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], 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

[*] Corresponding author

E-mail: nathaniel.street@umu.se

[¶] These authors contributed equally to this work

Abstract

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

buds undergoing bud flush in a natural population of *Populus tremula*. We performed expression Quantitative Trait Locus (eQTL) mapping and identified 164,290 significant eQTLs associating 6,241 unique genes (eGenes) with 147,419 unique SNPs (eSNPs). We found approximately four times as many local as distant eQTLs, 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 characteristics, gene expression genetic architecture and population genetics parameters: 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. Module core genes have experienced stronger selective constraint on coding and non-coding sequence, with connectivity being more strongly associated with selection signatures than expression level. 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.

41 While previous studies have examined the genetic architecture of gene expression variation, few were
42 performed in natural populations with fewer still integrating the co-expression network.

43 We undertook a study using a natural population of European aspen (*Populus tremula*), showing that
44 highly connected genes within the co-expression network are associated with polymorphisms of
45 lower than average effect size, suggesting purifying selection. Furthermore, the most highly
46 connected genes within co-expression network hubs were underrepresented for identified expression
47 quantitative trait loci, suggesting stabilizing selection. In contrast, genes in the periphery of the
48 network displayed signatures of relaxed selective constraint. Highly connected genes are therefore
49 buffered against large expression modulation, providing a mechanistic explanation of how network
50 robustness is established and maintained at the population level.

51 Introduction

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

62 abundance, metabolites etc.) and 'classical' phenotypes (physiognomy, morphology, physiology etc.)
63 represents an appealing alternative.

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

78 Advances in sequencing technologies have concordantly revolutionised transcriptomics studies,
79 particularly in non-model organisms. Following seminal work [3,4], numerous early studies in a range
80 of species established that there is a significant heritable component underlying natural variation of
81 gene expression levels among individuals within populations [5–17] and that this variation underlies a
82 number of phenotypes [18–25]. Given these findings, it became apparent that gene expression values
83 could be considered in the same way as any other quantitative phenotype and be subjected to linkage
84 or association mapping to identify polymorphisms contributing to expression level variation among

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

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

109 population differentiation are observed at the transcriptome level [53–58], and there is a lack of
110 consensus as to the extent to which selection acting on gene expression underlies adaptive
111 phenotypic trait variation [53]. There has also been little work to date examining eQTLs in the
112 context of the co-expression network that they induce. Co-expression networks are known to be
113 scale-free, having few highly connected nodes (genes) and many nodes with few connections. This
114 property imparts an inherent ability to buffer against single mutations of large negative effect [59–61]
115 as random mutation of an expression pattern (*i.e.* an eQTL) or coding sequence will more often
116 affect a network node of low connectivity. However, there has been little exploration of how
117 network properties relate the selection.

118 Species in the *Populus* genus have been established as a powerful model system for forest tree
119 genomics due to their relatively small genome, rapid growth, propensity for clonal propagation and
120 ease of genetic transformation [62]. *P. tremula* (European aspen) has many features that render it a
121 particularly useful model for population genetics and speciation studies [63,64], studies of which are
122 facilitated by availability of a draft *de novo* genome assembly [65] and population resequencing data
123 [66].

124 Here, we generated population-wide RNA-Seq data in a natural collection of *P. tremula*, assaying gene
125 expression in winter buds at the point of bud break. We performed eQTL mapping and constructed
126 a co-expression network, which was scale free and modular, with highly connected genes in the
127 module cores being under-represented for eQTLs and with eQTL effect size being negatively
128 correlated with gene connectivity. Patterns of polymorphism and divergence within genes in module
129 cores demonstrate that they are likely experiencing stronger selective constraint relative to genes in
130 the network periphery. Our results suggest that purifying selection plays an important role in
131 buffering the transcriptional network against large perturbations and that natural variation in gene

expression is more prevalent in genes of low network connectivity as a result of relaxed selective constraint.

Results

Population level gene expression similarity

We utilised the northern common garden (located at 63.9° N, near Umeå, Sweden) of the Swedish Aspen (SwAsp) collection [67], which comprises 116 *P. tremula* genotypes sampled across 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) [64,68] and minimal population structure [69]. Based on whole genome re-sequencing data aligned to a *de novo* assembly of the *P. tremula* genome (available at PopGenIE.org; [65]), 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, and 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

153 negative correlation with expression variance (Pearson r = -0.02; df = 29,670; $p < 4.5\times10^{-4}$) and a
154 positive correlation with median expression (Pearson r = 0.18; df = 29,670; $p < 2.2\times10^{-16}$). To
155 further examine whether population structure was apparent on the basis of expression variation
156 among genotypes, we performed hierarchical clustering of all individuals (Fig 1D) or genotypes (S1
157 Fig). We observed some evidence of clustering among genotypes (regions of blue in Fig 1D, S1 Fig)
158 but as genotypes (or individuals) did not cluster according to population of origin (as indicated by the
159 y axis color bars), this did not appear to be explained by population structure. We tested whether the
160 clustering could be used to predict the population of origin for genotypes by cutting the dendrogram
161 to produce 12 clusters that were used as the response in a multinomial logistic regression. The mean
162 accuracy for a 10-fold cross validation was 0.09. We performed permutation tests, which showed that
163 the mean Q_{ST} for all genes was significantly lower than expected by chance ($p < 0.001$).

164 We selected the 500 genes with the highest H^2 (0.88-1.0, 0.93 ± 0.03) and 500 genes with the highest
165 Q_{ST} (0.54-1.0, 0.71 ± 0.13) and subjected these to GO enrichment analysis to identify whether these
166 gene sets contained characteristic functional signatures (see S1 File for all results). Genes with high
167 H^2 were enriched for categories including protein phosphorylation (GO:0006468; $p = 3.7\times10^{-6}$),
168 while high Q_{ST} genes were enriched in terms including translation (GO:0006412; $p = 4.2\times10^{-20}$) and
169 gene expression (GO:0010467; $p = 4.6\times10^{-12}$). Likewise, we considered the genes with the lowest
170 values, which revealed enrichment of terms including cell wall modification (GO:0042545; p =
171 2.6×10^{-7}) for the 2,289 genes with an H^2 of zero and enrichment of terms including amino acid
172 activation (GO:0043038; p = 0.0012) among the 11,895 genes with a Q_{ST} of zero.

173 We performed a regression analysis to ascertain whether a set of geographic (latitude, longitude,
174 elevation), climatic (temperature, precipitation) or other (time since sample collection) factors
175 significantly explained the global patterns of gene expression similarity among genotypes (S2 Fig), as

176 identified by performing a PCA of the expression data. None of the gene expression principal
177 components (PCs) were significantly explained by these environmental factors, with the only
178 significant results found between PCs 2, 5 and 7 and the number of hours from collecting branches
179 from the field until bud sampling in the greenhouse for RNA extraction, which explained 6.6%,
180 3.2%, and total 2.1% expression variance, respectively.

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

197 A recent reanalysis [70] of two existing datasets assaying gene expression among natural accessions
198 of *A. thaliana* [71,72] observed that thousands of genes displayed clear present/absent expression

among accessions. In contrast, we did not find any genes displaying such a pattern of expression variation (S3 Fig), an observation that we also confirmed in an independent *P. tremula* dataset [73], albeit containing substantially fewer genotypes.

eQTL mapping

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. This distance threshold was 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*. Prior to mapping eQTLs, hidden confounders were removed from the gene expression data by removing variance attributable to the first nine principal components 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).

In total 164,290 eQTLs were identified at 5% empirical FDR: 131,103 local and 33,187 distant. These eQTLs represented pairwise associations between 6,241 unique genes (eGenes; 28% of genes considered) and 147,419 unique SNPs (eSNPs), with a mean of 21.0 local and 5.3 distant eSNPs per eGene, respectively. 4,091 genes had only local eQTLs, 1,050 had only distant eQTLS while 1,100 had both. The genomic context of eSNPs was determined by intersecting eSNP positions with gene annotations. After normalising for feature length, the majority of local eSNPs were located within untranslated regions (UTRs) and up- or down-stream (regulatory) regions of genes, with distinctly lower representation within exons than introns (Fig 2D). The genomic context distribution of local

222 and distant eSNPs was largely similar, although there were distinctly more eSNPs located within
223 intergenic regions for distant eQTLs. A local eSNP/eQTL can be located within the region (5'/3' 2
224 kbp flanking, 5'/3' UTR, exon, intron) of the associated eGene, within the region of a gene other
225 than the associated eGene or within an intergenic region. We found that approximately half of the
226 local eSNPs were located within the region of the eGene itself and half within another gene, with
227 relatively few local eSNPs located in intergenic regions. There was a clear tendency for a local eSNP
228 to be located proximal to the transcription start site (TSS) or the stop codon (S7 Fig). eGenes had
229 significantly higher heritability than non eGenes (median heritability difference was 0.16, permutation
230 test $p < 0.0001$) (Fig 2B), with this trend being slightly higher for local than distant eQTLs (S8 Fig).
231 eGenes similarly had higher expression variance than non-eGenes, both before (mean variance of
232 0.32 versus 0.24; permutation test $p < 0.0001$) and after (mean variance of 0.18 versus 0.084;
233 permutation test $p < 0.0001$) hidden confounder removal. In particular, eGenes represented 60 of
234 the 75 genes with variance > 2 (hypergeometric test, $p < 2.2\times10^{-16}$) and 178 of the 250 genes with
235 variance > 1 ($p < 2.2\times10^{-16}$) (S9 Fig). Before hidden confounder removal eGenes had marginally
236 higher mean expression than non-eGenes (mean expression 3.5 and 3.3, respectively; permutation
237 test p-value < 0.0001). There were no significant differences after hiden confounder removal,
238 regardless of whether the eQTLwas local or distant. eGenes with at least one local eQTL were
239 enriched for GO categories related to tRNA metabolic process (GO:0006399, $p = 1.5\times10^{-5}$), ncRNA
240 metabolic process (GO:0034660, $p = 2.6\times10^{-5}$) and organonitrogen compound biosynthetic process
241 (GO:1901566, $p = 2.2\times10^{-5}$), among others, while eGenes with at least one distant eQTL were
242 enriched for categories including protein phosphorylation (GO:0006468, $p = 0.0064$; see S1 File for
243 all results).

244 In contrast to a number of previous studies [20,35,37,48,51], we did not find evidence for any
245 distantly acting hotspots (Fig 2C, S10 Fig), which represent loci where large numbers of *trans*-acting
246 variants are co-located. The vast majority of eSNPs were associated with a single eGene (132,258
247 eSNPs) with a maximum of six eGenes associated with a single eSNP (S11A Fig). In contrast, only
248 1,248 of the 6,241 eGenes were associated with a single eSNP, with eGenes associated with up to
249 1,547 eSNPs (S11B Fig). To partially account for linkage, we fitted linear models between the
250 expression of each eGene and all the significant eSNPs for that gene, both local and distant. The use
251 of a linear model masks eSNPs that contain identical/redundant information and thus effectively
252 identifies haplotype blocks present in all individuals (which we refer to as 'unique eSNPs'), while also
253 producing a measure of how well the combination of eSNPs explains the expression of the
254 corresponding eGene (in terms of percentage variance explained, %VE). Of the 4,993 eGenes
255 associated with more than one eSNP, 4,703 were also associated to more than one unique eSNP of
256 which 4,210 genes were associated with at least one local eSNP and 1,203 were associated with at
257 least one distant eSNP. The adjusted %VE for the combination of eQTLs was, in general, higher
258 (mean %VE 51.1) than for single eSNPs (mean %VE 44.3). Local eSNPs explained significantly
259 more of the variance than distant eSNPs (local mean adjusted %VE = 51, distant mean adjusted
260 %VE = 47, Mann-Whitney $p < 2.2\times10^{-16}$, Fig 2A) and also had higher statistical significance (Mann-
261 Whitney $p = 6.9\times10^{-12}$, Fig 2E). As expected, there was a positive correlation between the maximum
262 %VE of the eSNPs associated with an eGene and gene expression H^2 (Pearson r = 0.47, df = 6,232,
263 $p < 2.2\times10^{-16}$). Of note, there were 2,780 non-eGenes with $H^2 > 0.5$, representing a population of
264 genes of relatively high expression heritability but with no identified eSNP associated with that
265 heritable variation.

Co-expression network

We used genotype mean gene expression values, after removal of hidden confounders, to calculate a co-expression network and subsequently considered the network characteristics for genes with and without a mapped eQTL (eGenes). In common with other biological networks, the network was scale-free (R^2 = 0.97). We compared the correlation and variance properties of our dataset to that of the *P. tremula* expression atlas (exAtlas; [65]), 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 (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, where these difference were virtually non-existent (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 $r = -0.08$; $df = 22{,}304$; $p < 2.2\times10^{-16}$; S9 Fig) and eGenes of the highest connectivity had lower effect size (based on the maximum effect size across all eQTLs associated to each eGene, Pearson $r = -0.15$; $df = 6{,}239$; $p < 2.2\times10^{-16}$, S14 Fig). Among the module cores, 28 contained at least one eGene, with 25 module cores being significantly under-represented for eGenes (permutation test, $p < 0.05$). This further emphasised that, in general, eGenes were not central in the network. There was at least one eSNP located within the region of a core gene (5'/3' flanking, 5'/3' UTR, exon, intron) for 32 of the 38 modules, with module cores also being under-

309 represented for eSNPs (permutation test, $p < 0.0001$). For a number of metrics, module 23 had
310 notable differences to the general pattern: It contained few core genes, all of which were eGenes; the
311 core genes had high variance compared to non-core genes; the module in general was enriched for
312 eGenes.

313 Subsequently, we tested whether the periphery of the network was over-represented for eGenes.
314 Sixty-four of 145 peripheral genes were eGenes, representing a significant enrichment (permutation
315 test, $p < 0.0001$). On the other hand, network module cores were enriched for both transcription
316 factors (permutation test, $p < 0.0001$), which had higher connectivity than non-transcription factors
317 (permutation test, $p < 0.0001$), and phylogenetically conserved genes (permutation test, $p < 0.0001$),
318 defined as genes with orthologs in *P. tremuloides*, *P. trichocarpa* and *A. thaliana* (20,318 genes; [74]).
319 Furthermore, *P. tremula*-specific genes, *i.e.* genes without orthologs in *P. tremuloides*, *P. trichocarpa* or *A.*
320 *thaliana* (1,614 genes), were slightly under-represented in network module cores (permutation test, p
321 = 0.009) while being slightly over-represented among eGenes (permutation test, $p = 0.0076$).

322 In addition to eGenes having generally lower connectivity, there was also a negative relationship
323 between eQTL effect size and co-expression connectivity for both local (Pearson $r = -0.15$; df =
324 5,189; $p < 2.2\times10^{-16}$) and distant eQTL (Pearson $r = -0.12$; df = 2,148; $p < 6.2\times10^{-8}$). eGenes had
325 higher heritability than non eGenes regardless of whether they were located within a module core or
326 not (mean difference of 0.10 and 0.15, respectively, permutation $p < 0.0001$; Fig 3C), and identified
327 eSNPs explained a substantial proportion of the heritable variance (blue boxes in Fig 3C). However,
328 the heritability of eGenes within the core was lower than for eGenes outside the core (mean
329 difference of 0.10, permutation test $p < 0.0001$, Fig 3C) and H^2 correlated negatively with
330 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. Module cores were underrepresented for eGenes, which is apparent by the low proportion of eGenes in cores compared to non-cores (Fig 3D). 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

In *P. tremula* 3,910 paralog pairs were detected [74], 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 $\hat{\beta} = -10.52$, $p < 2.2\times10^{-16}$). Excluding paralogs did not alter the fact that eGenes had significantly lower network connectivity than non-eGenes.

Gene expression evolution is associated with rate of sequence evolution

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,

353 irrespective of whether eQTLs were local or distant, eGenes exhibited significantly higher genetic
354 variation than non-eGenes (Fig 5A). Furthermore, although Tajima's D values are overall negative in
355 *P. tremula,* likely reflecting a historical range expansion [64], eGenes exhibited significantly higher
356 Tajima's D values compared to non-eGenes (Fig 5C), again regardless of whether the associated
357 eQTL was local or distant. These patterns were consistent for the different genetic contexts
358 considered, most likely reflecting local patterns of linkage (S15 Fig). One explanation for the lower
359 diversity at non-eGenes is that existing variation is segregating at lower frequencies and that non-
360 eGenes are experiencing stronger selective constraint compared to eGenes. To test this hypothesis
361 we looked for differences in selection efficacy on protein sequence evolution between eGenes and
362 non-eGenes by examining the ratios of intraspecific nonsynonymous to synonymous polymorphisms
363 ($\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and interspecific nonsynonymous to synonymous substitutions (d_N/d_S) across these
364 genes (using *P. trichocarpa* as an outgroup). Both $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ and d_N/d_S estimates were significantly
365 lower in non-eGenes compared to eGenes (Fig 5E,G), likely reflecting stronger purifying selection
366 acting on non-eGenes. As has been reported previously for *Capsella grandiflora* [28] and a wild
367 population of baboons [75], we observed a negative correlation between minor allele frequency and
368 eQTL effect size (S16 Fig), however permutation testing revealed that the effect of the winner's
369 curse was more prevalent in the permuted than in the original data (S17 Fig), which is in contrast to
370 the results in *Capsella.*

371 Likewise, we contrasted patterns of sequence diversity between genes located in module cores and
372 those not (core vs. non-core). Independent of whether a gene was an eGene or not, core genes had
373 significantly lower levels of genetic diversity and Tajima's D values (Fig 5B,D). In addition, core-
374 genes also had significantly reduced ratios of non-synonymous to synonymous polymorphisms ($\theta_{0\text{-}}$
375 $_{\text{fold}}/\theta_{4\text{-fold}}$) and substitutions ($d_N/d_S$) (Fig 5F,H). Again, these patterns were consistent across different

376 genomics contexts (S15 Fig). Taken together, these results suggest that genes in network module
377 cores experience reduced rates of molecular evolution due to stronger purifying selection and are
378 evolving under stronger selective constraint compared to non-core genes.

379 In order to examine and account for the non-independence among different measures of gene
380 expression, we performed PCA and regression analyses on representative gene expression measures
381 in addition to exploring the pairwise correlation among these factors (S18 Fig). The PCA considered
382 co-expression network connectivity, core vs. non-core, gene expression level and variance and eGene
383 vs. non-eGene and revealed that PC1, which explained 37.03% of the variation in these five
384 measures, was mainly dominated by the connectivity of genes in the co-expression network and
385 whether they are in network module cores or not (Fig 6A, S19A Fig), with a notably smaller
386 contribution of expression level. Gene expression variance and eGene status showed a strong
387 influence on PC2 (Fig 6A, S19A Fig) while expression level contributed largely to PC3 (Fig 6A, S19B
388 Fig). Logistic regression analyses for eGene status (S3 File) identified H^2, co-expression connectivity,
389 Tajima's D, expression level and core gene status, respectively, as significant predictors. In the case
390 of core gene status, whether a gene had an associated eQTL (eGene status), expression variance,
391 expression level and H^2 were the most predictive variables. Similarly, the regression model for gene
392 expression level identified eGene status, co-expression connectivity, core gene status, nucleotide
393 diversity and Tajima's D as the most predictive factors, respectively (S3 File).

394 We then tested for significant (Spearman's rank) correlations between sequence evolution rates and
395 PCs. The rates of sequence evolution over both short (represented by within-species diversity: θ_π,
396 Tajima's D and $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$) and long (represented by between-species divergence, d_N/d_S) timescales
397 showed significantly negative correlations with the connectivity and core status of genes in the co-
398 expression network (PC1) (Fig 6B-E). This indicates that the connectivity of genes within the

399 network both globally and within the local context of expression modules (core status) are key factor
400 associated with the rates of sequence evolution. PC2, which largely reflected the variance of gene
401 expression, showed significant positive correlations with patterns of genetic diversity within species
402 (θ_π and Tajima's D), and significant, although relatively weak, negative correlation with the rate of
403 protein sequence evolution (Fig 6B-E). In addition, gene expression level (largely represented by
404 PC3) showed significantly negative correlations with the rates of sequence evolution (Fig 6B-E).

405 Discussion

406 There have, to date, been relatively few studies assaying gene expression in natural, outbreeding
407 populations. Here, we used a collection of individuals of the outbreeding forest tree species *Populus*
408 *tremula* sampled across the species range in Sweden (the SwAsp collection), and grown in a clonally
409 replicated common garden experiment, to assay gene expression using RNA-Sequencing. We
410 identified prevalent heritability in gene expression levels, with 17% of the genes (5,924) having H^2 >
411 0.5, which is in line with observations in a number of species [5–17,53]. While many previous eQTL
412 studies were conducted using controlled, inter-specific crosses, these results attest the substantial
413 genetic variation contributing to gene expression in natural populations, highlighting the potential for
414 natural selection to modulate expression levels.

415 We identified more local than distant eQTLs, with local eQTLs explaining significantly more of the
416 variance in gene expression than distant eQTLs (Fig 2A), which is in agreement with previous studies
417 [35,48–51]. Although each eSNP was typically associated with only a single gene, many genes were
418 associated with more than one unique eSNP, with the combined variance explained by these eSNPs
419 being higher than for the individual eSNPs, both for local and distant eQTLs, and even when
420 accounting for potential linkage. For a large proportion of eGenes, the identified set of eSNPs

421 explained a relatively high proportion of the heritable expression variation (median 1.0 and s.d. 2.3;
422 S20 Fig). While it appeared that a single eSNP often accounted for a large proportion of the
423 explained heritable variance, due to linkage it is not possible to determine the relative contribution of
424 each eSNP within a haplotype block, with the apparent contribution of individual eSNPs largely
425 resulting from the order in which they are entered into the statistical model. While eGenes had
426 higher H^2 than non-eGenes, as is expected, there were, nonetheless, 2,780 non-eGenes with H^2 >0.5,
427 potentially reflecting that the expression variance of these genes is controlled by many eSNPs of
428 small effect size that we lacked the power to detect.

429 Local eQTLs were most frequently located in regulatory regions (Fig 2D), with UTRs having the
430 highest density of local eSNPs (~1.5 eSNP per kbp) followed by flanking regions (2 kbp up- and
431 downstream, ~1 per kbp) and introns (~0.75 per kbp), which is consistent with previous findings in
432 natural populations of humans [76,77], *Drosophila* [78] and *Capsella grandiflora* [28]. As UTRs are
433 currently not well annotated in *P. tremula* it should be noted that many SNPs currently classified as
434 being located in flanking regions may actually reside within UTRs. The global distribution of
435 genomic contexts for all investigated SNPs (regardless of whether they were eSNPs or not) was
436 similar to that of both local and distant eSNPs, suggesting no notable ascertainment bias for eSNPs.

437 We used the gene expression values to constructed a co-expression network. Compared to networks
438 more typical of systems biology studies, for example the *P. tremula* exAtlas network [65], where
439 samples originated from different tissues of a single genotype, the pairwise expression correlations
440 underlying our co-expression network were low. Despite this, the network displayed typical
441 characteristics, being scale-free with hubs and distinct modules [79]. The simplest explanation for the
442 presence of natural variation in gene expression would be that each unique expression pattern is
443 caused by a single eSNP: In network modules, where many genes share the same expression pattern,

444 a single eSNP modulating the expression of a central regulator would be sufficient to induce similar
445 expression variation (i.e. co-expression) in the set of connected genes within that network module. In
446 our data, we observed at least one eGene in 28 of the 38 module cores and one eSNP in 32 module
447 cores, thus potentially accounting for the co-expression structure of a majority of the modules. In
448 contrast, genes located in the network periphery have low connectivity and relatively unique
449 expression profiles. An eSNP modulating expression of such a gene would likely affect only the
450 peripheral eGene that it is associated with. Consistent with this, genes in module cores (the set of the
451 most highly connected genes within a module) were under-represented for eGenes and, respectively,
452 those in the network periphery were enriched for eGenes (Fig 3B). Furthermore, eGene connectivity
453 was negatively correlated with variance explained by the associated eSNPs (S14 Fig). It has similarly
454 been reported that hubs in human protein-protein interaction networks are less likely to be
455 associated with a detectable eQTL and that the effect size of eQTLs is negatively correlated with
456 connectivity in the protein-protein interaction network [80].

457 The Salicaceae lineage underwent a relatively recent (58 million years ago) whole-genome duplication
458 shared by all member species and that remains represented by a large number of paralogous gene
459 pairs in the genomes of *Populus* species [81]. If many of these duplicated genes are functionally
460 redundant or in the process of diverging, one would expect them to be overrepresented for eGenes
461 as sub- or neo-functionalisation requires derived SNPs to drive expression or coding divergence.
462 However, we saw an under-representation of eGenes in paralog pairs, suggesting that the SNPs that
463 initially induced expression divergence have reached fixation. Interestingly, we found progressively
464 lower expression correlation between paralog pairs containing zero, one or two eGenes (Fig 4),
465 indicating that a subset of paralogs are still undergoing sub- or neo-functionalisation, with their
466 associated eSNPs driving expression divergence.

467 A primary aim of our study was to integrate the results of eQTL mapping within the context of the
468 corresponding gene co-expression network. Regulatory networks are known to be redundant, scale-
469 free and, therefore, to be inherently buffered against single mutations of large negative effect [59–61].
470 In our case, genes with the potential to cause such large negative effects would be observed as hubs
471 in the co-expression network globally or as core genes within modules at the local scale. We would
472 therefore expect hub/core genes to be buffered against single polymorphisms of large effect size
473 and, in particular, the persistence of large negative effect mutations. As such, we could expect the
474 expression of highly connected genes to be subjected to stabilizing selection and the associated
475 eSNPs to be under purifying selection. In support of these expectations, our results indicate that
476 connectivity within the co-expression network is associated with the rate of sequence evolution and
477 that connectivity and core status are more influential than expression level (Fig 6, S3 File).
478 Connectivity and core status contributed primarily to independent PCs than expression level (Fig
479 6A), with connectivity being correlated negatively with signatures of sequence evolution at both
480 short (Fig 6B-C) and long (Fig 6D-E) timescales. In contrast, expression level exhibited low
481 correlation to sequence evolution at short timescales, but higher negative correlations over long
482 timescales (Fig 6B-E). Genes with high co-expression connectivity, mostly those within network
483 module cores, have significantly reduced genetic diversity and exhibit reduced rates of protein
484 evolution compared to more peripheral genes (Fig 5). This suggests that core genes, which represent
485 genes of high potential effect, have evolved under stronger evolutionary constraint than genes in the
486 periphery of the network. As genes in module cores generally have higher levels of pleiotropy and
487 lower levels of dispensability [59], they are consequently more constrained against changes in both
488 gene expression and protein sequence compared to genes in the periphery of the network [82].
489 Furthermore, we show that eGenes are overrepresented among non-core genes of the network and
490 that they have experienced lower levels of purifying selection relative to non-eGenes, corroborating a

491 relaxation of both expression and coding constraints for these genes. Relaxed selection of peripheral
492 eGenes is thus expected to result in an accumulation of weakly deleterious mutations that will
493 segregate as intra-specific polymorphisms. Consistent with these observations, we found that
494 conserved genes (*i.e.* old genes) were significantly over-represented in network cores while *P. tremula*
495 specific genes (*i.e.* young genes) were under-represented. This is in agreement with previous studies
496 that have shown that evolutionarily ancient genes tend to be more central in regulatory networks,
497 with increased constraints on expression and fewer associated eQTLs [83,84]. Taken together, these
498 results indicate how the co-expression network can be buffered against large perturbations via
499 constraint of core genes while enabling flexibility and adaptation by tolerating an accumulation of
500 mutations within the network periphery.

501 Conclusion

502 We identified substantial heritability of gene expression within a natural, out-breeding population, of
503 *P. tremula*. Polymorphisms associated with expression variance were most frequently located within
504 regulatory regions, suggesting that they act by inducing expression variance. The gene co-expression
505 network displayed typical characteristics, being scale-free and containing a small number of highly
506 connected genes, which were underrepresented by eQTLs and with a negative correlation between
507 eQTL effect size and network connectivity. Integration of the eQTL and population genetics
508 analyses with characteristics of the associated gene co-expression network highlight that the context
509 of a gene within the co-expression network appears to be an important determinant of the
510 evolutionary dynamics of transcribed loci. Our results point towards stronger selection acting on
511 network core genes compared to genes in the periphery of the co-expression network, with a
512 negative correlation between rates of sequence evolution and gene connectivity, suggesting that the
513 majority of segregating variation associated with gene expression is under purifying selection.

Materials and Methods

Samples

We collected branches form the SwAsp collection common garden in north of Sweden on 27th May 2012, before natural bud break but as close to the point of natural spring bud break as possible. Branches were placed in the greenhouse facility at the Umeå Plant Science Centre under conditions selected to induce rapid bud break (24 h light, temperature of 20 °C and humidity 50-70 %). At a defined point of emergence (S21 Fig) buds were harvested, flash frozen in liquid nitrogen and stored at -80 ° C until used for RNA isolation. Only terminal buds were sampled (i.e. no lateral buds were included). The time from the day branches were placed in the greenhouse until bud flush sampling ranged from one day to eight days (S22A Fig) and there was a high, positive correlation to bud flush date recorded in the field for the same year (S22B Fig; $r=0.776$, $p < 2.2\times10^{-16}$). As has previously been reported [85], 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 [86] 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 [86]. 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 [73]. 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, [87]) with accession number ERP014886.

RNA-Seq FASTQ-files were pre-processed and aligned to v1.0 of the *P. tremula* reference genome [74] as in [88]. In short, reads were quality and adapter trimmed using Trimmomatic v0.32 [89], rRNA matching reads were filtered using SortMeRNA v1.9 [90], reads were aligned to the v1.0 *P. tremula* reference genome using STAR 2.4.0f1 [91] and read counts were obtained using htseq-count from HTSeq [92]. FastQC [93] was used to track read quality throughout the process. Normalised gene expression values were obtained by applying a variance stabilising transformation (VST) to the raw counts from HTSeq, as implemented in the DESeq2 R package [94].

Gene expression H^2 and Q_{ST}

We calculated repeatability as an assumed upper bound estimate of broad sense heritability of gene expression (see [95] for discussion) from the variance estimates in our data according to the equation

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

555 where V_G is the genetic component of the variance calculated as the expression variance between
556 genotypes for a particular gene (*i.e.* variance among individual means) and V_P is the total phenotypic
557 variance calculated as the sum of V_G and V_E, where V_E is the environment component of the
558 variance calculated as the expression variance within genotypes for a particular gene (i.e. the mean
559 variance among clonal replicates). Point estimates of H^2 were obtained using the repeatability
560 function from the heritability v1.1 R package [96]. The significance of the broad sense heritability
561 was estimated by using an empirical null model where heritabilities were based on random genotype
562 assignments. For each gene, 1,000 permutations were performed and empirical p-values were
563 calculated using the empPvals function in the qvalue 2.6.0 R package.

564 Population differentiation (Q_{ST}; [97]) was calculated as

565 $$Q_{ST} = \frac{V_{between}}{V_{between} + 2V_{genetic}}$$

566 where $V_{between}$ is the variance among populations and $V_{genetic}$ is the genetic variance among genotypes
567 as computed using the lmer function from the lme4 v1.1.12 R package [98] using the formula

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

569 where *expression* is the expression of a gene, *population* is a factor representing the population of each
570 sample, and *clone* is a factor representing genotype replicates. As we use repeatability as an upper
571 bound estimate of H^2, our Q_{ST} estimates are conservative [99]. Similar to the broad sense heritability,
572 the significance of the Q_{ST} was estimated using permutation testing where genotype labels were
573 shuffled among the populations 1,000 times. Due to the characteristics of the data and the method,
574 many of the permutations resulted in undefined QST estimates (an average of 511 missing values per
575 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 [77,100,101]. The number of components to remove was determined by running the eQTL mapping with 0 to 20 PCs removed and selecting the number of components that yielded the largest number of significant eQTLs (Benjamini-Hochberg $p < 0.05$) (S5 Fig). This approach was based on the assumption that the number of identified eQTLs would increase if the removed PCs were removing unwanted, systematic variation (i.e. noise) rather than informative biological variation [77,100–103].

eQTL mapping

eQTL mapping was performed by associating genotype mean gene expression levels with biallelic SNPs using the R package Matrix eQTL v2.1.1 [104]. The corresponding raw resequencing data for all SwAsp genotypes is available at the NCBI SRA resource as BioProject PRJNA297202 (SRA: SRP065057) and the resultant Variant Call Format (VCF) file used for association mapping is available from [66]. Before performing association mapping, genes were filtered on variance so that only genes with a gene expression variance above 0.05 were included. SNPs were also filtered on minor allele frequency (MAF) and major genotype frequency (MGF); any SNPs with MAF < 0.1 or MGF > 0.9 were excluded to avoid spurious associations. A motivating example for the MGF filtering can be seen in S23 Fig. We also generated an LD-trimmed SNP set by removing one SNP from each pair of SNPs with a between SNP correlation coefficient (r^2) >0.2 in blocks of 50 SNPs using PLINK v1.9 [105], yielding 217,489 independent SNPs that were retained for analyses of population structure. The first genotype principal component based on the set of independent SNPs was used as a covariate in the linear model used by Matrix eQTL to account for the weak signature

599 of population structure. Permutation testing was used to determine eQTL significance whereby
600 genotype sample labels were permuted 1,000 times and the maximum absolute t-statistic from Matrix
601 eQTL was recorded for each expressed gene across all SNPs for each permutation, resulting in 1,000
602 random t-statistics being collected for each gene. Empirical p-values were calculated for each eQTL
603 in the observed data using the permuted t-statistics for the observed eGene with the empPvals
604 function in the qvalue v2.4.2 R package [106], and q-values (empirical FDR) were calculated with the
605 qvalue function in the same package.

606 When determining the genomic context of eSNPs, there were some cases where introns overlapped
607 exons as a result of overlapping gene model being present on the same strand. These 41 eSNPs were
608 discarded from the counting. Another type of overlap that was discarded were cases where an eSNP
609 overlapped a gene feature, but no sub-feature inside that gene (e.g. UTR, exon or intron). These
610 1961 eSNPs were excluded from the counting. Since many of the features overlap (e.g. exon and
611 untranslated regions), the priority for counting was untranslated region, exon/intron,
612 upstream/downstream and intergenic.

613 Permutation tests involving eGenes and non-eGenes were performed by shuffling eGene
614 assignments among the 22,306 genes that were considered for eQTL mapping. This was repeated
615 10,000 times and empirical p-values were calculated using the empPvals function in the qvalue v2.6.0
616 R package. The minor allele frequency (MAF) was compared between eSNPs (both local and distant)
617 and all SNPs throughout the genome. To test for the signature of selection on eQTLs, we also
618 estimated the correlation between minor allele frequency and eQTL effect size.

Co-expression

The R package WGCNA v1.51 [107] was used for constructing a co-expression network. The input gene expression values were per-gene genotype means corrected for hidden confounders (the same data used for eQTL mapping). We chose to use the unsigned network type for this study with the motivation that we did not want to discard negative relationships. By looking at this from an eQTL perspective, an eSNP can be positively associated with one gene while negatively associated with another. The relationship between these two genes would be missed if we used the signed or the signed-hybrid approaches. Using the unsigned approach, we assure that genes with strong negative correlation end up in the same network modules. A soft thresholding power of 5 was used to calculate adjacencies. The topological overlap matrix (TOM) was generated using the TOMsimilarity function with the signed approach in order to take negative edges into account (see [107] for details). In order to identify network modules, hierarchical clustering was applied to the TOM dissimilarity matrix (1 – TOM) and the resulting dendrogram was divided into modules using the cutreedynamic function. The connectivity of the network was then defined as the adjacency sum for each node, i.e. the weights of the edges that are connected to this node. This concept was applied to modules as well to obtain measures of intra- and inter-modular connectivity, i.e. the connectivity based on edges connecting the gene with other genes inside the same module, and connectivity based on edges connecting the gene with genes outside of the module.

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

Conserved genes and *P. tremula*-specific genes were inferred from gene family analysis based on

642 protein sequence similarity available on the PopGenIE FTP server [108]. In short, two rounds of
643 TribeMCL were run using BLASTP with an E-value threshold of 10^{-5} where the first round was all-
644 vs-all while the second was all-vs-all inside each gene family.

645 Transcription factors were identified by BLASTing the the *P. tremula* protein sequences against the
646 protein sequences of transcription factors in *P. trichocarpa* with an E-value threshold of 10^{-10} and
647 taking the best hit for each *P. trichocarapa* transcription factor. The *P. trichocarpa* transcription factor
648 annotations were taken from Plant Transcription Factor Database v3.0 [109].

649 All permutation tests based on the network were performed by computing the measure of interest on
650 the original network, and then performing the same computation on 10,000 networks were the gene
651 labels were shuffled. Empirical p-values were calculated using the empPvals function in the qvalue
652 v2.6.0 R package [106].

653 Population genetic analysis

654 To estimate nucleotide diversity (θ_π) and Tajima's D [110] among genes, we only used the reads with
655 mapping quality above 30 and the bases with quality score higher than 20 for the estimation of θ_π
656 and Tajima's D by ANGSD [111]. Genes with less than 50 covered sites left from previous quality
657 filtering steps (Wang et al. In prep) were excluded. In addition, θ_π and Tajima's D were also
658 calculated for seven site categories of each gene: 1kbp upstream of genes, 1kbp downstream of
659 genes, 3' UTR, 5' UTR, intron, 0-fold non-synonymous and 4-fold synonymous sites within genes.
660 We further compared the estimates of $\theta_{0\text{-fold}}/\theta_{4\text{-fold}}$ between different categories of genes. We used the
661 transcript with the highest content of protein-coding sites to categorize the seven genomic features
662 within each gene. Finally, we estimated the ratio of non-synonymous substitution and synonymous
663 substitution rate (d_N/d_S) using gKaKs pipeline [112] with codeml method [113] for a total of 33,039

664 orthologous genes that were determined by best BLASTp sequence homology searches between *P.*
665 *tremula* and *P. trichocarpa* [74,108]. Significance for the above statistical measurements between each
666 pair of gene categories was evaluated using Mann-Whitney tests.

667 Acknowledgements

668 We thank Dr Kathryn Robinson for her help sampling material, for providing environmental data
669 and for informative discussion. The authors also would like to acknowledge support from the
670 Science for Life Laboratory, the National Genomics Infrastructure (NGI), and the Uppsala
671 Multidisciplinary Center for Advanced Computational Science (UPPMAX) for providing assistance
672 in massive parallel sequencing and computational infrastructure.

673 References

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

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

678 3. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, et al. Regional and
679 strain-specific gene expression mapping in the adult mouse brain. Proc Natl Acad Sci U S A.
680 2000;97: 11038–43. doi:10.1073/pnas.97.20.11038

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

683 5. Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G. The contributions
684 of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat Genet.
685 2001;29: 389–95. doi:10.1038/ng766

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

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

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

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

698 10. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping
699 determinants of human gene expression by regional and genome-wide association. Nature.
700 2005;437: 1365–9. doi:10.1038/nature04244

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

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

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

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

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

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

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

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

722 19. Carroll SB. Endless forms: the evolution of gene regulation and morphological diversity. Cell.
723 2000;101: 577–580. doi:10.1016/S0092-8674(00)80868-5

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

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

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

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

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

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

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

739 26. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet.
740 2001;17: 388–91. doi:10.1016/S0168-9525(01)02310-1

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

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

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

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

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

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

755 33. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet. 2007;8:
756 206–16. doi:10.1038/nrg2063

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

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

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

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

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

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

776 40. Swanson-Wagner R a, DeCook R, Jia Y, Bancroft T, Ji T, Zhao X, et al. Paternal dominance
777 of trans-eQTL influences gene expression patterns in maize hybrids. Science. 2009;326: 1118–
778 1120. doi:10.1126/science.1178294

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

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

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

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

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

790 45. Kirst M, Basten CJ, Myburg AA, Zeng ZB, Sederoff RR. Genetic Architecture of Transcript-
791 Level Variation in Differentiating Xylem of a Eucalyptus Hybrid. Genetics. 2005;169: 2295–
792 2303. Available: http://www.genetics.org/cgi/content/abstract/169/4/2295

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

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

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

801 49. Hughes KA, Ayroles JF, Reedy MM, Drnevich JM, Rowe KC, Ruedi EA, et al. Segregating
802 Variation in the Transcriptome: Cis Regulation and Additivity of Effects. Genetics. 2006;173:
803 1347–1355. doi:10.1534/genetics.105.051474

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

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

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

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

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

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

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

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

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

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

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

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

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

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

838 62. Jansson S, Douglas CJ. *Populus*: A Model System for Plant Biology. Annu Rev Plant Biol.
839 2007;58: 435–458. doi:10.1146/annurev.arplant.58.032806.103956

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

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

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

849 66.

850 ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/VCF/SwAsp_94samples.filte

851 r.maf005.beagle.maf.recode.vcf.gz.

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

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

858 69. Hall D, Luquez V, Garcia VM, St Onge KR, Jansson S, Ingvarsson PK. Adaptive population
859 differentiation in phenology across a latitudinal gradient in European aspen (*Populus tremula*,
860 L.): a comparison of neutral markers, candidate genes and phenotypic traits. Evolution.
861 2007;61: 2849–60. doi:10.1111/j.1558-5646.2007.00230.x

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

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

867 72. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Paolo Casale F, et al. DNA
868 methylation in Arabidopsis has a genetic basis and shows evidence of local adaptation. Elife.
869 2015;4: e05255. doi:10.7554/eLife.05255

870 73. Robinson KM, Delhomme N, Mähler N, Schiffthaler B, Onskog J, Albrectsen BR, et al.
871 *Populus tremula* (European aspen) shows no evidence of sexual dimorphism. BMC Plant Biol.

872 2014;14: 276. doi:10.1186/s12870-014-0276-5

873 74. Sjödin A, Street NR, Sandberg G, Gustafsson P, Jansson S. The *Populus* Genome Integrative
874 Explorer (PopGenIE): a new resource for exploring the *Populus* genome. New Phytol.
875 2009;182: 1013–1025. doi:10.1111/j.1469-8137.2009.02807.x

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

878 76. Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, et al. Characterizing
879 the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.
880 Genome Res. 2014;24: 14–24. doi:10.1101/gr.155192.113

881 77. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding
882 mechanisms underlying human gene expression variation with RNA sequencing. Nature.
883 2010;464: 768–72. doi:10.1038/nature08872

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

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

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

892 81. Dai X, Hu Q, Cai Q, Feng K, Ye N, Tuskan GA, et al. The willow genome and divergent

893 evolution from poplar after the common genome duplication. Cell Res. Shanghai Institutes
894 for Biological Sciences, Chinese Academy of Sciences; 2014;24: 1274–1277.
895 doi:10.1038/cr.2014.83

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

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

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

905 85. Robinson KM, Ingvarsson PK, Jansson S, Albrectsen BR. Genetic variation in functional
906 traits influences arthropod community composition in aspen (*Populus tremula* L.). PLoS One.
907 2012;7: e37679. doi:10.1371/journal.pone.0037679

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

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

912 88. Delhomme N, Mähler N, Schiffthaler B, Sundell D, Mannepperuma C, Hvidsten TR, et al.
913 Guidelines for RNA-Seq data analysis. Epigenesys. 2014; Available:

914 http://www.epigenesys.eu/en/protocols/bio-informatics/1283-guidelines-for-rna-seq-data-
915 analysis

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

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

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

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

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

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

930 95. Dohm MR. Repeatability estimates do not always set an upper limit to heritability. Funct Ecol.
931 2002;16: 273–280. doi:10.1046/j.1365-2435.2002.00621.x

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

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

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

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

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

945 101. Mostafavi S, Battle A, Zhu X, Urban AE, Levinson D, Montgomery SB, et al. Normalizing
946 RNA-sequencing data by modeling hidden covariates with prior knowledge. PLoS One.
947 2013;8: e68141. doi:10.1371/journal.pone.0068141

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

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

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

955 105. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool

956 set for whole-genome association and population-based linkage analyses. Am J Hum Genet.
957 2007;81: 559–575. doi:10.1086/519795

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

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

962 108. ftp://plantgenie.org/Data/PopGenIE/Populus_tremula/v1.1/Annotation/DoubleMCL-tab-
963 annotation.txt.gz.

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

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

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

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

973 113. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24:
974 1586–1591. doi:10.1093/molbev/msm088

975

976 # Figures

977 **Fig 1. Population location and gene expression overview.** (A) Map of the original locations of
978 the SwAsp populations. The red arrow points to the location of the common garden used in this
979 study. (B) Distribution of gene expression heritability for all genes and for the subset of genes after
980 filtering to remove uninformative expression. (C) Distribution of gene expression Q_{ST} for all genes
981 and for the subset of genes after filtering to remove uninformative expression. (D) Sample clustering
982 based on all samples, including biological replicates. The heatmap represents the sample correlation
983 matrix based on the 500 genes with the highest expression variance. Darker colour indicates higher
984 correlation. The coloured bar represents the populations the samples belong to. The small clusters
985 on the diagonal correspond to biological replicates of each genotype.

986 **Fig 2. eQTL overview.** (A) Expression variance explained *(R^2)* for local and distant eQTLs. Box
987 plots show the maximum variance explained by a single eQTL for each gene and the total variance
988 explained by all eQTLs for each gene. The widths of the boxes are proportional to the number of
989 genes represented. The pairwise significiance of a Mann-Whitney test is indicated by asterisks: ***$p <$
990 0.001, **$p < 0.01$. (B) Genomic context of local and distant eSNPs. Context categories are
991 normalized for feature length. When an eSNP overlapped with features on both strands, both of
992 them were counted. For both local and distant eQTLs the features are based on the gene that is
993 closest to the eSNP, and furthermore, for local eQTLs, the features are divided into whether the
994 eSNP is located in or near the same gene that it is associated to ("associated gene") or not ("other
995 gene"). Flanking regions represent 2 kbp upstream and downstream from the gene. (C) Scatter plot
996 showing the positions of all significant eQTLs in this study. No evidence of eQTL hotspots can be
997 observed. Numbers indicate chromosome. (D) Broad sense heritability distributions for eGenes and

998 non-eGenes. (E) Manhattan plots for local eQTLs (upper) and distant eQTLs (lower). Each point
999 represents an eQTL.

1000 **Fig 3. Co-expression characteristics of eGenes and eSNPs.** (A) Gene expression variance
1001 distribution for all genes in the SwAsp data (before removal of hidden confounders) and the exAtlas
1002 data. (B) Distribution of co-expression connectivity for eGenes and non-eGenes. (C) Distributions
1003 of the proportion of total variance explained and heritability for eGenes and non-eGenes divided
1004 into core and non-core genes. (D) Genes having distantly, locally, or both distantly and locally acting
1005 eSNPs located within 2 kbp (or inside) the gene divided into core and non-core genes. (E) Genomic
1006 context of distantly, locally, or both distantly and locally acting eSNPs located within 2 kbp of an
1007 eGene divided into core and non-core eGenes. The eSNP counts are normalised for total feature
1008 length.

1009 **Fig 4. Correlation within paralog pairs as a function of the number of eGenes in the paralog**
1010 **pair.** The widths of the boxes are proportional to the number of genes in each set. The mean
1011 correlations for paralog pairs with 0, 1, or 2 eGenes were 0.17, 0.10, and 0.06, respectively. The
1012 pairwise significance of a Mann-Whitney test is indicated by asterisks: ***$p < 0.001$, **$p < 0.01$.

1013 **Fig 5. Measures of sequence diversity and divergence.** Nucleotide diversity (A,B), Tajima's D
1014 (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
1015 only distant eQTLs) and non-eGenes, as well as core and non-core genes from the gene expression
1016 network. Significance between each pair of gene categories was evaluated using Mann-Whitney tests
1017 and significance is indicated by asterisks above the boxplot: ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$, n.s.
1018 $p > 0.05$.

1019 **Fig 6. Associations between metrics of gene expression and sequence evolution in *Populus***
1020 ***tremula*.** (A) Percentage variance explained by five principal components comprising co-expression
1021 network connectivity, genes in network cores or not (Core vs. non-core genes), gene expression
1022 levels, gene expression variance and genes with eQTLs or not (eGenes vs. non-eGenes). Colour
1023 shadings depict the proportion contribution of each gene expression measure to each principal
1024 component. (B-E) Spearman's rank correlations between PCs and four metrics of sequence
1025 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
1026 shows the relative contribution of each gene expression measure to each PC. Asterisks indicate
1027 significant correlations, ***P < 1e-5, **P < 0.001,*P<0.05, ns=not significant (P >0.05).

1028 Supporting information

1029 **S1 File. GO enrichment analysis results.** Contains GO enrichment results for local and distant
1030 eGenes, network module cores, and subsets of genes with extreme H^2 and Q_{ST} values.

1031 **S2 File. Various gene statistics.** Contains statistics for all genes, as well as all genes included in the
1032 generation of the co-expression network and the eQTL mapping.

1033 **S3 File. HTML document outlining the regression models constructed.** Regression models
1034 were constructed in R using a number of starting variables. Uninformative variables were then
1035 removed by backward elimination using the stepAIC function from the MASS package resulting in a
1036 reduced model.

1037 **S1 Fig. Clustering of genotypes.** (Left) Heatmap of the sample correlation matrix based on the 500
1038 most variably expressed genes. Darker colour indicates higher correlation. The coloured bar
1039 represents the populations the genotypes belong to. (Right) The two first principal components from

1040 a principal component analysis (PCA) based on all genes. Again, colours represent the genotype
1041 population. The percentages in the axis labels indicate the amount of variance explained by each
1042 component.

1043 **S2 Fig. Correlations between gene expression principal components (PCs) and**
1044 **environmental variables.** The values in each tile represent the Pearson correlation between the
1045 gene expression PC (x-axis) and environmental variable (y-axis). Small asterisks represent a nominal
1046 p-value < 0.05 while large asterisks represent Benjamini-Hochberg (BH) adjusted p-values < 0.05.
1047 The only factor with significant correlations to expression PCs was "Hours to harvest", which is the
1048 number of hours into the sampling period that the buds were harvested. It was significantly
1049 associated with PC4 (BH-adjusted $p = 4.6\times10^{-6}$), PC7 (BH-adjusted $p = 0.030$) and PC2 (BH-
1050 adjusted $p = 0.033$).

1051 **S3 Fig. Relationship between gene expression ranks and the number of genotypes with**
1052 **detectable expression of the genes.** The number of genotypes that a gene was expressed in was
1053 determined by counting the number of genotypes with non-zero expression for each gene. The gene
1054 expression ranks were calculated by ranking the mean gene expression values where the mean was
1055 calculated only considering samples with non-zero expression.

1056 **S4 Fig. Physical distance between eSNPs and their associated genes for each of the 19**
1057 **chromosomes.**

1058 **S5 Fig. The number of eQTLs detected for different levels of hidden confounder removal.**
1059 The y-axis shows the number of eQTLs detected at 5% FDR (prior to the empirical FDR
1060 calculation). The x-axis shows the number of principal components regressed out of the gene

1061 expression data. Distant eQTLs are defined as eQTLs where the gene and the SNP are located on
1062 different scaffolds.

1063 **S6 Fig. Heatmap representing the sample clustering based on gene expression values after**
1064 **hidden confounder removal.** The 500 most variable genes in the original expression data were used
1065 for calculating the sample correlations (i.e. the same genes as in Fig 1D). The colour bar represents
1066 the population of the samples with the same colour scheme as in Fig 1D. The small clusters on the
1067 diagonal represent genotype replicates.

1068 **S7 Fig. Distribution of local eSNPs relative to the gene that they are associated to.** Up and
1069 downstream regions were divided into 1 kbp bins while the intergenic region was normalised to 20
1070 bins. In other words, up and downstream counts are not directly comparable to intragenic counts.

1071 **S8 Fig. Heritability distributions for four non-overlapping sets of genes:** genes with both local
1072 and distant eQTLs, genes with only distant eQTLs, genes with only local eQTLs, and genes with no
1073 significant eQTLs. The box widths are proportional to the number of genes in each set.

1074 **S9 Fig. Gene expression variance plotted against co-expression network connectivity.** eGenes
1075 are indicated by red points.

1076 **S10 Fig. Sliding window approach to detect eQTL hotspots.** The upper panel shows the
1077 number of eSNPs in genomic windows of 100 kb for each of the 19 chromosomes. The lower panel
1078 shows the number of unique genes that are associated to each genomic window (nominal eQTL p-
1079 value $< 1\times10^{-6}$).

1080 **S11 Fig. eSNPs per eGene and eGenes per eSNP.** (A) Bar plot of the number of associated
1081 eGenes per eSNP. (B) Histogram of the number of associated eSNPs per eGene. (C) The mean co-

1082 expression for genes associated with the same eSNP. The numbers on the x-axis represents the
1083 number of eGenes that a single eSNP is associated to, and the numbers above the boxes shows the
1084 total number of eSNPs that are represented, and in parentheses are the number of unique genes that
1085 they are associated with.

1086 **S12 Fig. The number of genes and the number of core genes in each network module.** Each
1087 bar is divided into eGene and non-eGene assignments. Module 0 contains genes not assigned to any
1088 module.

1089 **S13 Fig. Gene expression variance in module cores and non-cores for each of the 38**
1090 **modules.**

1091 **S14 Fig. Relationship between total variance explained by eQTL and co-expression**
1092 **connectivity.** The variance explained is the adjusted variance explained based on all eQTL
1093 associated with each gene.

1094 **S15 Fig. Comparisons of nucleotide diversity and Tajima's D for different genomic features.**
1095 Upstream and downstream are 1 kbp away from the gene start and end, respectively. 0-fold and 4-
1096 fold refers to 0-fold and 4-fold degenerate sites, respectively.

1097 **S16 Fig. Minor allele frequency of eSNPs.** (A) Comparison of minor allele frequency between all
1098 SNPs, distant eSNPs and local eSNPs. (B) The relationship between minor allele frequency and
1099 effect size (absolute value of beta) of eQTLs.

1100 **S17 Fig. Relationship between MAF and effect size in permuted eQTLs.** eQTL permutations
1101 were done by selecting 20 random subsets of 150,000 SNPs each. Each of these sets were associated
1102 with gene expression levels, and in addition, 50 permutations were performed for each set by

1103 shuffling the genotype labels of the SNP data, resulting in a total of 1,000 permuted eQTL sets. (A)
1104 The number of SNPs for the nominal p-value threshold 10^-6 (the same initial threshold that was
1105 used for the original associations). (B) The Spearman correlation between minor allele frequency and
1106 absolute effect size for the 20 sets and their corresponding permutations. (C) Allele frequency
1107 spectra for the 20 subsets and their permutations.

1108 **S18 Fig. Association between metrics of gene expression and sequence evolution.** Scatter
1109 plots (lower off-diagnal, the red-to-yellow-to-blue gradient indicates decreased density of observed
1110 events at a give location in the graph) and correlations with probability values (upper off-diagonal,
1111 spearman's rank correlations) for measures of gene expression and sequence evolution: kTotal, the
1112 total (global) connectivity in the network; is_core_gene, gene is part of module core or not;
1113 orig_expr_mean, mean expression before hidden confounder removal; orig_expr_var, expression
1114 variance before hidden confounder removal; is_egene, gene with eQTLs or not; tP, pairwise
1115 nucleotide diversity; tajD, Tajima's D; tP_0_4, the ratio of pairwise nucleotide diversity at zero-fold
1116 non-synonymous and four-fold synonymous sites; kaks, dN/dS.

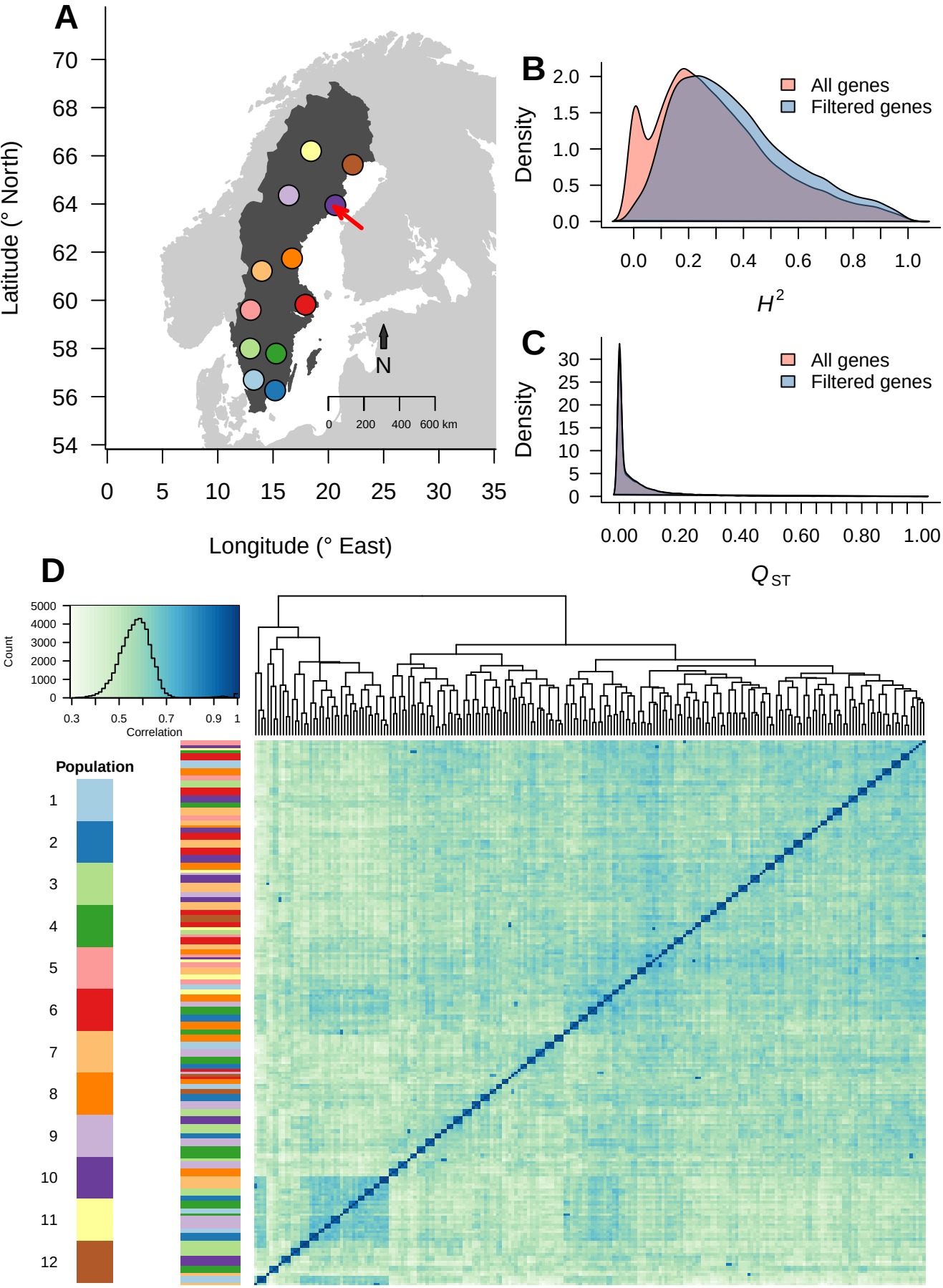
1117 **S19 Fig. Loadings of five gene expression measures on PC1 to PC5.** It comprises co-expression
1118 network connectivity, genes in network cores or not (Core vs. non-core genes), gene expression
1119 levels, gene expression variance and genes with eQTLs or not (eGenes vs. non-eGenes).

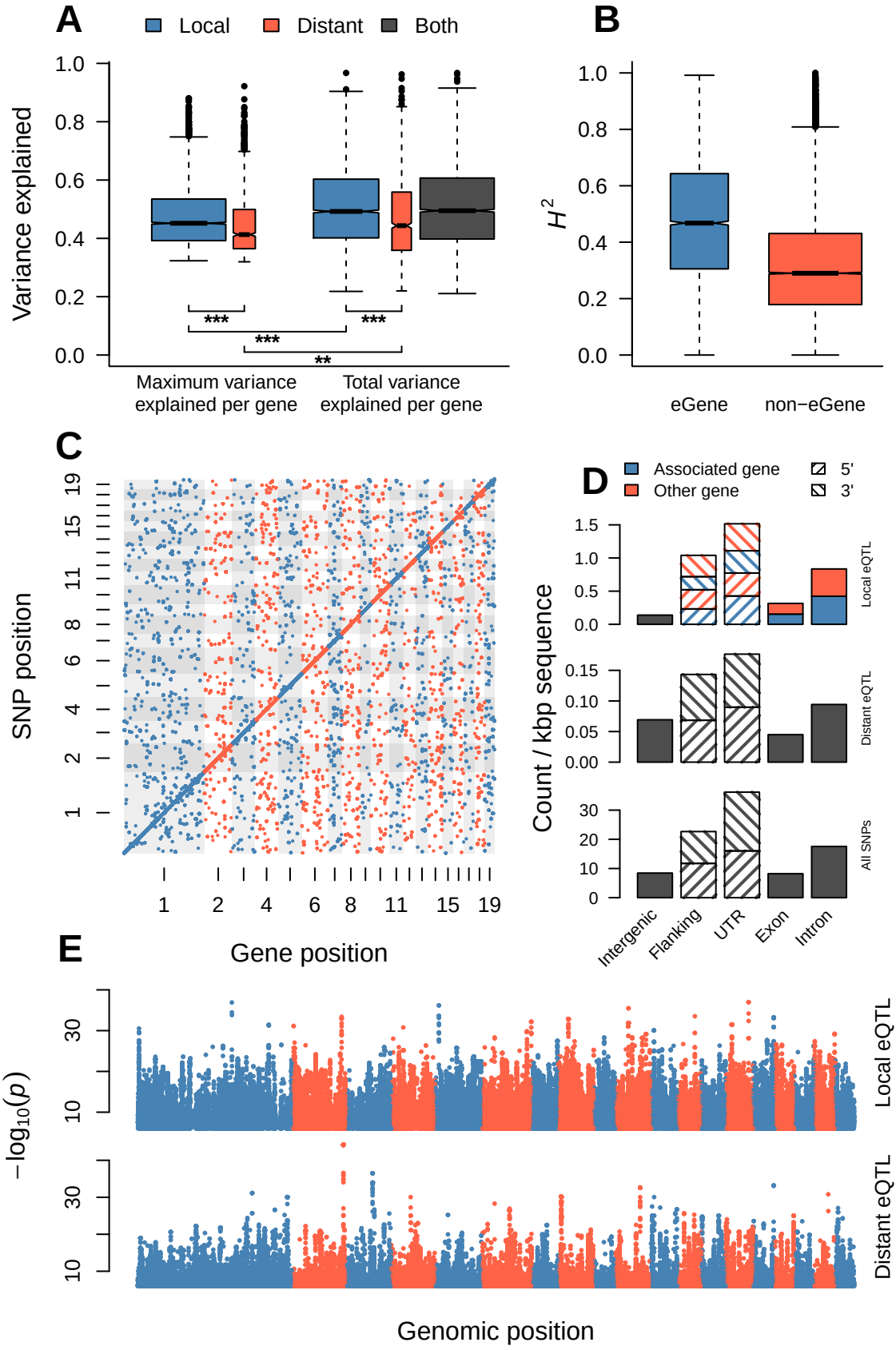
1120 **S20 Fig. Variance explained by eSNPs and heritability.** This figure assumes that eSNPs only
1121 account for heritable variation, explaining why some of the eSNPs explain more than 100% of the
1122 heritability. The x-axis has been truncated at 300%.

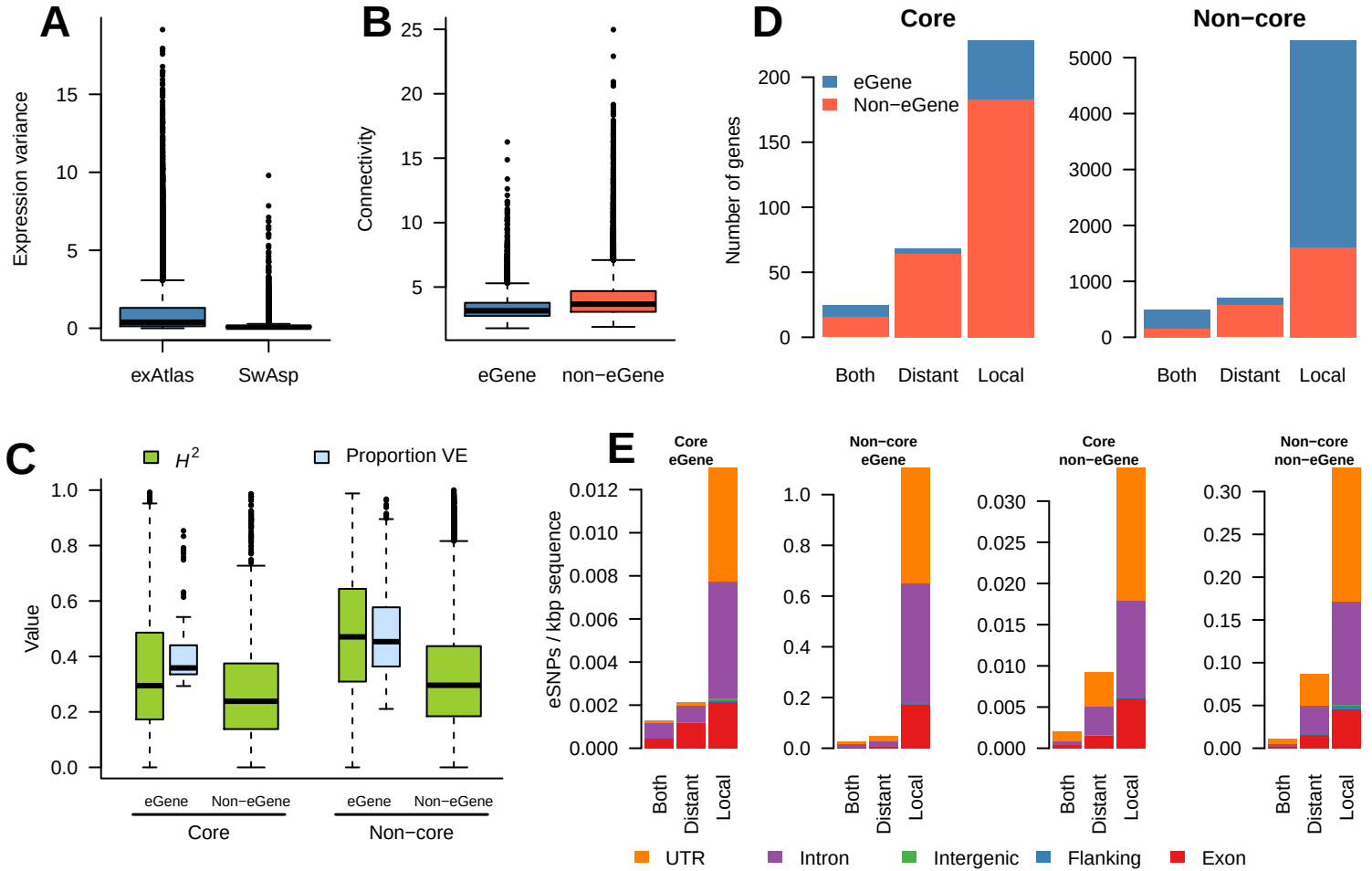
1123 **S21 Fig. Representative photo of the sampled bud flush stage.**

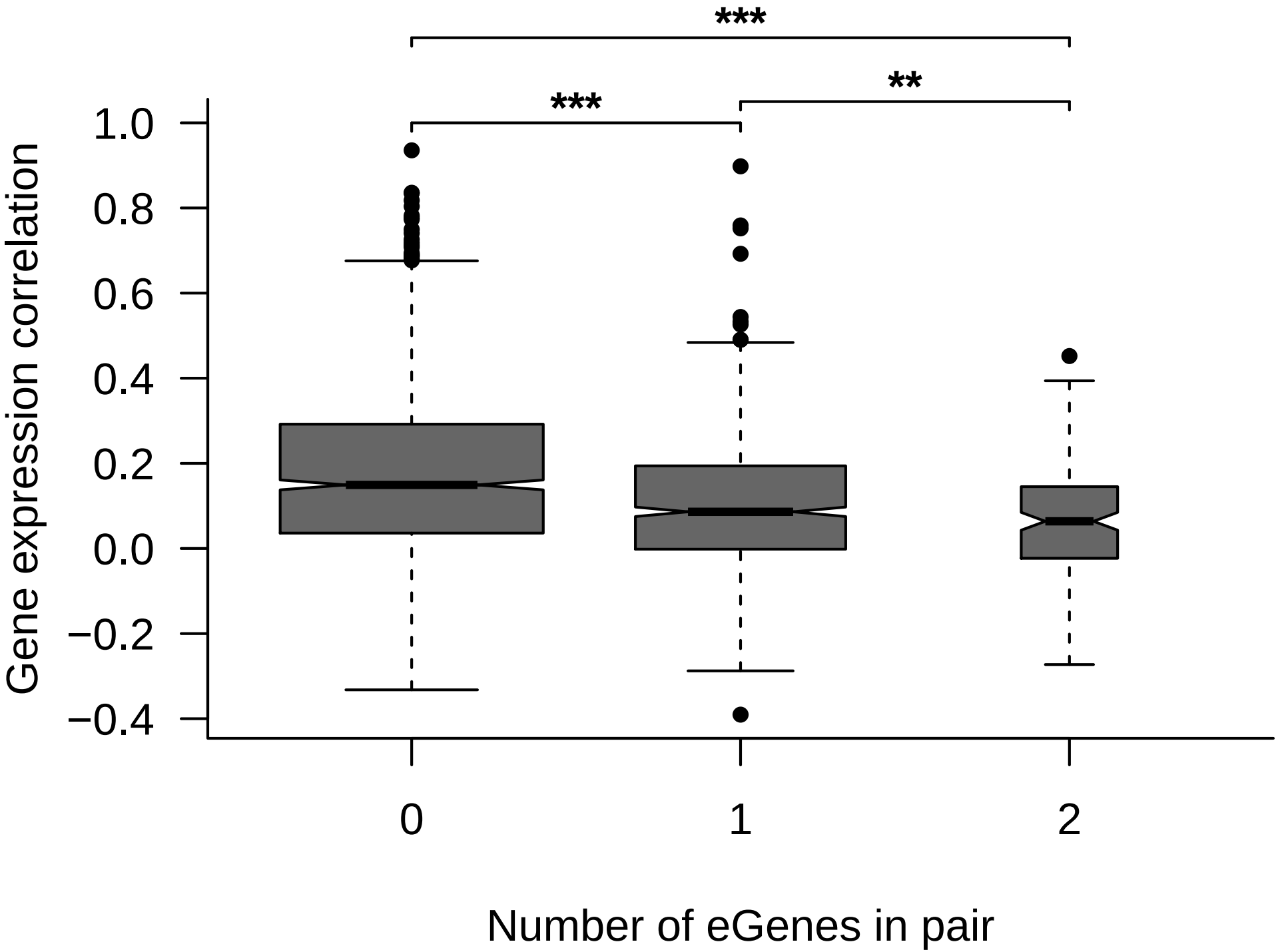
1124 **S22 Fig. Sampling time and day of bud flush.** (A) Box plot distributions of the Julian day of
1125 sampling for the SwAsp sub-populations. (B) The relationship between Julian day of bud flush for
1126 the greenhouse sampled buds and Julian day of bud flush in the field for the same year (2012).

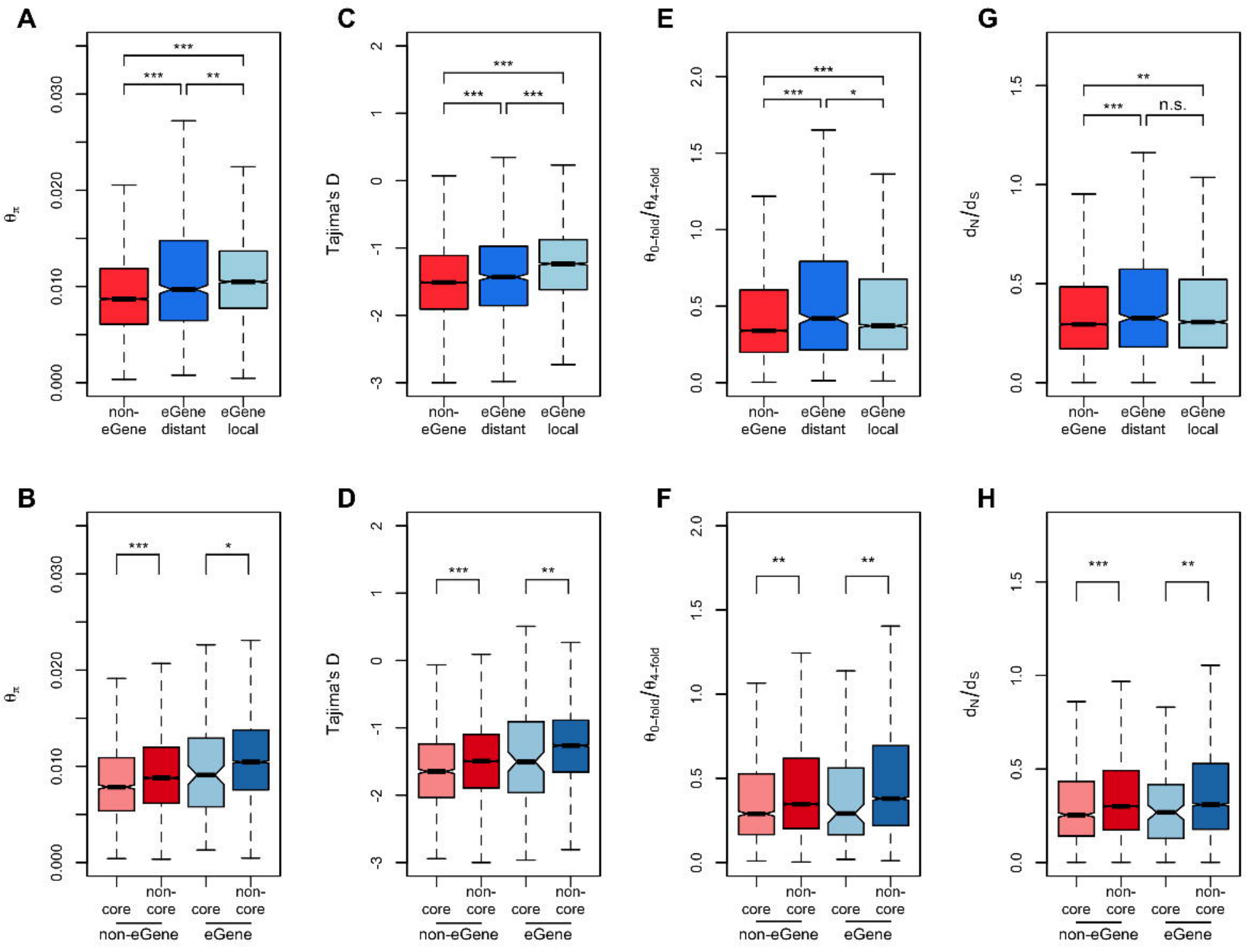
1127 **S23 Fig. Motivating example for major genotype filtering.** In this case the SNP is heterozygous
1128 in all samples but one, resulting in a minor allele frequency close to 0.5, but a major genotype
1129 frequency close to one. The resulting association turns out very significant, but is only supported by
1130 a single sample.

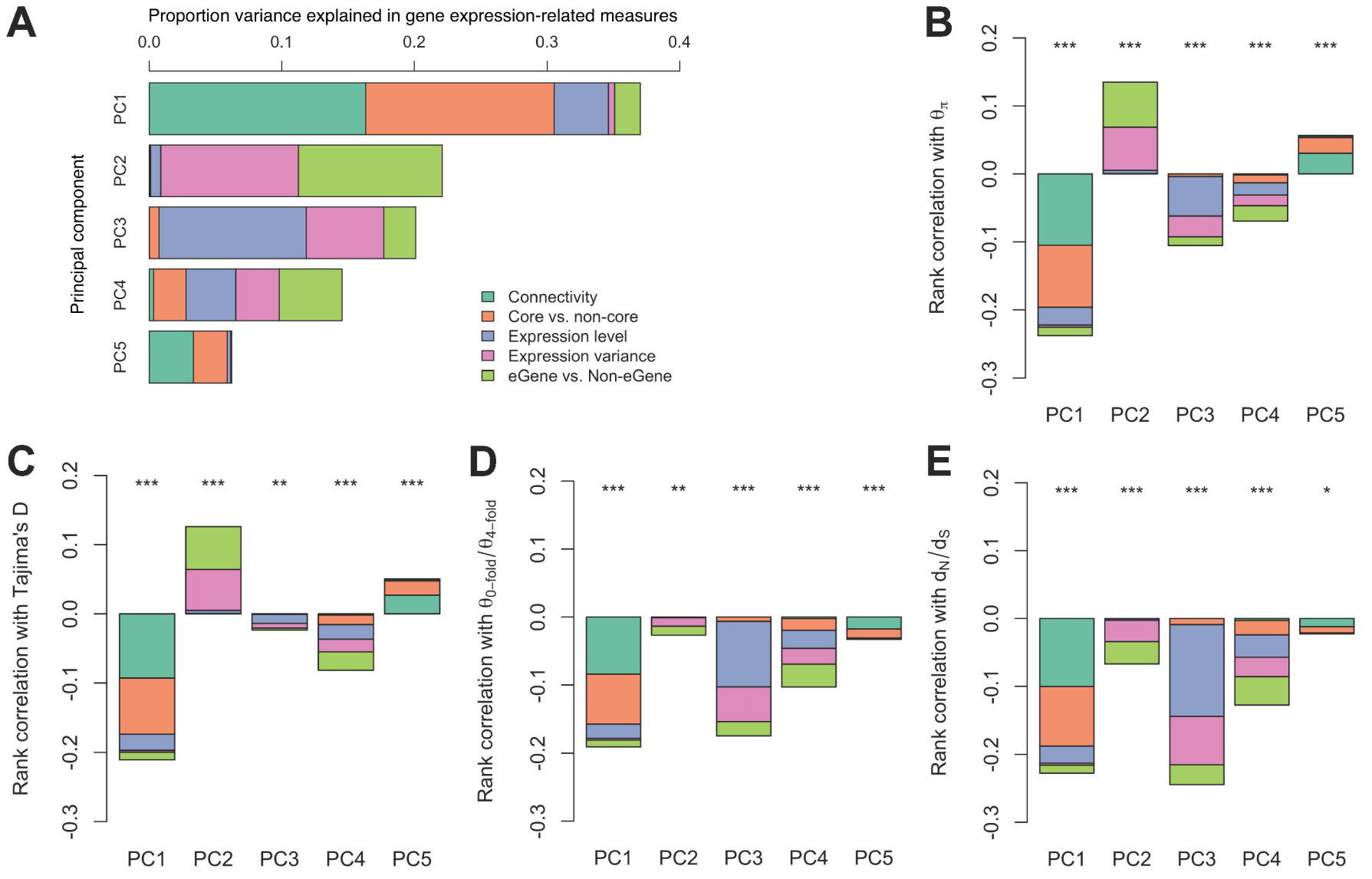

A
Latitude (° North)
Longitude (° East)
N
0 200 400 600 km
B
Density
All genes
Filtered genes
H^2
C
Density
All genes
Filtered genes
Q_{ST}
D
Count
Correlation
Population

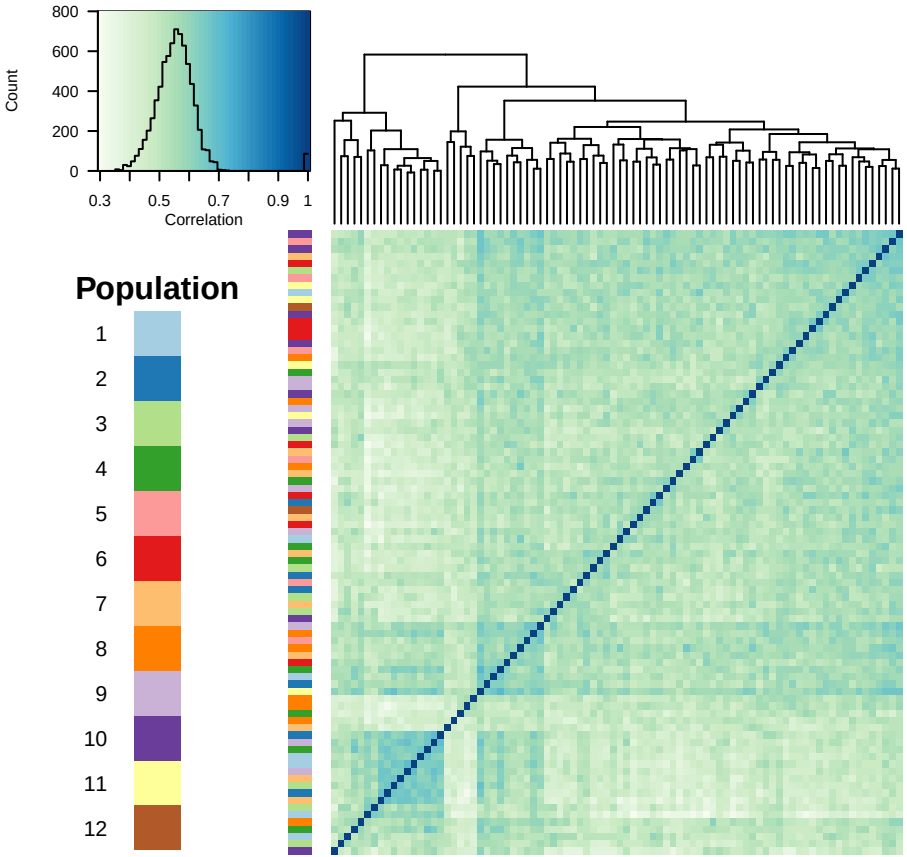

A
Local
Distant
Both
Variance explained
1.0
0.8
0.6
0.4
0.2
0.0
***
***
***
**
Maximum variance explained per gene
Total variance explained per gene
B
H^2
1.0
0.8
0.6
0.4
0.2
0.0
eGene
non−eGene
C
SNP position
1
2
4
6
8
11
15
19
Gene position
D
Associated gene
Other gene
5'
3'
Count / kbp sequence
Local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

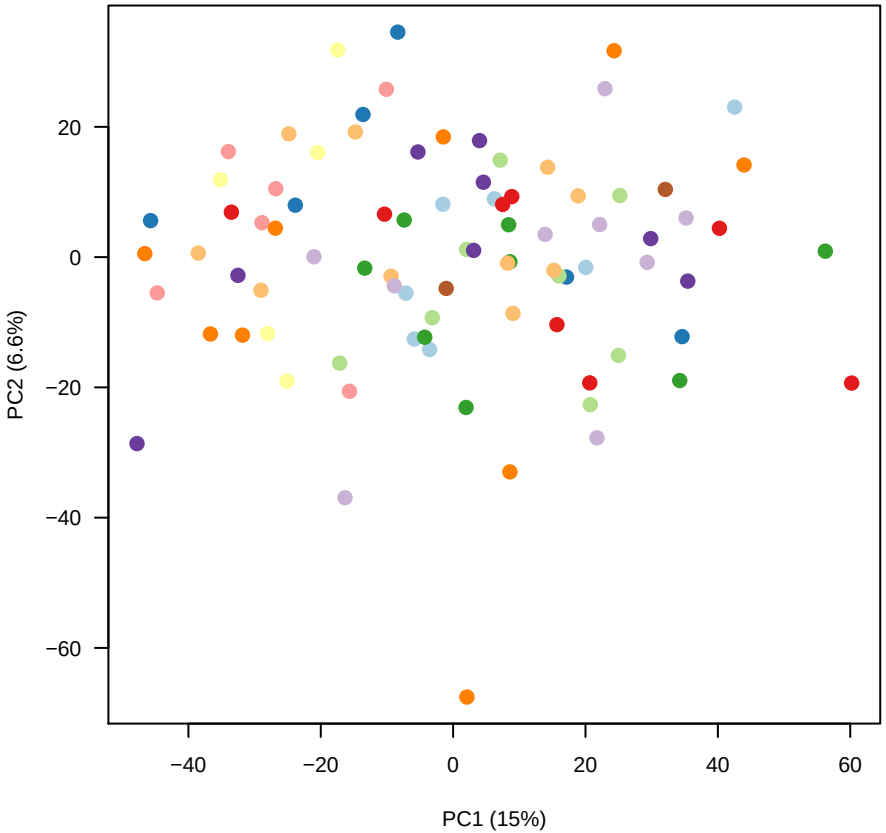

A
Expression variance
0
5
10
15
exAtlas
SwAsp
B
Connectivity
5
10
15
20
25
eGene
non−eGene
D
Core
Non−core
Number of genes
eGene
Non−eGene
0
50
100
150
200
0
1000
2000
3000
4000
5000
Both
Distant
Local
C
H^2
Proportion VE
Value
0.0
0.2
0.4
0.6
0.8
1.0
eGene
Non−eGene
Core
Non−core
E
Core eGene
Non−core 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

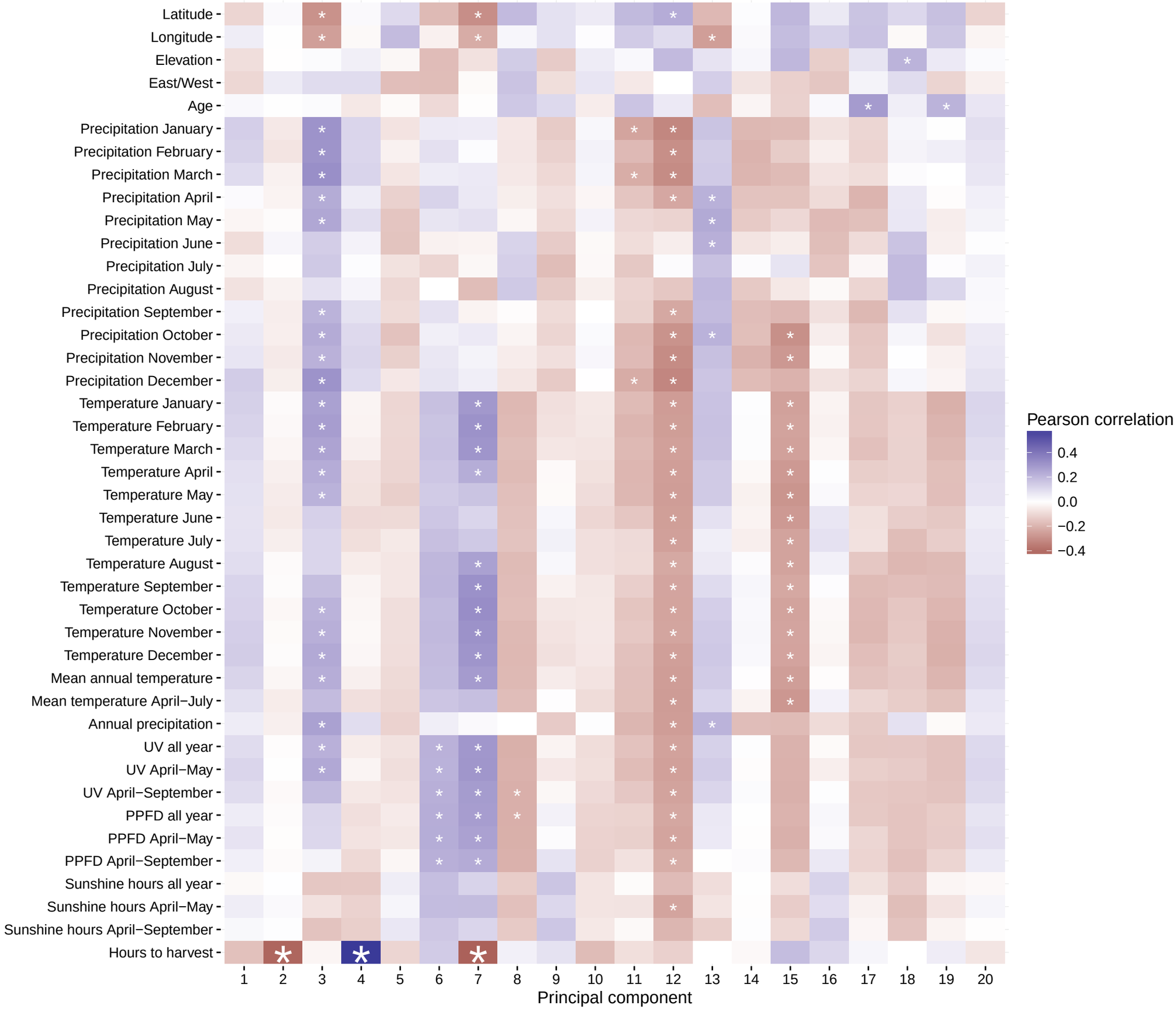

***
***
**
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
Gene expression correlation
0
1
2
Number of eGenes in pair

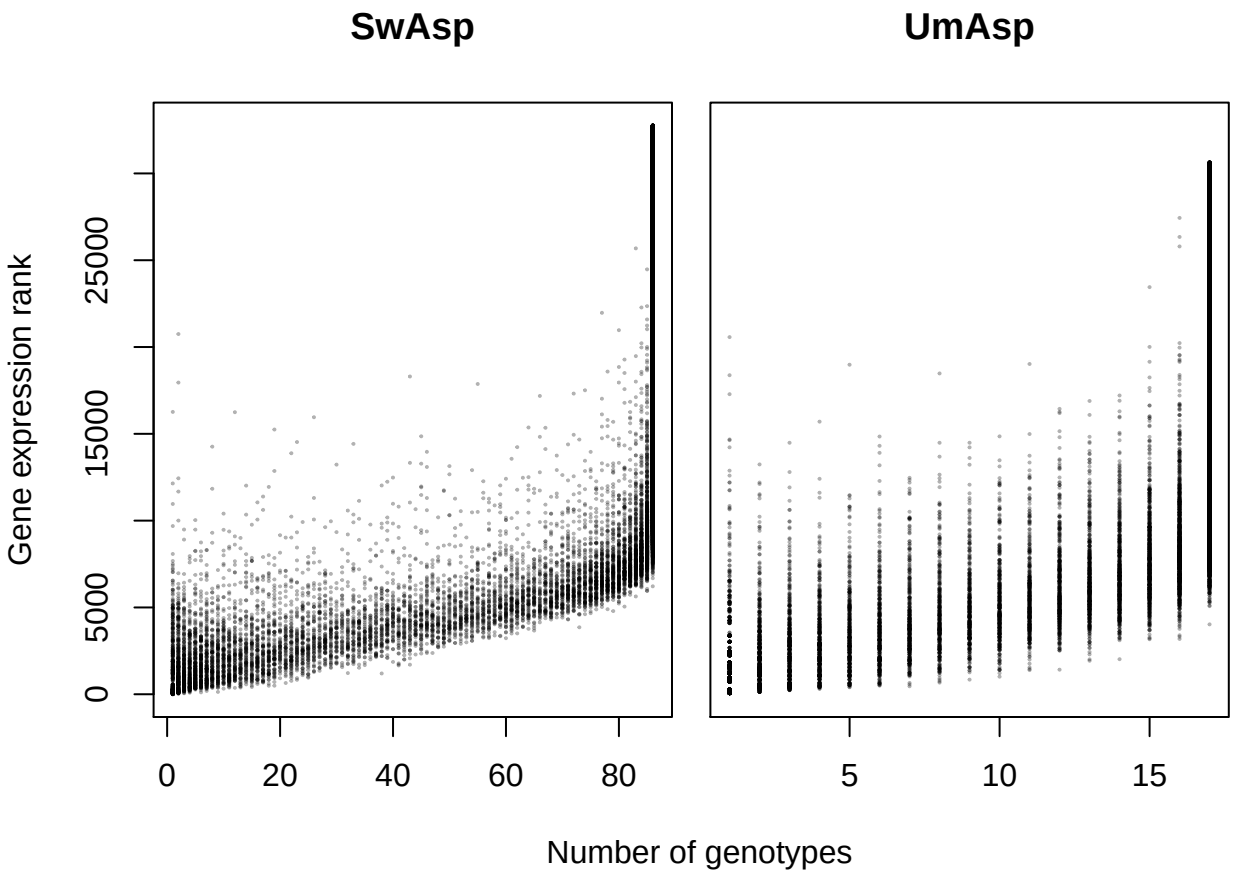

A
B
C
D
E
F
G
H
θπ
Tajima's D
θ0-fold/θ4-fold
dN/dS
non-eGene
eGene distant
eGene local
core
non-core
non-eGene
eGene
***
**
*
n.s.

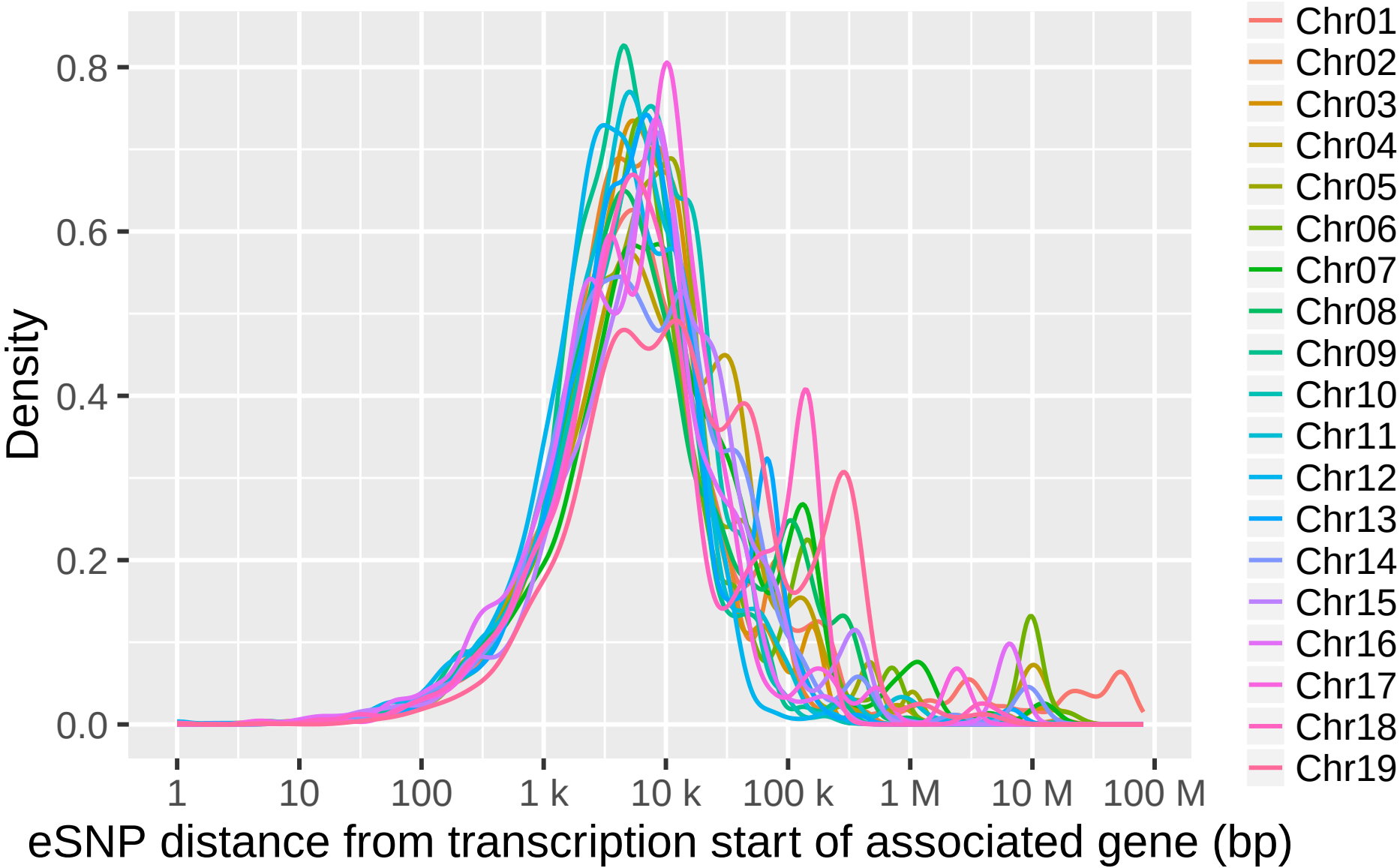

A
Proportion variance explained in gene expression-related measures
0.0
0.1
0.2
0.3
0.4
Principal component
PC1
PC2
PC3
PC4
PC5
Connectivity
Core vs. non-core
Expression level
Expression variance
eGene vs. Non-eGene
B
Rank correlation with θ_π
C
Rank correlation with Tajima's D
D
Rank correlation with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

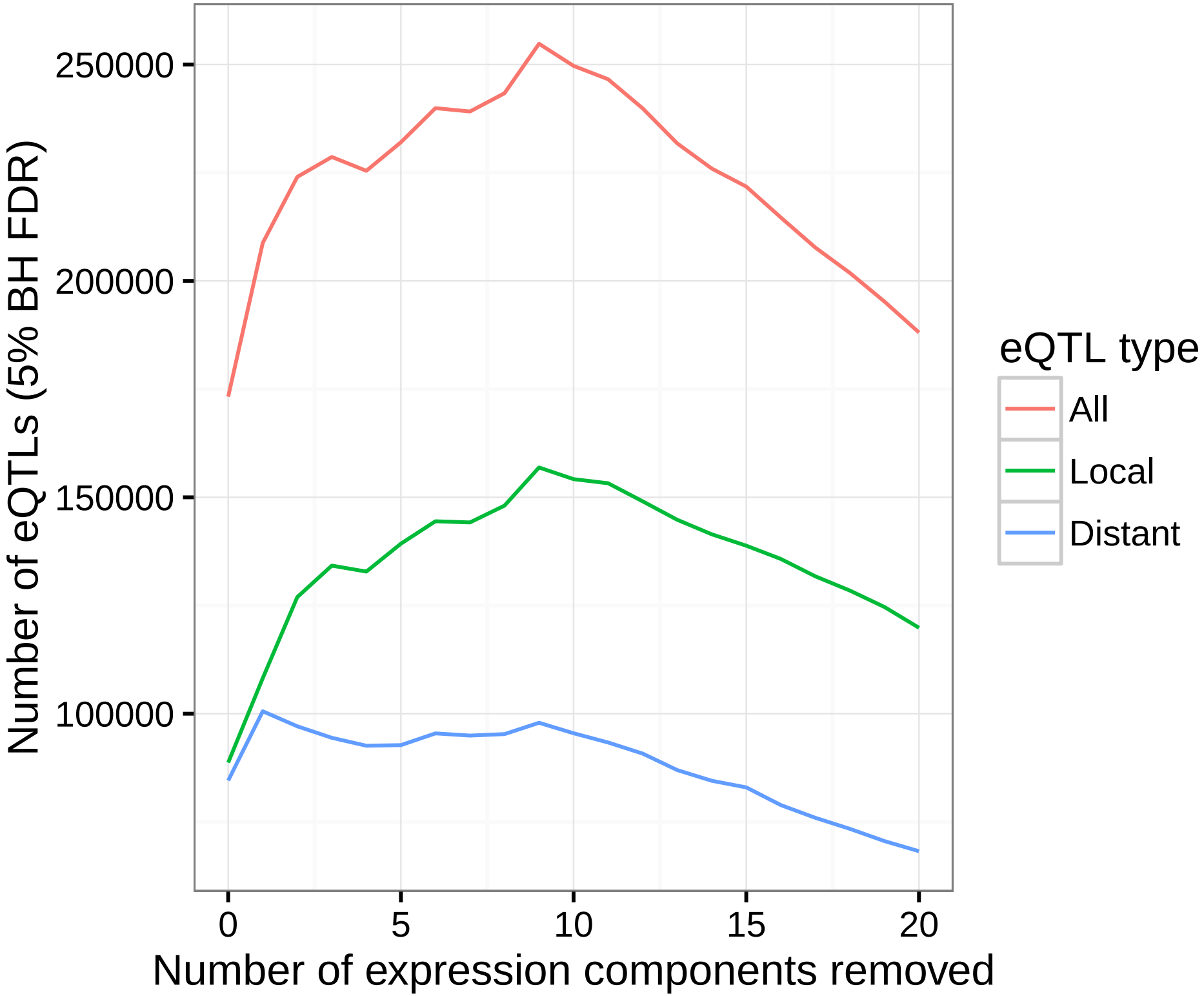

Count
800
600
400
200
0
0.3
0.5
0.7
0.9
1
Correlation
Population
1
2
3
4
5
6
7
8
9
10
11
12

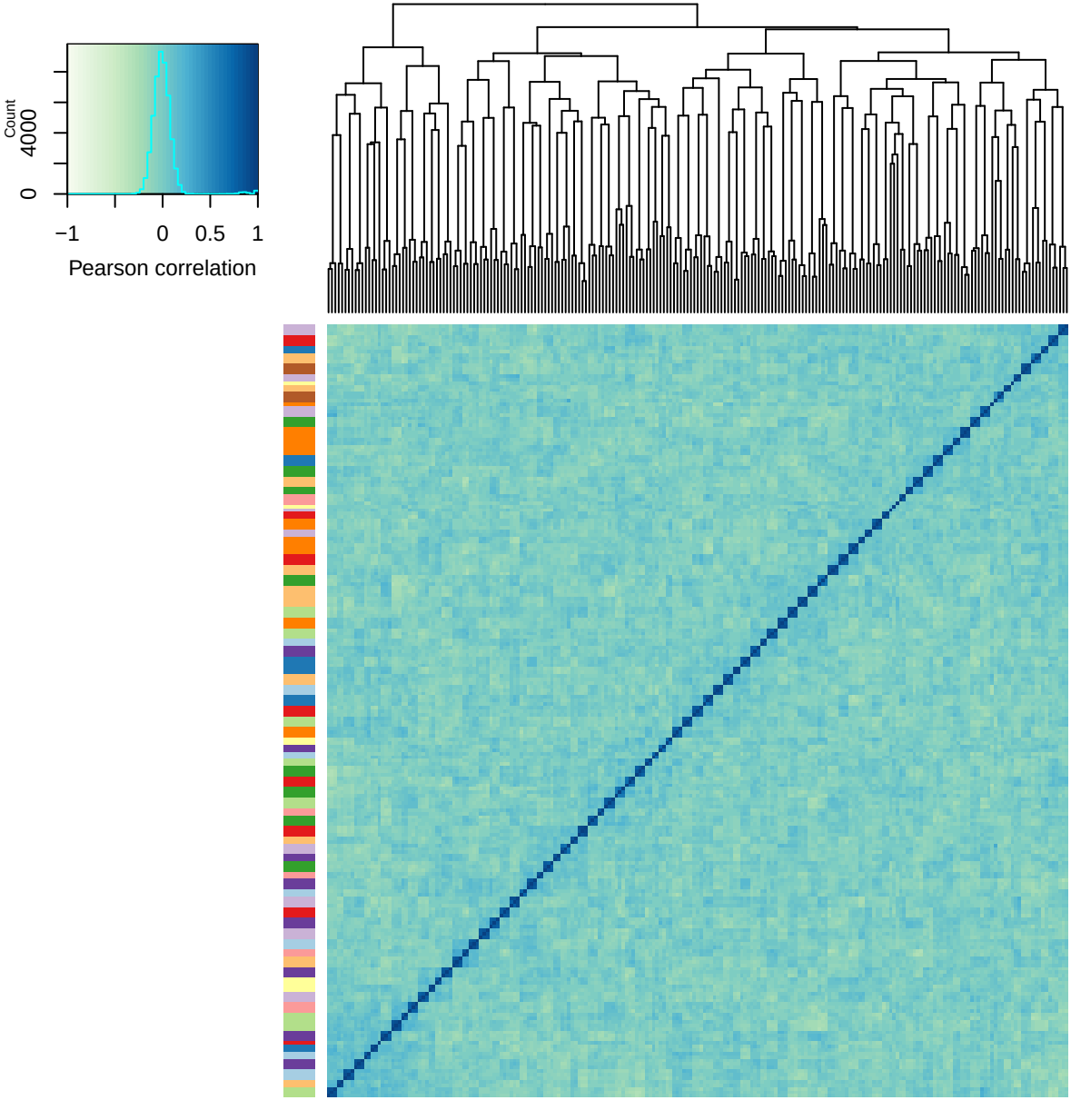

PC2 (6.6%)
20
0
−20
−40
−60
−40
−20
0
20
40
60
PC1 (15%)

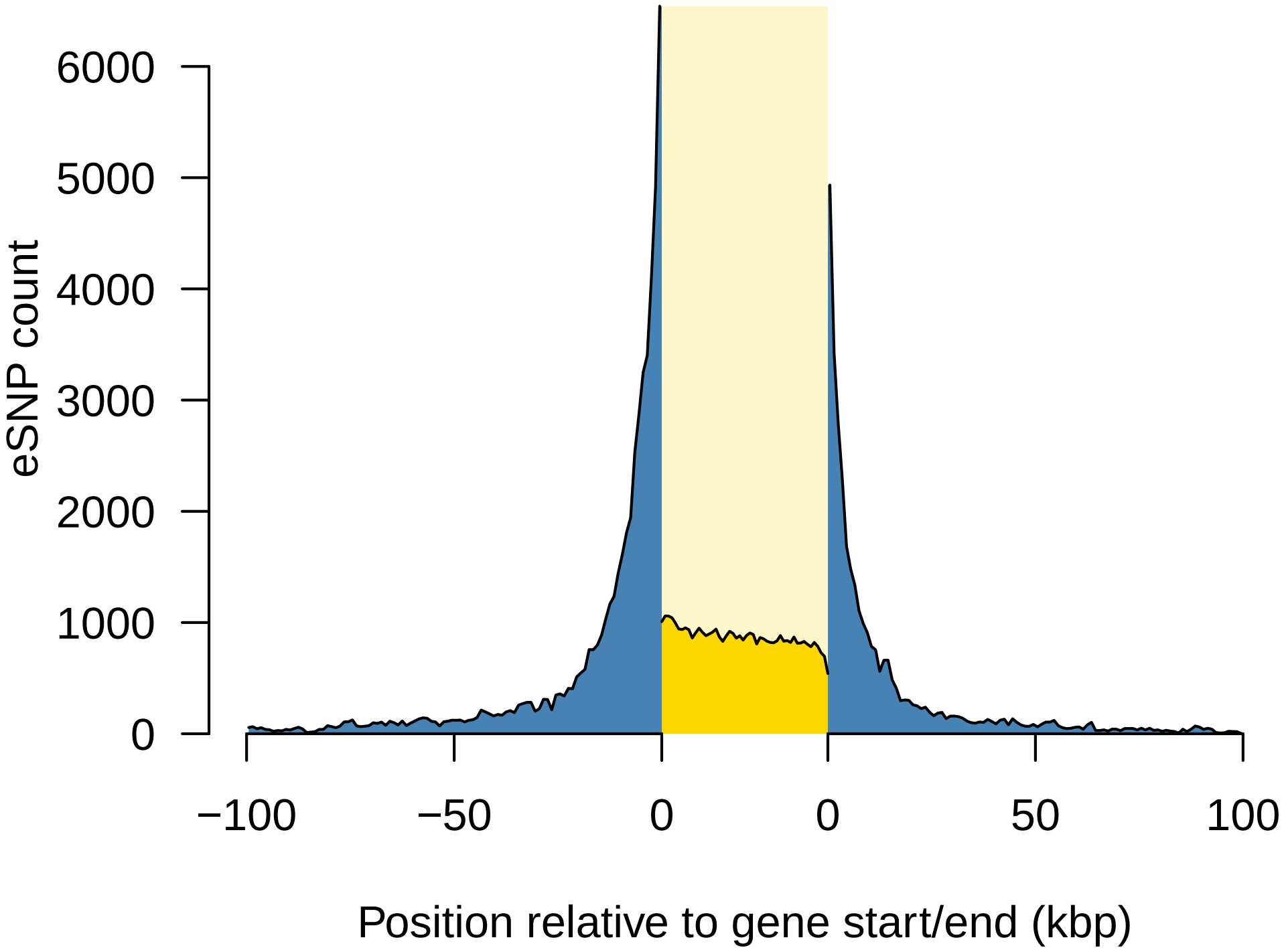

Latitude
Longitude
Elevation
East/West
Age
Precipitation January
Precipitation February
Precipitation March
Precipitation April
Precipitation May
Precipitation June
Precipitation July
Precipitation August
Precipitation September
Precipitation October
Precipitation November
Precipitation December
Temperature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

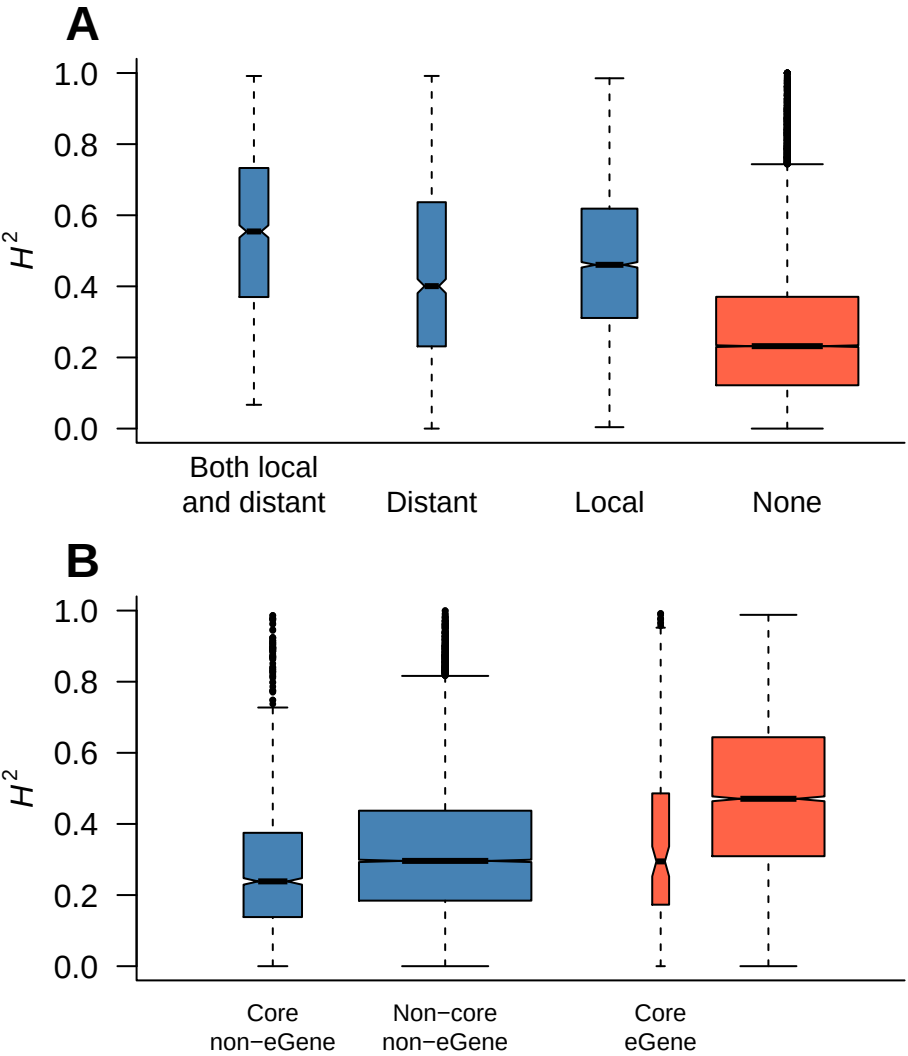

SwAsp
UmAsp
Gene expression rank
0
5000
15000
25000
0
20
40
60
80
5
10
15
Number of genotypes

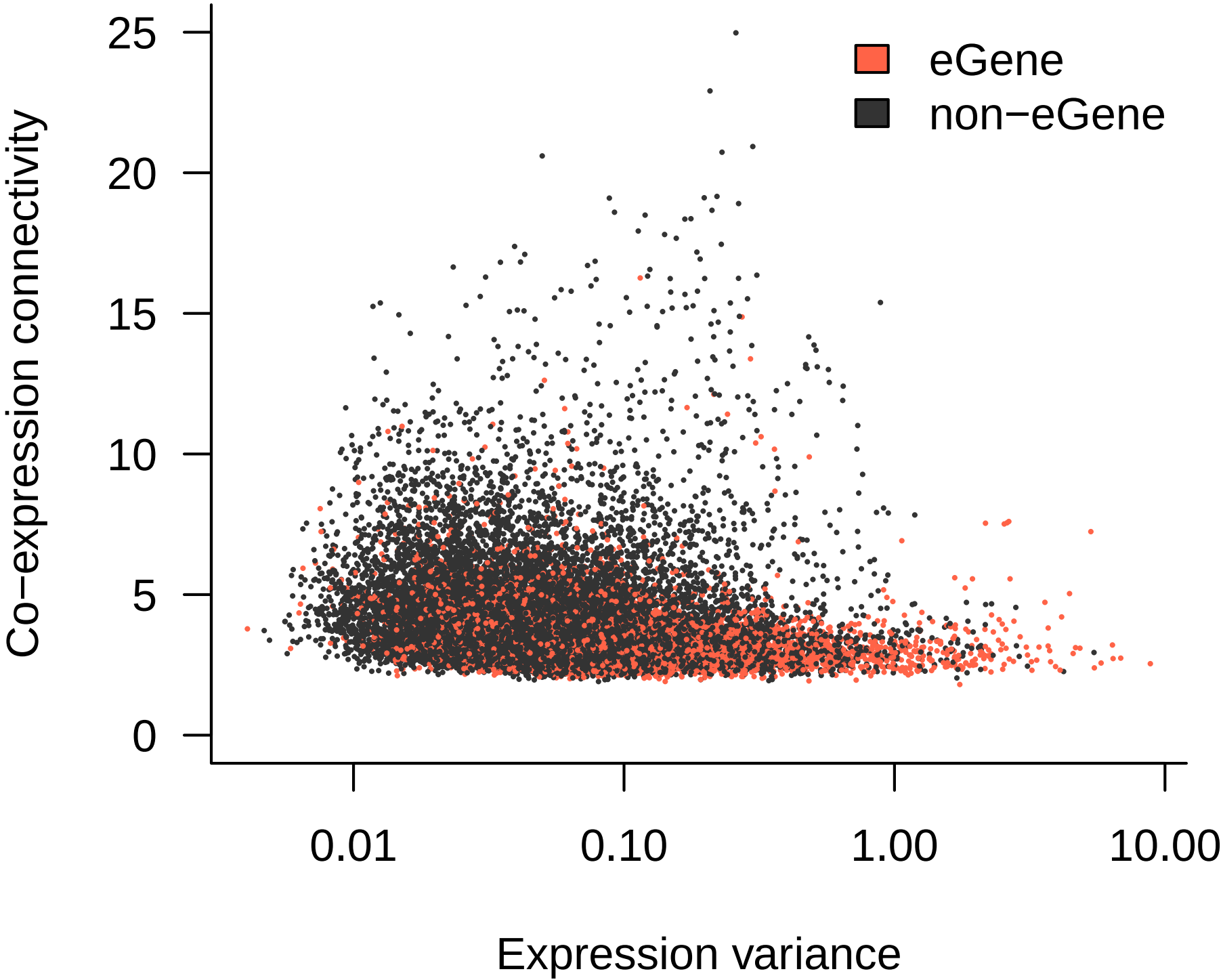

Density
0.8
0.6
0.4
0.2
0.0
1
10
100
1 k
10 k
100 k
1 M
10 M
100 M
eSNP distance from transcription start of associated gene (bp)
Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19

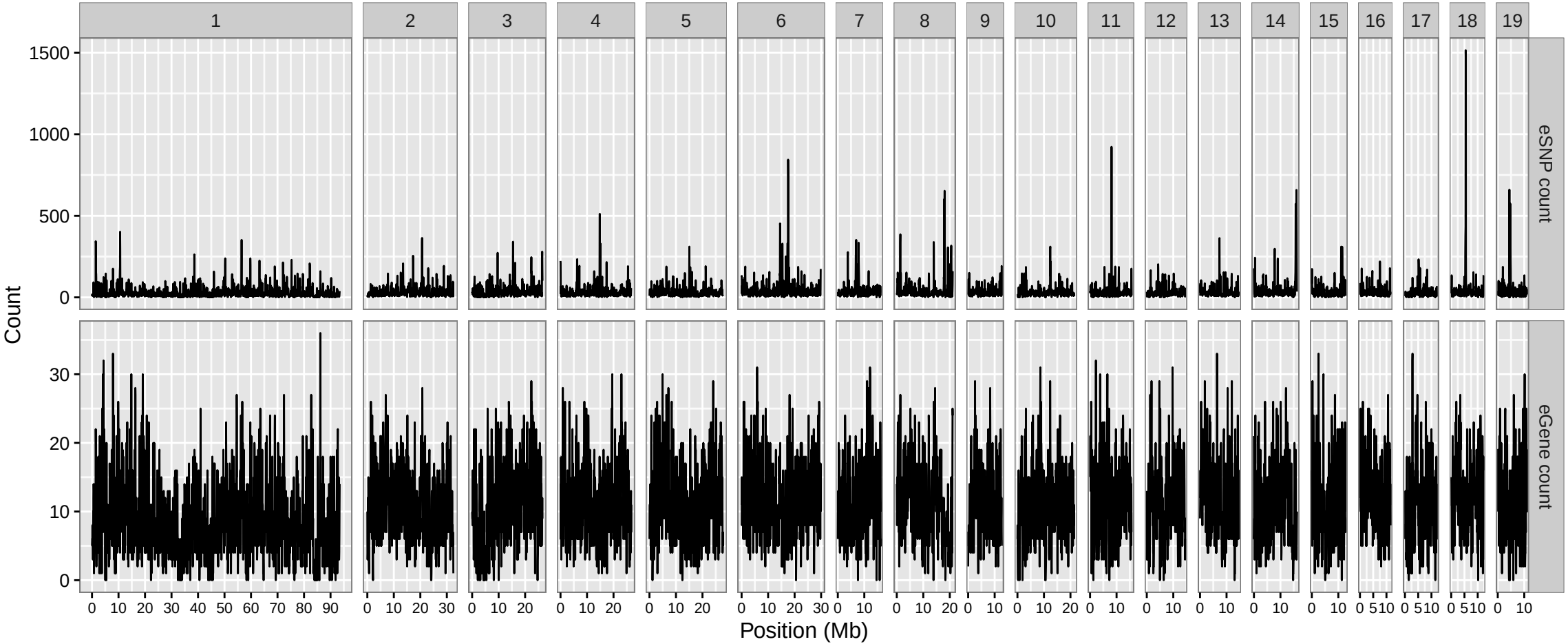

250000
200000
150000
100000
Number of eQTLs (5% BH FDR)
0
5
10
15
20
Number of expression components removed
eQTL type
All
Local
Distant

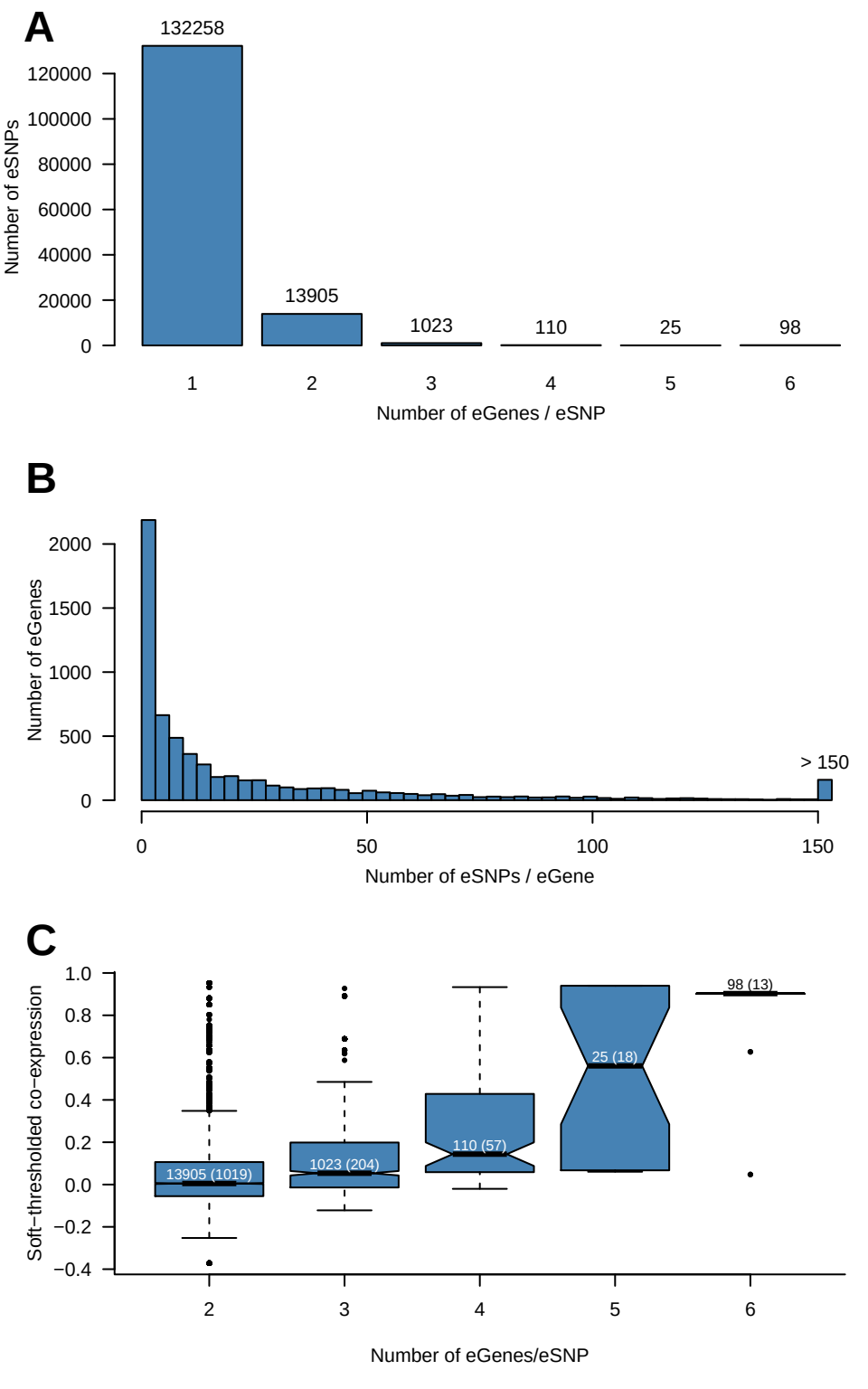

Count
4000
0
−1
0
0.5
1
Pearson correlation

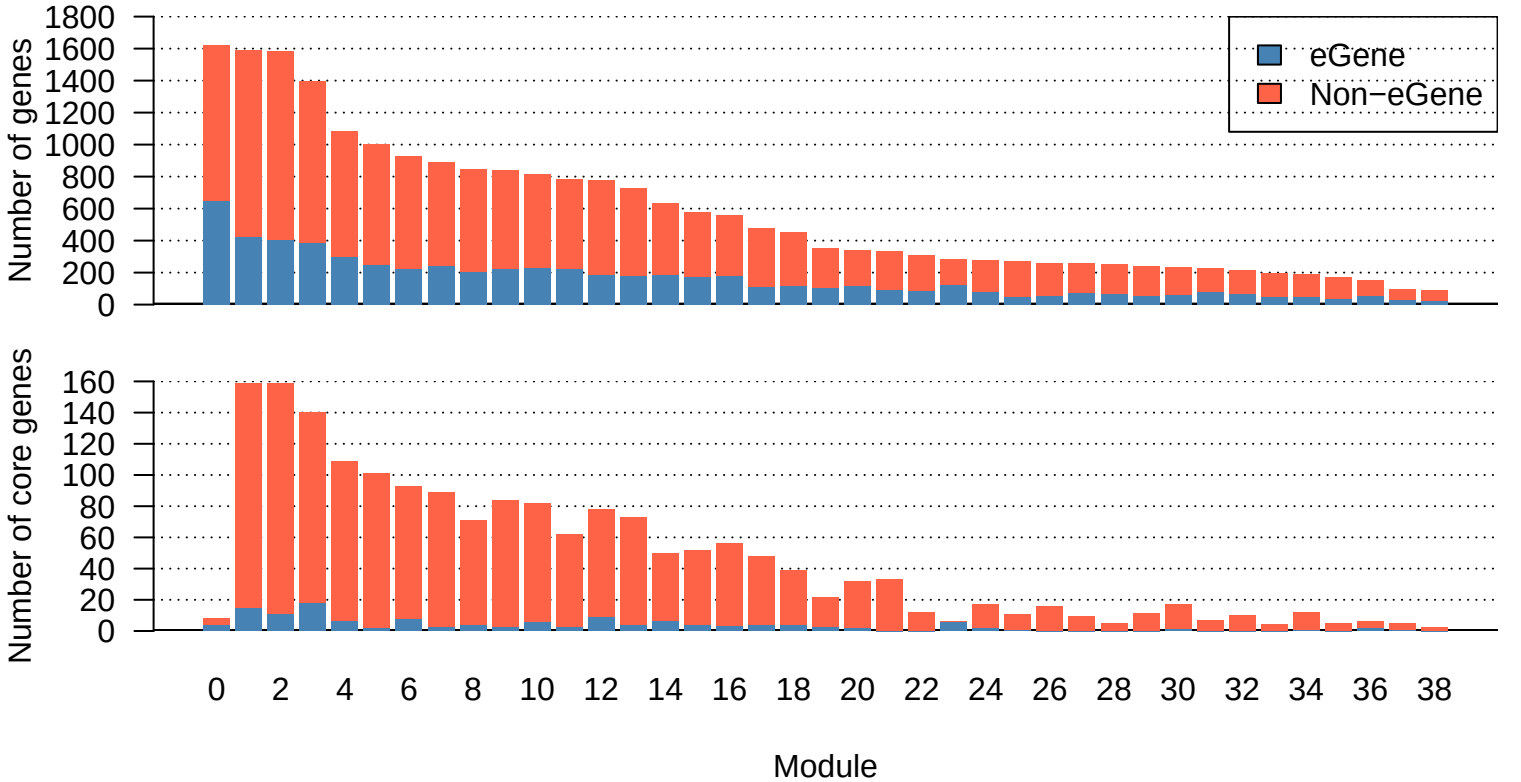

6000
5000
4000
3000
2000
1000
0
eSNP count
−100
−50
0
0
50
100
Position relative to gene start/end (kbp)

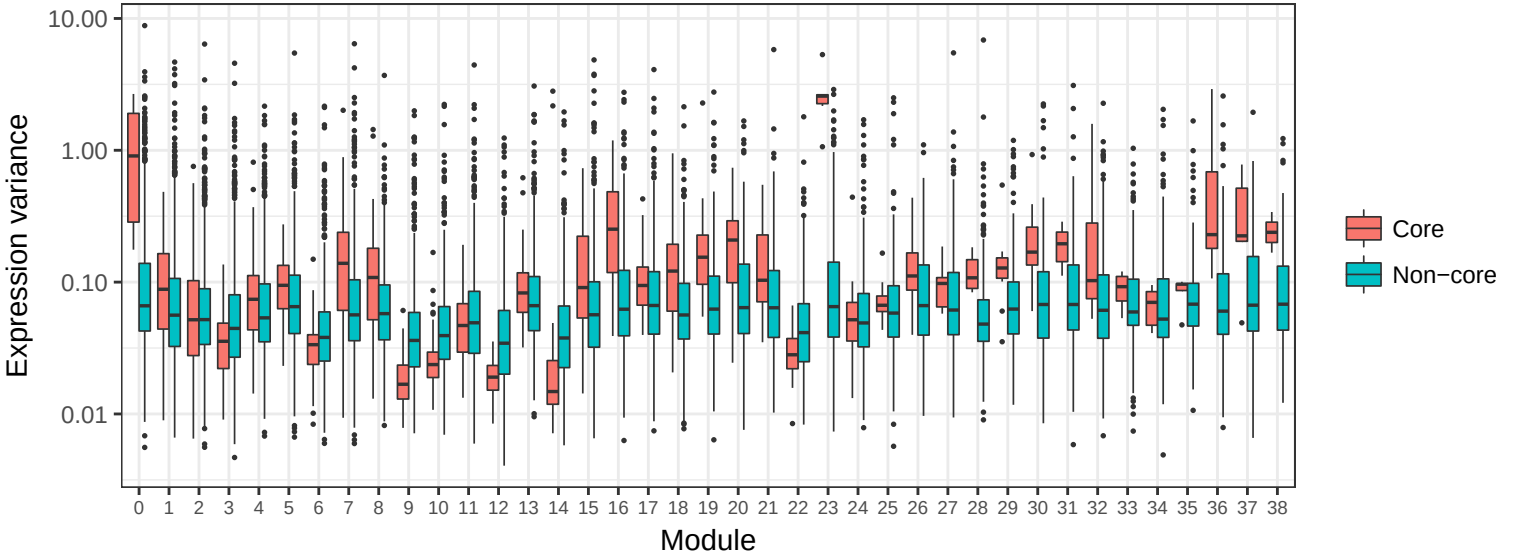

A
1.0
0.8
0.6
0.4
0.2
0.0
H^2
Both local and distant
Distant
Local
None
B
1.0
0.8
0.6
0.4
0.2
0.0
H^2
Core non−eGene
Non−core non−eGene
Core eGene

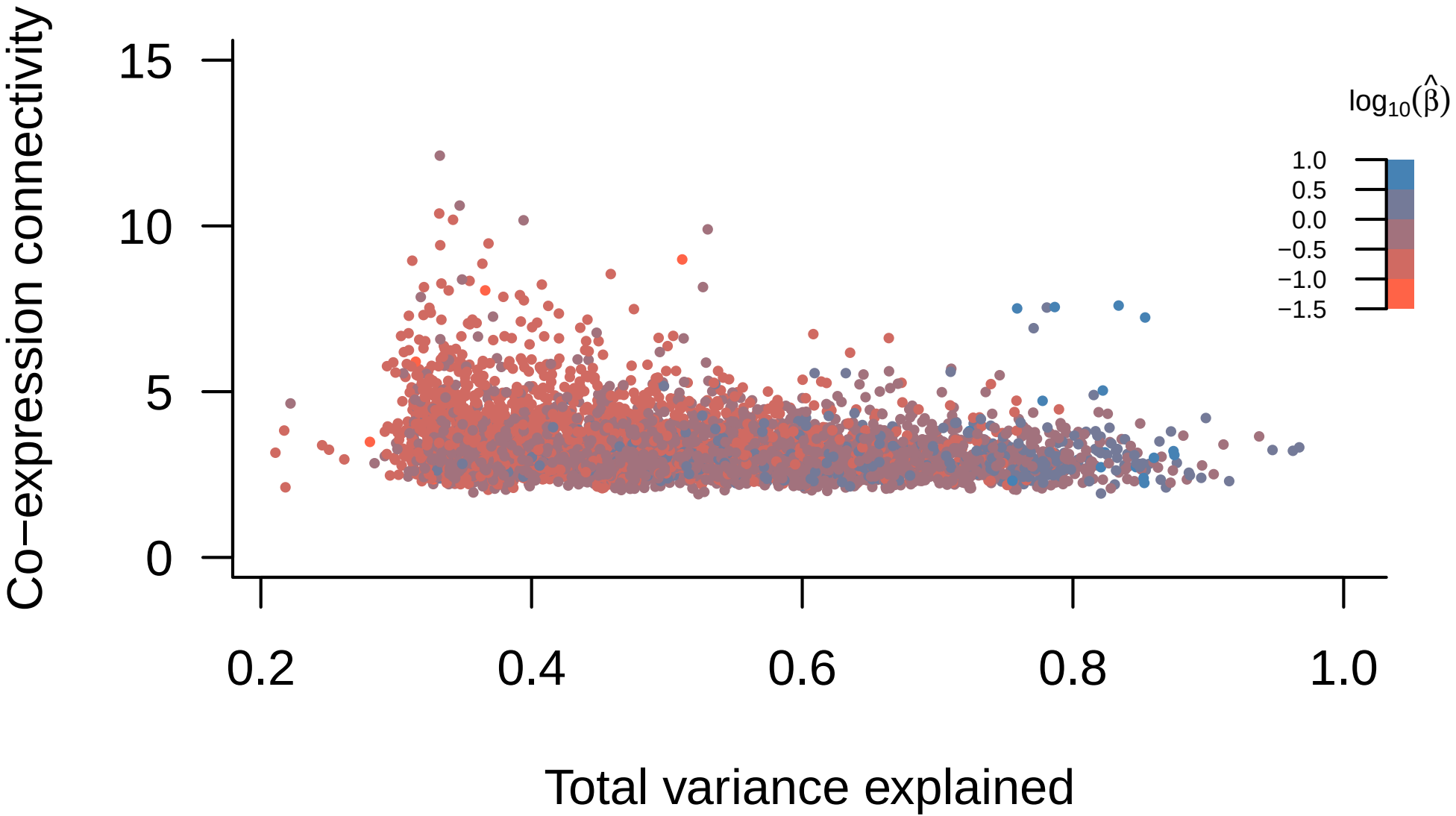

eGene
non−eGene
Co−expression connectivity
25
20
15
10
5
0
0.01
0.10
1.00
10.00
Expression variance

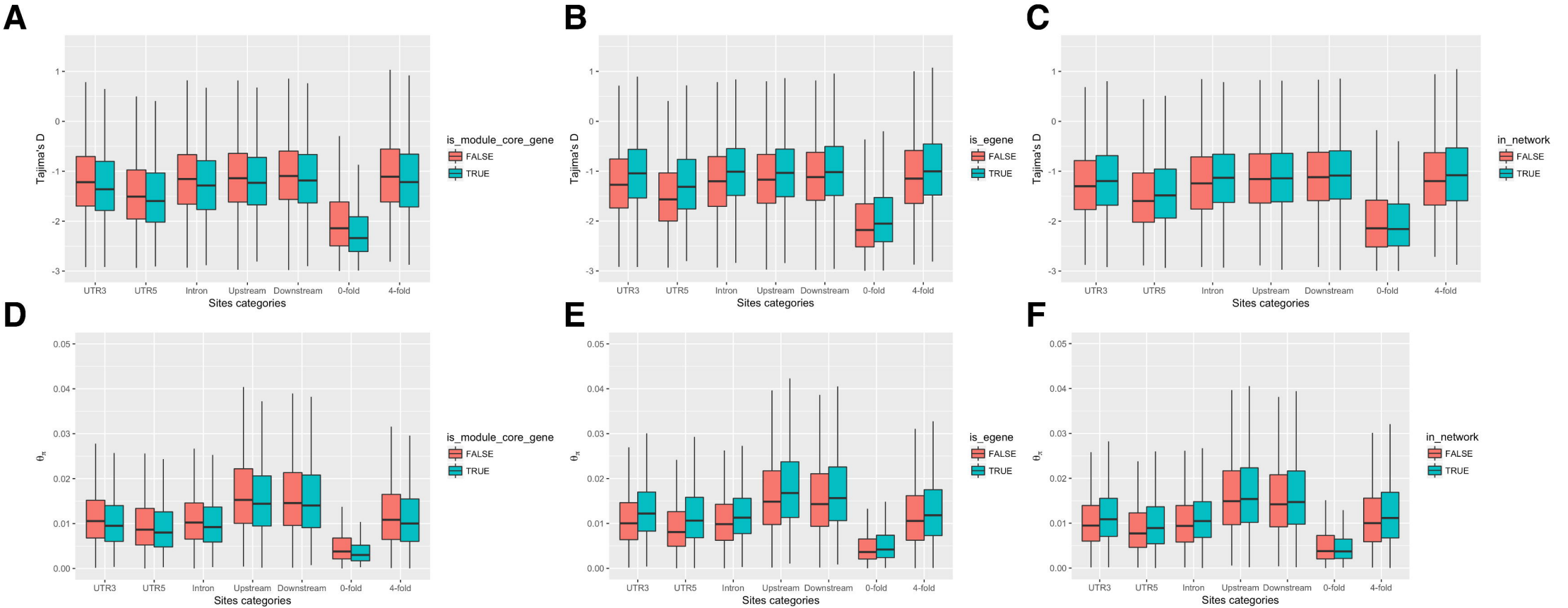

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
eSNP count
eGene count
Count
Position (Mb)

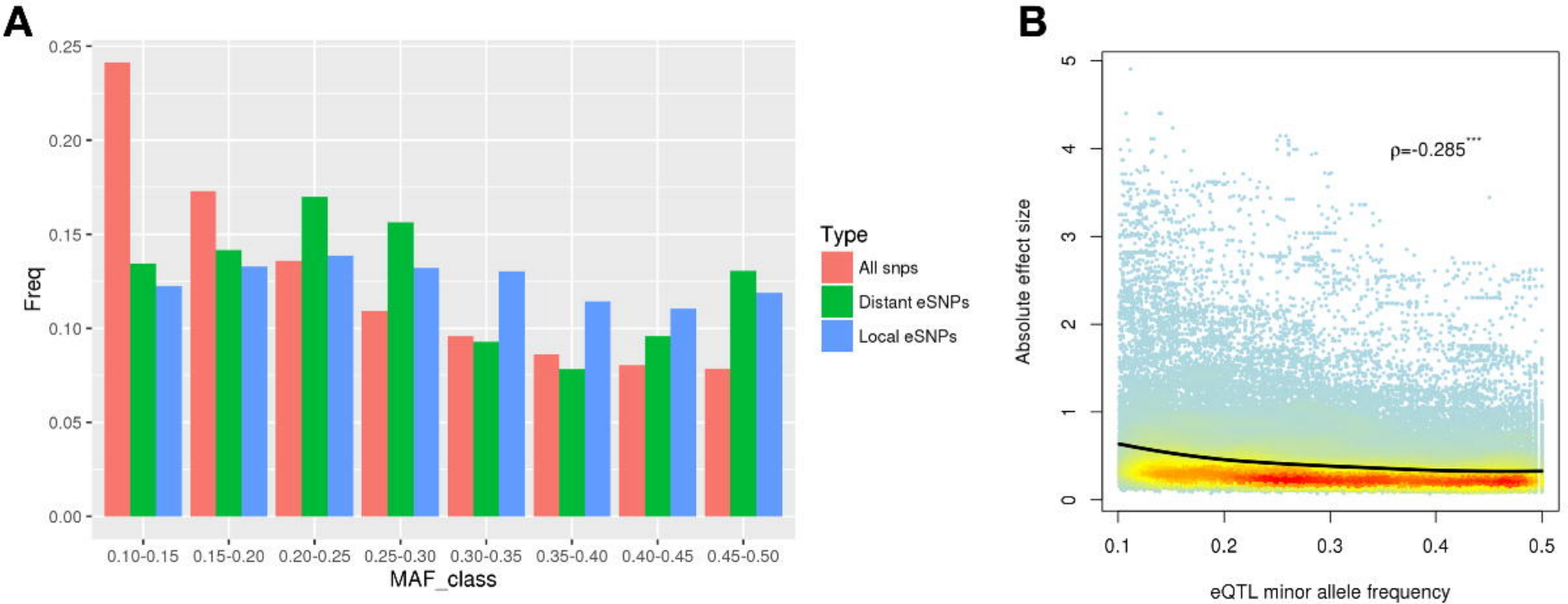

A
132258
13905
1023
110
25
98
Number of eSNPs
120000
100000
80000
60000
40000
20000
0
1
2
3
4
5
6
Number of eGenes / eSNP
B
Number of eGenes
2000
1500
1000
500
0
> 150
0
50
100
150
Number of eSNPs / eGene
C
Soft−thresholded co−expression
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
13905 (1019)
1023 (204)
110 (57)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

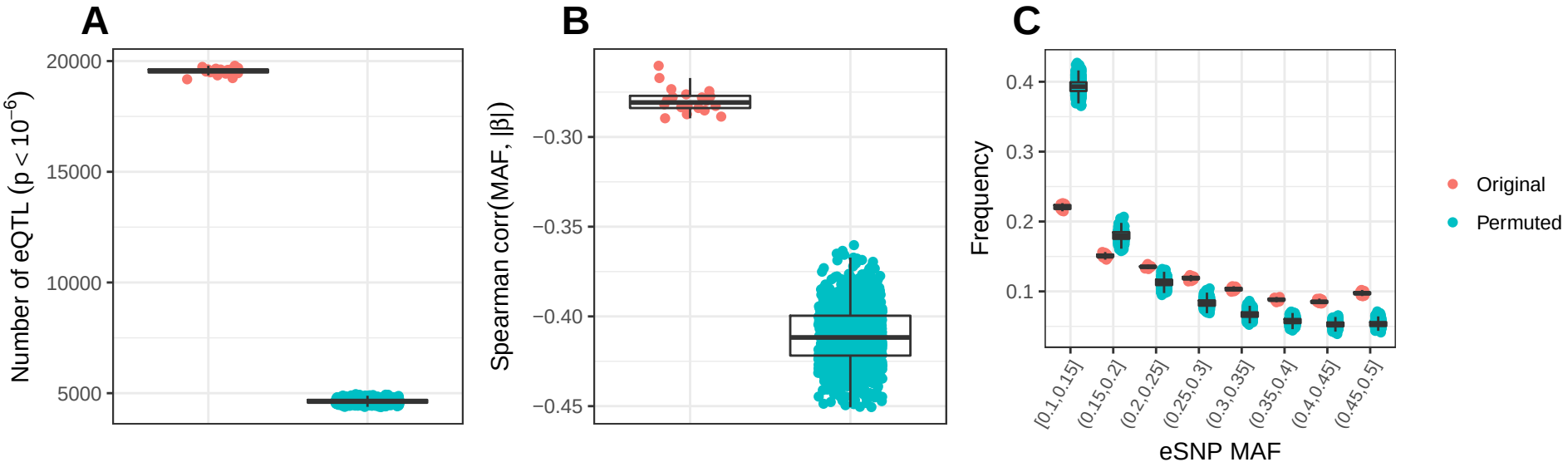

Number of genes
1800
1600
1400
1200
1000
800
600
400
200
0
eGene
Non−eGene
Number of core genes
160
140
120
100
80
60
40
20
0
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
Module

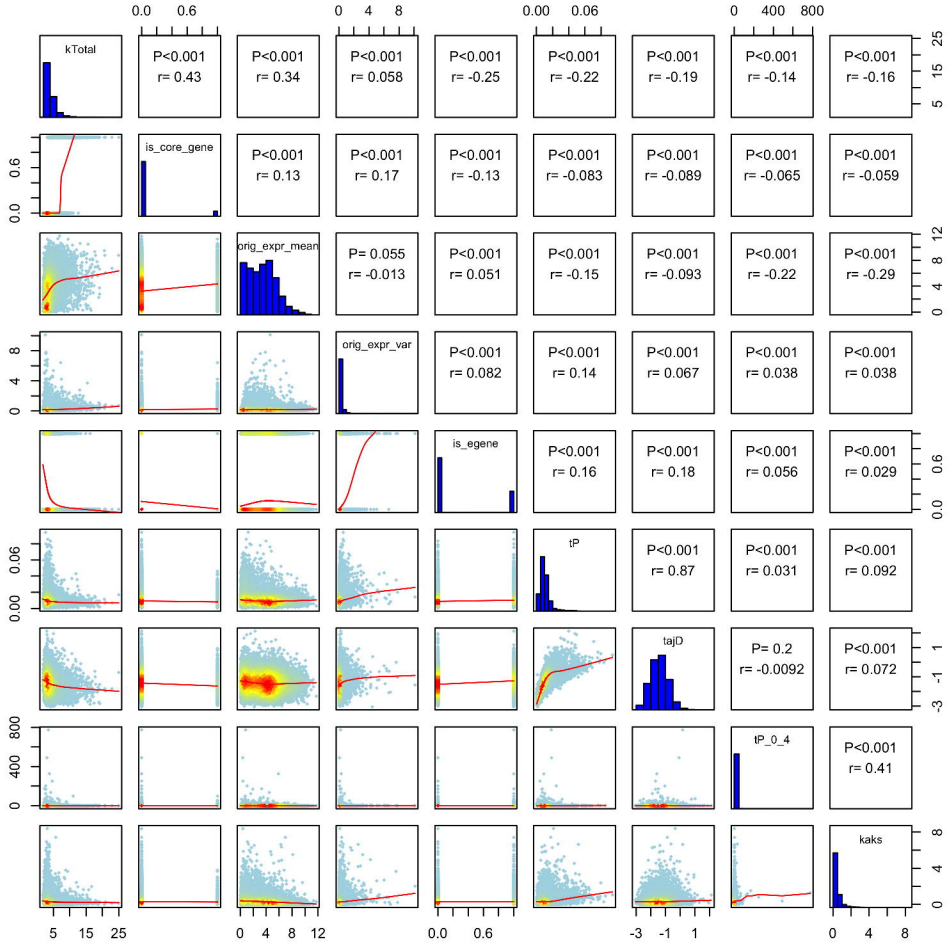

Expression variance
10.00
1.00
0.10
0.01
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
Module
Core
Non−core

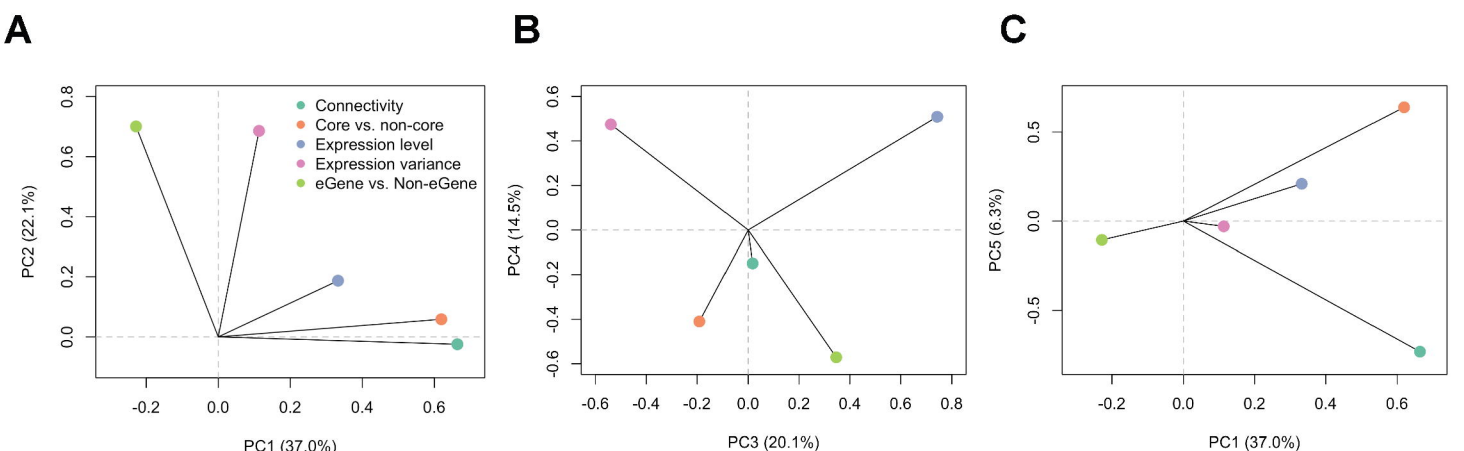

Co−expression connectivity
15
10
5
0
0.2
0.4
0.6
0.8
1.0
Total variance explained
$\log_{10}(\hat{\beta})$
1.0
0.5
0.0
−0.5
−1.0
−1.5

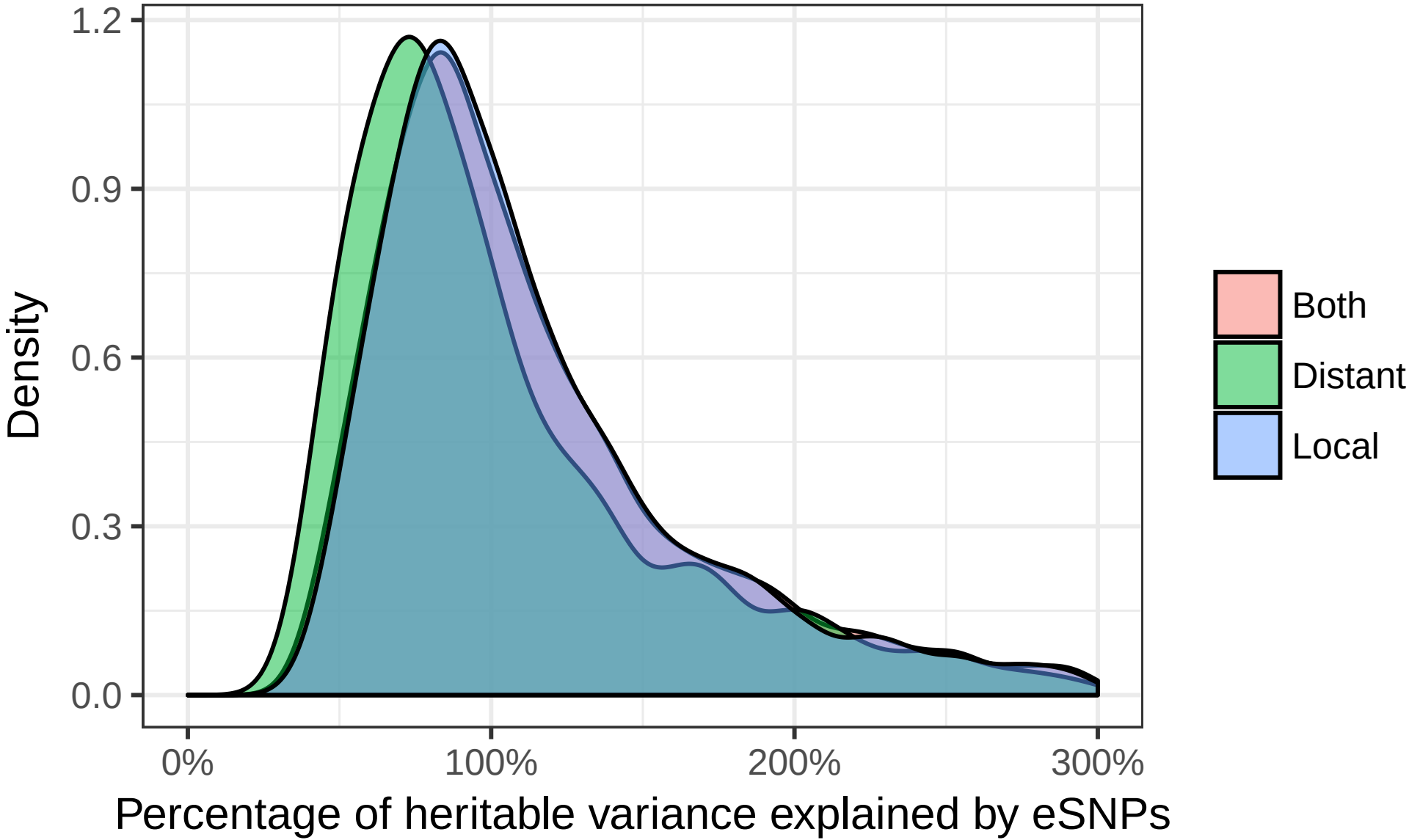

A
B
C
D
E
F
Tajima's D
θ_π
Sites categories
UTR3
UTR5
Intron
Upstream
Downstream
0-fold
4-fold
is_module_core_gene
is_egene
in_network
FALSE
TRUE


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

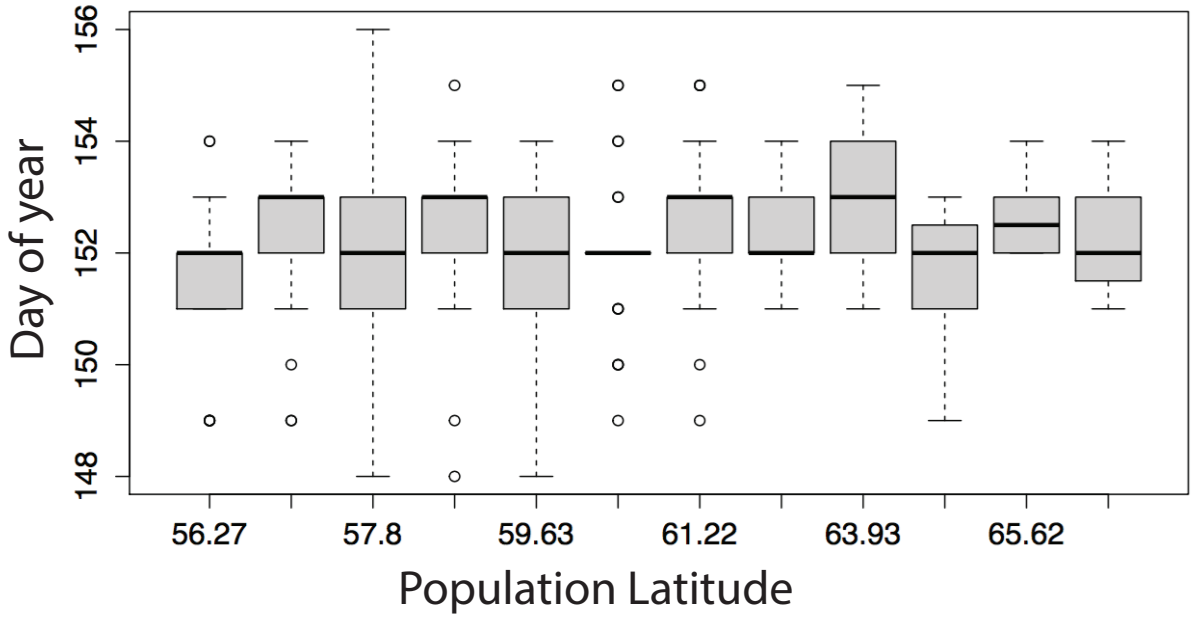

A
B
C
Number of eQTL (p < 10^-6)
20000
15000
10000
5000
Spearman corr(MAF, |β|)
−0.30
−0.35
−0.40
−0.45
Frequency
0.4
0.3
0.2
0.1
[0.1,0.15]
(0.15,0.2]
(0.2,0.25]
(0.25,0.3]
(0.3,0.35]
(0.35,0.4]
(0.4,0.45]
(0.45,0.5]
eSNP MAF
Original
Permuted


kTotal
P<0.001
r= 0.43
P<0.001
r= 0.34
P<0.001
r= 0.058
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<0.001
r= 0.17
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= 0.055
r= -0.013
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
orig_expr_var
P<0.001
r= 0.082
P<0.001
r= 0.14
P<0.001
r= 0.067
P<0.001
r= 0.038
P<0.001
r= 0.038
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

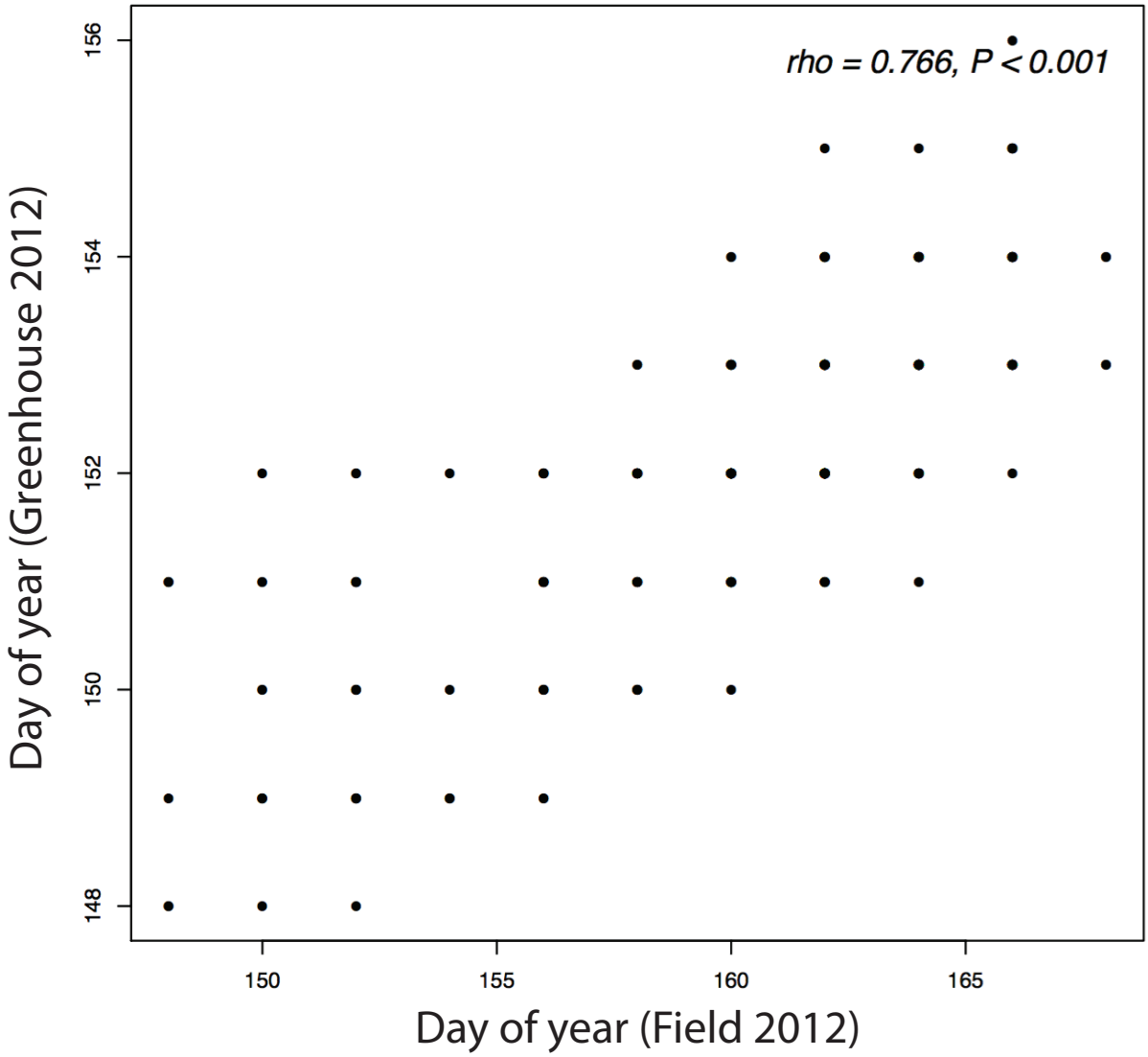

A
B
C
Connectivity
Core vs. non-core
Expression level
Expression variance
eGene vs. Non-eGene
PC2 (22.1%)
PC1 (37.0%)
PC4 (14.5%)
PC3 (20.1%)
PC5 (6.3%)
PC1 (37.0%)

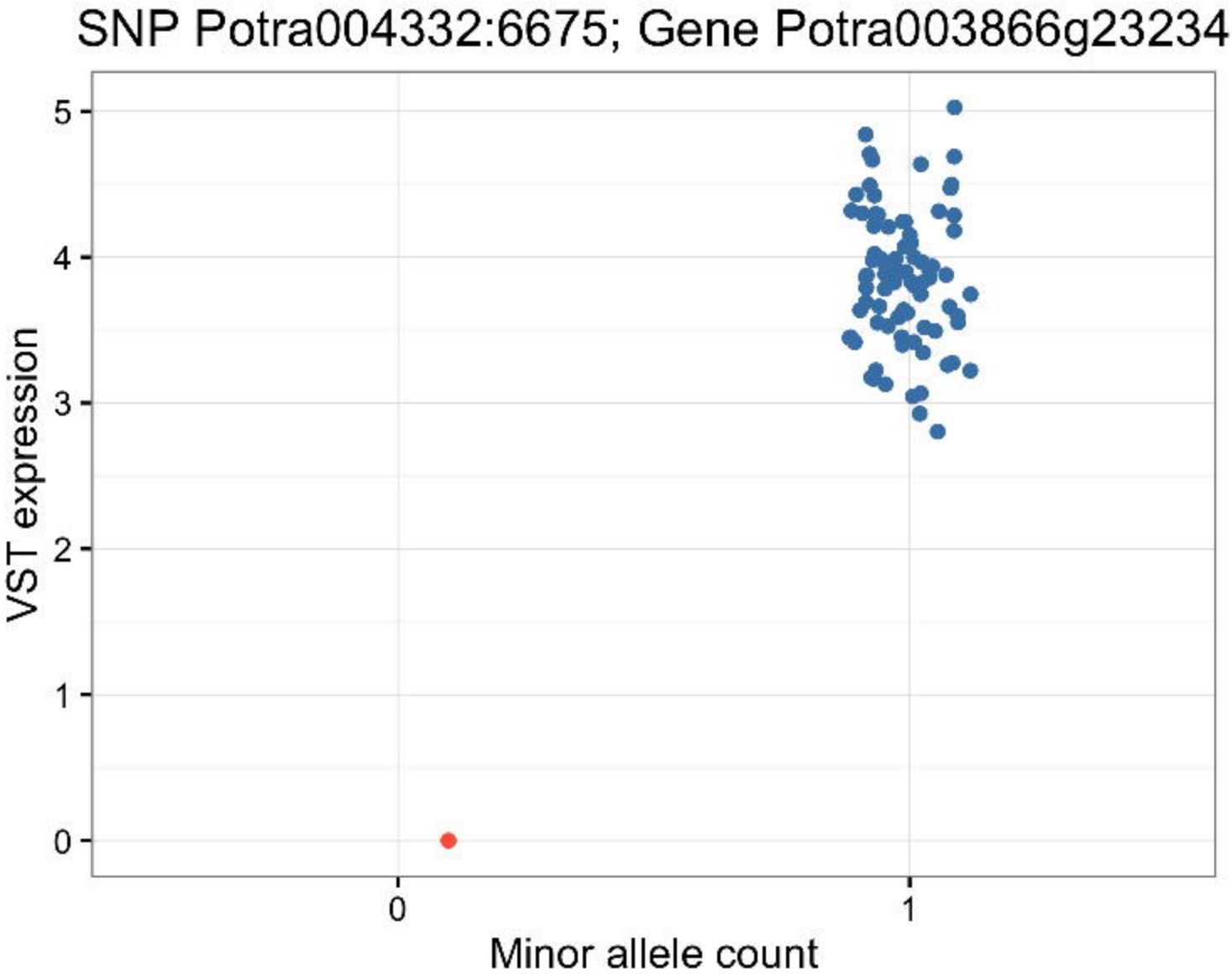

1.2
0.9
0.6
0.3
0.0
Density
0%
100%
200%
300%
Percentage of heritable variance explained by eSNPs
Both
Distant
Local

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