bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Modular environmental pleiotropy of genes involved in local adaptation to climate despite physical linkage

Authors

Katie E. Lotterhos[1,*], Kathryn A. Hodgins[2], Sam Yeaman[3], Jon Degner[4], Sally Aitken[4]

Affiliations

[1]Department of Marine and Environmental Sciences, Northeastern Marine Science Center, 430 Nahant Rd, Nahant, MA 01908

[2]School of Biological Sciences, Monash University, Wellington Rd, Clayton VIC 3800

[3]Department of Biological Sciences, University of Calgary, AB, T2N1N4

[4]Department of Forest and Conservation Sciences, 3041-2424 Main Mall, Vancouver, BC V6T 1Z4 Canada

* Corresponding Author, k.lotterhos@neu.edu

Keywords: landscape genomics, genetic-environment associations, genome-wide associations (GWAS), conifers, linkage disequilibrium, ion antiporters, auxin biosynthesis, flowering time

Running title: Co-association networks for landscape genomics

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Physical proximity among alleles shaped by different sources of selection is a fundamental aspect of genetic architectures critical for predicting their evolution. Theory predicts that evolution in complex environments should select for modular genetic architectures with limited pleiotropy among modules. However, limited data exist to test this hypothesis because the field lacks consensus for how to control for intercorrelated climate variables. We aim to characterize the genetic architecture of adaptation to climate, including the modularity of the architecture (number of distinct climate factors), overlap among modules, and physical linkage among loci. We introduce a co-association network analysis, which parses loci into groups based on differing environmental associations, and use it to study the genetic architecture of local adaptation to climate in lodgepole pine (*Pinus contorta*). We identified several non-overlapping modules of genes associated with environmental factors (aridity, freezing, geography), which supports the hypothesis of modular environmental pleiotropy. Notably, we found moderate physical linkage among some candidate loci in different modules, which may facilitate or hinder adaptation depending on the multivariate trajectory of climate change. Moreover, we show that associations with environmental principal components would have missed candidates and resulted in a limited interpretation regarding the selective environment. Finally, simulations revealed that the propensity of co-association modules to arise under neutrality increased with demographic complexity, but also that true causal loci are more highly-connected within the module, which may be useful for prioritizing candidates.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

Linkage and pleiotropy are fundamental aspects of genetic architecture (Hansen 2006). Pleiotropic genes that have effects on seemingly unrelated traits may influence the rate of adaptation (Orr 2000; Wang *et al.* 2010), and linkage among genes experiencing different kinds of selection can facilitate or hinder adaptation (Barton 2010; Aeschbacher & Bürger 2014; Reeve *et al.* 2016). Despite progress in understanding the underlying pleiotropic nature of phenotypes and the influence of pleiotropy on the rate of adaptation to specific conditions (Wagner & Zhang 2011), we have an incomplete understanding of the extent and magnitude of linkage and pleiotropy in local adaptation of natural populations to complex environments.

The use of the term ‘pleiotropy’ itself is fraught with controversy (Stearns 2010; Wagner & Zhang 2011; Paaby & Rockman 2013), indicating the need to carefully define how the term is being used. Here, we aim to characterize the number of separate components of the climate in which a mutation affects fitness, a form of ‘selectional pleiotropy’ (Paaby & Rockman 2013). A key feature of selectional pleiotropy is that traits are defined by the action of selection and not by the intrinsic attributes of the organism. In local adaptation to climate, an allele has been considered “antagonistically pleiotropic” for selection if it has different effects on fitness at different extremes of an environmental variable (e.g., positive effects on fitness in cold environments and negative effects in warm environments, (Savolainen *et al.* 2013)), which results in clinal relationships between alleles and environments (Haldane 1930, 1948; Slatkin 1973, 1978; Felsenstein 1976; Barton 1999). On the other hand, an allele or gene might be considered “environmentally pleiotropic” if it affects fitness in distinct environments (e.g., has effects on fitness in a cold environment and a dry environment) (Hancock *et al.* 2011). Here, we

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consider this latter case of environmental pleiotropy of individual genes in distinct climates. But conceptual issues arise from defining environments along the univariate axes that we measure, because “cold” and “dry” might be a single selective optimum (“cold-dry”) that a gene adapts to (Wagner & Zhang 2011). Moreover, climate variables such as temperature and precipitation are highly correlated across landscapes, and this correlation structure makes inferring pleiotropy from signals of selection to climate difficult. Indeed, in their study of adaptation of *Arabidopsis* to climate, Hancock et al. (2011) noticed that candidate loci showed signals of selection in multiple environmental variables, potentially indicating pleiotropic effects. However, they also found that a substantial proportion of this overlap was due to correlations among climate variables on the landscape, and as a result they were unable to fully describe pleiotropic effects.

An important aspect of pleiotropy is the extent to which individual mutations or genes are pleiotropic (e.g., the number of distinct traits or components of fitness that they affect). Genotype-phenotype mapping and coexpression studies have revealed that functional and developmental modules of genes are common, indicating limited pleiotropy (Wagner *et al.* 2007). These data support the Hypothesis of Modular Pleiotropy, which states that mutations are generally constrained to affect traits that are part of the same functional complex (Wagner *et al.* 2007; Paaby & Rockman 2013). Modular architectures are characterized by extensive pleiotropic effects among elements within a module, and a suppression of pleiotropic effects between different modules (Wagner 1996). Modular architectures are predicted to be favored when traits are under a combination of directional and stabilizing selection because modularity then allows adaptation to take place in one trait without undoing the adaptation achieved by another trait (Wagner 1996; Griswold 2006), and when genomes face complex spatial and temporal environments (Le Nagard *et al.* 2011). Adaptation to climate on a landscape fits these criteria, because traits are thought to be under stabilizing selection within populations but

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

directional selection among populations (Le Corre & Kremer 2003), and environmental variation among populations is complex with multiple abiotic and biotic challenges occurring at different spatial scales.

Although there is emerging agreement that organisms have modular organization of genes in their effects on phenotypes (Wagner *et al.* 2007; but see Boyle *et al.* 2017), understanding the ecological pressures and opportunities that favour modular architectures through the action of natural selection remains an open problem for several reasons. First, characterizing the genetic architecture of adaptation to different aspects of the multivariate environment has been elusive because of difficulty in characterizing the "climate" that a locus is adapting to and because of the spatial correlations among environments as mentioned above. The environmental variables that we are able to measure are univariate and may not be representative of the complex multivariate environment from the perspective of the organism. Similar issues occur with genome-wide association studies of phenotypes, as the phenotypic traits we are able to measure may not capture the complex multivariate phenotype coded for by a causal locus. Even when many variables are summarized with ordination such as principle components that are orthogonal, the axes that explain the most variation in physical environment don't necessarily correspond to the axes that cause selection. Secondly, the statistical methods widely used for inferring adaptation to climate are also univariate in the sense that they measure correlations between a single allele frequency and a single environmental variable (e.g., Frichot *et al.* 2013; Günther & Coop 2013; Gautier 2015). While some multivariate regression methods like redundancy analysis have been used to understand how multiple environmental factors shape genetic structure (Lasky *et al.* 2012; Benestan *et al.* 2016), they still rely on ordination and have not been used to identify distinct evolutionary modules of loci. Other issues arise in identifying distinct evolutionary modules because physical linkage will cause correlated response to

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection in neutral loci flanking a causal locus. Thus, large regions of the genome can share similar patterns of association to a given environment, such that many loci within a given candidate region are probably not causally responding to selection. Overall, current analytical techniques have given limited insight into the genetic architectures of adaptation to the multivariate environmental drivers of selection.

Here, we aim fill this gap by presenting a framework to characterize the genetic architecture of adaptation to the multivariate environment, including the selectional modularity of the architecture (e.g. the number of distinct climates that genes are adapting to), the extent of pleiotropic effects of genes in different modules (e.g. genes that contain SNPs associating with climates in different modules), physical linkage among genes within and among different modules, and the possibility for false signals arising from hitchhiking among physically linked genes within a module. The modules that we characterize can be thought of as “variational” modules (sensu Wagner *et al.* 2007), which are composed of features that vary together and are relatively independent of other such sets of features. Interactions between spatially varying selection and gene flow are expected to result in associations between allele frequencies and the selective environment (Hedrick *et al.* 1976; Barton 1983; Hedrick 1986). Here, we use these associations to determine the modularity of selectional pleiotropy in a co-association network analysis. In this approach, the genetic effects of loci on different traits (e.g., developmental/functional pleiotropy or modules) under selection are unknown, and we assume that each aspect of the multivariate environment selects for a trait or suite of traits that can be inferred by connecting candidate loci directly to the aspects of the environment that select for particular allele combinations. This approach identifies differences between groups of loci that

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

may covary with one environmental variable but covary in different ways with another, revealing patterns that are not evident through univariate analysis.

We apply these new approaches to characterize the genetic architecture of local adaptation to climate in lodgepole pine (*Pinus contorta*) using an exome capture dataset representing 281 populations of trees inhabiting a wide range of environments varying in the strength and nature of stresses (Hodgins *et al.* 2016; Suren *et al.* 2016; Yeaman *et al.* 2016b). Furthermore, we compare the approach to associations based on principal components and evaluate the approach using simulated landscape genetic data in order to better understand its benefits and caveats. Lodgepole pine is a coniferous species inhabiting a wide range of environments in northwestern North America. Previous work based on reciprocal transplants and common garden experiments have shown extensive local adaptation (Illingworth 1978; Liepe *et al.* 2016; Yeaman *et al.* 2016b). Conifers, like other common species of *Pinus,* may face numerous types of selective environments across their range, including freezing temperatures, precipitation, and aridity (Howe *et al.* 2003; Eckert *et al.* 2010b; a; Alberto *et al.* 2013). Understanding local adaptation to climatic factors in this species will improve projections of climate change impacts and facilitate forest management.

Previously, we used comparative genomics to discover candidates for convergent adaptation to freezing between lodgepole pine and the interior spruce complex (*Picea glauca* x *Picea engelmannii*) (Hodgins *et al.* 2016; Suren *et al.* 2016; Yeaman *et al.* 2016b). However, the comparative approach is limited to discovering parallel patterns between species, and overlooks selective factors unique to one species. As in most other systems, the genomic architecture in pine underlying adaptation to the multivariate environment has not been well characterized. For each individual, the climate of its source population was characterized using 22 environmental

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variables. First, we identified top candidate exome contigs for local adaptation to climate as those that contain more outliers for genotype-environment associations and genotype-phenotype associations than expected by chance (Yeaman *et al.* 2016b). Within these top candidate contigs we identified top candidate SNPs whose frequencies were associated with one or more environmental variables more strongly than expected by chance, using a criterion that excluded false positives in simulated data described below. To this set of top candidate SNPs, we performed a co-association network analysis to characterize pleiotropy and linkage of the architecture.

Methods

Sampling and climate

We obtained 281 seedlots of lodgepole pine from available operational reforestation collections from natural populations. Seedlots were selected to represent the full range of climatic and ecological conditions within the species range in British Columbia and Alberta based on ecosystem delineations. Seedlot origins were characterized climatically by estimating climate normals for 1961-1990 from geographic coordinates using the software package ClimateWNA (Wang *et al.* 2012). The program extracts and downscales the moderate spatial resolution generated by PRISM (Daly *et al.* 2008) to scale-free and calculates many climate variables for specific locations based on latitude, longitude and elevation. The downscaling is achieved through a combination of bilinear interpolation and dynamic local elevational adjustment. We obtained 19 climatic and 3 geographical variables (latitude, longitude, and elevation). Geographic variables may correlate with some unmeasured environmental variables that present selective pressure to populations (e.g., latitude correlates with day length). Many of these variables were correlated with each other on the landscape (Figure 1a).

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sequencing, bioinformatics, and annotation

DNA from frozen needle tissue was purified using a Macherey-Nagel Nucleospin 96 Plant II Core kit automated on an Eppendorf EpMotion 5075 liquid handling platform. One microgram of DNA from each individual tree was made into a barcoded library with a 350 bp insert size using the BioO NEXTflex Pre-Capture Combo kit. Six individually barcoded libraries were pooled together in equal amounts before sequence capture. The capture was performed using custom Nimblegen SeqCap probes (see Suren *et al.* 2016; Yeaman *et al.* 2016b for more details) and the resulting captured fragments were amplified using the protocol and reagents from the NEXTflex kit. All sample preparation steps followed the recommended protocols provided.

After capture, each pool of six libraries was combined with another completed capture pool and the 12 individually barcoded samples were then sequenced, 100 base pair paired-end, on one lane of an Illumina HiSeq 2500 (at the McGill University and Genome Quebec Innovation Centre).

Sequenced reads were filtered and aligned to the loblolly pine genome (Neale *et al.* 2014) using bwa mem (Li & Durbin 2009) and variants were called using GATK Unified Genotyper (DePristo *et al.* 2011), with steps included for removal of PCR duplicates, realignment around indels, and base quality score recalibration (DePristo *et al.* 2011; Yeaman *et al.* 2016b). SNPs calls were filtered to eliminate variants that did not meet the following cutoffs: quality score >= 20, map quality score >= 45, FisherStrand score <= 33, HaplotypeScore <= 7, MQRankSumTest <= -12.5, ReadPosRankSum >-8, and allele balance < 2.2, minor allele frequency > 5%, and genotyped successfully in >10% of individuals. Ancestral alleles were coded as a 0 and derived alleles coded as a 1 for data analysis.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We used the annotations developed for pine in (Yeaman *et al.* 2016b). Briefly, we performed a BLASTX search against the TAIR 10 protein database and identified the top blast hit for each transcript contig (e-value cut-off was 10^{-6}). We also performed a BLASTX against the nr database screened for green plants and used Blast2GO (Conesa & Götz 2008) to assign GO terms and enzyme codes (see Yeaman *et al.* 2014, 2016b for details). We also assigned GO terms to each contig based on the GO *A. thaliana* mappings and removed redundant GO terms. To identify if genes with particular molecular function and biological processes were over-represented in top candidates, we performed a GO enrichment analysis using topGO (Alexa & Rahnenführer 2009). All GO terms associated with at least two candidate genes were analyzed for significant over-representation within each cluster and in all candidate loci (FDR 5%).

Top Candidate SNPs

Top candidate exome contigs were obtained from (Yeaman *et al.* 2016b). Briefly, exome contigs with unusually strong signatures of association from multiple association tests (uncorrected genotype-phenotype and genotype-environment correlations, for details see Yeaman et al. 2016) were identified as those with more outlier SNPs than expected by random with a probability of $P < 10^{-9}$, which is a very restrictive cutoff (note that due to non-independence among SNPs in the same contig, this *P*-value is an index, and not an exact probability). Thus, the subsequent analysis is limited to loci that we have the highest confidence are associated with adaptation as evidenced by a large number of significant SNPs (not necessarily the loci with the largest effect sizes).

Next, we identified top candidate SNPs within the set of top candidate contigs. These “top candidate SNPs” had genetic-environment associations with (i) *P*-values lower than the

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bonferroni cutoff for the uncorrected Spearman's ρ (~10^{-8} = 0.05/(number of SNPs times the number of environmental variables) and (ii) $\log_{10}(BF) > 2$ for the structure-corrected Spearman's ρ (Bayenv2, for details see below). The resulting set of candidate SNPs reject the null hypothesis of no association with the environment with high confidence. In subsequent analyses we interpret the results both before and after correction for population structure, to ensure that structure correction does not change our overall conclusions. Note that because candidate SNPs are limited to the top candidate contigs in order to reduce false positives in the analysis, these restrictive cutoffs may miss many true positives.

For uncorrected associations between allele frequencies and environments, we calculated the non-parametric rank correlation Spearman's ρ between allele frequency for each SNP and each environmental variable. For structure-corrected associations between allele frequencies and environments, we used the program Bayenv2 (Günther & Coop 2013). Bayenv2 is implemented in two steps. In the first step the variance-covariance matrix is calculated from allelic data. Using the set of non-coding SNPs, we calculated the variance-covariance matrix from the final run of the MCMC after 100,000 iterations, with the final matrix averaged over 3 MCMC runs. In the second step, the variance-covariance matrix is used to control for evolutionary history in the calculation of test statistics for each SNP. For each SNP, Bayenv2 outputs a Bayes factor (a value that measures the strength of evidence in favor of a linear relationship between allele frequencies and the environment after population structure is controlled for) and Spearman's ρ (the non-parametric correlation between allele frequencies and environment variables after population structure is controlled for). Previous authors have found that the stability of Bayes factors is sensitive to the number of iterations in the MCMC (Blair *et al.* 2014). We ran 3 replicate chains of the MCMC with 50,000 iterations, which we found produced stable results.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bayes factors and structure-corrected Spearman's ρ were averaged over these 35 replicate chains and these values were used for analysis.

Co-association networks

We first organized the associations into a matrix with SNPs in columns, environments in rows, and the specific SNP-environment association in each cell. These data were used to calculate pairwise Euclidean distances between SNPs based on their associations, and this distance matrix was used to cluster SNP loci with Ward's hierarchical clustering using the hclust package in R. As described in the results, this resulted in 4 main clusters in the data. For each of these main clusters, we used undirected graph networks to visualize submodules of SNPs. Nodes (SNPs) were connected by edges if they had a pairwise Euclidean distance less than 0.1 from the distance matrix described above. We found that the results were not very sensitive to this distance threshold. Co-association networks were visualized using the igraph package in R v 1.0.1 (Csardi & Nepusz 2006).

Linkage disequilibrium

Linkage disequilibrium was calculated among pairwise combinations of SNPs within exome contigs (genes). Mean values of Pearson's correlation coefficient squared (r^2) were estimated across across all SNPs annotated to each pair of individual genes, excluding SNPs genotyped in fewer than 250 individuals (to minimize the contribution of small sample sizes to the calculation of gene-level means).

Recombination rates

An Affymetrix SNP array was used to genotype 95 full-sib offspring from a single cross of two parents. Individuals with genotype posterior probabilities of > 0.001 were filtered out. This array yielded data for 13,544 SNPs with mapping-informative genotypes. We used the package

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"onemap" in R with default settings to estimate recombination rates among pairs of loci, retaining all estimates with LOD scores > 3 (Margarido *et al.* 2007). This dataset contained 2760 pairs of SNPs that were found together on the same genomic contig, separated by a maximum distance of 13k base pairs. Of these 7,617,600 possible pairs, 521 were found to have unrealistically high inferred rates of recombination (r > 0.001), and are likely errors. These errors probably occurred as a result of the combined effect of undetected errors in genotype calling, unresolved paralogy in the reference genome that complicates mapping, and differences between the reference loblolly genome that was used for SNP design and the lodgepole pine genomes. As a result, recombination rates that were low (r < 0.001) were expected to be relatively accurate, but we do not draw any inferences about high recombination estimates among loci.

Associations with principal components of environments

To compare inference from co-association networks to another multivariate approach, we conducted a principal components analysis of environments using the function prcomp() in R. Then, we used Bayenv2 to test associations with PC axes as described above and used BF > 2 as criteria for significance of a SNP on a PC axis. Note that this criterion is less conservative than that used to identify candidates for the network analysis (because it did not require the additional criteria of a significant Bonferroni-corrected *P*-value), so it should result in greater overlap between PC candidates and network candidates.

Enrichment of co-expressed genes

To determine if adaptation cluster members had similar gene functions, we examined their gene expression patterns in response to seven growth chamber climate treatments using previously published RNAseq data (Yeaman *et al.* 2014). We used a Fisher's exact test to determine if

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

genes with a significant treatment effect were over-represented in each cluster and across all adaptation candidates relative to the other sequenced and expressed genes. We also examined if co-expressed gene networks that were previously identified using the same experimental data were over-represented in the adaptation clusters relative to the other sequenced and expressed genes.

Galaxy biplots

To give insight into how the species has evolved to inhabit multivariate environments relative to the ancestral state, we visualized the magnitude and direction of associations between the derived allele frequency and environmental variables. Allelic correlations with any pair of environmental variables can be visualized by plotting the value of the non-parametric rank correlation Spearman's ρ of the focal allele with variable 1 against the value with variable 2. Spearman's ρ can be calculated with or without correction for population structure. Note also that the specific location of any particular allele in a galaxy biplot depends on the way alleles are coded. SNP data were coded as 0, 1, or 2 copies of the loblolly reference allele. If the reference allele has positive Spearman's ρ with temperature and precipitation, then the alternate allele has a negative Spearman's ρ with temperature and precipitation. For this reason, the alternate allele at a SNP should be interpreted as a reflection through the origin (such that Quadrants 1 and 3 are symmetrical and Quadrants 2 and 4 are symmetrical if the reference allele is randomly chosen).

A prediction ellipse was used to visualize the genome-wide pattern of covariance in allelic effects on a galaxy biplot. For two variables, the 2 x 2 variance-covariance matrix of $Cov(\rho(f, E_1), \rho(f, E_2))$, where f is the allele frequency and E_x is the environmental variable, has a geometric interpretation that can be used to visualize covariance in allelic effects with

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ellipses. The covariance matrix defines both the spread (variance) and the orientation (covariance) of the ellipse, while the expected values or averages of each variable (E[E_1] and E[E_2]) represent the centroid or location of the ellipse in multivariate space. The geometry of the two-dimensional (1 - α) x 100% prediction ellipse on the multivariate normal distribution can then be approximated by:

$$l_j = \sqrt{\lambda_j \chi^2_{df=2,\alpha}},$$

where l_j= {1, 2} represents the lengths of the major and minor axes on the ellipse, respectively, λ_j represents the eigenvalues of the covariance matrix, and $\chi^2_{df=2,\alpha}$ represents the value of the χ^2 distribution for the desired α value (Titterington 1976; Pison *et al.* 1999; Kaufman & Rousseeuw 2009). In the results, we plot the 95% prediction ellipse (α = 0.05) corresponding to the volume within which 95% of points should fall assuming the data is multivariate normal, using the function `ellipsoidPoints()` in the R package `cluster`. This approach will work when there is a large number of unlinked SNPs in the set being visualized; if used on a candidate set with a large number of linked SNPs and/or a small candidate set with non-random assignment of alleles (i.e., allele assigned according to a reference), the assumptions of this visualization approach will be violated.

Visualization of allele frequencies on the landscape

ESRI ArcGIS v10.2.2 was used to visualize candidate SNP frequencies across the landscape. Representative SNPs having the most edges within each sub-network were chosen and plotted against climatic variables representative of those clusters. Mean allele frequencies were calculated for each sampled population and plotted using ESRI ArcGIS v10.2.2. Climate data and 1 km resolution rasters were obtained using ClimateWNA v5.40 (Wang *et al.* 2012) and

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

shaded with colour gradients scaled to the range of climates across the sampling locations. The climates for each sampling location were also plotted, as some sampling locations were at especially high or low elevations relative to their surrounding landscapes. For clarity, only sampling locations containing at least two sampled individuals were plotted.

Simulations

We used individual-based simulations to study the potential caveats of the co-association network analysis by comparing the connectedness of co-association networks arising from false positive neutral loci vs. a combination of false positive neutral loci and true positive loci that had experienced selection to an unmeasured environment. Specifically, we used simulations with random sampling designs from three replicates across three demographic histories: (i) isolation by distance at equilibrium, and non-equilibrium range expansion from a (ii) single refuge or from (iii) two refugia. These landscape simulations were similar to lodgepole pine in the sense that they simulated large effective population sizes and resulted in similar F_{ST} across the landscape as that observed in pine ((Lotterhos & Whitlock 2014, 2015), F_{ST} in simulations ~ 0.05, vs. F_{ST} in pine ~ 0.016 (Yeaman *et al.* 2016b)). To explore the how the allele frequencies that evolved in these simulations might yield spurious patterns under the co-association network analysis, we overlaid the 22 environmental variables used in the lodgepole pine dataset onto published landscape genomic simulations (Lotterhos & Whitlock 2014, 2015) with different demographic histories. To simulate the unmeasured environment, a small proportion of SNPs (1%) were subject to computer-generated spatially varying selection along a weak latitudinal cline (Lotterhos & Whitlock 2014, 2015). We assumed that 22 environmental variables were measured, but not the “true” selective environment; our analysis thus represents the ability of co-association networks to correctly cluster selected loci even when the true selective environment was unmeasured, but a number of other environmental variables were measured

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(correlations between the selective environment and the other variables ranged from 0 to 0.2). Note that the simulations differ from the empirical data in at least two ways: (i) there is only one selective environment (so we can evaluate whether a single selective environment could result in multiple clusters of SNPs in the data given the correlation structure of observed environments), and (ii) loci were unlinked.

For each of the 22 environmental variables for lodgepole pine populations, we used interpolation to estimate the value of the variable at the simulated locations. This strategy preserved the correlation structure among the 22 environmental variables. For each of the 22 variables, we calculated the uncorrected rank correlation (Spearman's ρ) between allele frequency and environment. The 23rd computer-generated environment was not included in analysis, as it was meant to represent the hypothetical situation that there is a single unmeasured (and unknown) environmental variable that is the driver of selection. The 23rd environment was correlated from 0-0.2 with the other 22 variables.

We compared two thresholds for determining which loci were retained for co-association network analysis, keeping loci with either: (i) a *P*-value lower than the Bonferroni correction (0.05/(# environments * # simulated loci)) and (ii) a log-10 Bayes Factor greater than 2 (for at least one of the environmental variables). Using both criteria is more stringent and both were used in the lodgepole pine analysis. In the simulations, however, we found that using both criteria resulted in no false positives in the outlier list (see Results); therefore we used only the first of these two criteria so that we could understand how false positives may affect interpretation of the co-association network analysis. For a given set of outliers (e.g., only false positives or false positives and true positives), hierarchical clustering and undirected graph networks were built in the same manner as described for the lodgepole pine data.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Top candidates

Our "top candidate" approach identified a total of 117 candidate exome contigs out of a total of 86,566 contigs. These contigs contained 801 top-candidate SNPs (out of 1,098,930 SNPs) that were strongly associated with these environments and likely either causal or tightly linked to a causal locus. These top candidate SNPs were enriched for X^TX outliers (Supplemental Figure 1: X^TX is an analog of F_{ST} that measures differentiation in allele frequencies across populations). To elucidate patterns of multivariate association, we apply the co-association network analysis and galaxy biplots to these 801 top candidate SNPs.

Co-association networks

Hierarchical clustering of top candidate SNPs revealed a large number of modules in the network. For the purposes of presentation, we grouped SNPs into 4 main groups, each with several submodules, classified according to the kinds of environmental variables that are most strongly associated with them: Aridity, Freezing, Geography, and an assorted group we bin as "Multi" (Figure 1A, B). Interestingly, this clustering by association signatures does not closely parallel the underlying correlation structure of the environmental variables themselves. For example, TD, DD_0, and LAT are all relatively strongly correlated, but the "Freezing" SNPs are strongly correlated with TD and DD_0 but not LAT (Figure 1A, 1B).

These four groups of SNPs varied considerably in the modularity of their underlying genetic architecture (Figure 1C-F). The "Multi" group stands for multiple environments because these SNPs showed associations with 19 to 21 of the 22 environmental variables. This group consisted of 60 top candidate SNPs across just 3 exome contigs and undirected graph networks

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

revealed 2 sub-modules within this group (Figure 1c, g, Supplementary Figure 2). The “Aridity” group consisted of 282 SNPs across 28 exome contigs and showed associations with climate moisture deficit, annual heat:moisture index, mean summer precipitation, and temperature variables excluding frost (Figure 1b). All these SNPs were very similar in their patterns of association and grouped into a single network module (Figure 1d, Supplementary Figure 3). The “Freezing” group consisted of 176 SNPs across 21 exome contigs and showed associations with freezing variables including number of degree-days below 0^{o}C, mean coldest month temperature, and variables related to frost (Figure 1b). SNPs from eight of the exome contigs in this group formed a single module, with the remaining SNPs mainly clustering by contig (Figure 1e, Supplementary Figure 4). The final group, “Geography,” consisted of 282 SNPs across 28 exome contigs that showed consistent associations with the geographical variables elevation and longitude, but variable associations with climate variables (Figure 1b). This was a loosely connected network consisting of several submodules with representation from 1 to 9 exome contigs (Figure 1f, Supplementary Figure 6). Even if the candidate loci that we studied consist of both true positives and false positive SNPs, our results suggest that there are largely non-overlapping genomic architectures underlying adaptation to these four primary aspects of the multivariate environment (Figure 1g). Network analysis using structure-corrected associations between allele frequency and the environmental variables resulted in broadly similar patterns, although the magnitude of the correlations was reduced (Supplemental Figure 6).

To determine if co-association networks correspond to associations driven by linkage disequilibrium (LD), we calculated mean LD among all the top candidate contigs (see *Methods: Linkage Disequilibrium*) and found that the co-association network visualization captured patterns of LD among the contigs through their common associations with environmental

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variables (Figure 2 lower diagonal, Supplementary Figure S7). There was higher than average LD within the submodules of the Multi, Aridity, and Freezing clusters, and very low LD between the Aridity cluster and the other groups (Figure 2 lower diagonal, Supplementary Figure S7). The LD among the other three groups (Multi, Freezing, and Geography) was small, but higher with each other than with Aridity. Thus, our approach captures the same information as simple LD-based clustering with the important additional benefit of linking LD clusters to likely environmental drivers of selection.

The high LD observed within the four main climate modules could arise via selection by the same aspect of the multivariate environment, via physical linkage on the chromosome, or both. We used a mapping population to disentangle these two hypotheses, by calculating recombination rates among the top candidates (see *Methods: Recombination rates*). Of the 117 top candidate contigs, 66 had SNPs that were represented in our mapping population. The recombination data revealed that all the genes in the Aridity cluster have strong LD and are physically linked (Figure 2). Within the other three clusters, we found physical proximity for only a few genes (but note that our mapping analysis does not have high power to infer recombination rate when loci are physically unlinked; see Methods). Interestingly, low recombination rates were estimated among some genes belonging to different environmental modules, even though there was little LD among these genes (Figure 2).

Comparison to conclusions based on principal components of environments

We compared the results from the co-association network analysis to associations with principal components (PC) of the environments. We used the same criteria ($\log_{10}$ BF > 2 in bayenv2) to determine if a locus was a significant outlier and compared (i) overlap with top candidates based on outliers from associations with environments, and (ii) interpretation of the selective

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

environment based on loadings of environments to PC axes. The first three PC axes explained 44% (PC1), 22% (PC2), and 15% (PC3) of the variance in the environments (80% total). Overall, 80% of the geography SNPs, 75% of the Freezing SNPs, 20% of the Aridity SNPs, and 10% of the Multi SNPs were not outliers along the first 10 PC axes and would have been missed by a study based on PC axes. Below, we outline whether interpretation of selective environment based on PCs is consistent with that based on associations with environments.

Some of the temperature and frost variables (MAT: mean annual temperature, EMT: extreme minimum temperature, DD0: degree days below 0C, DD5: degree days above 5C, bFFP: begin frost-free period, FFP: frost free period, eFFP: end frost free period, labels in Figure 1A) had the highest loadings for PC1 (Supplementary Figure S8). Almost all of the SNPs in the Multi cluster (90%) and 19% of SNPs in the freezing cluster were outliers along this axis (Supplementary Figure 9, less than 2% of candidate SNPs in the other clusters were outliers). For PC1, interpretation of the selective environment (e.g., MAT, DD0, FFP, eFFP, DD5) is somewhat consistent with the co-association network analysis (both Multi SNPs and Freezing SNPs show associations with all these variables, Figure 1B). However, the Multi SNPs and Freezing SNPs had strong associations with other variables (e.g., Multi SNPs showed strong associations with Latitude and Freezing SNPs showed strong associations with Longitude, Figure 1B) that did not load strongly onto this axis, and would have been missed in an interpretation based on associations with principal components.

For PC2, many precipitation and aridity variables loaded strongly onto this axis, including mean annual precipitation, annual heat:moisture index, climate moisture deficit, and precipitation as snow (Supplementary Figure 8). However, few top candidate loci were outliers along this PC

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

axis: only 13% of Freezing SNPs, 10% of Aridity SNPs, and less than 3% of Multi or Geography SNPs were outliers (Supplementary Figure 9).

For PC3, latitude, elevation, and two frost variables (beginning frost-free period and frost-free period) had the highest loadings (Supplementary Figure 9). The majority (78%) of the Aridity SNPs were outliers in PC3 (Supplementary Figure 10). Based on the PC association, this would lead one to conclude that the Aridity SNPs show associations with latitude, elevation, and frost-free period. While the Aridity SNPs do have high associations with latitude (5th row in Figure 1B), they show very low associations with the beginning of frost-free period, elevation, and frost-free period (3rd, 4th, and last row in Figure 1B, respectively). Thus, interpretation of the environmental drivers of selection based on associations with PC3 would have been very different from the univariate associations.

Visualization of multivariate allele associations

While the network visualization gave insight into patterns of LD among loci, it does not give insight into the patterns of allele frequency change on the landscape, relative to the ancestral state. As illustrated above, principal components would not be useful for this latter visualization. Instead, we accomplished this by plotting the association of a derived allele with one environmental variable against the association of that allele with a second environmental variable (example in Figure 3). Note that when the two environmental variables themselves are correlated on the landscape, an allele with a larger association in one environment will also have a larger association with a second environment, regardless of whether or not selection is shaping those associations. We can visualize (i) the expected genome-wide covariance using shading of quadrants and (ii) the observed genome-wide covariance using a 95% prediction ellipse (Figure 3, see *Methods: galaxy biplots*). Since alleles were coded according to their

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

putative ancestral state in loblolly pine (*Pinus taeda*), the location of any particular SNP in the plot represents the bivariate environment in which the derived allele is found in higher frequency than the ancestral allele (Figure 3). Visualizing the data in this way allows us to understand the underlying correlation structure of the data, as well as to develop testable hypotheses about the true selective environment and the fitness of the derived allele relative to the ancestral allele.

We overlaid the top candidate SNPs, colored according to their grouping in the co-association network analysis, on top of this genome-wide pattern. We call these plots galaxy biplots because of the characteristic patterns we observed when visualizing data this way (Figure 4). Galaxy biplots revealed that the Aridity group showed associations with hot/dry versus cold/wet environments, while the Multi and Freezing groups showed patterns of associations with hot/wet versus cold/dry environments (Figure 4a). These outlier patterns became visually more extreme for some SNPs after correcting the associations for population structure (Figure 4b). For the most part, the Freezing group showed associations with elevation but not latitude, while the Multi group showed associations with latitude but not elevation (Figure 4c, e). The structure correction polarized these patterns somewhat, suggesting that the structure-corrected allelic associations become more extreme when their pattern of allele frequency went against the background population structure (Figure 4d, f).

We also visualized the patterns of allele frequency on the landscape for a few representative SNPs, which were chosen because they had the highest degree (i.e., the highest number of connections) in their submodule (and more likely to be true positives, see *Simulated datasets*). An example of the allele frequency of a SNP located in the contig with the greatest number of top candidates from the Multi cluster (Contig #1 from Figure 1) is shown in Figure 5a, which illustrates why this SNP had significant associations with latitude and mean annual temperature.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Similarly, an example of a SNP in the Aridity cluster (Contig #8 from Figure 1) that had significant associations with annual heat:moisture index and latitude is shown in Figure 5b. These landscapes reveal the complex environments that may be selecting for particular combinations of genotypes despite potentially high gene flow in this widespread species.

Candidate gene annotations

Of the 47 candidate genes identified by Yeaman et al. 2016 as undergoing convergent evolution in lodgepole pine with the interior spruce hybrid complex, 10 were retained in our stringent criteria for top candidates. All of these contigs grouped into the Freezing and Geography clusters (shown by "*" in Figure 1G), which were the two clusters that had many SNPs with significant associations with elevation. This is consistent with the pattern of local adaptation in the interior spruce hybrid zone, whereby Engelmann spruce is adapted to higher elevations and white spruce is adapted to lower elevations (De La Torre *et al.* 2015).

Although many of the candidate genes were not annotated, as is typical for conifers, the genes underlying adaptation to these environmental gradients had diverse putative functions. The top candidate SNPs were found in 3' and 5' untranslated regions and open reading frames in higher proportions than in the entire dataset (Supplemental Figure S10). A gene ontology (GO) analysis using previously assigned gene annotations (Yeaman *et al.* 2014, 2016b) found that a single molecular function, solute:cation antiporter activity, was over-represented across all top candidates (Supplemental Table S1). In the "Aridity" and "Geography" clusters, annotated genes included sodium or potassium ion antiporters (one in the "Aridity"cluster, a KEA4 homolog, and two in the "Geography" cluster, NHX8 and SOS1 homologs), suggestive of a role in drought, salt or freezing tolerance (Blumwald *et al.* 2000). Genes putatively involved in auxin biosynthesis were also identified in the "Aridity" (YUCCA 3) and "Geography" (Anthranilate synthase

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

component) clusters (Supplemental Table S2), suggestive of a role in plant growth. In the "Freezing" and "Geography" clusters, several flowering time genes were identified (Ahlfors *et al.* 2004) including a homolog of CONSTANS (Amasino & Michaels 2010) in the "Freezing" cluster and a homolog of FY, which affects FCA mRNA processing, in the "Geography" cluster (Amasino & Michaels 2010) (Supp Table 2). In addition, several putative drought/stress response genes were identified, such as DREB transcription factor (Singh & Laxmi 2015) in the "Freezing" cluster, and an RCD1-like gene found with outlier SNPs in the "Geography" and "Freezing" clusters (Supp Table 2). RCD-1 is implicated in hormonal signaling and in the regulation of several stress-responsive genes in *Arabidopsis thaliana* (Ahlfors *et al.* 2004). In the "Multi" cluster the only gene that was annotated functions in acclimation of photosynthesis to the environment in *A. thaliana* (Walters *et al.* 2003).

Enrichment of co-expressed genes

To further explore if adaptation clusters have similar gene functions, we examined their gene expression patterns in response to climate treatments using previously published RNAseq data (Yeaman *et al.* 2014). We found that the "Freezing" cluster had an over-representation of the P2 co-regulated gene expression cluster ($P < 0.05$) with five (23%) of the "Freezing" genes found within the P2 expression network, revealing coordinated expression in response to climate conditions. Homologs of all five were present in *A. thaliana*, and four of these genes consisted of transcription factors involved in abiotic stress response (*DREB* transcription factor), flowering time (*CONSTANS*, pseudoresponse regulator) or floral development (floral homeotic protein *HUA1*). No other significant over-representation of gene expression class was identified for the four adaptation clusters or for all adaptation candidates.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Simulated datasets

On the simulated datasets, the *P*-value and Bayes factor criteria for choosing top candidate SNPs in the empirical data produced no false positives (Supplemental Figure 11), although using these criteria also reduced the proportion of true positives. Therefore, we used less stringent criteria to analyze the simulations so that we could also better understand patterns created by unlinked, false positive neutral loci.

We found that selected loci generally formed a single tightly connected co-association network even though they were unlinked, and that the degree of connectedness of selected loci was greater than among neutral loci (Figure 6). Thus, a single co-association module typically resulted from adaptation to a single selective environment. This occurred because of non-random associations in allele frequencies among selected loci due to selection by a common environmental factor. The propensity of neutral loci to form tightly-clustered co-association networks increased with the complexity of the demographic history: the false positive neutral loci from the two refugia model forming tightly connected networks (Figure 6 right column), despite the fact that all simulated loci were unlinked. This occurred because of non-random associations in allele frequency due to a shared demographic history. In some cases, selected loci formed separate or semi-separate modules according to their strengths of selection (e.g. Figure 6a, Supplementary Figure 12).

Discussion

Genetic architecture of adaptation: modular pleiotropy and linkage

Co-association networks provided a valuable framework for interpreting the genetic architecture of adaptation to a multivariate environment in lodgepole pine. We observed that SNPs from most genes associated with only a single climate module, which may be interpreted as limited

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

pleiotropic effects of genes on different aspects of climate adaptation. Within climate modules, we observed associations between genes and several environmental variables, which may or may not be interpreted as extensive pleiotropic effects within a climate module (depending on whether univariate environmental variables are considered distinct climates or collectively represent a single multivariate optimum). These observations are consistent with the Hypothesis of Modular Pleiotropy, and give insight into the ecological pressures that favor the evolution of modules by natural selection (Wagner *et al.* 2007). Interestingly, we found limited correspondence between co-expression modules (inferred by co-expression analysis) and the co-association modules detected here that are putatively favored by natural selection. This might be interpreted in support of the idea that the developmental/functional modularity of the genotype to phenotype map may not correspond to the modularity of the genotype to fitness map. However, power of the analysis could be low due to stringent statistical cutoffs and the pattern warrants further investigation.

Interestingly, we also observed physical linkage between genes that were associated with different climate modules (Figure 2). This is somewhat unexpected from a theoretical perspective, as selection would be expected to disfavour linkage and increase recombination between genes adapting to selection pressures with different spatial patterns of variation (Lenormand & Otto 2000; Guillaume 2011; Chebib & Guillaume 2017). Interestingly, while the linkage map suggests that these loci are sometimes located relatively close together on a single chromosome, this does not seem to be sufficient physical linkage to also cause a noticeable increase in LD. Thus it possible that the amount of physical linkage sometimes observed between genes in different modules is not strong enough to constrain adaptation to these differing gradients. Additional research and improved genetic maps will be required to explore these questions in greater depth. If this finding is robust and not compromised by false positives,

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

physical linkage among genes adapting to different climates could either facilitate or hinder a rapid evolutionary response as the multivariate environment changes (Aeschbacher & Bürger 2014; Reeve *et al.* 2016).

Within modules, we observed varying patterns of physical linkage among genes. The Aridity cluster, in particular, consisted of several tightly linked genes on a chromosome that may have arisen for a number of different reasons. Clusters of physically linked genes such as this may act as a single large-effect QTL (Christians & Senger 2007) and may have evolved due to competition among alleles or genomic rearrangements (Yeaman 2013, although these are rare in conifers), increased establishment probability due to linked adaptive alleles (Aeschbacher & Bürger 2014), or divergence within inversions (Kirkpatrick 2006). Alternatively, if the Aridity region was one of low recombination, a single causal variant could create the appearance of linked selection (Charlesworth *et al.* 1997), or a widespread false positive signal may have arisen due to genomic variation such as background selection and increased drift (Charlesworth *et al.* 1993; Charlesworth 2012; Hoban *et al.* 2016), or a widespread false signal may have arisen due to a demographic process such as allele surfing (Klopfstein *et al.* 2006; Hofer *et al.* 2009).

Physiological adaptation of lodgepole pine to climate

Our multivariate approach highlights the need to disentangle the physiological effects and importance of freezing versus drought in local adaptation in conifers. We found distinct groups of candidate loci along an axis of warm/wet to cold/dry (the Freezing and Multi modules), and another distinct group along an axis of cold/wet to warm/dry (the Aridity module). Selection by drought conditions in winter may occur through extensive physiological remodeling that allows

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cells to survive intercellular freezing by desiccating protoplasts - but also results in drought stress at the cellular level (Yeaman *et al.* 2014). Another type of winter injury in lodgepole pine - red belt syndrome - is caused by warm, often windy events in winter, when foliage desiccates but the ground is too cold for roots to be able to supply water above ground (Bella & Navratil 1987). This may contrast with drought selection in summer, when available soil water is lowest and aridity highest. The physiological and cellular mechanisms of drought and freezing response have similarities but also potentially important differences that could be responsible for the patterns that we have observed.

Our results provide a framework for developing hypotheses that will disentangle the specific drivers of selection and provide genotypes for assisted gene flow in reforestation (Aitken & Whitlock 2013). While climate change is expected to increase average temperatures across this region, some areas are experiencing more precipitation than historic levels and others experiencing less (Mbogga *et al.* 2009). Tree mortality rates are increasing across North America due to increased drought and vapour pressure deficit for tree species including lodgepole pine, and associated increased vulnerability to damaging insects, but growth rates are also increasing with warming temperatures and increased carbon dioxide (Hember *et al.* 2017a; b). Hot, dry valleys in southern BC are projected to have novel climates emerge that have no existing analogues in North America (Mahony *et al.* 2017). The considerable standing adaptive variation we observe here involving many genes could facilitate adaptation to new temperature and moisture regimes, or could hinder adaptation if novel climates are at odds with the physical linkage among alleles adapted to different climate stressors.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Limitations of associations with principal components

For these data, a PC-based association analysis would have led to a very limited interpretation of the environmental drivers of selection because the ordination is not biologically informed as to what factors are driving divergent selection. First, many putative candidates in the Freezing and Geography groups would have been missed. Second, strong associations between the Multi SNPs and environmental variables that did not load strongly onto PC1, such as latitude, would have also been missed. Finally, many Aridity SNPs were outliers in a PC axis (PC3) that was strongly correlated with variables that the Aridity SNPs did not have any significant associations with. This occurred because no single variable loaded strongly onto PC3 (the maximum loading of any single variable was 0.38) and many variables had moderate loadings on to PC3, such that no single environmental variable explained the majority of the variance (the maximum variance explained by any one variable was 15%). Thus, associations with higher PC axes become increasingly difficult to interpret when the axis itself explains less variance of the multivariate environment (PC3 explained 15%) and the environments loading onto that axis explain similar amounts of variance in that axis. While principal components will capture which environments tend to covary most together, this may have nothing to do with the combination of environmental factors that drive local adaptation to climate and it needlessly adds a layer of complexity to the analysis that may not represent anything biologically important. Rather, co-association networks highlight which combinations of environments are biologically important in partitioning the genes likely involved in adaptation.

Benefits and caveats of co-association networks

Co-association networks provide an intuitive framework to understand patterns of associations across many environment variables. By parsing loci into different groups based on their associations with environmental variables, this framework offers a more informative approach

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

than grouping loci according to their associations with single environmental variables. In practice, causal loci that show associations in a single climate variable such as temperature are not equal with respect to the way they adapt to climate as a whole. Nevertheless, correlation among environmental variables on the landscape, as well as unmeasured selective forces, will make it difficult to infer the exact conditions that select for particular allelic combinations. Results from the framework applied here make it easier, however, to generate hypotheses that can be tested with future experiments.

The analysis of simulated data shows that investigators should consider demographic history and choose candidates with caution for data analysis to exclude false positives, as we have attempted here. Co-association networks can arise among unlinked neutral loci by chance, and it is almost certain that some proportion of the "top candidates SNPs" in this study are false positives due to linkage with causal SNPs or due to demographic history. The simulated data also showed, however, that causal SNPs tend to have a higher degree in their co-association network than neutral loci, and this might help to prioritize SNPs for follow up experiments, SNP arrays, and genome editing.

Conclusions

As the climate changes, the evolutionary response will be determined by the extent of physical linkage among alleles selected by climate, in combination with the strength of selection and phenotypic optima across the environmental gradient, the scale and pattern of environmental variation, and the details of migration and demographic fluctuations across the landscape. While theory has made strides to provide a framework for predicting the genetic architecture of local adaptation under divergence with gene flow to a single environment (Bürger & Akerman 2011; Yeaman & Whitlock 2011; Kremer & Le Corre 2012; Le Corre & Kremer 2012; Yeaman 2013;

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Flaxman *et al.* 2013; Aeschbacher & Bürger 2014; Yeaman *et al.* 2016a), as well as the evolution of correlated traits under different directions and/or strengths of selection when those traits have a common genetic basis (Guillaume 2011; Chebib & Guillaume 2017), how genetic architectures evolve on complex heterogeneous landscapes has been understudied. Furthermore, it has been difficult to test theory because the field still lacks a framework for evaluating empirical observations of adaptation in many dimensions. Here, we have attempted to develop a framework toward understanding adaptation to several complex environments with different spatial patterns, which may also be useful for understanding the genetic basis of multivariate phenotypes from genome-wide association studies. This could lay the foundation for future studies to study modularity across the genotype-phenotype-fitness continuum.

Data Accessibility

The datasets and code used to create the results in this manuscript will be archived on Dryad upon acceptance of publication.

Author Contributions

KEL conceived of the analysis, conducted analyses, and lead writing of the manuscript. KH and SY did the bioinformatics and various specific analyses. JD created the allele frequency landscape plots. SA led the AdapTree project. All authors contributed to writing of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Acknowledgements

Mike Whitlock provided valuable advice and feedback on various aspects of the research. We thank Jeremy Yoder for organizing the SNP chip data used for calculating the recombination rates. Pia Smets, Connor Fitzpatrick and Sarah Markert assembled and grew genetic materials, and Kristin Nurkowski prepared sequence capture libraries. Tongli Wang and Andreas Hamann selected populations based on climatic distribution of species. Seeds were kindly donated by 63 forest companies and agencies in Alberta and British Columbia (listed at http://adaptree.forestry.ubc.ca/seed-contributors/).

References

Aeschbacher S, Bürger R (2014) The effect of linkage on establishment and survival of locally beneficial mutations. *Genetics*, **197**, 317–336.

Ahlfors R, Lång S, Overmyer K *et al.* (2004) *Arabidopsis* RADICAL-INDUCED CELL DEATH1 belongs to the WWE protein-protein interaction domain protein family and modulates abscisic acid, ethylene, and methyl jasmonate responses. *Plant Cell*, **16**, 1925–1937.

Aitken SN, Whitlock MC (2013) Assisted gene flow to facilitate local adaptation to climate change. *Annu. Rev. Ecol. Evol. Syst.*, **44**, 367–388.

Alberto FJ, Aitken SN, Alía R *et al.* (2013) Potential for evolutionary responses to climate change - evidence from tree populations. *Glob. Chang. Biol.*, **19**, 1645–1661.

Alexa A, Rahnenführer J (2009) Gene set enrichment analysis with topGO.

Amasino RM, Michaels SD (2010) The timing of flowering. *Plant Physiol.*, **154**, 516–520.

Barton NH (1983) MULTILOCUS CLINES. *Evolution*, **37**, 454–471.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Barton NH (1999) Clines in polygenic traits. *Genet. Res.*, **74**, 223–236.

Barton NH (2010) Genetic linkage and natural selection. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **365**, 2559–2569.

Bella IE, Navratil S (1987) Growth losses from winter drying (red belt damage) in lodgepole pine stands on the east slopes of the Rockies in Alberta. *Canadian Journal of Forest Research.* , **17**, 1289–1292.

Benestan L, Quinn BK, Maaroufi H *et al.* (2016) Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in American lobster (*Homarus americanus*). *Mol. Ecol.*, **25**, 5073–5092.

Blair LM, Granka JM, Feldman MW (2014) On the stability of the Bayenv method in assessing human SNP-environment associations. *Hum. Genomics*, **8**, 1.

Blumwald E, Aharon GS, Apse MP (2000) Sodium transport in plant cells. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, **1465**, 140–151.

Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: From polygenic to omnigenic. *Cell*, **169**, 1177–1186.

Bürger R, Akerman A (2011) The effects of linkage and gene flow on local adaptation: A two-locus continent–island model. *Theor. Popul. Biol.*, **80**, 272–288.

Charlesworth B (2012) The effects of deleterious mutations on evolution at linked sites. *Genetics*, **190**, 5–22.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res.*, **70**, 155–174.

Chebib J, Guillaume F (2017) What affects the predictability of evolutionary constraints using a G-matrix? The relative effects of modular pleiotropy and mutational correlation. *Evolution*.

Christians JK, Senger LK (2007) Fine mapping dissects pleiotropic growth quantitative trait locus into linked loci. *Mamm. Genome*, **18**, 240–245.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conesa A, Götz S (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int. J. Plant Genomics*, **2008**, 619832.

Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**, 1–9.

Daly C, Halbleib M, Smith JI *et al.* (2008) Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology*, **28**, 2031–2064.

De La Torre A, Ingvarsson PK, Aitken SN (2015) Genetic architecture and genomic patterns of gene flow between hybridizing species of Picea. *Heredity*, **115**, 153–164.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Eckert AJ, Bower AD, González-Martínez SC *et al.* (2010a) Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology*, **19**, 3789–3805.

Eckert AJ, van Heerwaarden J, Wegrzyn JL *et al.* (2010b) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda* L., Pinaceae). *Genetics*, **185**, 969–982.

Felsenstein J (1976) The theoretical population genetics of variable selection and migration. *Annual Review of Genetics*, **10**, 253–280.

Flaxman SM, Feder JL, Nosil P (2013) Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. *Evolution*, **67**, 2577–2591.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

Griswold CK (2006) Pleiotropic mutation, modularity and evolvability. *Evolution & Development*, **8**, 81–93.

Guillaume F (2011) Migration-induced phenotypic divergence: the migration-selection balance of

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlated traits. *Evolution*, **65**, 1723–1738.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Haldane JBS (1930) A mathematical theory of natural and artificial selection (Part VI, Isolation). *Mathematical Proceedings of the Cambridge Philosophical Society*, **26**, 220.

Haldane JBS (1948) The theory of a cline. *Journal of Genetics*, **48**, 277–284.

Hancock AM, Brachi B, Faure N *et al.* (2011) Adaptation to climate across the *Arabidopsis thaliana* genome. *Science*, **334**, 83–86.

Hansen TF (2006) The evolution of genetic architecture. *Annual Review of Ecology, Evolution, and Systematics*, **37**, 123–157.

Hedrick PW (1986) Genetic polymorphism in heterogeneous environments: a decade later. *Annual Review of Ecology and Systematics*, **17**, 535–566.

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual review of ecology and systematics*, **7**, 1–32.

Hember RA, Kurz WA, Coops NC (2017a) Increasing net ecosystem biomass production of Canada's boreal and temperate forests despite decline in dry climates. *Global Biogeochemical Cycles*, **31**, 2016GB005459.

Hember RA, Kurz WA, Coops NC (2017b) Relationships between individual-tree mortality and water-balance variables indicate positive trends in water stress-induced tree mortality across North America. *Global Change Biology*, **23**, 1691–1710.

Hoban S, Kelley JL, Lotterhos KE *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *The American Naturalist*, **188**, 379–397.

Hodgins KA, Yeaman S, Nurkowski KA, Rieseberg LH, Aitken SN (2016) Expression divergence Is correlated with sequence evolution but not positive selection in conifers. *Molecular Biology and Evolution*, **33**, 1502–1516.

Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

selection. *Annals of Human Genetics*, **73**, 95–108.

Howe GT, Aitken SN, Neale DB *et al.* (2003) From genotype to phenotype: unraveling the complexities of cold adaptation in forest trees. *Canadian Journal of Botany*, **81**, 1247–1266.

Illingworth K (1978) Study of lodgepole pine genotype-environment interaction in B.C. In: *Proceedings International Union of Forestry Research Organizations (IUFRO) Joint Meeting of Working parties: Douglas-fir provenances, Lodgepole Pine Provenances, Sitka Spruce Provenances and Abies Provenances*, pp. 151–158. Vancouver, British Columbia, Canada.

Kaufman L, Rousseeuw PJ (2009) *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kirkpatrick M (2006) Chromosome inversions, local adaptation and speciation. *Genetics*, **173**, 419–434.

Klopfstein S, Currat M, Excoffier L (2006) The fate of mutations surfing on the wave of a range expansion. *Molecular Biology and Evolution*, **23**, 482–490.

Kremer A, Le Corre V (2012) Decoupling of differentiation between traits and their underlying genes in response to divergent selection. *Heredity*, **108**, 375–385.

Lasky JR, Des Marais DL, McKay JK *et al.* (2012) Characterizing genomic variation of *Arabidopsis thaliana*: the roles of geography and climate. *Molecular Ecology*, **21**, 5512–5529.

Le Corre V, Kremer A (2003) Genetic variability at neutral markers, quantitative trait land trait in a subdivided population under selection. *Genetics*, **164**, 1205–1219.

Le Corre V, Kremer A (2012) The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, **21**, 1548–1566.

Le Nagard H, Chao L, Tenaillon O (2011) The emergence of complexity and restricted pleiotropy in adapting networks. *BMC Evolutionary Biology*, **11**, 326.

Lenormand T, Otto SP (2000) The evolution of recombination in a heterogeneous environment. *Genetics*, **156**, 423–438.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* , **25**, 1754–1760.

Liepe KJ, Hamann A, Smets P, Fitzpatrick CR, Aitken SN (2016) Adaptation of lodgepole pine and interior

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

spruce to climate: implications for reforestation in a warming world. *Evolutionary Applications*, **9**, 409–419.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, **23**, 2178–2192.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular ecology*, **24**, 1031–1046.

Mahony CR, Cannon AJ, Wang T, Aitken SN (2017) A closer look at novel climates: new methods and insights at continental to landscape scales. *Global Change Biology*.

Margarido GRA, Souza AP, Garcia AAF (2007) OneMap: software for genetic mapping in outcrossing species. *Hereditas*, **144**, 78–79.

Mbogga MS, Hamann A, Wang T (2009) Historical and projected climate data for natural resource management in western Canada. *Agricultural and Forest Meteorology*, **149**, 881–890.

Neale DB, Wegrzyn JL, Stevens KA *et al.* (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biology*, **15**, R59.

Orr HA (2000) Adaptation and the cost of complexity. *Evolution*, **54**, 13–20.

Paaby AB, Rockman MV (2013) The many faces of pleiotropy. *Trends in Genetics*, **29**, 66–73.

Pison G, Struyf A, Rousseeuw PJ (1999) Displaying a clustering with CLUSPLOT. *Computational Statistics & Data Analysis*, **30**, 381–392.

Reeve J, Ortiz-Barrientos D, Engelstädter J (2016) The evolution of recombination rates in finite populations during ecological speciation. *Proceedings Biological Sciences / The Royal Society*, **283**.

Savolainen O, Lascoux M, Merilä J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Singh D, Laxmi A (2015) Transcriptional regulation of drought response: a tortuous network of transcriptional factors. *Frontiers in Plant Science*, **6**, 895.

Slatkin M (1973) Gene flow and selection in a cline. *Genetics*, **75**, 733–756.

Slatkin M (1978) Spatial patterns in the distributions of polygenic characters. *Journal of Theoretical Biology*, **70**, 213–228.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Stearns FW (2010) One hundred years of pleiotropy: a retrospective. *Genetics*, **186**, 767–773.

Suren H, Hodgins KA, Yeaman S *et al.* (2016) Exome capture from the spruce and pine giga-genomes. *Molecular Ecology Resources*, **16**, 1136–1146.

Titterington DM (1976) Algorithms for computing D-optimal design on finite design spaces. *Proceedings of the 1976 Conference on Information Science and Systems*, 213–216.

Wagner GP (1996) Homologues, natural kinds and the evolution of modularity. *American Zoologist*, **36**, 36–43.

Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. *Nature Reviews Genetics*, **8**, 921–931.

Wagner GP, Zhang J (2011) The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, **12**, 204–213.

Walters RG, Shephard F, Rogers JJM, Rolfe SA, Horton P (2003) Identification of mutants of *Arabidopsis* defective in acclimation of photosynthesis to the light environment. *Plant Physiology*, **131**, 472–481.

Wang T, Hamann A, Spittlehouse DL, Murdock TQ (2012) ClimateWNA—high-resolution spatial climate data for western North America. *Journal of Applied Meteorology and Climatology*, **51**, 16–29.

Wang Z, Liao B-Y, Zhang J (2010) Genomic patterns of pleiotropy and the evolution of complexity. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 18034–18039.

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E1743–51.

Yeaman S, Aeschbacher S, Bürger R (2016a) The evolution of genomic islands by increased establishment probability of linked alleles. *Molecular Ecology*, **25**, 2542–2558.

Yeaman S, Hodgins KA, Lotterhos KE *et al.* (2016b) Convergent local adaptation to climate in distantly related conifers. *Science*, **353**, 1431–1433.

Yeaman S, Hodgins KA, Suren H *et al.* (2014) Conservation and divergence of gene expression plasticity following c. 140 million years of evolution in lodgepole pine (*Pinus contorta*) and interior spruce (*Picea glauca*×*Picea engelmannii*). *The New Phytologist*, **203**, 578–591.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Yeaman S, Whitlock MC (2011) The genetic architecture of adaptation under migration-selection balance. *Evolution*, **65**, 1897–1911.

Figure Legends

Figure 1. Co-association network analysis for *Pinus contorta*. A) Correlations among environments. B) Hierarchical clustering of associations among allele frequencies (of SNPs in columns) with environments (in rows). C-F) Undirected graph networks of the 4 major groups of SNPs from the hierarchical clustering. Each node is a SNP and is labeled with a number and color according to its exome contig. For the the larger submodules, the number of independent exome contigs is indicated. The color coding of the nodes is shown along the x-axis in the bottom graph (G), which is a histogram of the number of top candidate SNPs in each contig, according which of the 4 major groups each SNP clusters with. Contigs previously identified as undergoing convergent evolution with spruce by Yeaman et al. 2016 are indicated with "*".

Figure 2. Comparison of linkage disequilibrium (measured as correlation in allele frequencies, lower diagonal) and recombination rates (upper diagonal) for exome contigs in the mapping population. Rows and column labels correspond to Figure 1G. Darker areas represent either high physical linkage (low recombination) or high statistical linkage disequilibrium.

Figure 3. Overview of galaxy biplots. The association between allele frequency and one variable is plotted against the association between allele frequency and a second variable. The Spearman's ρ correlation between the two variables (mean annual temperature or MAT and mean annual precipitation or MAP in this example) is shown in the lower right corner. When the

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

two variables are correlated, genome-wide covariance is expected to occur in the direction of their association (shown with quadrant shading in light grey). The observed genome-wide distribution of allelic effects is plotted in dark grey and the 95% prediction ellipse is plotted as a black line. Because derived alleles were coded as 1 and ancestral alleles were coded as 0, the location of any particular SNP in bivariate space represents the type of environment that the derived allele is found in higher frequency, whereas the location of the ancestral allele would be a reflection through the origin (note only derived alleles are plotted).

Figure 4. Galaxy biplots for different environmental variables for regular (left column) and structure-corrected (right column) associations. Top candidate SNPs from the 4 groups are highlighted against the genome-wide background. The internal color of each point corresponds to the contig that SNP is located within (as shown on the x-axis in Figure 1G), while the outline color of each point corresponds to the multivariate environment that SNP clusters with. Top row: mean annual temperature (MAT) vs. mean annual precipitation (MAP), middle row: MAT and Elevation, bottom row: MAT and latitude (LAT).

Figure 5. Pie charts represent the frequency of a derived allele across the landscape, overlain on top of an environment that the SNP shows significant associations with. The mean environment for each population is shown by the color of the outline around the pie chart. A) Allele frequency pattern for a SNP from contig 1 in the Multi cluster from Figure 1. The derived allele had negative associations with temperature but positive associations with latitude. B) Allele frequency pattern for a SNP from contig 8 in the Aridity cluster. The derived allele had negative associations with annual:heat moisture index (and other measures of aridity) and positive associations with latitude. SNPs were chosen as those with the highest degree in their submodule.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 6. Comparison of co-association networks resulting from simulated data for 3 demographies: A) isolation by distance (IBD), B) range expansion from a single refuge, and C) range expansion from two refugia. All SNPs were simulated unlinked and 1% of SNPs were simulated under selection to an unmeasured weak latitudinal cline. Boxplots of degree of connectedness of a SNP as a function of its strength of selection, across all replicate simulations (top row). Examples of networks formed by datasets that were neutral-only (middle row) or neutral+selected (bottom row) outlier loci.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Tables

Table S1. Results from GO analysis for all top candidates and for each cluster. The top 5 processes are shown for each category. P represents the P-value from parent-child Fisher test, while "fdr" represents significance after correction for false discovery rate.

Table S2. Top candidate exome contigs and their annotations. For each contig the following information is indicated: the number of outlier SNPs in each cluster (“Multi”, “Aridity”, “Freezing”, or “Geography”), the color used for plotting (“col”), the cluster the contig is assigned to according to the majority of outlier SNPs (“cluster”), whether or not its homolog shows convergent signals of adaptation with spruce (“is.covergent”), TAIR ID (“tair”), and putative gene function (“Annotations”).

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Figures

Figure S1. Histogram of X^TX estimated from Bayenv2 for all SNPs (top) and for top candidate SNPs (bottom).

Figure S2. Undirected graph network for the Multi group (enlarged version of Figure 1C).

Figure S3. Undirected graph network for the Aridity group (enlarged version of Figure 1D).

Figure S4. Undirected graph network for the Freezing group (enlarged version of Figure 1E).

Figure S5. Undirected graph network for the Geography group (enlarged version of Figure 1F).

Figure S6. Heatmap of structure-corrected allele associa- tions with the environment, analogous to Figure 1B in the main paper. Note that although the pattern is very similar, the magnitude of allele correlations is smaller in the structure-corrected data.

Figure S7. Mean correlation among allele frequencies between top candidate contigs. Contigs are ordered the same as Figure 1G in the main paper.

Figure S8. The length and direction of each vector represents the scaled loading of that environmental variable onto the PC axis. The color of each vector represents the mean proportion of variance explained by that environment in the two axes plotted.

Figure S9. The distribution of Bayes Factors for the association between SNPs and environments along the first three PC axes. Colored points correspond to the candidate described in the main paper: Aridity (orange), Multi (green), Freezing (blue), and Geography (yellow). Vertical and horizontal lines represent criteria for significance. Note that candidate SNPs all had BF > 2 with at least one univariate environmental variable.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure S10. Proportion of SNPs falling into various categories for genomic features in the entire dataset compared to in the top candidate list. 3primeFLANK: 3' flanking region; 3primeUTR: 3' untranslated region; 5primeFLANK: 5' flanking region; 5primeUTR: 5' untranslated region; non-tcontig: not located in a transcriptomic contig (intergenic); nonsyn: non-synonymous substitution; unk-adj: unknown adjacent region; unk-flank: unknown flanking region; UNKNOWN-ORF: unknown open reading frame.

Figure S11. Error rates from the simulations given a less stringent criteria (Bonferroni, left) and a more stringent criteria (Bonferroni and Bayes Factors from bayenv2, right). The less stringent criteria was used for the simulations because it had some false positives (A), while the more stringent criteria was used for the empirical data because it didn't have any false positives (B). While using the more stringent criteria resulted in no false positives, it also reduced the number of true positives (compare C and D), with the most severe reduction under isolation by distance.

Figure S12. The simulated datasets were nested within randomly generated selective environments, such that different demographic histories were simulated on the same environmental landscape. For this randomly generated environment, loci simulated under stronger selection had a propensity to cluster differently than loci simulated under weaker selection.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figures

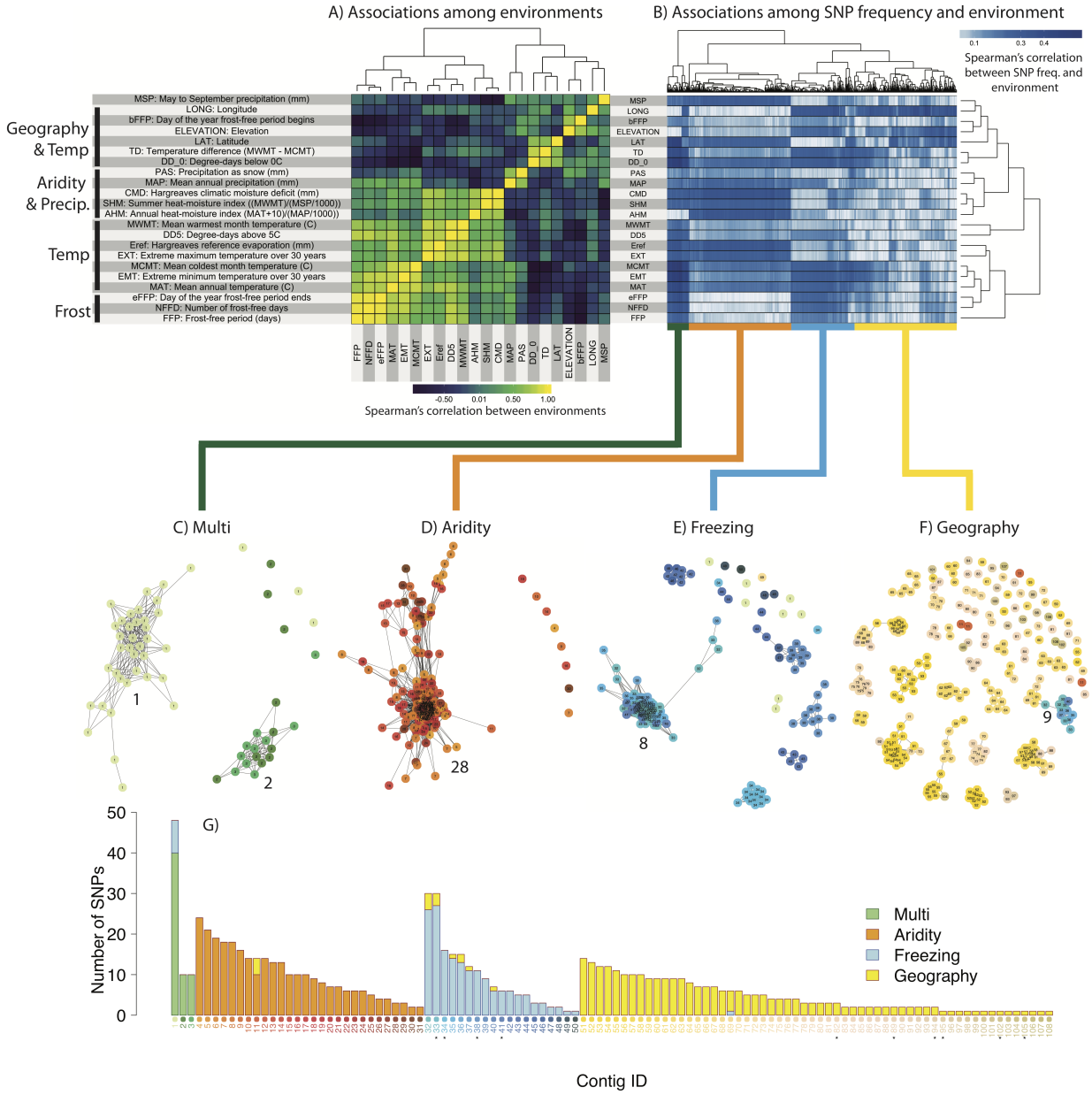

FIGURE 1. Co-association network analysis for *Pinus contorta*. A) Correlations among environments. B) Hierarchical clustering of associations among allele frequencies (of SNPs in columns) with environments (in rows). C-F) Undirected graph networks of the 4 major groups of SNPs from the hierarchical clustering. Each node is a SNP and is labeled with a number and color according to its exome contig. For the the larger submodules, the number of independent exome contigs is indicated. The color coding of the nodes is shown along the x-axis in the bottom graph (G), which is a histogram of the number of top candidate SNPs in each contig, according which of the 4 major groups each SNP clusters with. Contigs previously identified as undergoing convergent evolution with spruce by Yeaman et al. 2016 are indicated with "*".

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

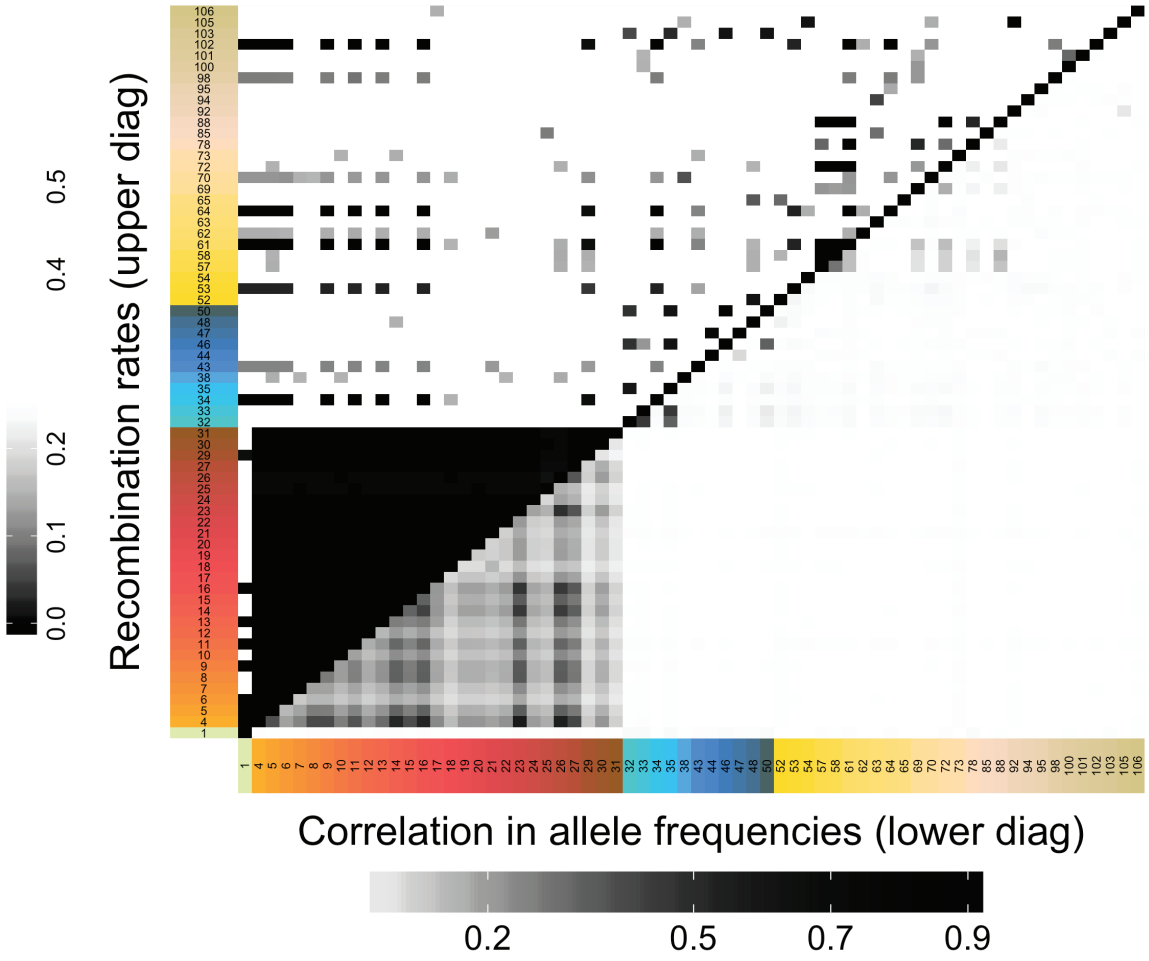

FIGURE 2. Comparison of linkage disequilibrium (measured as correlation in allele frequencies, lower diagonal) and recombination rates (upper diagonal) for exome contigs in the mapping population. Rows and column labels correspond to Figure 1G. Darker areas represent either high physical linkage (low recombination) or high statistical linkage disequilibrium.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

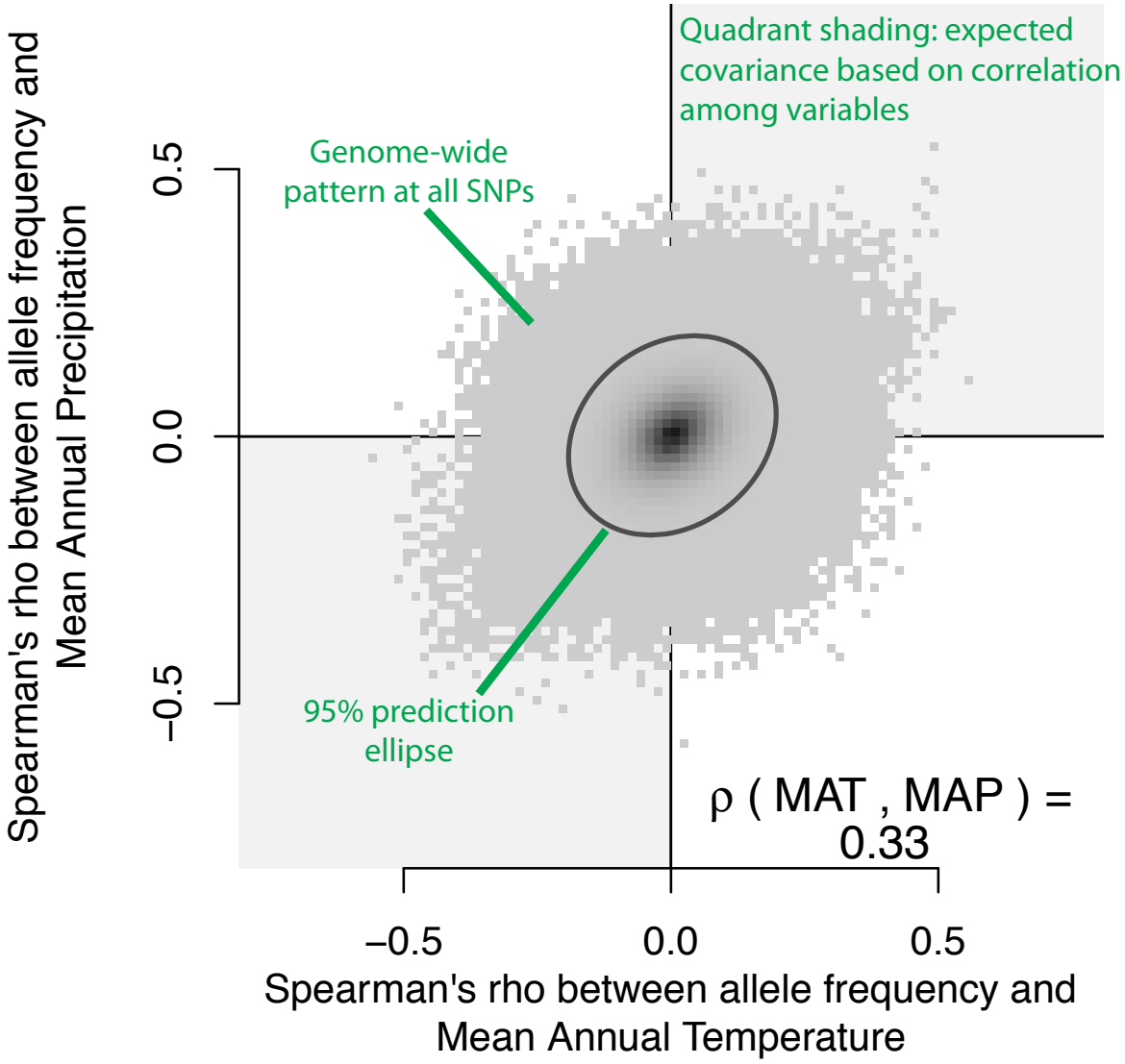

FIGURE 3. Overview of galaxy biplots. The association between allele frequency and one variable is plotted against the association between allele frequency and a second variable. The Spearmans ρ correlation between the two variables (mean annual temperature or MAT and mean annual precipitation or MAP in this example) is shown in the lower right corner. When the two variables are correlated, genome-wide covariance is expected to occur in the direction of their association (shown with quadrant shading in light grey). The observed genome-wide distribution of allelic effects is plotted in dark grey and the 95% prediction ellipse is plotted as a black line. Because derived alleles were coded as 1 and ancestral alleles were coded as 0, the location of any particular SNP in bivariate space represents the type of environment that the derived allele is found in higher frequency, whereas the location of the ancestral allele would be a reflection through the origin (note only derived alleles are plotted).

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

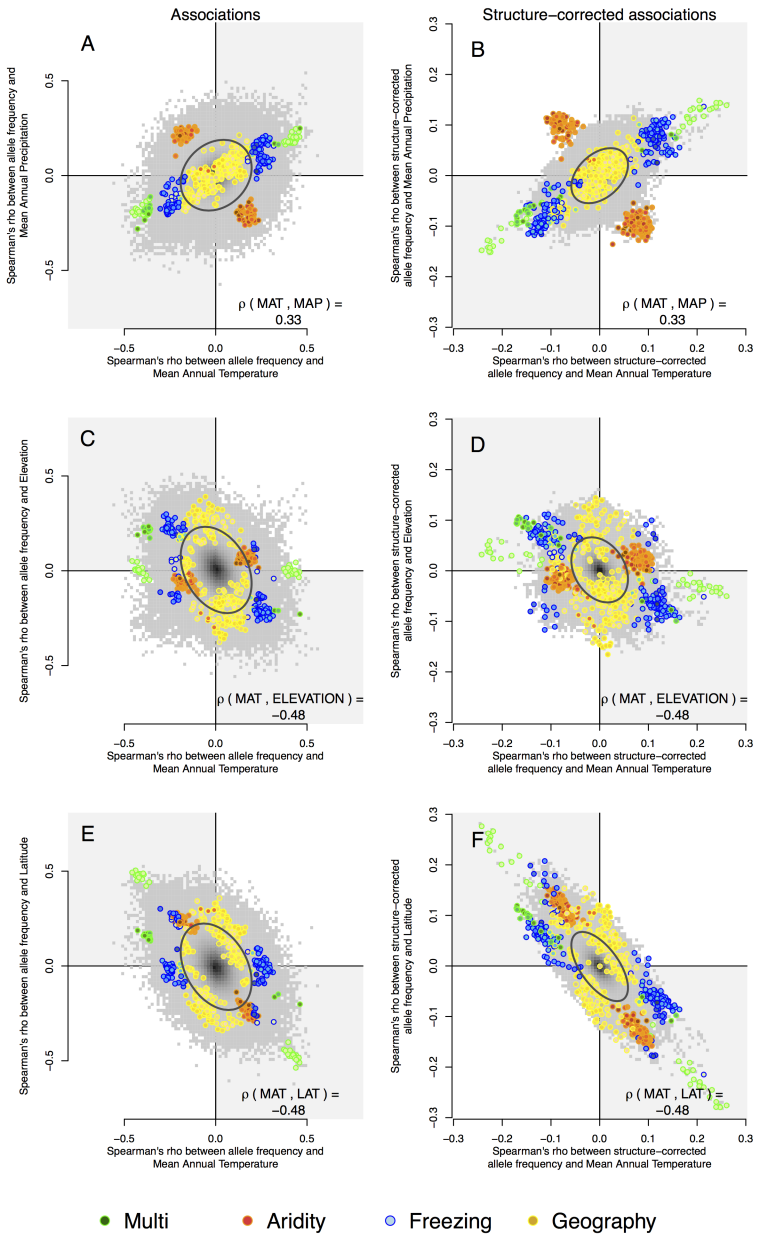

FIGURE 4. Galaxy biplots for different environmental variables for regular (left column) and structure-corrected (right column) associations. Top candidate SNPs from the 4 groups are highlighted against the genome-wide background. The internal color of each point corresponds to the contig that SNP is located within (as shown on the x-axis in Figure 1G), while the outline color of each point corresponds to the group that SNP clusters with. Top row: mean annual temperature (MAT) vs. mean annual precipitation (MAP), middle row: MAT and Elevation, bottom row: MAT and latitude (LAT).

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

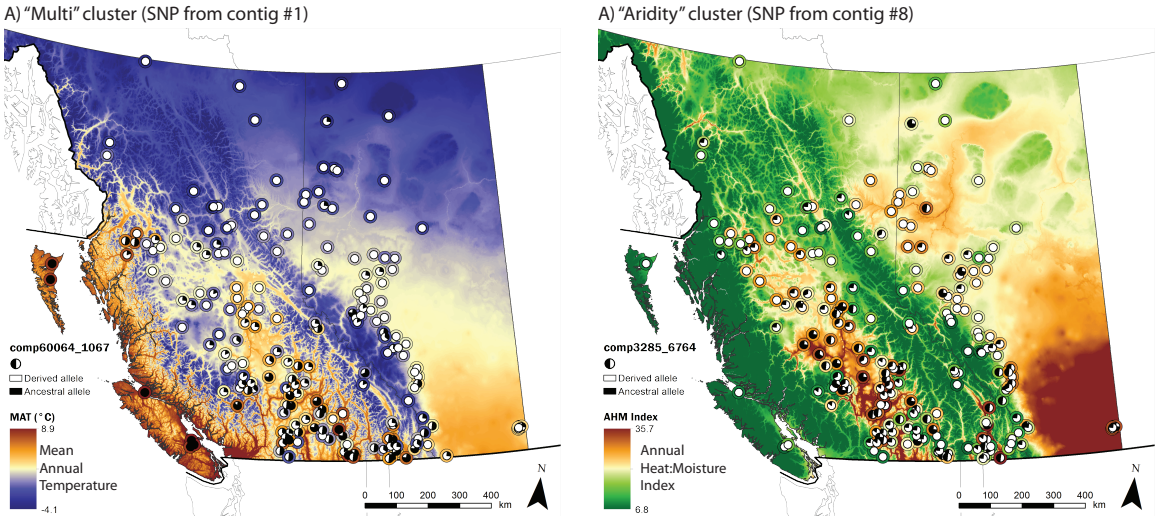

FIGURE 5. Pie charts represent the frequency of a derived allele across the landscape, overlain on top of an environment that the SNP shows significant associations with. The mean environment for each population is shown by the color of the outline around the pie chart. A) Allele frequency pattern for a SNP from contig 1 in the Multi cluster from Figure 1. The derived allele had negative associations with temperature but positive associations with latitude. B) Allele frequency pattern for a SNP from contig 8 in the Aridity cluster. The derived allele had negative associations with annual:heat moisture index (and other measures of aridity) and positive associations with latitude. SNPs were chosen as those with the highest degree in their submodule.

bioRxiv preprint doi: https://doi.org/10.1101/202481; this version posted October 13, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

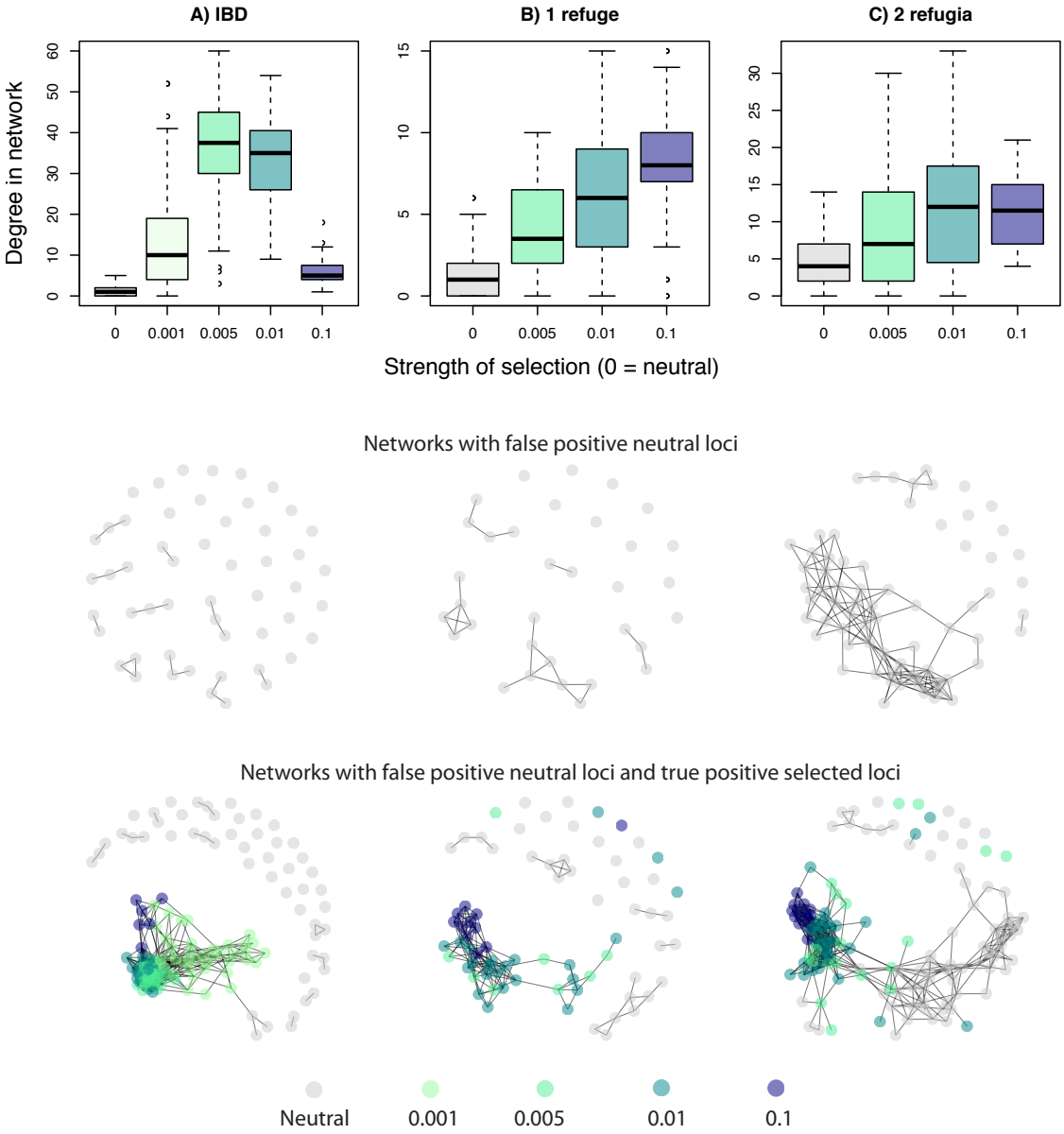

FIGURE 6. Comparison of co-association networks resulting from simulated data for 3 demographies: A) isolation by distance (IBD), B) range expansion from a single refuge, and C) range expansion from two refugia. All SNPs were simulated unlinked and 1% of SNPs were simulated under selection to an unmeasured weak latitudinal cline. Boxplots of degree of connectedness of a SNP as a function of its strength of selection, across all replicate simulations (top row). Examples of networks formed by datasets that were neutral-only (middle row) or with neutral+selected (bottom row) outlier loci.