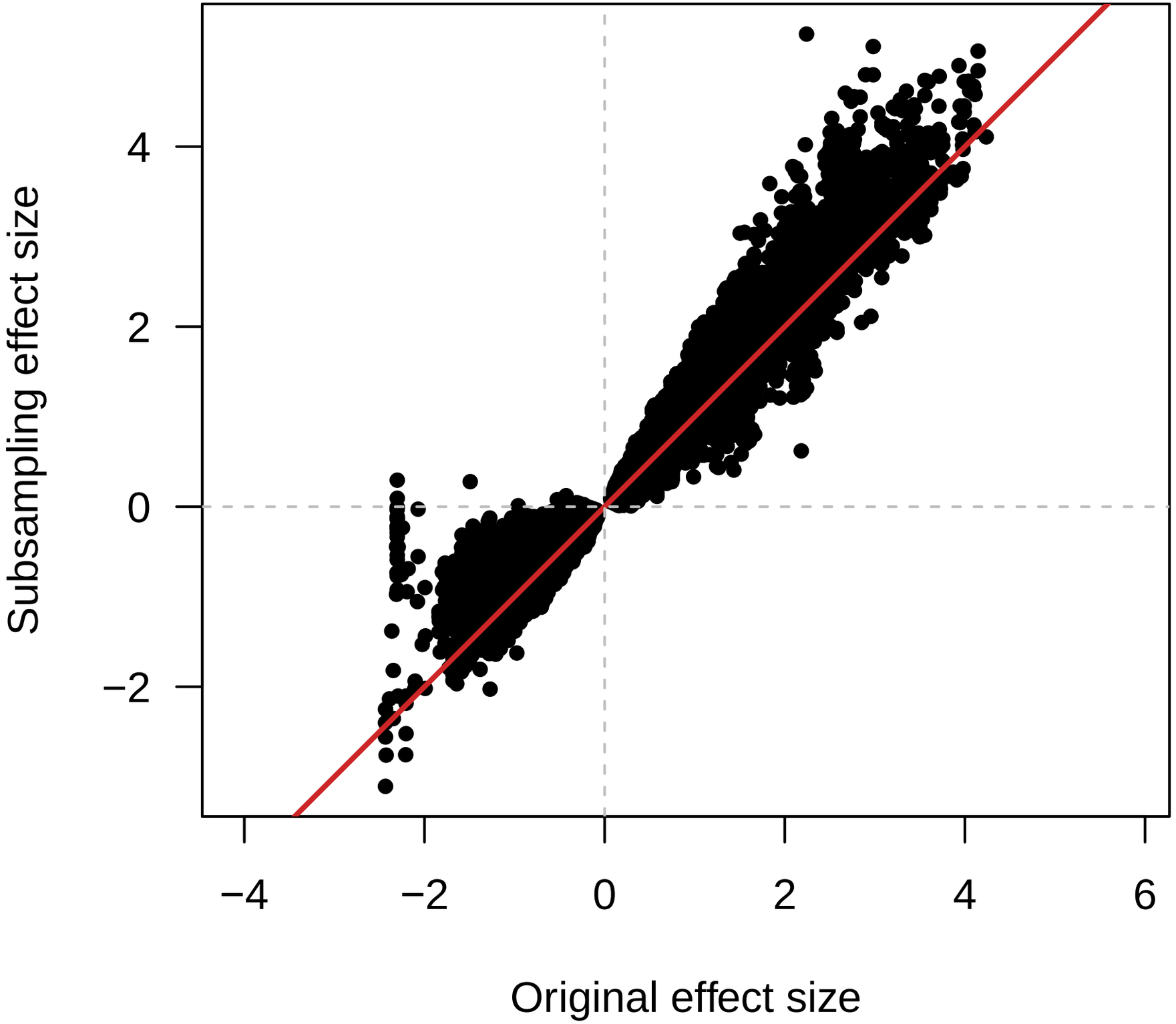

eGene
non−eGene
Co−expression connectivity
25
20
15
10
5
0
0.01
0.10
1.00
10.00
Expression variance

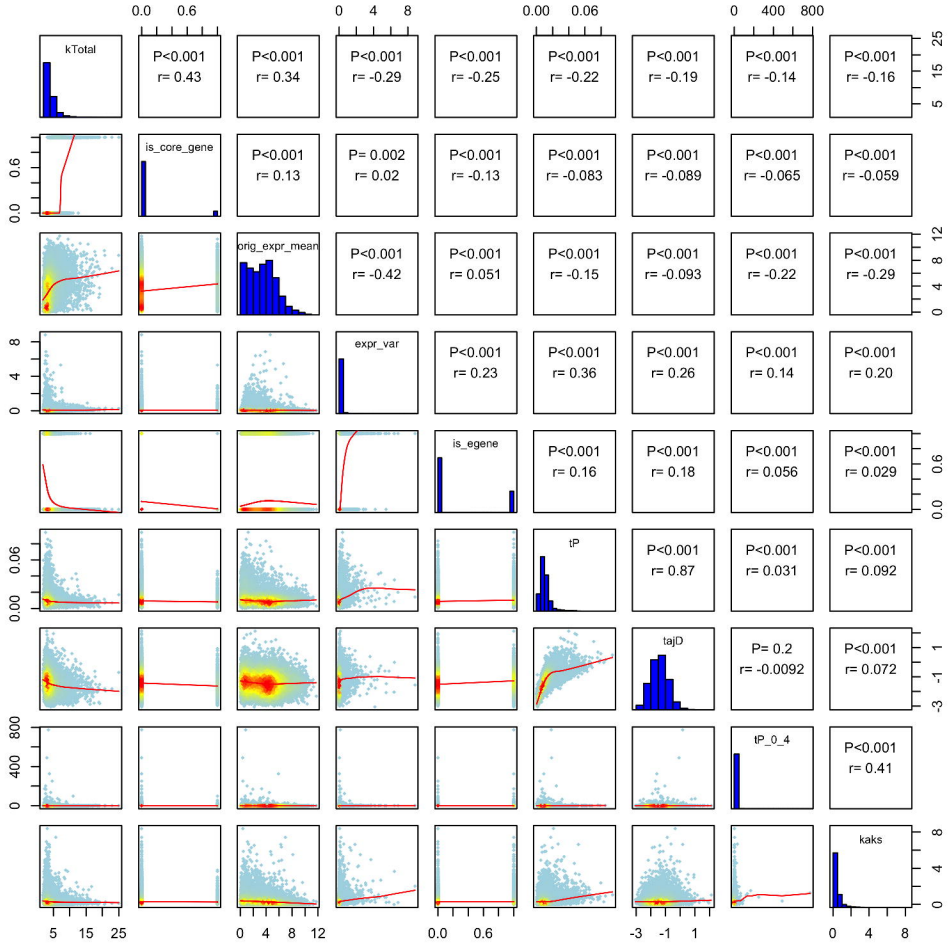

Co−expression connectivity
15
10
5
0
0.2
0.4
0.6
0.8
1.0
Total variance explained
$\log_{10}(\hat{\beta})$
1.0
0.5
0.0
−0.5
−1.0
−1.5

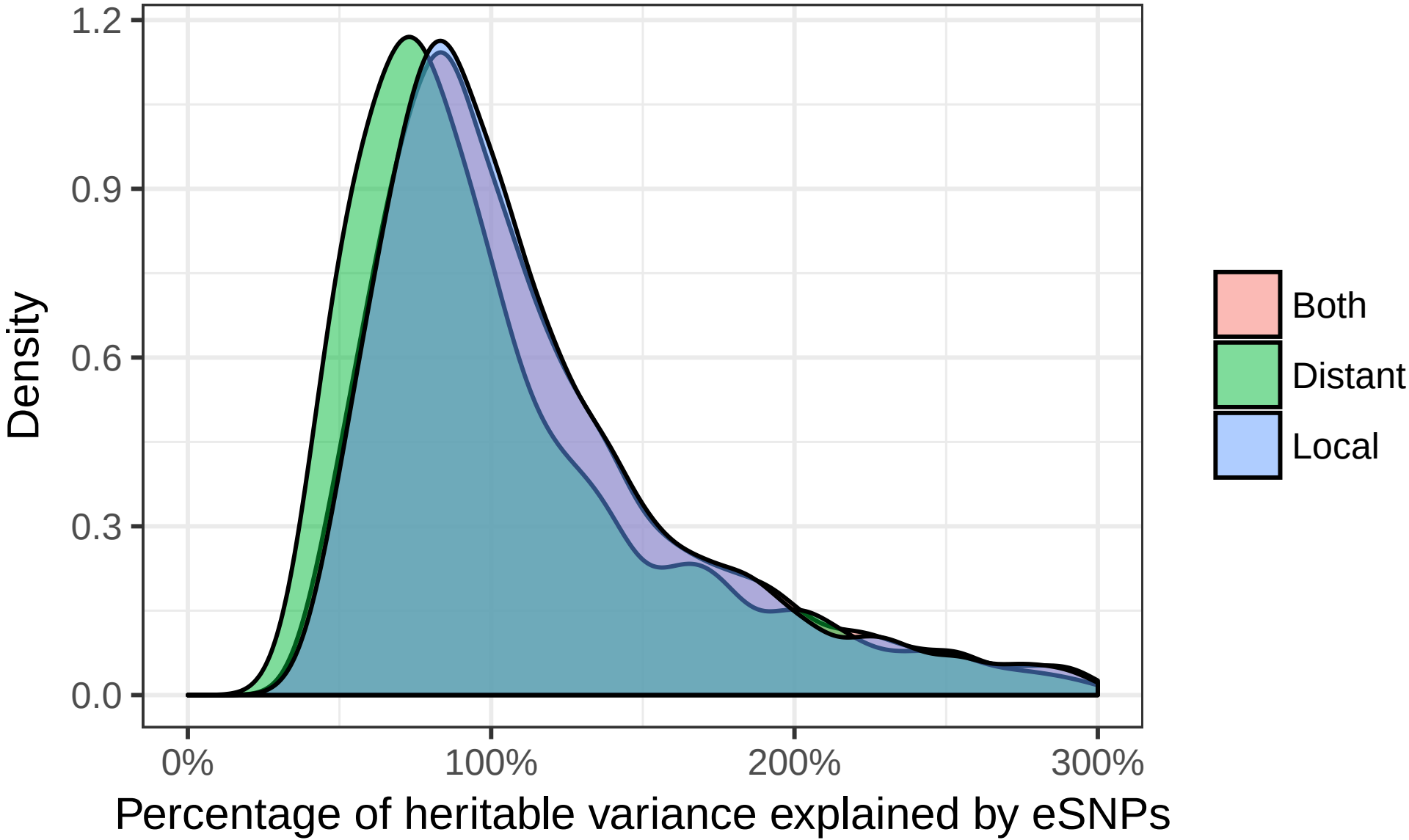

A
B
C
D
E
F
Tajima's D
θ_π
Sites categories
UTR3
UTR5
Intron
Upstream
Downstream
0-fold
4-fold
is_module_core_gene
is_egene
in_network
FALSE
TRUE


A
Freq
0.25
0.20
0.15
0.10
0.05
0.00
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
Absolute effect size
0
1
2
3
4
5
ρ=-0.285***
0.1
0.2
0.3
0.4
0.5
eQTL minor allele frequency

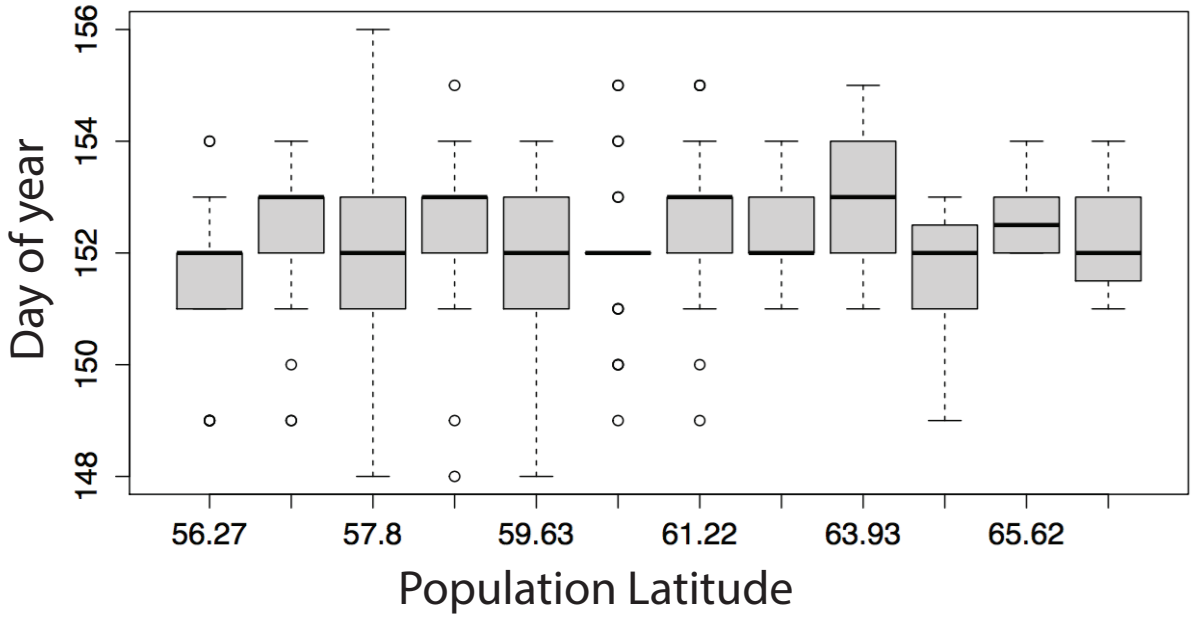

A
Number of eQTL (p < 10^-6)
20000
15000
10000
5000
B
Spearman corr(MAF, |β|)
−0.30
−0.35
−0.40
−0.45
C
Frequency
0.4
0.3
0.2
0.1
[0.1,0.15]
(0.15,0.2]
(0.2,0.25]
(0.25,0.3]
(0.3,0.35]
(0.35,0.4]
(0.4,0.45]
(0.45,0.5]
eSNP MAF
Original
Permuted


Subsampling effect size
Original effect size
4
2
0
−2
−4
−2
0
2
4
6

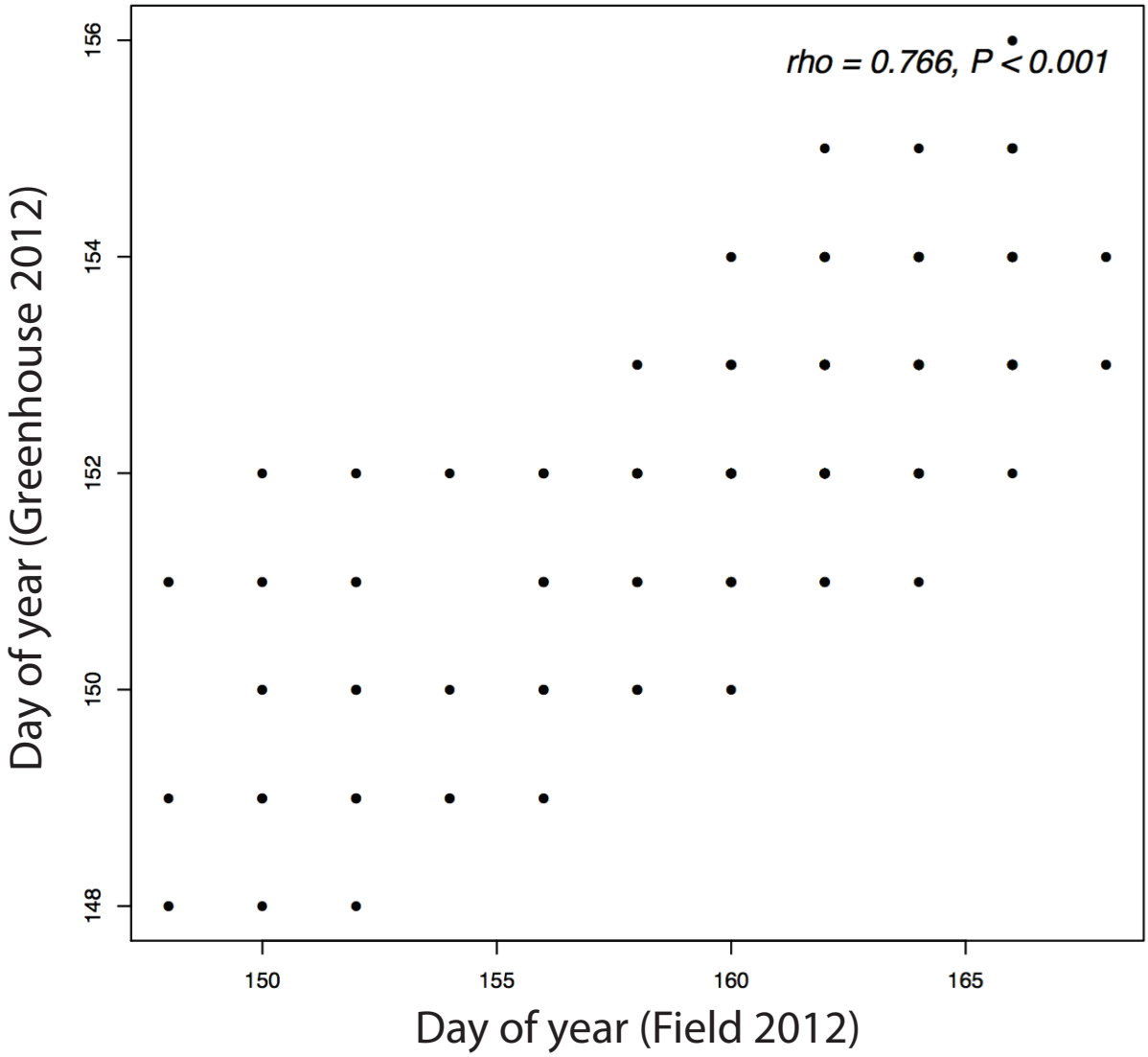

kTotal
P<0.001
r= 0.43
P<0.001
r= 0.34
P<0.001
r= -0.29
P<0.001
r= -0.25
P<0.001
r= -0.22
P<0.001
r= -0.19
P<0.001
r= -0.14
P<0.001
r= -0.16
is_core_gene
P<0.001
r= 0.13
P= 0.002
r= 0.02
P<0.001
r= -0.13
P<0.001
r= -0.083
P<0.001
r= -0.089
P<0.001
r= -0.065
P<0.001
r= -0.059
orig_expr_mean
P<0.001
r= -0.42
P<0.001
r= 0.051
P<0.001
r= -0.15
P<0.001
r= -0.093
P<0.001
r= -0.22
P<0.001
r= -0.29
expr_var
P<0.001
r= 0.23
P<0.001
r= 0.36
P<0.001
r= 0.26
P<0.001
r= 0.14
P<0.001
r= 0.20
is_egene
P<0.001
r= 0.16
P<0.001
r= 0.18
P<0.001
r= 0.056
P<0.001
r= 0.029
tP
P<0.001
r= 0.87
P<0.001
r= 0.031
P<0.001
r= 0.092
tajD
P= 0.2
r= -0.0092
P<0.001
r= 0.072
tP_0_4
P<0.001
r= 0.41
kaks

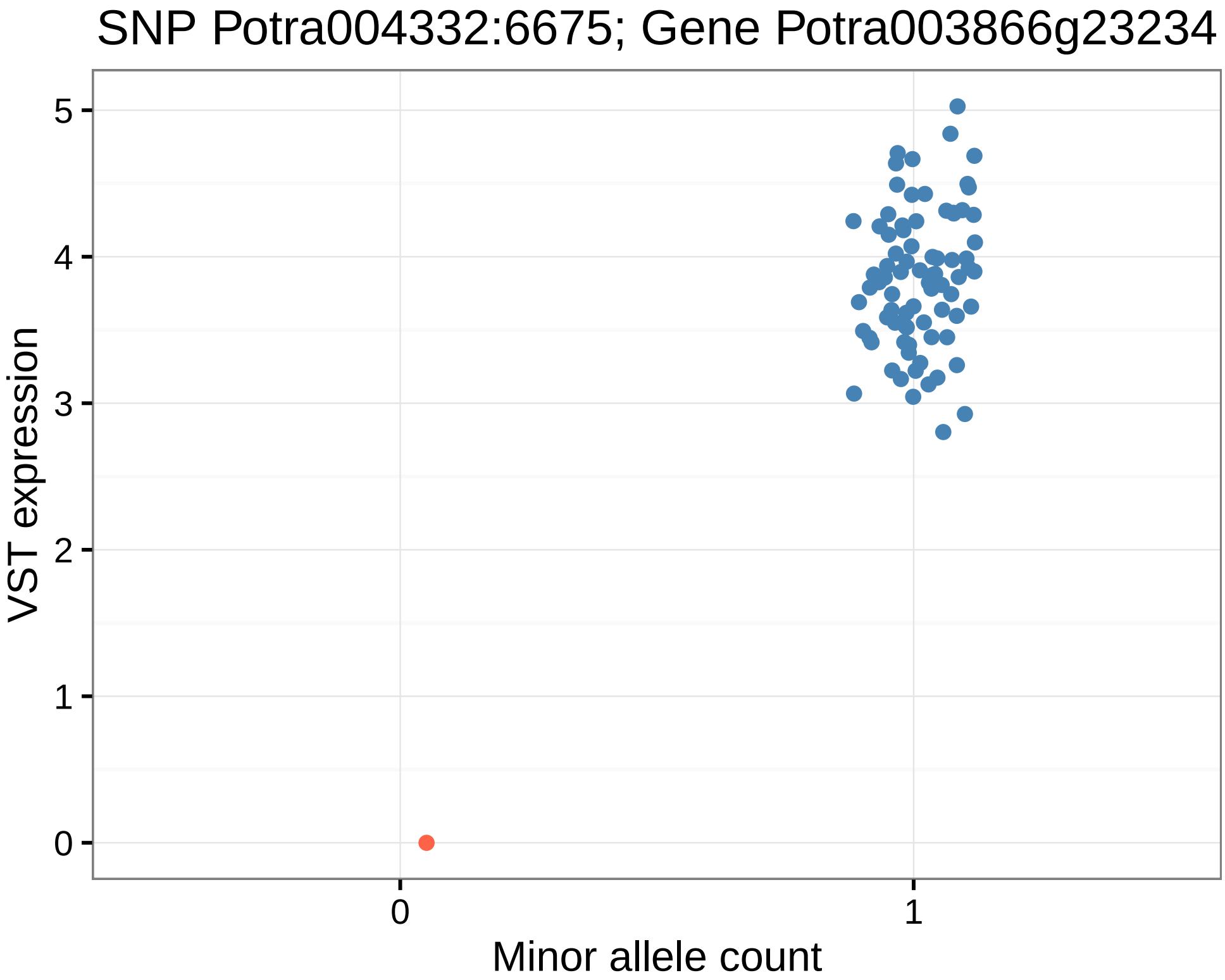

1.2
0.9
0.6
0.3
0.0
Density
0%
100%
200%
300%
Percentage of heritable variance explained by eSNPs
Both
Distant
Local

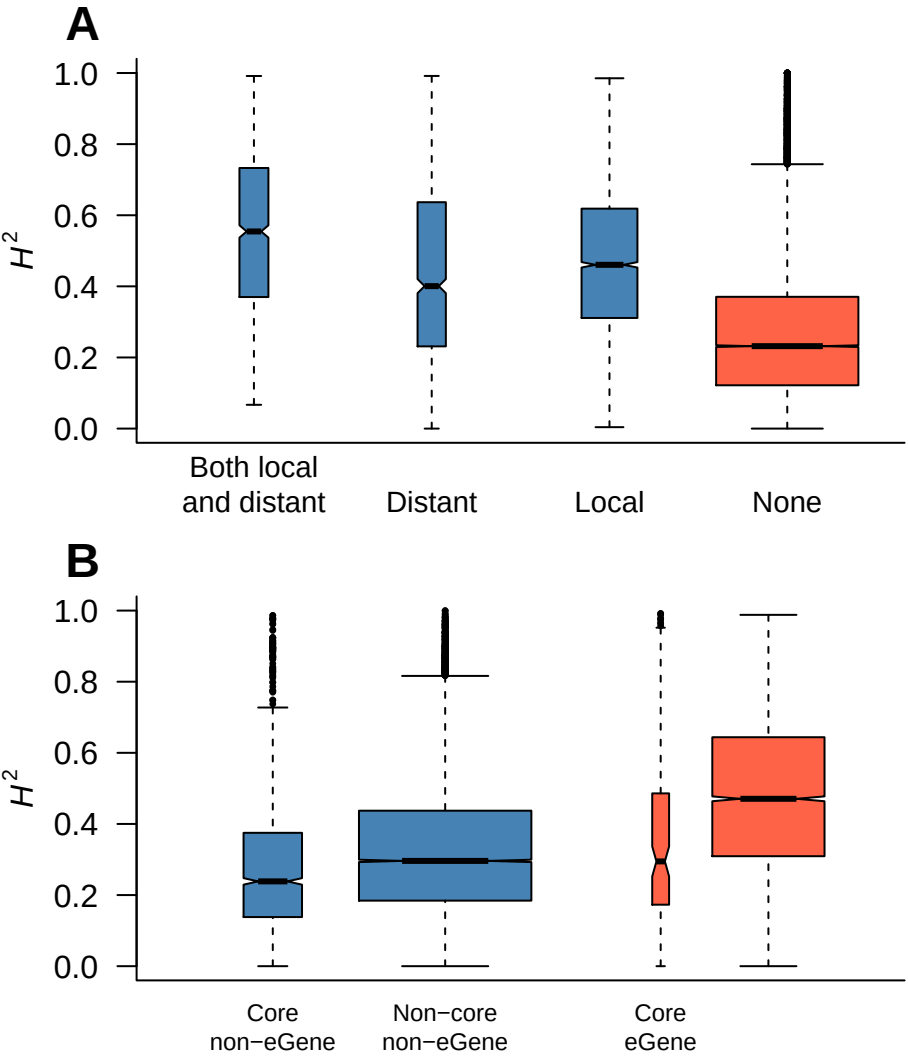
A


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)


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


A
1.0
0.8
0.6
0.4
0.2
0.0
H^2
Both local and distant
Distant
Local
None
B
1.0
0.8
0.6
0.4
0.2
0.0
H^2
Core non−eGene
Non−core non−eGene
Core eGene