Comprehensive understanding of population structure and adaptation through graphical representation of gene–environment–trait associations

Reiichiro Nakamichi[1*], Shuichi Kitada[2], and Hirohisa Kishino[3*]

[1]Japan Fisheries Research & Education Agency, Yokohama 236-8648, Japan; [2]Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan; [3]Graduate School of Agriculture and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan.

*Correspondence: nakamichi@affrc.go.jp; kishino@lbm.ab.a.u-tokyo.ac.jp

ABSTRACT

A variable environment affects the physiological states of individuals and, in the long run, modifies their shapes. These changes, together with geographic barriers, generate population structure. Here, we propose a graphical representation of significant associations between genes, environments, and traits. A unique feature of the graph is the node of genome F_{ST}. The subnetwork around this node suggests the cause and the effects of population structure and segregation. A global structure of the graph enables to grasp a comprehensive picture of adaptation to the environment. Focused look at the neighbors of the environmental factors identifies the adaptive traits and the genetic background that supported the adaptation of the traits. Isolated nodes express genetic differentiations that are not explained by the population structure, implying the presence of some unrecognized environmental factor. We show the potential usefulness of our graphical representation by a detailed analysis of public dataset of wild poplar.

KEYWORDS

F_{ST}; environmental adaptation; genome scan; SNPs

Living organisms are adapted to their environment. This environmental adaptation can create significant differences in phenotypes and traits among populations of a species. For example, populations of sockeye salmon exhibit diversity in regards to life history traits such as spawning time and habitat, adult body size and shape, rearing time in freshwater and seawater, and adaptation to local spawning and rearing habitats within complex lake systems (Hilborn *et al.* 2003). Populations of walking-stick insects have diverged in body size, shape, host preference and behavior in parallel with the divergence of their host-plant species (Nosil *et al.* 2002). Aridity gradients may be the cause of geographically structured populations of Poaceae characterized by cytotype segregation of diploids and allotetraploids (Manzaneda *et al.* 2012). When correlated with variation in environmental factors over local populations, such variation in traits and phenotypes can offer an opportunity for understanding natural selection processes (Coop *et al.* 2010). Adaptation to environmental factors can change traits and phenotypes of a species, thereby creating population structure. Geographical isolation, which can lead to reproductive isolation and consequent differences in allele frequencies, also contributes to population structuring (Wright 1965). Population structure needs to be considered when analyzing correlations among genes, traits and environmental factors across population samples taken from a wide range of geographical regions.

Genome-wide association studies (GWASs) are widely used to identify associations between genes and traits/environments (Visscher *et al.* 2017). When data are obtained from a metapopulation exhibiting population structure, the effect of genotypes can be inferred by eliminating population structure effects (Devlin and Roeder 1999) to avoid spurious associations (Pritchard and Rosenberg 1999). One representative software program,

TASSEL (Yu *et al.* 2006; Bradbury *et al.* 2007), performs this type of analysis using a unified mixed model. Alternatively, associations can be tested in a Hardy-Weinberg population that has been decomposed from a structured population (Pritchard *et al.* 2000). Future challenges for large-scale GWASs from wild populations (wild GWASs) include the development of methods that take population structure into account (Santure and Garant 2018).

So-called "genome scan methods" consider geographically structured populations and detect SNPs related to environmental variables, traits and phenotypes (De Mita *et al.* 2013; De Villemereuil *et al.* 2014). For instance, BayeScan (Foll and Gaggiotti 2008) detects SNPs that create major differentiation in terms of global F_{ST} over a metapopulation. As illustrated in Figure 1A (top), 16 outliers were detected out of 281 SNPs in Atlantic herring in one study (Limborg *et al.* 2012); these outliers included a SNP in a heat-shock protein (HSP70) whose allele frequency was negatively correlated with mean sea surface salinities in spawning grounds (Figure 1A, bottom). As another example, Bayenv (Coop *et al.* 2010) and the latent factor mixed model (LFMM (Frichot *et al.* 2013)) can detect SNPs that are highly correlated with environmental factors and traits.

To obtain a comprehensive picture of population structure and adaptation using related genes, we propose a novel graphical representation of gene–environment–trait associations. The graph consists of a set of nodes and edges that connect pairs of nodes with significant association. Our graph describes correlations among allele frequencies of SNPs, states of traits, and environmental and location factors. The unique feature of our method is the use of a genome-wide population differentiation node, which enables inference of the

84 determinants of population structure. Environmental factor nodes around this node may be
85 the causal force for the population structure, whereas the among-locality variation of
86 nearby traits may be the result of population differentiation, or vice versa.
87
88 In the conceptual figure of Figure 1B, a location factor, L1, is correlated with E1, an
89 environmental factor that is correlated in turn with genome F_{ST}. Two traits, T1 and T2, are
90 affected by this environmental factor. G4 and G6 are the candidate genes behind the
91 differentiation of T1. Likewise, G9, G10 and G11 are the candidate genes for T2.
92 Population structure (genome F_{ST}) may have differentiated according to some unknown
93 traits related to G1, G2 and G3, as well as trait T2. By examining the functions of genes G7
94 and G8, inference of the traits selected by environmental factor E1 may be possible. Of
95 interest, the hidden factor that differentiates gene G5 can be investigated by plotting the
96 allele frequencies of G5 relative to location factor L2. In this way, our method provides a
97 comprehensive perspective for understanding the genetic and ecological mechanisms of
98 environmental adaptation of a species.
99

100 ## Materials and Methods

101 ***Significance of gene–environment–trait associations***

102 The node of genome F_{ST} is in the form of a distance matrix between pairs of local
103 populations. Likewise, all other nodes of SNPs, traits and environmental factors are
104 represented by matrices whose elements are the differences between pairs of local
105 populations (Supplementary Figure S1). Consequently, the correlation between a pair of
106 nodes is the correlation between the between-population distance matrices. The

significance of correlation between a pair of nodes is measured by simple linear regression analysis. Here, the dependent variable is a distance matrix of a node, and the explanatory variable is the distance matrix of the other node. To take account of correlations in the error term, we carried out bootstrap resampling of populations and individuals. For each of the bootstrap datasets, we calculated among-population distance matrices for each node pair, and obtained the regression coefficient for each node pair. The z value is the ratio of the original regression coefficient to its bootstrap standard deviation. By applying the Benjamini-Hochberg method (Benjamini and Hochberg 1995) to these *p*-values, we selected significant correlations with a false discovery rate (FDR) of 0.01. A node pair with a significant correlation was connected by an edge. We note that these edges represent the total associations of direct and indirect effects.

Estimation of pairwise F_{ST}

A locus pairwise F_{ST} at a single marker is the normalized difference of the allele frequencies and measures the genetic differentiation between a pair of local populations. To capture the fine-scale population structure even under high gene flow, we adopted an empirical Bayes estimator using EBFST function of the R package FinePop (Kitada *et al.* 2017). By averaging both numerators and denominators over multiple markers, we obtained genome F_{ST}. Genome F_{ST} indicates the magnitude of population differentiation over the genome, while locus F_{ST} indicates the contribution of each gene to population differentiation.

Application to wild poplar data

We demonstrate how the graphical representation provides a comprehensive picture of

131 population differentiation and environmental adaptation by analyzing a publicly available
132 data. It contains genetic and trait information of 445 individuals of wild poplar (*Populus*
133 *trichocarpa*), which were collected from various regions over a range of 2,500 km, near the
134 Canadian–US border at a latitude of 44′ to 59′ N, a longitude of 121′ to 138′ W, and an
135 altitude of 0 to 800 m (McKown *et al.* 2014). The data included genotypes of 34,131
136 SNPs (3,516 genes) and values of stomatal anatomy, leaf tannin, ecophysiology,
137 morphology and disease. Here, we focused on the four traits: adaxial stomata density
138 (ADd), abaxial stomata density (ABd), and leaf rust disease morbidity (AUDPC) measured
139 in 2010 and 2011 (DP10 and DPC11, respectively). Each sampling location was described
140 by 11 environmental/geographical variables: latitude (lat), longitude (lon), altitude (alt),
141 longest day length (DAY), frost-free days (FFD), mean annual temperature (MAT), mean
142 warmest month temperature (MWMT), mean annual precipitation (MAP), mean summer
143 precipitation (MSP), annual heat-moisture index (AHM) and summer heat-moisture index
144 (SHM).
145
146 We performed a clustering analysis using the geographical distribution and divided the 445
147 individuals into subpopulations. We applied model-based clustering (Fraley and Raftery
148 2016) with three types of spatial information—latitude, longitude and altitude—as the
149 explanation variables. Using the Bayesian information criterion based on the mclustBIC
150 function in the R package mclust (Scrucca *et al.* 2016) under the VEV model (ellipsoidal,
151 equal shape), we obtained 22 subpopulations: 5 in northern British Colombia (NBC), 11 in
152 southern British Colombia (SBC), 3 in inland British Colombia (IBC) and 3 in Oregon
153 (ORE).

154

155 ***Marker screening before analysis of the graphical representation***

156 Because our major concern was identifying correlations between among-population
157 differentiations of genes, traits and environmental factors, we selected the SNP with the
158 highest global F_{ST} value over 22 populations, designated as the tag SNP, from each of the
159 3,516 gene regions. Out of the 3,516 tag SNPs, only those that were differentiated among
160 populations were subjected to the graphical representation analysis. We note that the scaled
161 global F_{ST} values, calculated as:

162 $$\widetilde{F_{ST}} = \frac{F_{ST}}{\frac{median(F_{ST})}{k - \frac{2}{3} + \frac{4}{27k} - \frac{8}{729k^2}}}$$

163 approximately follow a chi-squared distribution with degree of freedom k (= the number of
164 populations – 1) (Weir and Hill 2002). We performed chi-squared tests on the 3,516 genes
165 and identified 507 tag SNPs with significant differentiation among populations ($p < 0.05$).
166 Therefore, we used a total of 523 variables: 4 traits, 11 location/environmental factors,
167 genome F_{ST} and 507 genes (Supplementary Table S1).

168

169 ## Results and Discussion

170 ***Global structure of the network***

171 Our generated network then identified relationships between genome F_{ST}, 8 environmental
172 and 2 location factors, 4 traits and 317 genes (Figure 2A, Supplementary Table S2). The
173 network consisted of a large cluster centered around genome F_{ST} along with several
174 isolated small clusters (Figure 2A, Supplementary Table S2). The location and
175 environmental factors lat, lon, MAT and DAY were directly connected to genome F_{ST} in the

176 estimated network, whereas alt was not included in the graph. In contrast, four water-
177 related factors, MAP, MSP, AHM and SHM, were several steps away from genome F_{ST}.
178 Several isolated clusters of genes were present at the boundary of the network. Although
179 these clusters were differentiated between local populations, the absence of a significant
180 correlation with genome F_{ST} implies that the diversity of these traits was not simply the
181 result of population differentiation, but was instead due to adaptation to the local
182 environment.
183
184 ***Determinants of population structure***
185 The radius-one neighborhood of genome F_{ST} suggested that temperature and day length
186 were the main environmental factors causing population structure, with the observation that
187 the edges of the graph collected significant correlations (Supplementary Figure S2).
188 Assisted by the Entrez summaries and GO terms of any genes shown in the graph window,
189 we found that many genes related to fertility were affected by the population structure
190 (Supplementary Figure S2, Table S3). An example was SHT (spermidine
191 hydroxycinnamoyl transferase), which is related to pollen development and pollen exine
192 formation (Grienenberger *et al.* 2009). The scatter plot visualized in the graph window
193 provided information that correlated genome F_{ST} with the SHT gene. Other fertility-related
194 genes included MYB5 (myb domain protein 5), HAB1 (hyper sensitive to ABA1) and
195 ACT7 (actin 7) functioning in seed germination (Li *et al.* 2009; Saez *et al.* 2006; Gilliland
196 *et al.* 2003), AT3G08640 (alphavirus core family protein, DUF3411) and HOG1 (S-
197 adenosyl-L-homocysteine hydrolase) involved in embryo development (Rocha *et al.* 2005),
198 LUG (transcriptional corepressor LEUNIG) and VRN1 (AP2/B3-like transcriptional factor
199 family protein) related to flower development (Conner and Liu 2000; Levy *et al.* 2002) and

200 REV (homeobox-leucine zipper family protein/lipid-binding START domain-containing
201 protein) associated with flower morphogenesis (Talbert *et al.* 1995).
202
203 ***Daylight, latitude, stomatal density and disease***
204 Consistent with McKown et al. (McKown *et al.* 2014), our network confirmed a strong
205 connection between ADd and disease progress (DP10 and DP11) (Figure 2B). In contrast,
206 ABd was not directly connected to DPs, but exhibited a strong connection to DAY, as did
207 DPs. All these nodes were directly connected to genome F_{ST}.
208
209 Average ADd was constant in the southern region up to 50° N, but increased with latitude
210 in the northern region (Figure 3A). In contrast, average ABd decreased with latitude in the
211 northern region (Figure 3B). DAY, which occurs in early summer, increased monotonically
212 with latitude (Figure 3C), while MAT decreased on average with latitude (Figure 3D). This
213 result indicates that poplar trees in northern populations experience longer and weaker
214 sunshine in summer but drop their leaves earlier. Interestingly, the pore size of abaxial
215 stomata was larger at lower values of ABd (Figure 4A), which demonstrates that northern
216 populations had larger abaxial stomata, but their density was lower because the leaf area
217 was limited. In contrast, the large variation in the pore size of adaxial stomata displayed no
218 relationship with ADd (Figure 4B). The presence of larger stomata causes leaves to have a
219 lower stomatal density but a greater photosynthetic efficiency (Lawson and Blatt 2014).
220 These results suggest that northern populations must increase photosynthetic efficiency to
221 adapt to an environment with weak sunshine and a shorter period before leaf shed. The
222 increased ADd of northern populations suggests that adaxial stomata compensate for the
223 decrease in abaxial stomata. Stomatal closure is part of the innate immune response to

bacterial invasion (Melotto *et al.* 2006). An increase in abaxial stomata size and adaxial stomata density might increase the risk of disease invasion. Our results suggest that wild poplar can expand its habitat northward by increasing photosynthetic capacity while heightening its risk of disease, although the latter is less significant in northern areas (McKown *et al.* 2014). This ecological trade-off may be a cause of the population structure of wild poplar.

Photosynthesis and circadian rhythm in response to day length

Geraldes et al. (Geraldes *et al.* 2014) have identified a large number of F_{ST} outliers that are overrepresented in genes involved in circadian rhythm and response to red/far-red light. In our graph, the allele frequencies of genes related to photosynthesis and the circadian cycle were found to be influenced by day length (Figure 5, Supplementary Table S4). For example, ACT7 (actin 7) is related to response to light stimulus (McDowell *et al.* 1996), and its allele frequencies were negatively correlated with DAY and lon (Figure 5, lower left). Geographical mapping of ATC7 allele frequencies and day length confirmed this correlation (Figure 6A). Other genes included PRR7 (pseudo-response regulator 7) and TOC1 (CCT motif-containing response regulator protein) related to circadian rhythm (Farré *et al.* 2005; Alabadí *et al.* 2001), APX2 (ascorbate peroxidase 2) associated with response to high light intensity and response to oxidative stress (Karpinski *et al.* 1997), EXPA1 (expansin A1) involved in response to red light (Esmon *et al.* 2006) and SUS4 (sucrose synthase 4) related to the carbon assimilation process (Bieniawska *et al.* 2007). These results suggest that day length is the most important factor controlling photosynthesis and that latitude causes differentiation of photosynthetic genes. Finally, the allele frequencies of SYP121 (syntaxin of plant 121) related to stomatal movement (Bassham and Blatt 2008)

248 and PIP3 (plasma membrane intrinsic protein 3) participating in response to abscisic acid
249 and water channel activity (Weig *et al.* 1997) were also significantly correlated with day
250 length (figures not shown).
251
252 ***Damage response, the circadian system and stomata related to disease susceptibility***
253 Morbidity due to leaf rust disease (DP10 and DP11) showed a close relationship to adaxial
254 stomatal density (ADd) and day length (DAY) (Figs. 2B and 5, Supplementary Table S5).
255 Genes closely connected to DAY, such as ACT7 (related to response to wounding), PRR7,
256 APX2 and PIP3, were also closely linked to morbidity. Other genes, namely, FHY3 (far-red
257 elongated hypocotyls 3) related to circadian rhythm (Allen *et al.* 2006) and DRT100 (DNA-
258 damage repair/toleration 100) functioning in DNA repair (Pang *et al.* 1993), also were
259 involved in this cluster. Because the DAY-related genes control stomatal opening and
260 closing, our subgraph (Figure 5) suggests that fungal invasion into tissues occurs through
261 stomata (Melotto *et al.* 2006). SHT (spermidine hydroxycinnamoyl transferase) was closely
262 connected to leaf rust disease morbidity. As described above, SHT is related to pollen
263 development and connected to genome F_{ST}. In addition, spermidine is known as a
264 modulator of the immune process (Theoharides 1980). This result thus implies that the
265 functions of SHT in immune and reproduction play important roles in population
266 differentiation and adaptation through disease resistance. DRT100 allele frequencies were
267 negatively correlated with DP11 (figure not shown), and the geographical gradients of
268 DRT100 allele frequencies and DP11 well explained the correlation (Supplementary Figure
269 S3A). This result suggests that leaf rust disease affects fertility and promotes population
270 differentiation. Principal component analysis using these genes, which were neighbors of
271 ADd, DP10 and DP11, clearly revealed differences in morbidity between locations from

north to south (Supplementary Figure S4). This result implies that the phenotypes controlled by circadian and light-responsive genes have adapted to local environments according to latitude and day length and are responsible for the morbidity-related population differentiation.

Body growth affected by temperature

Genes in the subgraph around MAT and FFD were those involved in shoot development (Supplementary Figure S5, Table S6), such as LAS (lateral suppressor, GRAS family transcription factor) related to secondary shoot formation (Greb *et al.* 2003) and REV (homeobox-leucine zipper family protein/lipid-binding START domain-containing protein) linked to primary shoot apical meristem specification and leaf morphogenesis. LAS allele frequencies were negatively correlated with MAT (Supplementary Figure S5, lower left). The geographical gradients of LAS allele frequencies and MAT supported this correlation (Supplementary Figure S3b). These results imply that temperature strongly supports body growth of poplar.

Drought stress resistance depends on water conditions

The environmental factors MAP, MSP, AHM and SHM exhibited no direct connection to genome F_{ST} (Figure 2A, Supplementary Figure S6, Table S7). An indirect connection was apparent, however, through genes with functions related to water stress. These genes were XERICO (RING/U-box superfamily protein), HK2 (histidine kinase 2) and ABA1 (ABA deficient 1, zeaxanthin epoxidase) involved in response to osmotic stress and response to salt stress (Ko *et al.* 2006; Tran *et al.* 2007; Xiong *et al.* 2002), CBF4 (C-repeat binding factor 4) related to response to drought (Haake *et al.* 2002) and AGP14 (arabinogalactan

296 protein 14) participating in root hair elongation (Lin *et al.* 2011). A close examination of
297 CBF4, directly connected to AHM, revealed that its allele frequencies were negatively
298 correlated with AMH and clustered by geographical groups (Supplementary Figure S6,
299 lower left). CBF4 allele frequencies were particularly differentiated in IBC where AHM
300 was high (Figure 6B). This result suggests that the CBF4 gene has differentiated as an
301 adaptive response to dry weather. The apparent weak relationship between water stress and
302 F_{ST} may be a consequence of the relatively small differences in water conditions in this
303 dataset.
304
305 ***Vernalization depends on some unknown environmental conditions***
306 Several isolated gene clusters, which were unconnected to genome F_{ST}, environmental
307 factors or traits, appeared in the global network (Figure 2A). Each cluster contained genes
308 whose functions were strongly related. For example, the largest isolated cluster consisted of
309 vernalization genes (Supplementary Figure S7, Table S8), such as FUS6 related to
310 regulation of flower development and seed germination (Chory *et al.* 1996), GA3OX1
311 associated with response to gibberellin and response to red light (McGinnis *et al.* 2003) and
312 VRN1 linked to vernalization response and regulation of flower development. Although
313 these genes may not be directly responsible for population structure, the appearance of the
314 isolated cluster in the network implies a latent relationship between vernalization and
315 population differentiation. In regards to the geographical distribution of their allele
316 frequencies, FUS6 and GA3OX1 had similar, complicated patterns (Figure 6C,
317 Supplementary Figure S3C). Populations in SBC and eastern IBC had a similar pattern of
318 allele frequencies, while those in northern NBC, southern ORE and western IBC displayed
319 a different pattern. This result may imply an adaptation to a microenvironment not observed

320 in this data. For example, the direction of a mountain slope can create different habitats
321 with different daylight conditions.
322
323 ***Gene ontology enrichment analysis***
324 No significant GOs were predicted by gene ontology enrichment analysis (Subramanian *et*
325 *al.* 2005; Alexa *et al.* 2006) for the union of the set of 317 genes selected for the graph and
326 the sets of neighboring genes mentioned above relative to the other complementary gene
327 sets. To obtain a comprehensive picture based on solid evidence, we focused on
328 geographically differentiating SNPs and selected pairs of nodes by controlling FDR. As a
329 consequence, we may have diminished the ability to identify differences between the two
330 sets of genes. Alternatively, mutations in a few members of the relevant pathways may have
331 enabled adaptation to the variable environments.
332
333 Our method identifies genes related to environmental adaptation with a FDR of 1% and
334 visualizes their network, including genome F_{ST}, environmental and location factors, and
335 traits. Our example using wild poplar has revealed the potential of our graphical model
336 representation to aid comprehensive understanding of ecological and genetic mechanisms
337 underlying environmental adaptation and population structuring. While conventional
338 GWAS and genome scanning effectively search for genes related to some given factors or
339 traits, our method captures the overall picture of the relationship among genes,
340 environmental factors and traits in association with population structure. By following the
341 sub-network of genes around target environmental factors and traits, we can obtain a
342 detailed understanding of the relationship of genes behind environmental adaptation and

343 population differentiation. In particular, detection of collaboratively adapted gene clusters,
344 which are not directly associated with the given environment/trait factors, is an advantage
345 of our graphical representation. Our R software module GET.graph aids this process by
346 displaying subgraphs and scatter plots of allele frequencies of genes vs. environmental
347 factors/traits. GET.graph retains the biological functions of genes retrieved from public
348 databases, such as GO and ENTREZ, and helps us smoothly interpret the graph. Through
349 this process, we can reach comprehensive understanding of population structure and
350 adaptation by characterizing the sub-networks of the graph (Figure 2a).
351
352 Our graph collects significant correlations that sum up both direct and indirect
353 relationships, while partial correlations extract direct relationships (Kishino and Waddell
354 2000; De La Fuente *et al.* 2004; Liu 2013). Collection of significant partial correlations in
355 this setting is left for future study. Finally, we must be aware of computational feasibility.
356 The calculation load greatly increases depending on the number of variables and is roughly
357 proportional to the square of the number of variables. The analysis for this paper took
358 several hours on an Intel Core i7 (6 core) workstation. The above step of prescreening
359 variables is therefore indispensable. As a final remark, the data, especially genomic data,
360 often include missing values. Our graphical representation method describes relationships
361 between population means of allele frequencies, trait values and environmental/location
362 factors; therefore, like Bayenv (Coop *et al.* 2010), it analyzes sample means among
363 measured data. As long as the means of measured allele/environmental/trait variables are
364 unbiased estimates of the corresponding sample means, the procedure is also unbiased.
365
366 The R software module (GET.graph) that implements the network analysis described in this

paper is available in the FinePop package at CRAN (https://CRAN.R-project.org/package=FinePop).

Acknowledgements

This study was supported by the Japan Society for the Promotion of Science Grant-in-Aid for Scientific Research 25280006 and 16H02788 to HK and 18K05781 to SK.

Literature Cited

Alabadí, D., T. Oyama, M. J. Yanovsky, F. G. Harmon, P. Más *et al.*, 2001 Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. Science 293: 880–883. https://doi.org/10.1126/science.1061320

Alexa, A., J. Rahnenführer, and T. Lengauer, 2006 Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22: 1600–1607. https://doi.org/10.1093/bioinformatics/btl140

Allen, T., A. Koustenis, G. Theodorou, D. E. Somers, S. A. Kay *et al.*, 2006 Arabidopsis FHY3 specifically gates phytochrome signaling to the circadian clock. Plant Cell 18: 2506–2516. https://doi.org/10.1105/tpc.105.037358

Bassham, D. C., and M. R. Blatt, 2008 SNAREs: cogs and coordinators in signaling and development. Plant Physiol. 147: 1504–1515. https://doi.org/10.1104/pp.108.121129

390 Benjamini, Y., and Y. Hochberg, 1995 Controling the false discovery rate: a
391 practical and powerful approach to multiple testing. J. Royal Stat. Soc. 57:
392 289–300. https://doi.org/10.2307/2346101
393 Bieniawska, Z., D. H. Paul Barratt, A. P. Garlick, V. Thole, N. J. Kruger *et al.*,
394 2007 Analysis of the sucrose synthase gene family in Arabidopsis. Plant J.
395 49: 810–828. https://doi.org/10.1111/j.1365-313X.2006.03011.x
396 Bradbury, P. J., Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss *et al.*,
397 2007 TASSEL: software for association mapping of complex traits in
398 diverse samples. Bioinformatics 23: 2633–2635.
399 https://doi.org/10.1093/bioinformatics/btm308
400 Chory, J., M. Chatterjee, R. K. Cook, T. Elich, C. Fankhauser *et al.*, 1996 From
401 seed germination to flowering, light controls plant development via the
402 pigment phytochrome. Proc. Natl. Acad. Sci. 93: 12066-12071.
403 https://doi.org/10.1073/pnas.93.22.12066
404 Conner, J., and Z. Liu, 2000 LEUNIG, a putative transcriptional corepressor
405 that regulates AGAMOUS expression during flower development. Proc.
406 Natl. Acad. Sci. 97: 12902-12907. https://doi.org/10.1073/pnas.230352397
407 Coop, G. D., D. Witonsky, A. Di Rienzo, and K. J. Pritchard, 2010 Using
408 environmental correlations to identify loci underlying local adaptation.
409 Genetics 185: 1411–1423. https://doi.org/10.1534/genetics.110.114819
410 De La Fuente, A., N. Bing, I. Hoeschele, and P. Mendes, 2004 Discovery of
411 meaningful associations in genomic data using partial correlation

412 coefficients. Bioinformatics 20: 3565–3574.
413 https://doi.org/10.1093/bioinformatics/bth445
414 De Mita, S., A. C. Thuillet, L. Gay, N. Ahmadi, S. Manel *et al.*, 2013 Detecting
415 selection along environmental gradients: analysis of eight methods and
416 their effectiveness for outbreeding and selfing populations. Mol. Ecol. 22:
417 1383–1399. https://doi.org/10.1111/mec.12182
418 De Villemereuil, P., É. Frichot, É. Bazin, O. François, and O. E. Gaggiotti, 2014
419 Genome scan methods against more complex models: when and how much
420 should we trust them? Mol. Ecol. 23: 2006–2019.
421 https://doi.org/10.1111/mec.12705
422 Devlin, B., and K. Roeder, 1999 Genomic control for association studies.
423 Biometrics 55: 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x
424 Esmon, C. A., A. G. Tinsley, K. Ljung, G. Sandberg, L. B. Hearne *et al.*, 2006 A
425 gradient of auxin and auxin-dependent transcription precedes tropic
426 growth responses. Proc. Natl. Acad. Sci. 103: 236–241.
427 https://doi.org/10.1073/pnas.0507127103
428 Farré, E. M., H. S. L, F. Harmon, M. J. Yanovsky, and S. A. Kay, 2005
429 Overlapping and distinct roles of PRR7 and PRR9 in the Arabidopsis
430 circadian clock. Curr. Biol. 15: 47–54.
431 https://doi.org/10.1016/j.cub.2004.12.067
432 Foll, M., and O. Gaggiotti, 2008 A genome scan method to identify selected loci
433 appropriate for both dominant and codominant markers: a Bayesian

perspective. Genetics 180: 977-993. https://doi.org/10.1534/genetics.108.092221

Fraley, C., and A. E. Raftery, 2016 Model-based clustering, discriminant analysis and density estimation. BMC Bioinformatics 17: 287. https://doi.org/10.1198/016214502760047131

Frichot, E., S. D. Schoville, G. Bouchard, and O. François, 2013 Testing for associations between loci and environmental gradients using latent factor mixed models. Mol. Biol. Evol. 30: 1687–1699. https://doi.org/10.1093/molbev/mst063

Geraldes, A., N. Farzaneh, C. J. Grassa, A. D. McKown, R. D. Guy *et al.*, 2014 Landscape genomics of Populus trichocarpa the role of hybridization limited gene flow and natural selection in shaping patterns of population structure. Evolution 68: 3260–3280. https://doi.org/10.1111/evo.12497

Gilliland, L. U., L. C. Pawloski, M. K. Kandasamy, and R. B. Meagher, 2003 Arabidopsis actin gene ACT7 plays an essential role in germination and root growth. Plant J. 33: 319–328. https://doi.org/10.1046/j.1365-313X.2003.01626.x

Greb, T., O. Clarenz, E. Schäfer, D. Müller, R. Herrero *et al.*, 2003 Molecular analysis of the LATERAL SUPPRESSOR gene in Arabidopsis reveals a conserved control mechanism for axillary meristem formation. Genes Dev. 17: 1175–1187. https://doi.org/10.1101/gad.260703

Grienenberger, E., S. Besseau, P. Geoffroy, D. Debayle, D. Heintz *et al.*, 2009 A BAHD acyltransferase is expressed in the tapetum of Arabidopsis anthers

457 and is involved in the synthesis of hydroxycinnamoyl spermidines. Plant J.
458 58: :246–259. https://doi.org/10.1111/j.1365-313X.2008.03773.x
459 Haake, V., D. Cook, J. L. Riechmann, O. Pineda, M. F. Thomashow *et al.*, 2002
460 Transcription factor CBF4 is a regulator of drought adaptation in
461 Arabidopsis. Plant Physiol. 130: 639–648. https://doi.org/10.1104/pp.006478
462 Hilborn, R., T. P. Quinn, D. E. Schindler, and D. E. Rogers, 2003 Biocomplexity
463 and fisheries sustainability. Proc. Natl. Acad. Sci. 100: 6564–6568.
464 https://doi.org/10.1073/pnas.1037274100
465 Karpinski, S., C. Escobar, B. Karpinska, G. Creissen, and P. M. Mullineaux,
466 1997 Photosynthetic electron transport regulates the expression of
467 cytosolic ascorbate peroxidase genes in Arabidopsis during excess light
468 stress. Plant Cell 9: 627–640. https://doi.org/10.1105/tpc.9.4.627
469 Kishino, H., and P. J. Waddell, 2000 Correspondence analysis of genes and
470 tissue types and finding genetic links from microarray data. Genome
471 Inform Ser Workshop Genome Inform. 11: 83–95.
472 Kitada, S., R. Nakamichi, and H. Kishino, 2017 The empirical Bayes estimators
473 of fine-scale population structure in high gene flow species. Mol. Ecol.
474 Resour. 17: 1210–1222. https://doi.org/10.1111/1755-0998.12663
475 Ko, J. H., S. H. Yang, and K. H. Han, 2006 Upregulation of an Arabidopsis
476 RING-H2 gene, XERICO, confers drought tolerance through increased
477 abscisic acid. Plant J. 47: 343–355. https://doi.org/10.1111/j.1365-
478 313X.2006.02782.x

479 Lawson, T., and M. R. Blatt, 2014 Stomatal size, speed, and responsiveness
480 impact on photosynthesis and water use efficiency. Plant Physiol. 164:
481 1556–1570. https://doi.org/10.1104/pp.114.237107
482 Levy, Y. Y., S. Mesnage, J. S. Mylne, A. R. Gendall, and C. Dean, 2002 Multiple
483 roles of Arabidopsis VRN1 in vernalization and flowering time control.
484 Science 297: 243–246. https://doi.org/10.1126/science.1072147
485 Li, S. F., O. N. Milliken, H. Pham, R. Seyit, R. Napoli *et al.*, 2009 The
486 Arabidopsis MYB5 transcription factor regulates mucilage synthesis, seed
487 coat development, and trichome morphogenesis. Plant Cell 21: 72–89.
488 https://doi.org/10.1105/tpc.108.063503
489 Limborg, M. T., S. J. Helyar, M. De Bruyn, M. I. Taylor, E. E. Nielsen *et al.*,
490 2012 Environmental selection on transcriptome - derived SNPs in a high
491 gene flow marine fish, the Atlantic herring (Clupea harengus). Mol. Ecol.
492 21: 3686–3703. https://doi.org/10.1111/j.1365-294X.2012.05639.x
493 Lin, W. D., Y. Y. Liao, T. J. Yang, C. Y. Pan, T. J. Buckhout *et al.*, 2011
494 Coexpression-based clustering of Arabidopsis root genes predicts functional
495 modules in early phosphate deficiency signaling. Plant Physiol. 155: 1383–
496 1402. https://doi.org/10.1104/pp.110.166520
497 Liu, W., 2013 Gaussian graphical model estimation with false discovery rate
498 control. Ann. Stat. 41: 2948–2978. https://doi.org/10.1214/13-AOS1169
499 Manzaneda, A. J., P. J. Rey, J. M. Bastida, C. Weiss-Lehman, E. Raskin *et al.*,
500 2012 Environmental aridity is associated with cytotype segregation and

polyploidy occurrence in Brachypodium distachyon (Poaceae). New Phytol. 193: 797–805. https://doi.org/10.1111/j.1469-8137.2011.03988.x

McDowell, L. M., Y. An, S. Huang, E. C. McKinney, and R. B. Meagher, 1996 The Arabidopsis ACT7 actin gene is expressed in rapidly developing tissues and responds to several external stimuli. Plant Physiol. 111: 699–711. https://doi.org/10.1104/pp.111.3.699

McGinnis, K. M., S. G. Thomas, J. D. Soule, L. C. Strader, J. M. Zale *et al.*, 2003 The Arabidopsis SLEEPY1 gene encodes a putative F-Box subunit of an SCF E3 ubiquitin ligase. Plant Cell 15: 1120–1130. https://doi.org/10.1105/tpc.010827

McKown, A. D., R. D. Guy, L. Quamme, J. Klápště, J. La Mantia *et al.*, 2014 Association genetics, geography and ecophysiology link stomatal patterning in Populus trichocarpa with carbon gain and disease resistance trade-offs. Mol. Ecol. 23: 5771–5790. https://doi.org/10.1111/mec.12969

Melotto, M., W. Underwood, J. Koczan, K. Nomura, and S. Y. He, 2006 Plant stomata function in innate immunity against bacterial invasion. Cell 126: 969–98. https://doi.org/10.1016/j.cell.2006.06.054

Nosil, P., B. J. Crespi, and C. P. Sandoval, 2002 Host-plant adaptation drives the parallel evolution of reproductive isolation. Nature 417: 440–443. https://doi.org/10.1038/417440a

Pang, Q., J. B. Hays, I. Rajagopal, and T. S. Schaefer, 1993 Selection of Arabidopsis cDNAs that partially correct phenotypes of Escherichia coli DNA-damage-sensitive mutants and analysis of two plant cDNAs that

524 appear to express UV-specific dark repari activities. Plant Mol. Biol. 22:
525 411–426. https://doi.org/10.1007/BF00015972
526 Pritchard, J. K., and N. A. Rosenberg, 1999 Use of unlinked genetic markers to
527 detect population stratification in association studies. Am. J. Hum. Genet.
528 65: 220–228. https://doi.org/10.1086/302449
529 Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly, 2000
530 Association mapping in structured populations. Am. J. Hum. Genet. 67:
531 170–181. https://doi.org/10.1086/302959
532 Rocha, P. S., M. Sheikh, R. Melchiorre, M. Fagard, S. Boutet *et al.*, 2005 The
533 Arabidopsis HOMOLOGY-DEPENDENT GENE SILENCING1 gene codes
534 for an S-adenosyl-L-homocysteine hydrolase required for DNA
535 methylation-dependent gene silencing. Plant Cell 17: 404–417.
536 https://doi.org/10.1105/tpc.104.028332
537 Saez, A., N. Robert, M. H. Maktabi, J. I. Schroeder, R. Serrano *et al.*, 2006
538 Enhancement of abscisic acid sensitivity and reduction of water
539 consumption in Arabidopsis by combined inactivation of the protein
540 phosphatases type 2C ABI1 and HAB1. Plant Physiol. 141: 1389–1399.
541 https://doi.org/10.1104/pp.106.081018
542 Santure, A. W., and D. Garant, 2018 Wild GWAS - association mapping in
543 natural populations. Mol. Ecol. Resour. 18: 729–738.
544 https://doi.org/10.1111/1755-0998.12901

545 Scrucca, L., M. Fop, T. B. Murphy, and A. E. Raftery, 2016 mclust 5: clustering,
546 classification and density estimation using Gaussian finite mixture
547 models. R J. 8: 205–233.
548 Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.*,
549 2005 Gene set enrichment analysis: a knowledge-based approach for
550 interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102:
551 15545–15550. https://doi.org/10.1073/pnas.0506580102
552 Talbert, P. B., H. T. Adler, D. W. Parks, and L. Comai, 1995 The REVOLUTA
553 gene is necessary for apical meristem development and for limiting cell
554 divisions in the leaves and stems of Arabidopsis thaliana. Development
555 121: 2723–2735.
556 Theoharides, T. C., 1980 Polyamines spermidine and spermine as modulators of
557 calcium-dependent immune processes. Life Sci. 27: 703–713.
558 https://doi.org/10.1016/0024-3205(80)90323-9
559 Tran, L. S., T. Urao, F. Qin, K. Maruyama, T. Kakimoto *et al.*, 2007 Functional
560 analysis of AHK1/ATHK1 and cytokinin receptor histidine kinases in
561 response to abscisic acid, drought, and salt stress in Arabidopsis. Proc.
562 Natl. Acad. Sci. 104: 20623–20628.
563 https://doi.org/10.1073/pnas.0706547105
564 Visscher, P. M., N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy *et al.*, 2017 10
565 years of GWAS discovery: biology, function, and translation. Am. J. Hum.
566 Genet. 101: 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

567 Weig, A., C. Deswarte, and M. J. Chrispeels, 1997 The major intrinsic protein
568 family of Arabidopsis has 23 members that form three distinct groups with
569 functional aquaporins in each group. Plant Physiol. 114: 1347–1357.
570 https://doi.org/10.1104/pp.114.4.1347
571 Weir, B. S., and W. G. Hill, 2002 Estimating F-statistics. Annu. Rev. Genet. 36:
572 721–750. https://doi.org/10.1146/annurev.genet.36.050802.093940
573 Wright, S., 1965 The interpretation of population structure by F - statistics
574 with special regard to systems of mating. Evolution 19: 395–420.
575 https://doi.org/10.2307/2406450
576 Xiong, L., H. Lee, M. Ishitani, and J. K. Zhu, 2002 Regulation of osmotic stress-
577 responsive gene expression by the LOS6/ABA1 locus in Arabidopsis. J.
578 Biol. Chem. 277: 8588-8596. https://doi.org/10.1074/jbc.M109275200
579 Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki *et al.*, 2006 A unified
580 mixed-model method for association mapping that accounts for multiple
581 levels of relatedness. Nat. Genet. 38: 203–208.
582 https://doi.org/10.1038/ng1702
583

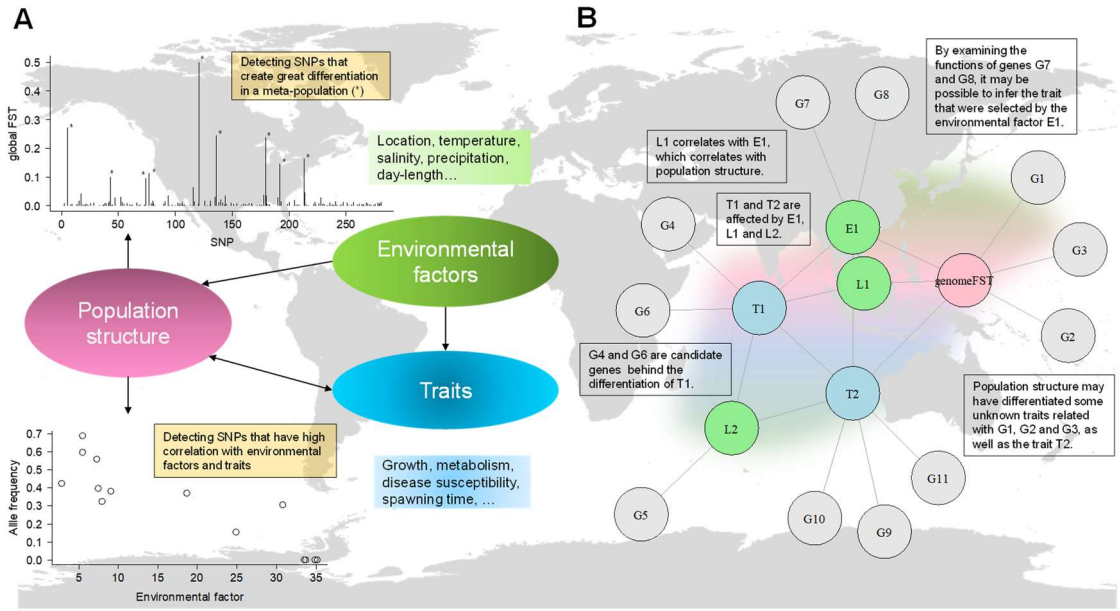

Figure 1 Conceptual diagram of the detection of genes controlling environmental adaptation. (A) Representative genome scan methods. (B) A network-based method enabling comprehensive understanding of environmental adaptation of traits and genes that leads to population structure. E, environmental factors, such as temperature, daylength, precipitation and salinity. L, location factors, such as longitude, latitude, altitude and geographical distance. T, traits such as height, size, metabolism, disease susceptibility and reproductive season.

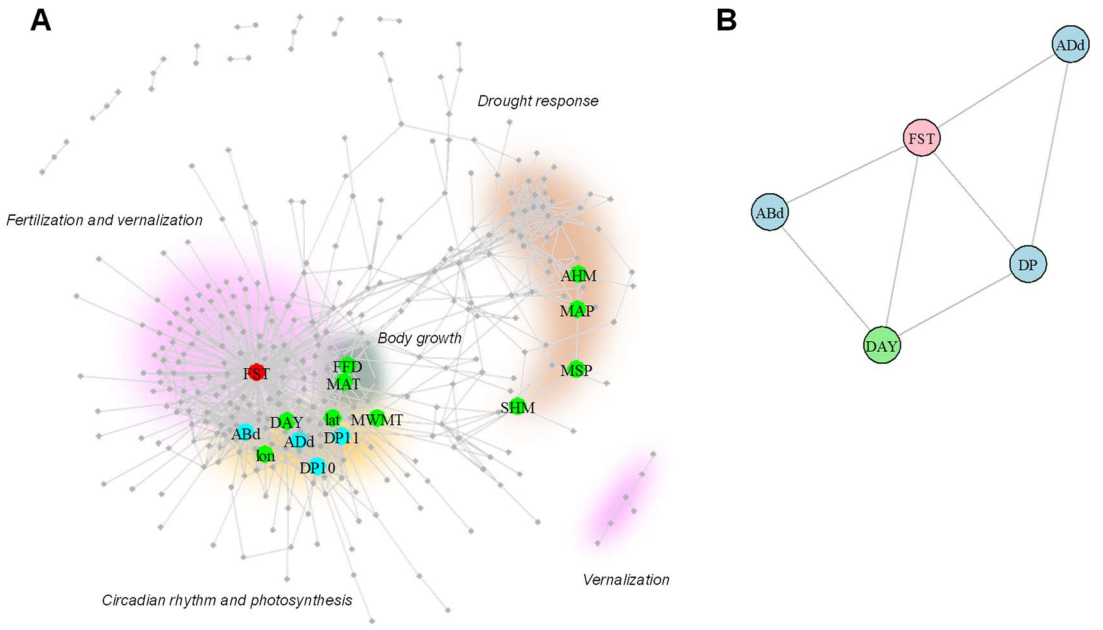

Figure 2 Whole network graph of the environmental adaptation of wild poplar. The red circle indicates population differentiation in terms of genome F_{ST}; green circles are environmental and location factors, blue ones are traits, and gray dots are genes. (A) Global structure of the estimated network. Colored clouds show clusters of genes with similar functions. (B) Relationship between F_{ST}, abaxial/adaxial stomata density (ABd and ADd), day length (DAY) and morbidity (DP10 and DP11).

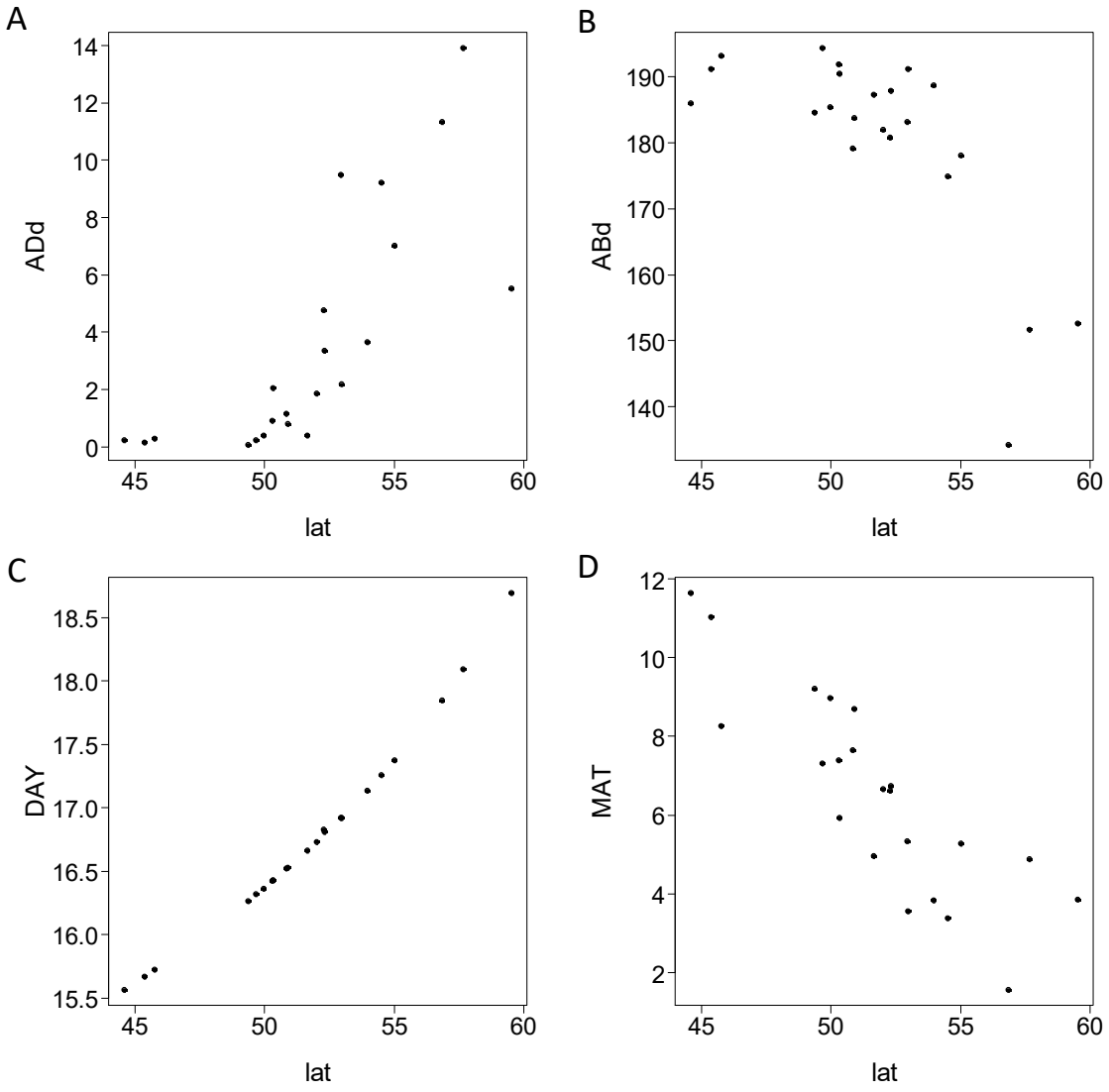

Figure 3 Geographic distribution of stomatal density of wild poplar, day length and temperature. (A) Latitude (lat) vs. adaxial stomata density (ADd). (B) lat vs. abaxial stomata density (ABd). (C) lat vs. longest day length (DAY). (D) lat vs. mean annual temperature (MAT).

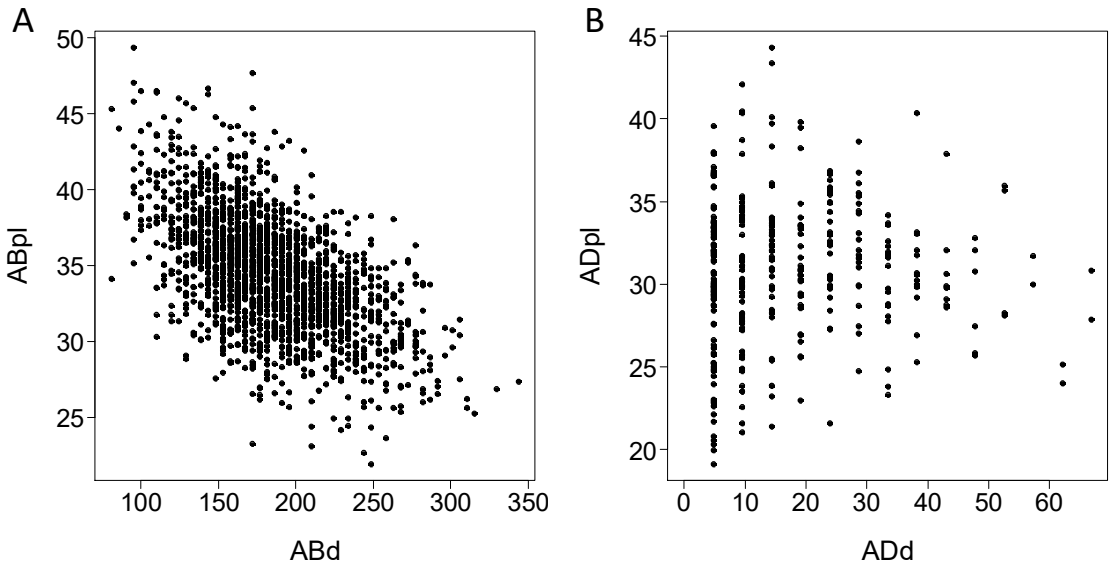

Figure 4 Stomatal density and pore size of wild poplar. (A) Abaxial stomata density (ABd) vs pore length (ABpl). (B) Adaxial stomata density (ADd) vs pore length (ADpl).

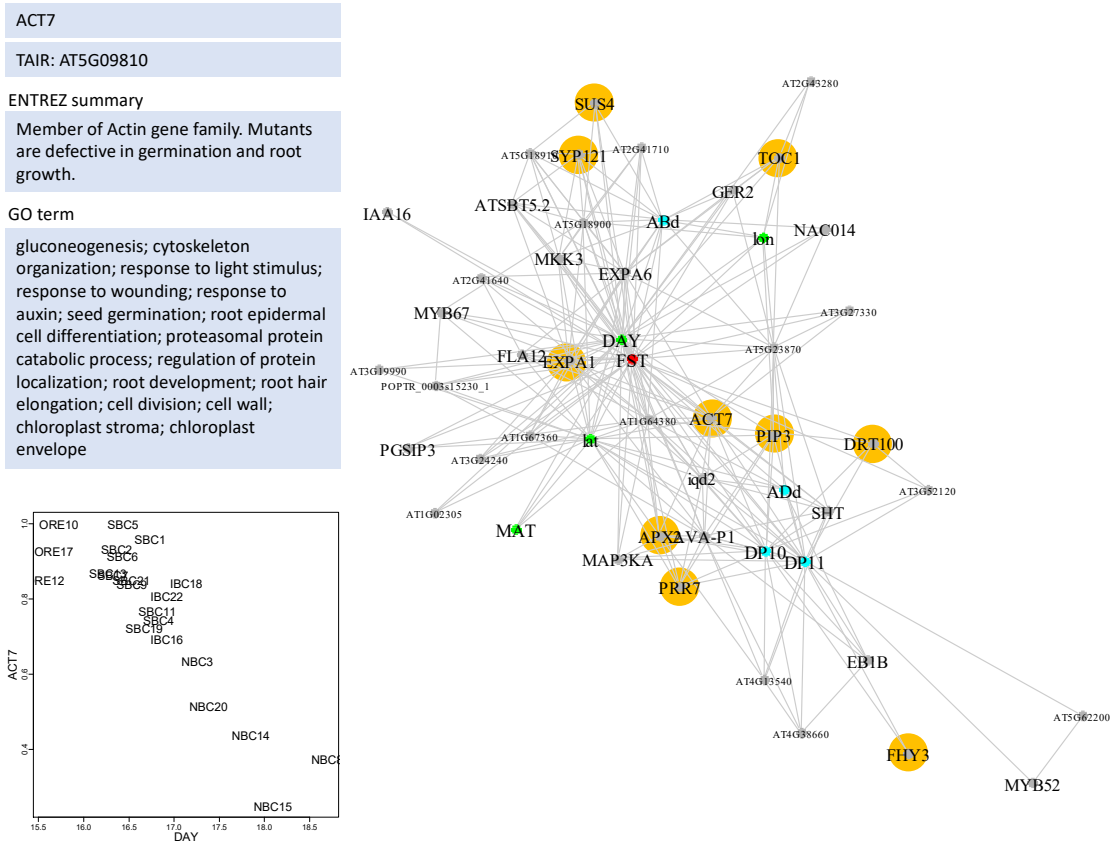

Figure 5 Radius-one neighborhood of longest day length (DAY) and disease progress (DP10 and DP11). The biological functions (upper left) of genes in the neighborhood and their correlations (lower left) with the center node can be seamlessly examined by entering a gene name in the upper left text box. In this example, ACT7 was examined (left), and genes related to photosynthesis and circadian rhythm are colored in orange (see text).

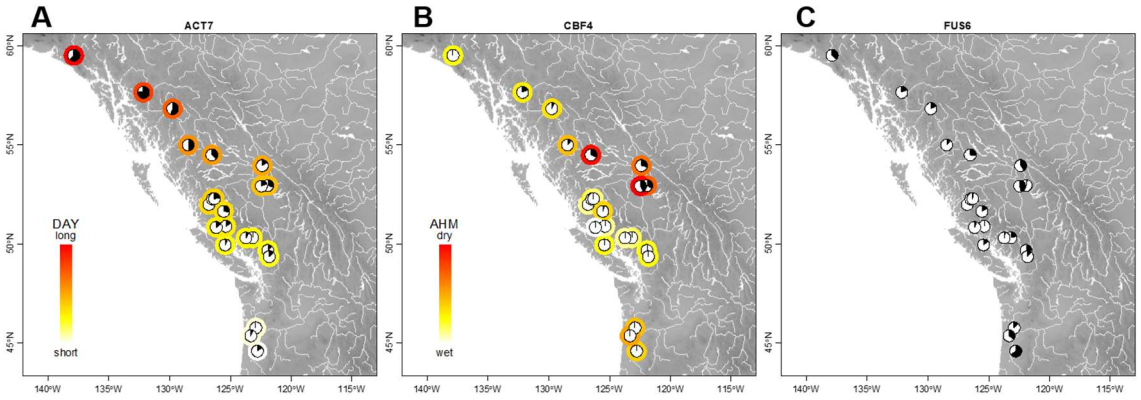

Figure 6 Values of environmental factors and allele frequencies of differentiated genes superimposed on a geographical map. Pie charts show allele frequencies of genes (black: minor allele; white: major allele). Heat colors are used to illustrate gradients of the environmental factors and traits. (A) Day length (DAY) and a light-response gene (ACT7). (B) Humidity (AHM) and a drought-response gene (CBF4). (C) A gibberellin-response gene (FUS6).

Original Data

	E	T	G_1	G_2	G_3
pop1	E_{pop1}	T_{pop1}	$AF_{pop1}^{G_1}$	$AF_{pop1}^{G_2}$	$AF_{pop1}^{G_3}$
pop2	E_{pop2}	T_{pop2}	$AF_{pop2}^{G_1}$	$AF_{pop2}^{G_3}$	$AF_{pop2}^{G_3}$
pop3	E_{pop3}	T_{pop3}	$AF_{pop3}^{G_1}$	$AF_{pop3}^{G_4}$	$AF_{pop3}^{G_3}$
pop4	E_{pop4}	T_{pop4}	$AF_{pop4}^{G_1}$	$AF_{pop4}^{G_5}$	$AF_{pop4}^{G_3}$

Distance Data for Graph

	genome F_{ST}	E	T	G_1	G_2	G_3
pop1 vs 2	$F_{ST}^{genome}{}_{pop1,pop2}$	$d(E_{pop1,pop2})$	$d(T_{pop1,pop2})$	$F_{ST}^{G_1}{}_{pop1,pop2}$	$F_{ST}^{G_2}{}_{pop1,pop2}$	$F_{ST}^{G_3}{}_{pop1,pop2}$
pop1 vs 3	$F_{ST}^{genome}{}_{pop1,pop3}$	$d(E_{pop1,pop3})$	$d(T_{pop1,pop3})$	$F_{ST}^{G_1}{}_{pop1,pop3}$	$F_{ST}^{G_2}{}_{pop1,pop3}$	$F_{ST}^{G_3}{}_{pop1,pop3}$
pop1 vs 4	$F_{ST}^{genome}{}_{pop1,pop4}$	$d(E_{pop1,pop4})$	$d(T_{pop1,pop4})$	$F_{ST}^{G_1}{}_{pop1,pop4}$	$F_{ST}^{G_2}{}_{pop1,pop4}$	$F_{ST}^{G_3}{}_{pop1,pop4}$
pop2 vs 3	$F_{ST}^{genome}{}_{pop2,pop3}$	$d(E_{pop2,pop3})$	$d(T_{pop2,pop3})$	$F_{ST}^{G_1}{}_{pop2,pop3}$	$F_{ST}^{G_2}{}_{pop2,pop3}$	$F_{ST}^{G_3}{}_{pop2,pop3}$
pop2 vs 4	$F_{ST}^{genome}{}_{pop2,pop4}$	$d(E_{pop2,pop4})$	$d(T_{pop2,pop4})$	$F_{ST}^{G_1}{}_{pop2,pop4}$	$F_{ST}^{G_2}{}_{pop2,pop4}$	$F_{ST}^{G_3}{}_{pop2,pop4}$
pop3 vs 4	$F_{ST}^{genome}{}_{pop3,pop4}$	$d(E_{pop3,pop4})$	$d(T_{pop3,pop4})$	$F_{ST}^{G_1}{}_{pop3,pop4}$	$F_{ST}^{G_2}{}_{pop3,pop4}$	$F_{ST}^{G_3}{}_{pop3,pop4}$

Figure S1 Data matrices. The original dataset is a matrix of environmental factor mean (E_i), trait value mean (T_i) and minor allele frequency of the *k*th gene (AF_i^k) in the *i*th population. Distance data for our graphical representation consist of a matrix of pairwise genome *F*ST ($F_{ST}{}_{i,j}^{genome}$), the pairwise difference of environmental factors ($d(E_i, E_j)$), the pairwise difference of trait values ($d(T_i, T_j)$) and pairwise locus *F*ST of the *k*th gene ($F_{ST}{}_{i,j}^{k}$) between *i*th and *j*th populations.

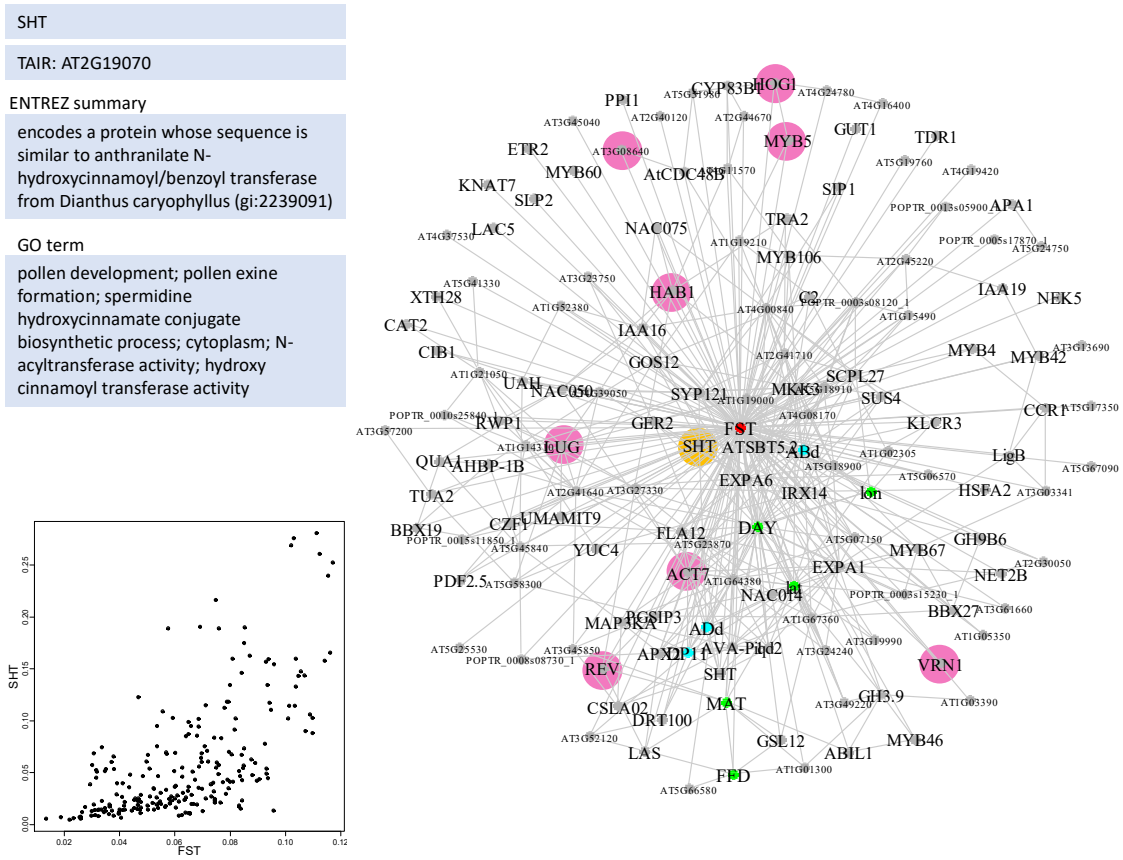

Figure S2 Radius-one neighborhood of genome F_{ST}. Any subgraph can be drawn given any subgraph origins, radii and genes. The Entrez summary, GO term and a scatter plot between any two nodes can be shown in the same graph window. SHT, colored in orange, was examined in this example. The scatter plot shows the correlation between genome F_{ST} and the SHT gene. Genes related to reproduction are colored in pink (see text).

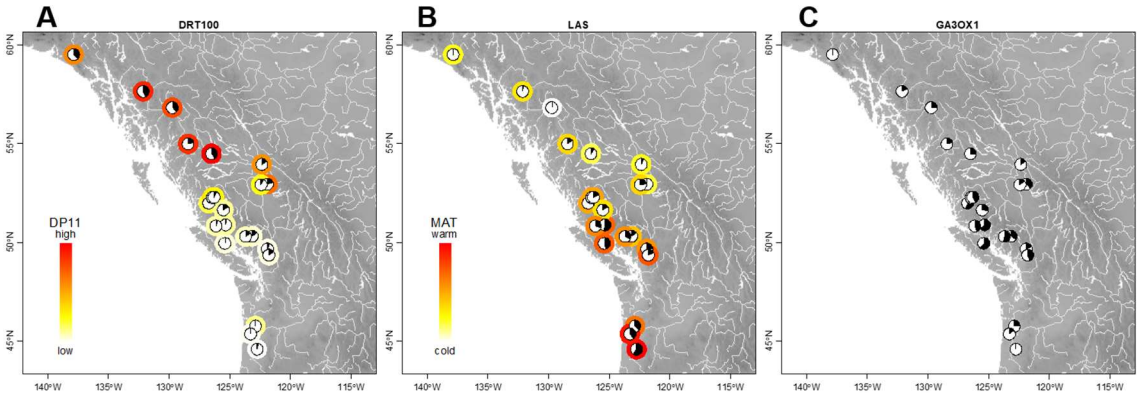

Figure S3 Values of environmental factors and allele frequencies of differentiated genes superimposed on a geographical map. Pie charts show allele frequencies of genes (black: minor allele; white: major allele). Heat colors are used to display gradients of the environmental factors and traits. (A) Disease progress (DP11) and a DNA repair gene (DRT100). (B) Annual temperature (MAT) and a lateral control gene (LAS). (C) A flower development gene (GA3OX1).

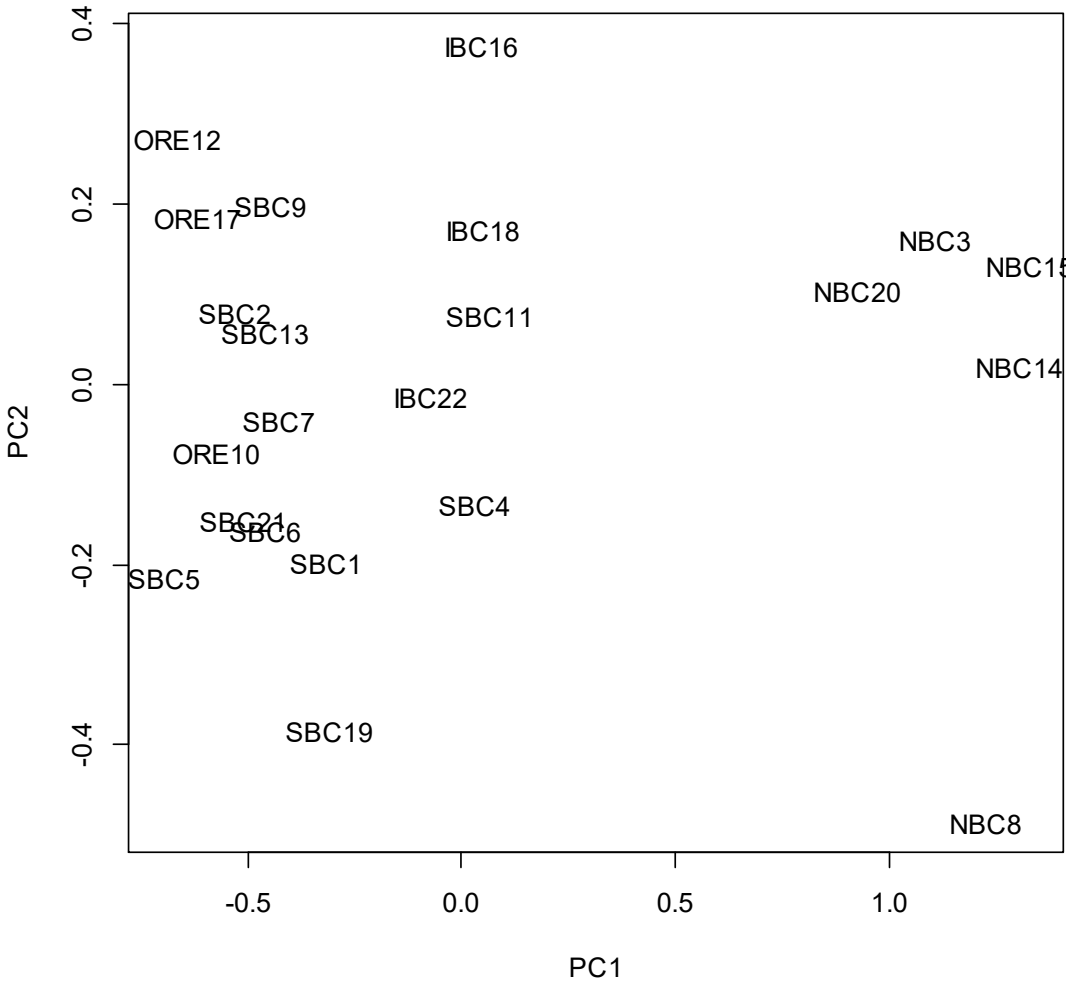

Figure S4 Principal component analysis plot of genes neighboring adaxial stomata density (ADd) and morbidity (DP10, DP11) in 22 populations of wild poplar. NBC, northern British Colombia; SBC, southern British Colombia; IBC, inland British Colombia; ORE, Oregon.

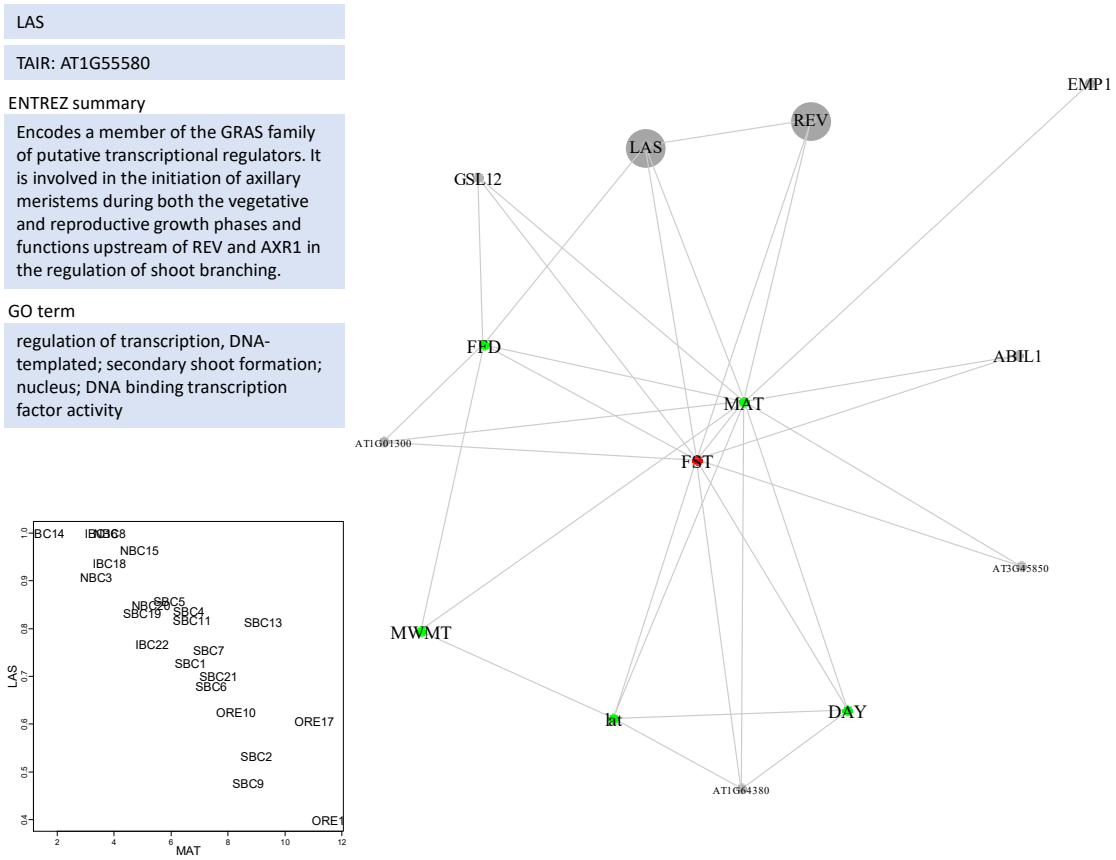

Figure S5 Radius-one neighborhood of mean annual temperature (MAT). The Entrez summary/GO term for LAS and the scatter plot for MAT and allele frequencies of LAS are shown in the graph window. Genes related to body growth are colored in gray (see text).

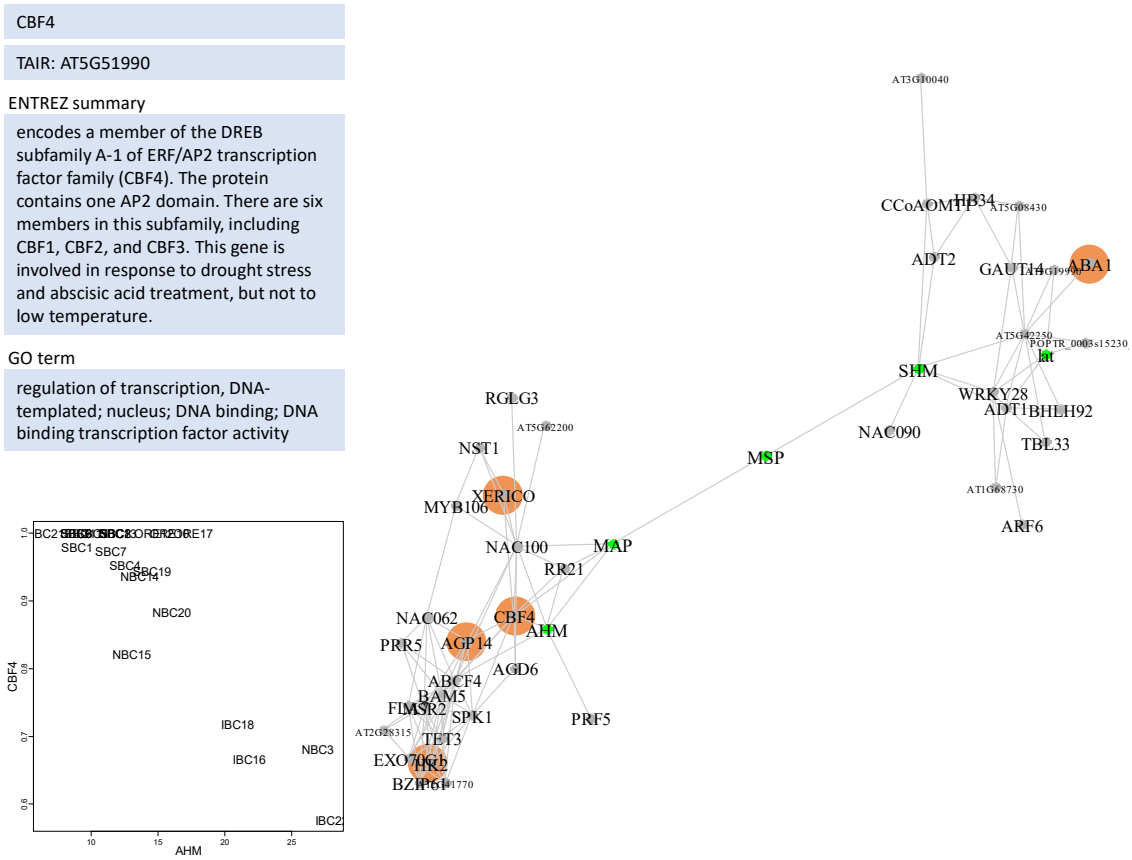

Figure S6 Neighborhood of precipitation (MAP and MSP) and moisture (AHM and SHM). The Entrez summary/GO term for CBF4, the scatter plot for AHM and allele frequencies of CBF4 are shown. Genes related to drought stress response are colored in brown (see text).

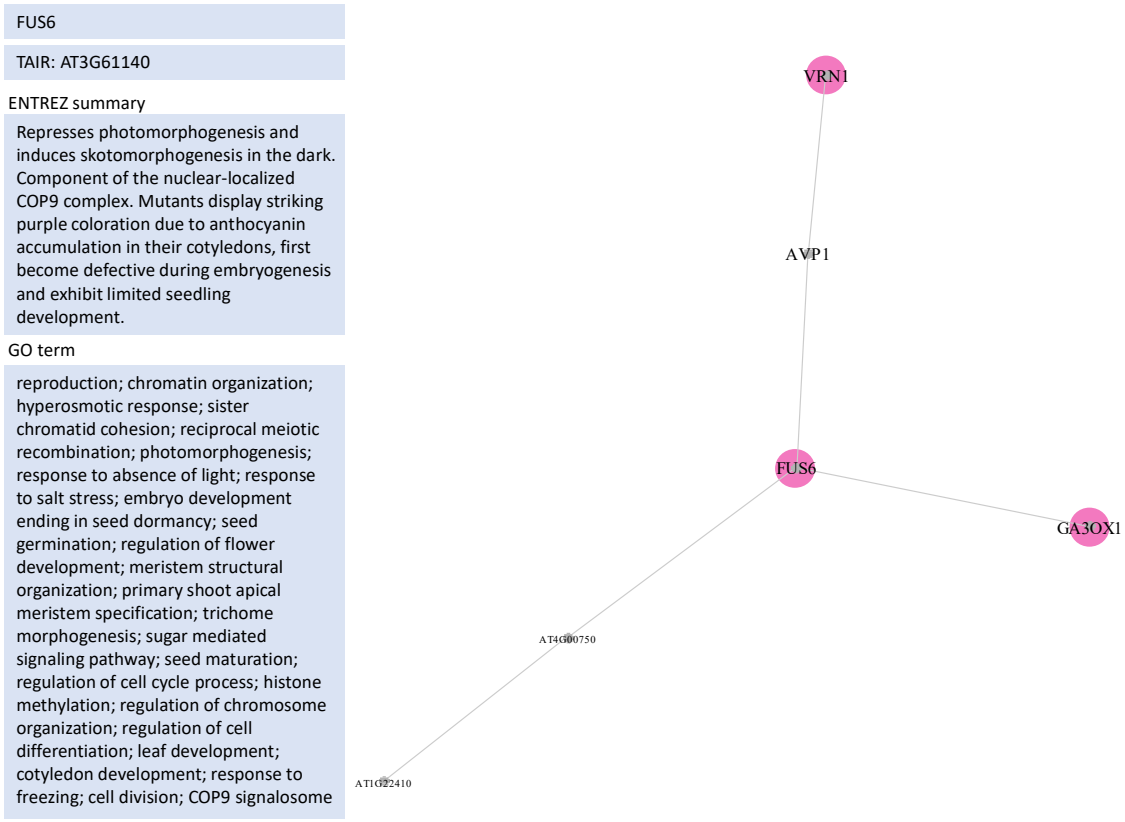

Figure S7 Isolated cluster of vernalization genes from Figure 2A. The Entrez summary/GO term for FUS6 is shown in the graph window. Genes related to vernalization are colored in pink (see text).