NOTE: I am going to start using Paper Pile for references. You can find it under “Add-ons”

Thoughts on where to submit:

- PNAS (if I can make it short and people think it is worthy) impact factor 9.4
- Evolution letters 10.772
- Plos genetics 6.6
- Molecular Ecology 5.9
- Genetics 5.9
-

Abstract (ROUGH) **5**

Introduction **5**

Methods (possible to just cite Yeaman?) **6**

Populations Sampled 6

Characterization of the Environment 6

Sequencing 7

Bioinformatics 8

Identification of top candidate contigs 8

Identification of candidate SNPs 9

Visualization of allelic effects: introducing galaxy plots 9

Significance of patterns at candidate SNPs in galaxy plots 13

Annotation 13

Other data visualization? 14

Results **14**

Candidate SNPs 14

Visualization with galaxy plots 14

Annotation 21

Other landscape visualization (mapping?) 21

Discussion 22

Data Accessibility 23

Author Contributions 23

Acknowledgements 23

References 23

Multidimensional data visualization provides novel insights into adaptation to multivariate environments

Authors (“?” to be discussed with Sally):

Katie Lotterhos

Kay Hodgins

Sam Yeaman

Jon Denger (?)

Haktan Suren (?)

Kristin Nurkowski (?)

Pia Smets (?)

Simon Nadeau (?)

Tongli Wang (?)

Andreas Hamann (?)

Jason Holliday (?)

Mike Whitlock

Loren Rieseberg (?)

Sally Aitken

Keywords:

Running title:

Abstract (ROUGH)

A basic line of inquiry in evolutionary genetics aims to understand the genetic basis of local adaptation, and reverse ecology has become a common approach used to fulfill this goal. The limitation of the reverse ecology approach, as it has been widely presented in the literature, is that little can be learned about evolutionary processes by listing specific loci as outliers relative to the genome-wide distribution. Here, we take reverse ecology analyses to the next level by employing a novel data visualization and testing approach that identifies different types of behavior in the candidate loci across multiple, correlated environments. We apply this new approach to an exome capture dataset across 281 populations of lodgepole pine (*Pinus contorta*), and identify two groups of loci that are associated with the multivariate environment in different ways. The first group had one allele adapting to a hot/wet environment and the second allele adapting to a cold/dry environment, while the second group had one allele adapting to a hot/dry environment and the other allele adapting to a cold/wet environment. These different groups suggest that different combinations of environmental pressures have been shaping adaptation via responses in suites of genes, which would not have been discovered without our novel data visualization and testing method. This novel visualization approach, which we call galaxy plots, can be used for any combination of correlated variables (environment and/or phenotype) and is implemented in an R package.

Introduction

A basic line of inquiry in evolutionary genetics aims to understand the genetic basis of local adaptation (Stinchcombe & Hoekstra 2008). Local adaptation is the evolution by natural selection of increased fitness to local conditions, and is crucial to our understanding of the process of evolution (Lenormand 2002; Blanquart et al. 2013). The extent of adaptation to local conditions can determine the geographic range of a population, its relative fitness compared to other species, its response to environmental change, and perhaps even the probability of extinction (Kawecki & Ebert 2004, Savolainen et al 2014).

Recently, the ‘reverse ecology’ approach (sensu Li et al. 2008) has become a popular means to find the targets of adaptive natural selection, because there is only limited knowledge of the specific traits responsible for adaptation across environmental gradients. Genetic-environment associations (GEAs) have become a major statistical approach applied in reverse ecology studies to identify putative targets of selection (REFS, reviews). GEAs measure the association between allele frequencies and environments, thus capturing associations for unmeasured traits that are responsible for adaptation across environmental gradients.

The limitation of the reverse ecology approach, as it has been widely presented in the literature, is that little can be learned about ecological factors driving adaptation by listing specific loci as outliers relative to the genome-wide distribution. Moreover, many of the phenotypic and environmental variables that are used in GEAs are highly correlated with each other, and it isn’t always clear how to take into account this correlation structure when identifying candidates.

While the identification of candidate loci is an important step towards understanding the genetic basis of adaptation, it is possible that key pieces of the evolutionary puzzle are being missed by ignoring the direction that specific alleles evolve in response to multifaceted environmental stress.

Here, we take reverse ecology analyses to the next level by visualizing the direction of correlation of specific alleles with multiple environmental or phenotypic variables, and testing whether the observed patterns are strongly differentiated from patterns in the genomic background. Our strategy was to first identify genes or genomic contigs that had strong evidence of being involved in adaptation because they possessed more outliers than expected by random chance, and then to identify candidate SNPs within these genes or contigs. Using this set of candidates, we then employed a novel data visualization approach that allowed us to identify different types of behavior in the candidate alleles across multiple, correlated environments and phenotypes.

Study system - local adaptation and climate change and trees (Sally)

- more sections to set up focus questions and results
- Freezing and aridity as selective pressure on pine

Methods

Populations Sampled

We obtained 281 seedlots of lodgepole pine from from available operational reforestation seedlots collected in natural populations (Figures/Pine_samples_tongli.png). Each contained seed bulked from at least 10 seed parents in British Columbia, and at least 30 seed parents in Alberta. 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).

Characterization of the Environment

Climate variables were generated using ClimateWNA (Wang et al. 2012). The program extracts and downscales 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 analyzed genetic patterns for 19 climatic and 3 geographical variables (latitude, longitude, and elevation) (Table-Environment). Geographic variables may correlate with some unmeasured climate 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-Envi-Dendro).

Abbreviation	**Variable**
LAT	Latitude
LONG	Longitude
ELEVATION	Elevation
MAT	Mean annual temperature (C)
MWMT	Mean warmest month temperature (C)
MCMT	Mean coldest month temperature (C)
TD	Temperature difference (MWMT - MCMT)
MAP	Mean annual precipitation (mm)
MSP	May to September precipitation (mm)
AHM	Annual heat-moisture index (MAT+10)/(MAP/1000))
SHM	Summer heat-moisture index ((MWMT)/(MSP/1000))
DD_0	Degree-days below 0C
DD5	Degree-days above 5C
NFFD	Number of frost-free days
bFFP	Day of the year frost-free period begins
eFFP	Day of the year frost-free period ends
FFP	Frost-free period (days)
PAS	Precipitation as snow (mm)
EMT	Extreme minimum temperature over 30 years
EXT	Extreme maximum temperature over 30 years
Eref	Hargreaves reference evaporation (mm)
CMD	Hargreaves climatic moisture deficit (mm)

Table-Environment: Abbreviations of environmental variables used in this study.

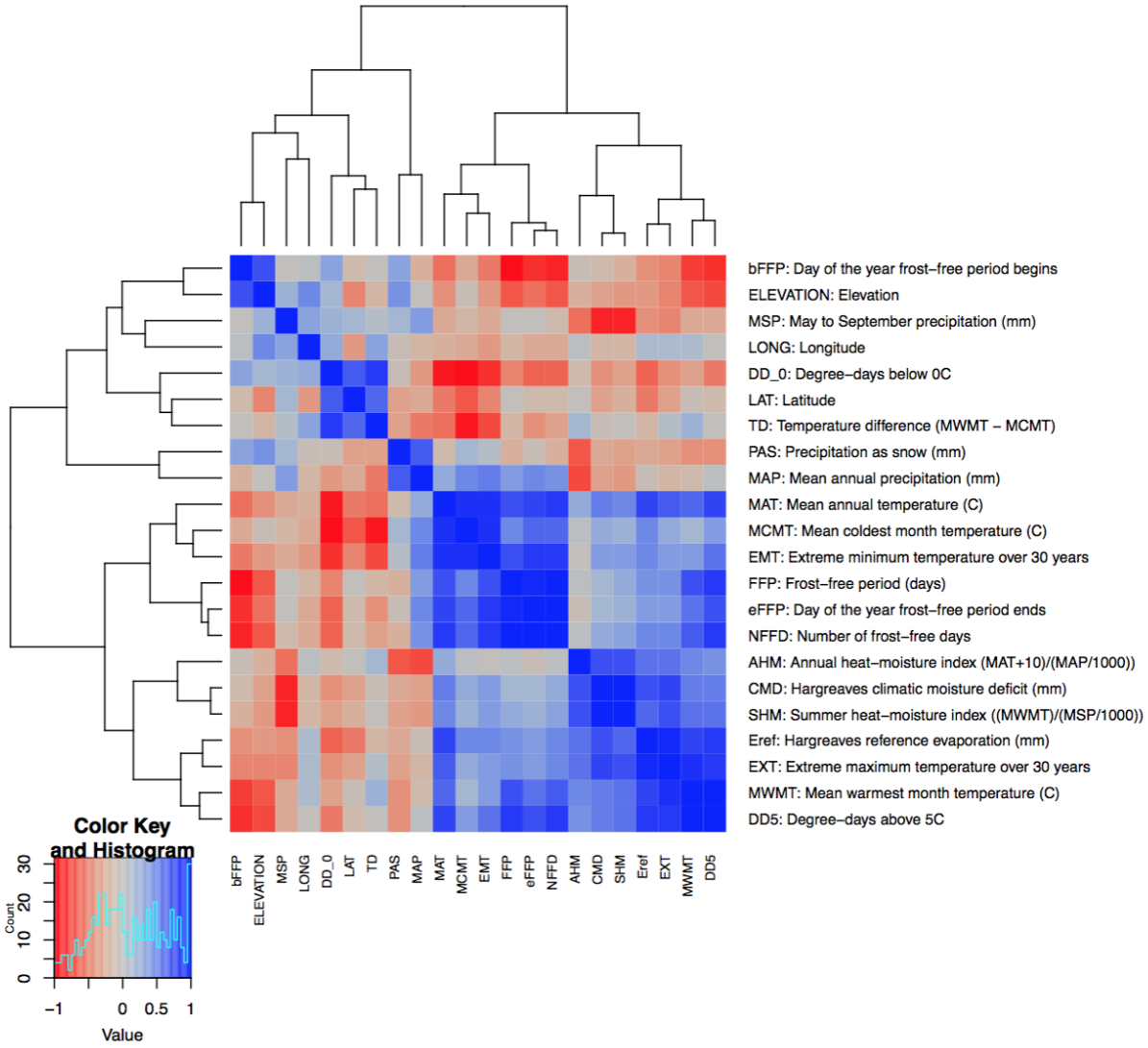


Figure-Envi-Dendro. Correlations among environmental variables used in this study. (Note that I made this for Sam's science paper and this is a subset of the entire figure that is in the supplement of that paper)

Sequencing

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 for more details) and then 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, the pool of six libraries was pooled with another completed capture pool for a total of 12 individually barcoded samples. The group of 12 was then sequenced, 100 base pair paired-end, on one lane of an Illumina HiSeq 2500™ instrument by the McGill University and Genome Quebec Innovation Centre.

Bioinformatics

Sequenced reads were filtered and aligned to the loblolly pine genome (Neale et al. 2014) using bwa mem (Li and Durban 2009) and variants were called using GATK unified genotyper, with steps included for removal of PCR duplicates, realignment around indels, and base quality score recalibration (De Pristo et al. 2011; see Yeaman et al. 2016 for more details). SNPs calls were filtered to eliminate variants that did not meet all of 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%, genotyped successfully in >10% of individuals.

Identification of top candidate contigs

Maybe we need to say here something about the contigs being mostly transcriptome genes?

The goal of this analysis was to identify contigs that had signatures of selection at more SNPs than would be expected at random. First, we used several methods to study signatures of selection at SNPs, both with and without correction for population structure. Our rationale for including both was that an association without a population structure correction may have high power but also a high false positive rate, while a method with a population structure correction should have a lower false positive rate but may also lose power if the adaptive landscape correlates strongly with population structure.

I haven;t seen a justification for using the uncorrected in the text. If an allele is correlated with an environment because of spatial autocorrelation it is very likely to be correlated with a second environment that is correlated with the first. I don't think you have corrected for this in any way.

Signals of selection with correction for population structure

The method that we used that corrected for population structure was designed as a genetic-environment association, in which genotype is modeled as a function of an environmental variable. We used the program Bayenv2 (Günther and Coop 2013) for this analysis, which corrects for population structure with a variance-covariance matrix of allele frequencies (Günther and 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 the test statistics for each SNP. The test statistics output by Bayenv2 for each SNP are 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 (REF). We ran 3 replicate chains of

where did 35 come from?

the MCMC with 50,000 iterations, which we found produced stable results. Bayes factors and Spearman's ρ were averaged over these 35 replicate chains and these values were used for analysis.

Signals of selection without correction for population structure

We also calculate the non-parametric rank correlation Spearman's ρ between allele frequency and any variable.

I wouldn't call these spurious correlations -- they result from selection nearby. Just delete the word.

Binomial test for identification of top candidate contigs

Top candidate contigs were obtained from the results of Yeaman et al. (2016). Briefly, SNPs with unusually strong signatures of association were first identified as those found in the bottom 1% of the distribution of *p*-values (for uncorrected associations or for a genome-wide association) or the top 1% of the distribution of Bayes factors (bayenv2), which we refer to as "outlier SNPs" (this threshold was calculated twice, once across all environmental tests of association, and once across all phenotypic tests of association in Yeaman et al.). We then searched for contigs that had an unusually large number of outlier SNPs, which would be expected when strong selection generates spurious associations at neutral SNPs that flank a causal locus. To identify these "top candidate contigs", we calculated a binomial expectation for the number of outlier SNPs per contig, as a function the number of SNPs per contig and the overall probability of a SNP being an outlier (this deviates from 1%, as some environmental variables or traits were more strongly or weakly associated overall). The top candidate contigs were then identified as those with more outlier SNPs than expected from this binomial test, with a "top candidate cutoff" of $p < 10^{-9}$, which is a very restrictive cutoff. 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).

Identification of candidate SNPs

The goal of this analysis was to identify top candidate SNPs from the set of top candidate contigs. From this set of top candidate contigs, we then identified "candidate SNPs" as those with *P*-values lower than the Bonferroni cutoff ($P < \sim 10^{-7}$) for the uncorrected Spearman's rho association between allele frequency and any single environmental variable. These candidates are SNPs that we have high confidence in rejecting the null hypothesis of no association with the environment. We previously found that population structure correction missed some important adaptive SNPs that correlated with genetic structure (as was documented for this dataset in Yeaman et al. 2016). While we chose candidate SNPs based on a statistic that had not been corrected for population structure, we interpret the results both before and after correction for population structure (see below). Note that because candidate SNPs are limited to be discovered in the top candidate contigs, we expect false positives due to not correcting for structure to be lower compared to choosing candidate SNPs across all contigs.

Visualization of allelic effects: introducing galaxy plots

We introduce the concept of galaxy plots as a data visualization strategy that can be used to determine whether a set of candidate SNPs have a different behavior in multidimensional space relative to the rest of the genome-wide distribution. We can statistically analyze the observed patterns of candidate SNPs in the galaxy plots based on the expectation that candidate loci should show a covariance in allelic effects that is significantly different from the genome-wide covariance in allelic effects across two variables, and that the magnitude and direction of the covariance can be used to inform how the candidate SNPs are adapting to the combination of those two variables. While analysis is focused solely on environmental variables, the same principle applies to any combination of variables (such as phenotype and environment, or phenotype and phenotype).

Allelic correlations with any variables (geographic or environmental) were 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 has the desirable properties that it is positive if the allele frequency increases with the environmental variable, it is negative if the allele frequency decreases with the environmental variable, and it is zero if there is no relationship. Note that Spearman's ρ can be calculated with or without correction for population structure (as described in the "Signatures of selection…" methods) and we compare both in this study.

Our goal was to visualize how specific alleles reflect adaptation to more than one environmental variable, while understanding the relationships between the variables themselves. For instance, an allele at a SNP may be found at higher frequency in hot and wet environments than in cold and wet environments. However, temperature and precipitation may be correlated on the landscape such that warmer environments also experience more precipitation. Even for a SNP that is only adapting to one of these variables, this correlation between variables affects the associations we expect to see between a SNP and the second variable. An example is shown in Figure X for two variables that vary from uncorrected, to positively correlated, to negatively correlated. If these two variables are uncorrelated, then one would expect to see a spherical genome-wide distribution in the allelic correlations between two variables (Figure XA). When two variables themselves are correlated on the landscape, this makes interpretation more difficult because an allele that has a non-zero Spearman's ρ with one variable will also tend to have a non-zero Spearman's ρ with the second variable, even if that association occurs by chance (false positive). Therefore for two positively-correlated variables, the genome-wide covariance in Spearman's ρ is also expected to be positive (Figure XB), while the opposite is true for two negatively-correlated variables (Figure XC).

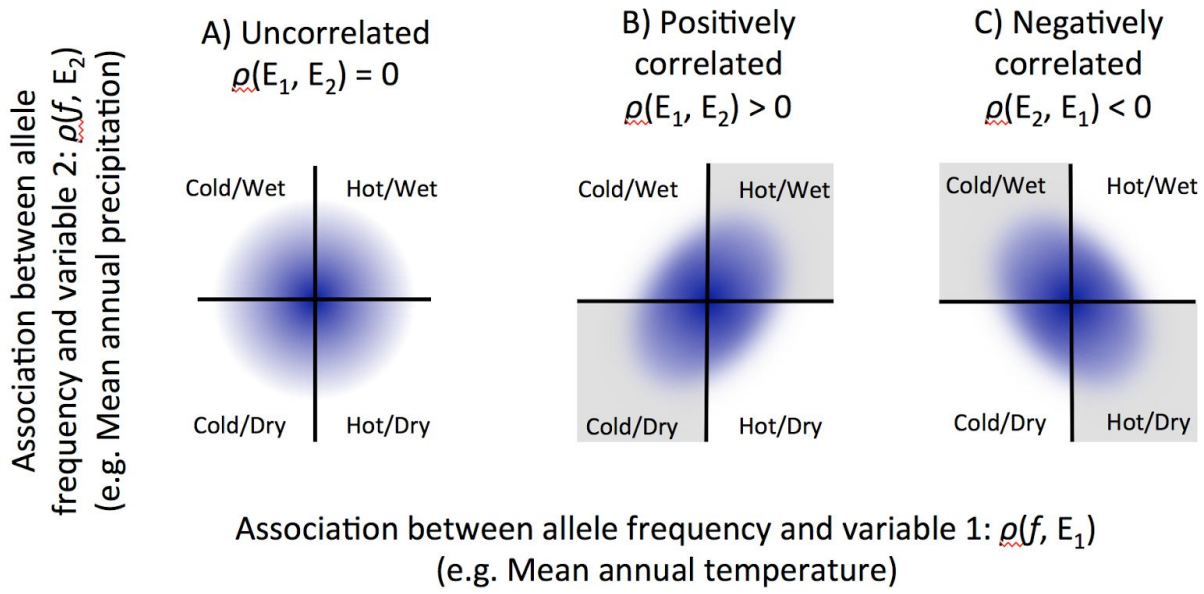


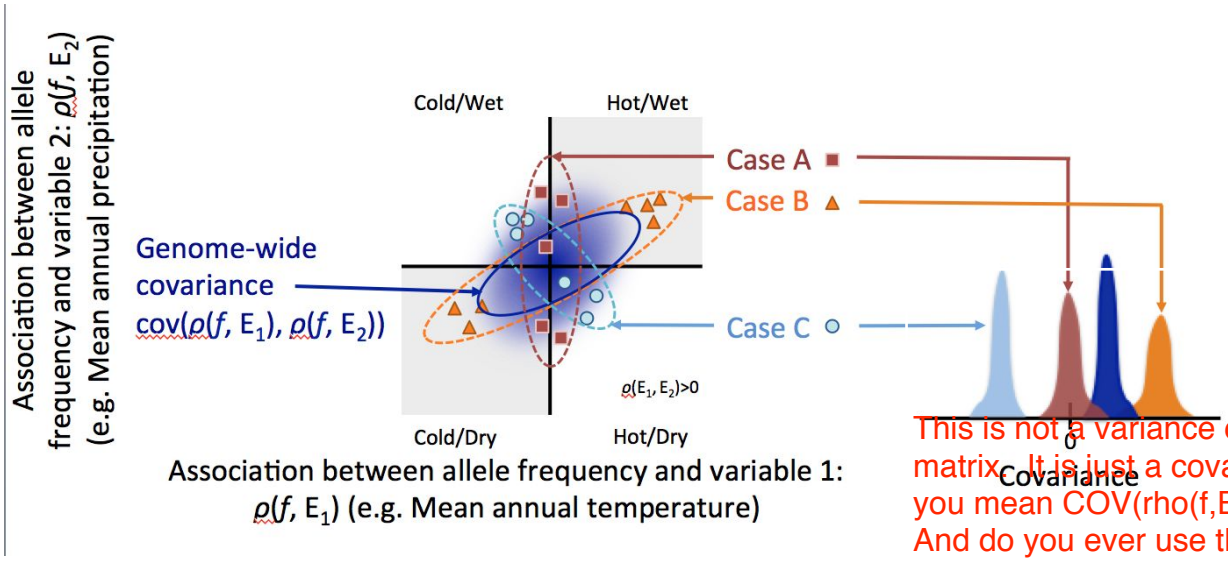
Figure X. Galaxy plots showing three examples of correlations between reference allele frequency and two variables, shown with temperature and precipitation as an example. A) Both variables are uncorrelated. B) Expected genome-wide pattern if both variables are positively correlated. C) Expected genome-wide pattern if both variables are negatively correlated. In B and C, the shading indicates the quadrants that the genome-wide pattern is expected to follow.

Note that the specific location of any particular allele in a galaxy plot depends on the way alleles are coded. SNP data is typically coded as 0, 1, or 2 copies of the reference allele (homozygous alternate allele, heterozygous, or homozygous reference allele, respectively). 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).

The goal of the galaxy plot is to visualize the effects of candidate SNPs compared to the genome-wide pattern. In this context, one could take two approaches toward understanding the genetic basis of adaptation: (i) identify outliers *a posteriori* relative to this genome-wide expectation in multidimensional space, or (ii) using a set of *a priori* candidate SNPs from candidate genes, study whether the patterns at those SNPs are different from the genome-wide expectation. In this study, we take the second approach because it should reduce the number of false positives and reveal patterns at candidate genes.

A set of candidate SNPs overlaid onto a galaxy plot can reveal the allelic effects of those SNPs in the bivariate environment. The magnitude and direction of the covariance in allelic effects of a group of candidate SNPs can be used to inform how the candidate SNPs are adapting to the combination of those two variables, when compared to the genome-wide covariance. In multivariate space, the covariance can be visualized as an ellipsoid encompassing a group of

SNPs. In Figure Galaxy-concept we illustrate this in two dimensions with three cases. Case A shows a group of alleles that have associations with precipitation but no associations with temperature (Figure Galaxy-concept). While the covariance among allelic effects in these two environments is zero, it is still different than the genome-wide pattern in this example, which makes it an interesting pattern. Case B shows a group of alleles that are adapting in the same direction as the genome-wide pattern, but the allelic effects are larger than the genome-wide pattern; this results in a larger covariance than the genome-wide pattern. Case C shows a group of alleles that are adapting in an opposing direction to the genome-wide pattern, resulting in negative covariance among allelic effects in this example.



This is not a variance covariance matrix. It is just a covariance. Do you mean COV(rho(f,Ei),rho(f,Ej))? And do you ever use the variance terms at all? If not, why mention?

Figure Galaxy-concept. Examples of patterns of allelic effects that could be observed in galaxy plots for the 3 cases described in the main text. The location of individual SNP effects are indicated by respective symbols for the three cases, while the covariance in allelic effects among those SNPs is visualized on the left plot by an ellipsoid in multivariate space. The covariance is calculated between Spearman's $\rho(f, E_1)$ and Spearman's $\rho(f, E_2)$, where *f* is the allele frequency and E_x is the environmental variable. The figure on the right shows hypothetical sampling distribution for covariance if unlinked SNPs from a group were randomly resampled many times and the covariance was recalculated.

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 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:

give reference

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

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. In reality the data will violate this assumption because of non-independence among linked SNPs, but we use ellipses only as a visualization tool. 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`.

I think this is the problem. This is not actually the null hypothesis that tell yu anything new. You already know that these loci are different (they're outliers). What you want to know is are they more correlated than you would expect given that they are outliers, and I don;t think your method does this test.

Significance of patterns at candidate SNPs in galaxy plots

The goal of the following analysis is to quantify whether the covariance in allelic effects at a candidate set of SNPs is significantly different from the genome-wide (null) distribution, while controlling for the fact that there were multiple candidate SNPs from the same gene or genomic contig. In other words, we tested the null hypothesis that the covariance in allelic effects of a focal group of SNPs was equal to the genome-wide covariance. We accomplished this by calculating 95% confidence intervals on the covariance in allelic effects for two environmental variables ($Cov(\rho(f, E_1), \rho(, E_2))$) of any particular group of SNPs using a bootstrap approach. Because in some cases a single contig contained multiple candidate SNPs, we needed to account for the potential bias caused by gametic disequilibrium between these SNPs when performing the bootstrap. We repeated the following analysis for two estimates of Spearman's *ρ*: (i) the raw correlation uncorrected for population structure; (ii) after correction for population structure in the program Bayenv2.

I had a hard time following this,and I'm not sure it is a bootstrap. Your test is about the relationship of info from one SNP compared to all the other SNPs.

I had a hard time following this paragraph. It is not a bootstrap as described.

For a variable pair, one replicate of the bootstrap to compare the observed distribution for a group of SNPs to the genome-wide (null) distribution consisted of 3 steps: (i) randomly sample one SNP from each contig in a focal group of SNPs without replacement (for contigs containing only one candidate SNP, the same SNP would be chosen for each replicate bootstrap); (ii) for each randomly selected candidate SNP, randomly sample a matching SNP from a different contig in the dataset, choosing one with an allele frequency within +- 0.025 of the frequency of the candidate SNP; (iii) calculate the covariance in allelic effects ($Cov(\rho(f, E_1), \rho(f, E_2))$) for the candidate bootstrap and the genome-wide matching bootstrap. For each variable pair, this process was repeated 100 times and used to create a sampling distribution of covariance. In assessing significance for a two-tailed test, we determined whether the 95% confidence intervals on covariance for the focal group of SNPs overlapped with the 95% confidence intervals on the random group of SNPs (the 95% confidence intervals were calculated from the 0.025 and 0.975 quantiles of the sampling distribution of covariance). When these confidence intervals did not overlap, then we rejected the null hypothesis of equal covariance in allelic effects for a focal group of SNPs and the genome-wide pattern.

The main problem with this is that is doesn't address the root problem that the high correlation between f/env correlations is likely higher on average by the null model with outliers.

Annotation

We used the annotations developed for pine in Yeaman et al. 2016. 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 *et al.* 2005; Conesa & Gotz 2008) to assign GO terms and enzyme codes (see Yeaman et al. 2016? 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 *et al.* 2006). All GO terms associated with at least five top candidate genes were analyzed for significant over-representation (FDR 5%).

Other data visualization?

(I also think a redundancy analysis could be informative here, especially if applied separately to the Temperature and Precipitation groups below).

Results

Candidate SNPs

First, we identified contigs that contained more outlier SNPs than expected by chance based on a binomial test given the length of the contig and a p-value of 10^{-9}. From this set of 12731 SNPs in 321 contigs, we then identified candidate SNPs as those with significant Bonferroni-corrected *P*-values, which resulted in 1714 SNPs in 167 contigs. These SNPs are strong local adaptation candidates and that the next step is determine how allele frequencies of these candidate SNPs behave in relation to multiple environmental variables.

Visualization with galaxy plots

Preliminary analyses revealed two groups of candidate SNPs that showed different behavior in galaxy plots. The most striking pattern exhibited by these two groups was in the MAT-MAP comparison (mean annual temperature vs. mean annual precipitation) (Supplementary Figure 1), which showed clusters of SNPs in the hot/wet, cold/wet, cold/dry, and hot/dry quadrants of the plot, both with and without population-structure correction. Because the entire candidate set did not covary across all variables, this limited our ability to test for significance of covariance for the entire set.

Data visualization revealed that certain environmental variables had outliers that fell into the 1st and 3rd quadrants of MAT-MAP (hot/wet and cold/dry environments, respectively), while other variables had outliers that fell into the 2nd and 4th quadrants of MAT-MAP (cold/wet and hot/dry environments, respectively) (Supplementary Figure 2). Therefore, we identified the environmental variables that could explain the different patterns of the SNPs in the galaxy plot. For each environmental variable, we counted the number of outlier SNPs in the 1st and 3rd quadrants vs. the 2nd and 4th quadrants of the MAT-MAP comparison. Next, we tested the null

hypothesis that 50% of outlier SNPs from any particular variable fall into the 1st/3rd quadrants and 50% fall into the 2nd/4th quadrants. We tested this null hypothesis with a binomial test, with the number of trials *n* equal to the the number of outlier SNPs in that variable, the number of successes equal to the number of SNPs that fall into the 1st/3rd quadrants (hot/wet and cold/dry), and the probability of success $p = 0.5$. Because we conducted 44 tests (22 for uncorrected and 22 for structure-corrected data), we used Bonferroni corrected *P*-values to test for significance. While we recognize that the assumptions of the binomial test are violated due to linkage among SNPs in the same contig, we use this only as a tool to assign SNPs to groups.

Using the binomial test we identified the environmental variables that explained the two groups of SNPs. The first group of SNPs were outliers in LONG, Elevation, MAT, MWMT, MCMT, TD, DD0, DD5, NFFD, eFFP, FFP, and EMT and showed signatures of selection to hot/wet and cold/dry environments greater than the binomial expectation (Supplementary Table S1). Outlier SNPs in these variables were assigned to the “Freezing” Group because these environmental variables largely measure the extent of cold experienced by populations (plotted in blue in the subsequent graphs). The second group of SNPs were outliers in MAP, MSP, AHM, SHM, EXT, Eref, and CMD, and showed signatures of selection to cold/wet and hot/dry environments greater than the binomial expectation (Supplementary Table S1). Outlier SNPs in these variables were assigned to the “Aridity” Group because these environmental variables largey measure the amount of heat relatively to the amount of precipitation experienced by populations (plotted in orange in subsequent graphs). It should be noted that Latitude had a large number of outlier SNPs (452), but interestingly these SNPs showed both groups of behavior and so the binomial test was not significant (Supplementary Table S1, Supp Figure S3).

A summary of the number of candidate SNPs in each variable is shown in Figure Awesome-Barplot. The height of the bars represents the number of outlier SNPs for each contig within each variable. For each contig we calculated the difference between total number of outlier SNPs across the Freezing variables and the number of outlier SNPs across the Aridity variables and used this as a score to represent color for that contig (Figure Awesome-Barplot). For the correlated variables within each group, there was a lot of overlap among outliers and contigs.

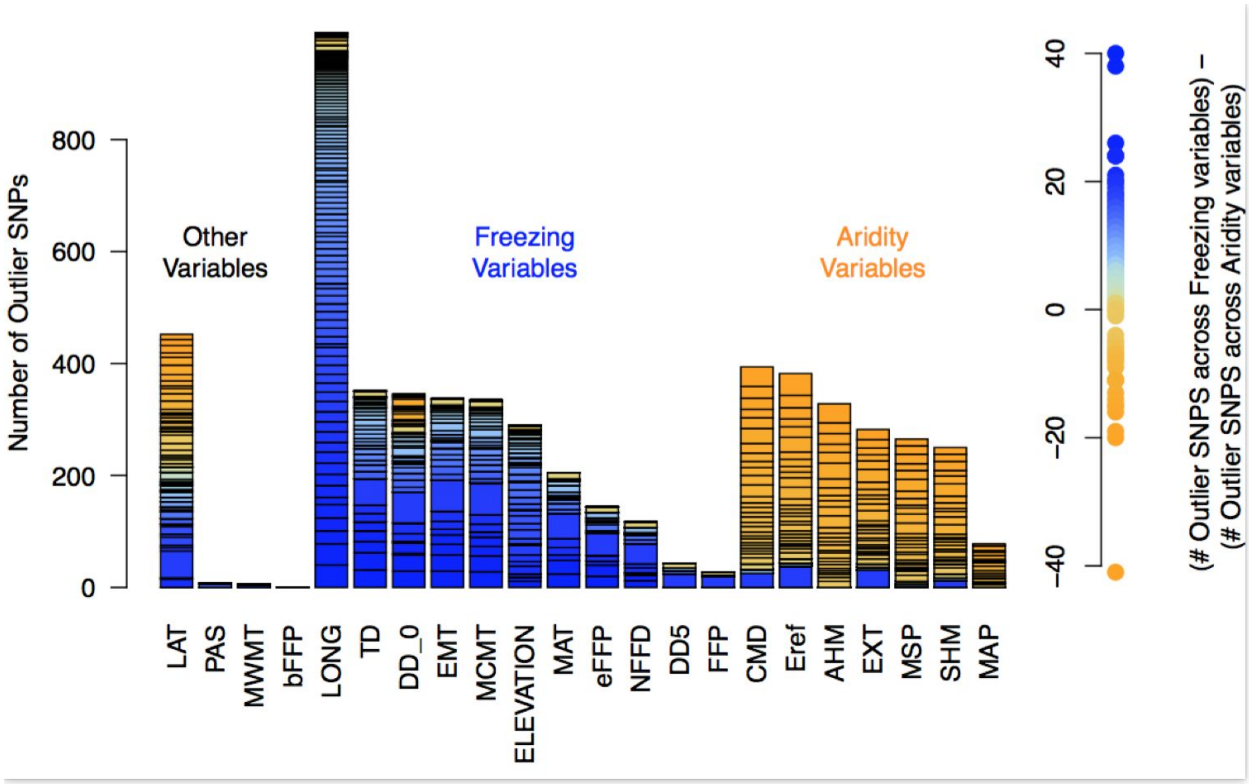


Figure Awesome-Barplot. The height of the bars represents the number of outlier SNPs from each contig in an environmental variable. Contigs were colored according to a score that depended on the number of outlier SNPs across Freezing variables minus the number of outliers across Aridity variables.

A total of 1316 candidate SNPs in 153 transcriptome contigs were identified for the Freezing group, and a total of 425 candidate SNPs in 31 transcriptome contigs were identified for the Aridity group. There was some overlap of the Aridity group with the Freezing Group: 113 SNPs in 23 contigs belonged to both Groups (Figure Pine Venn). In the Freezing group, 11 contigs were previously identified as having an ortholog in spruce that was undergoing convergent adaptation to climate (Yeaman 2016 top 47). None of the contigs in the Aridity group were previously identified as having an ortholog in spruce that was undergoing convergent adaptation to climate.

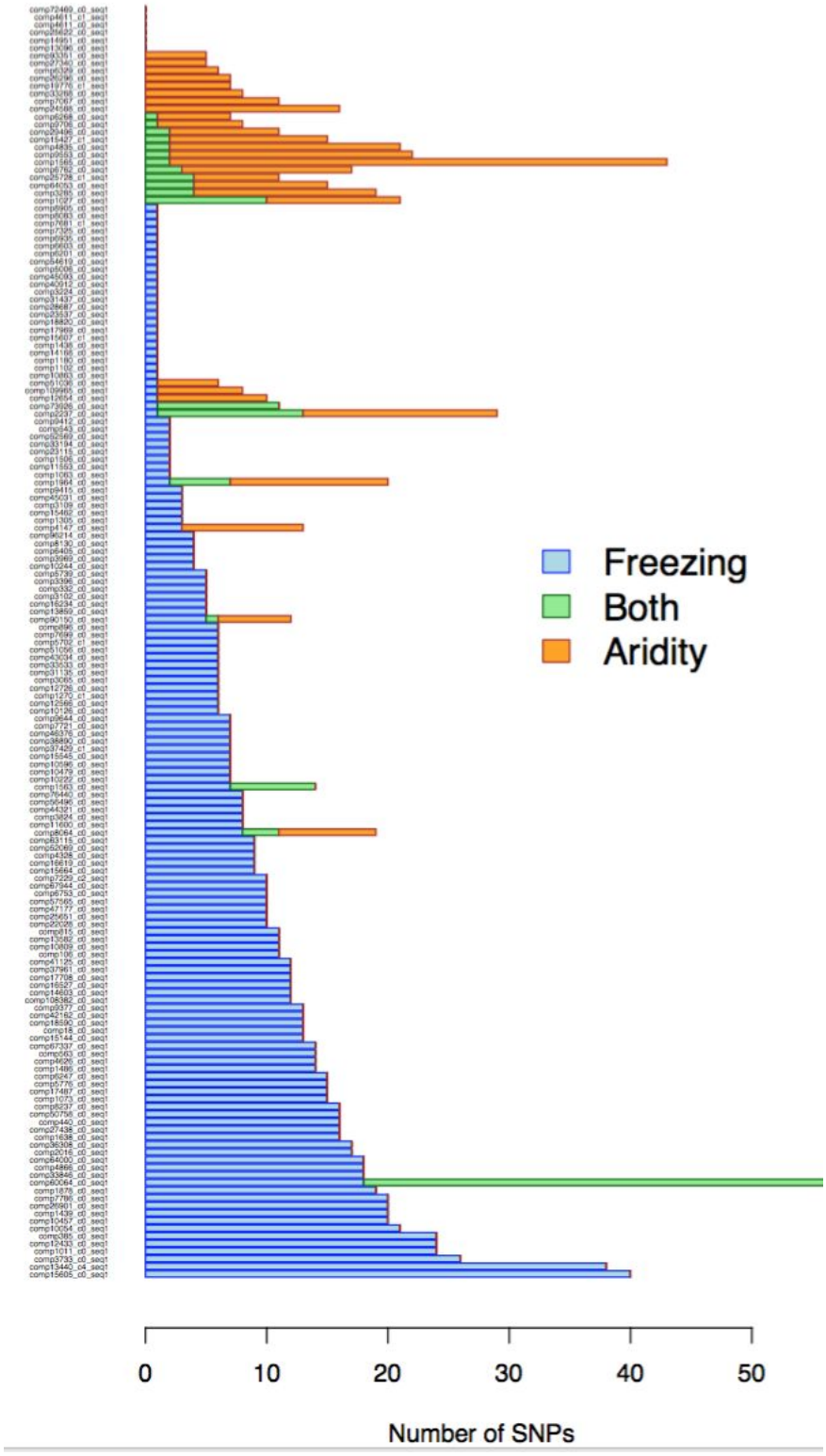
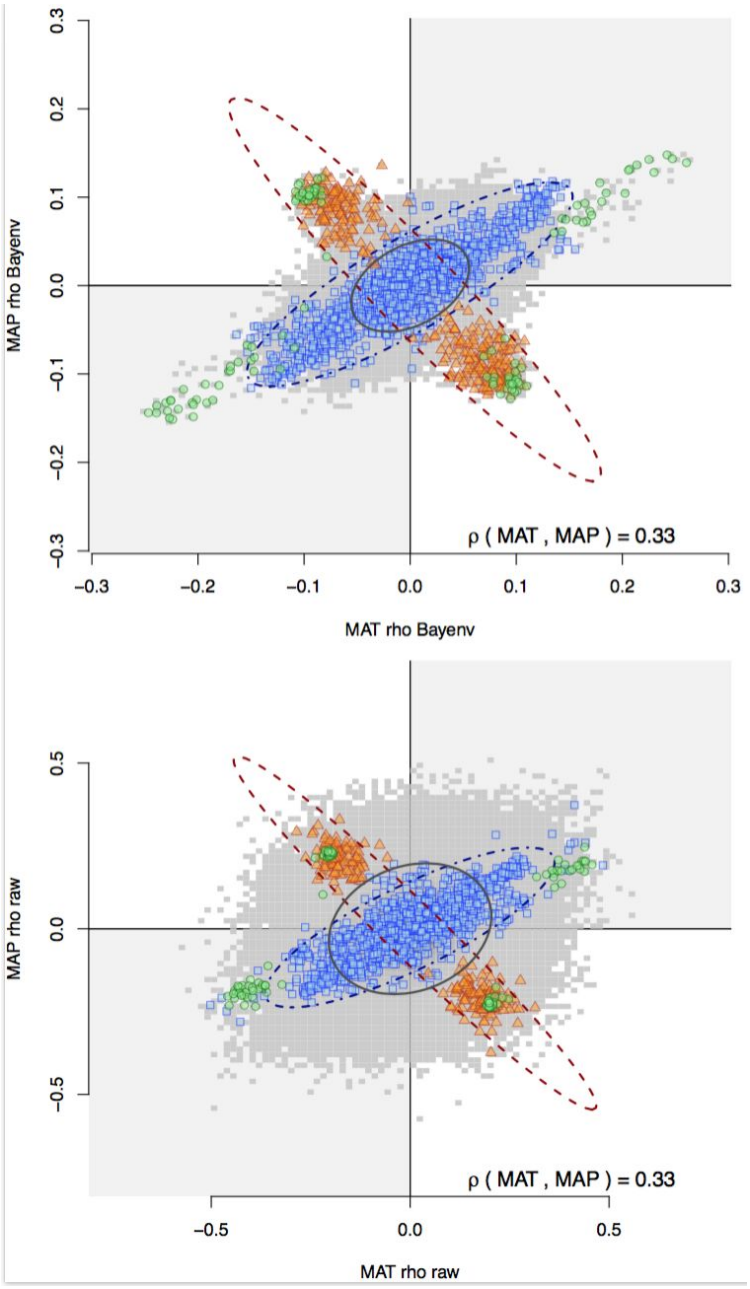


Figure PINE VENN: Frequency distribution of SNPs with contigs for the Freezing and Aridity groups. Each row represents a distinct genomic contig. SNPs were assigned to the Freezing or Aridity group based on whether they were an outlier in an Freezing or Aridity environmental variable.

In candidate SNPs of the Freezing group, one allele was found in higher frequency in hot/moist environments (Quadrant 1 of Figure MAIN RESULT), while the alternate allele was found in higher frequency in cold/dry environments (Quadrant 3 of Figure MAIN RESULT). In loci of the Aridity group, one allele was found in higher frequency in cold/moist environments (Quadrant 2 of Figure MAIN RESULT), while the alternate allele was found in higher frequency in warm/dry environments (Quadrant 4 of Figure MAIN RESULT). Whether uncorrected or structure-corrected allele frequencies are used in the analysis gave the same result, but note that some of the SNPs become more “outlierly” while others have a reduced effect size after correction for population structure (Figure MAIN RESULT).



NOTE: Figure MAIN RESULT is in the process of being updated with the Galaxy 2.0 format:

Galaxy 2.0: Top is structure-corrected associations and bottom is uncorrected associations. Blue squares are the Freezing Group, orange triangles are the Aridity Group, and green circles are outliers in environmental variables in both groups. The solid grey line represents the 95% confidence ellipse on the genome-wide data, while the 95% confidence ellipses on SNPs that belong only to one group or the other are shown with dashed lines.

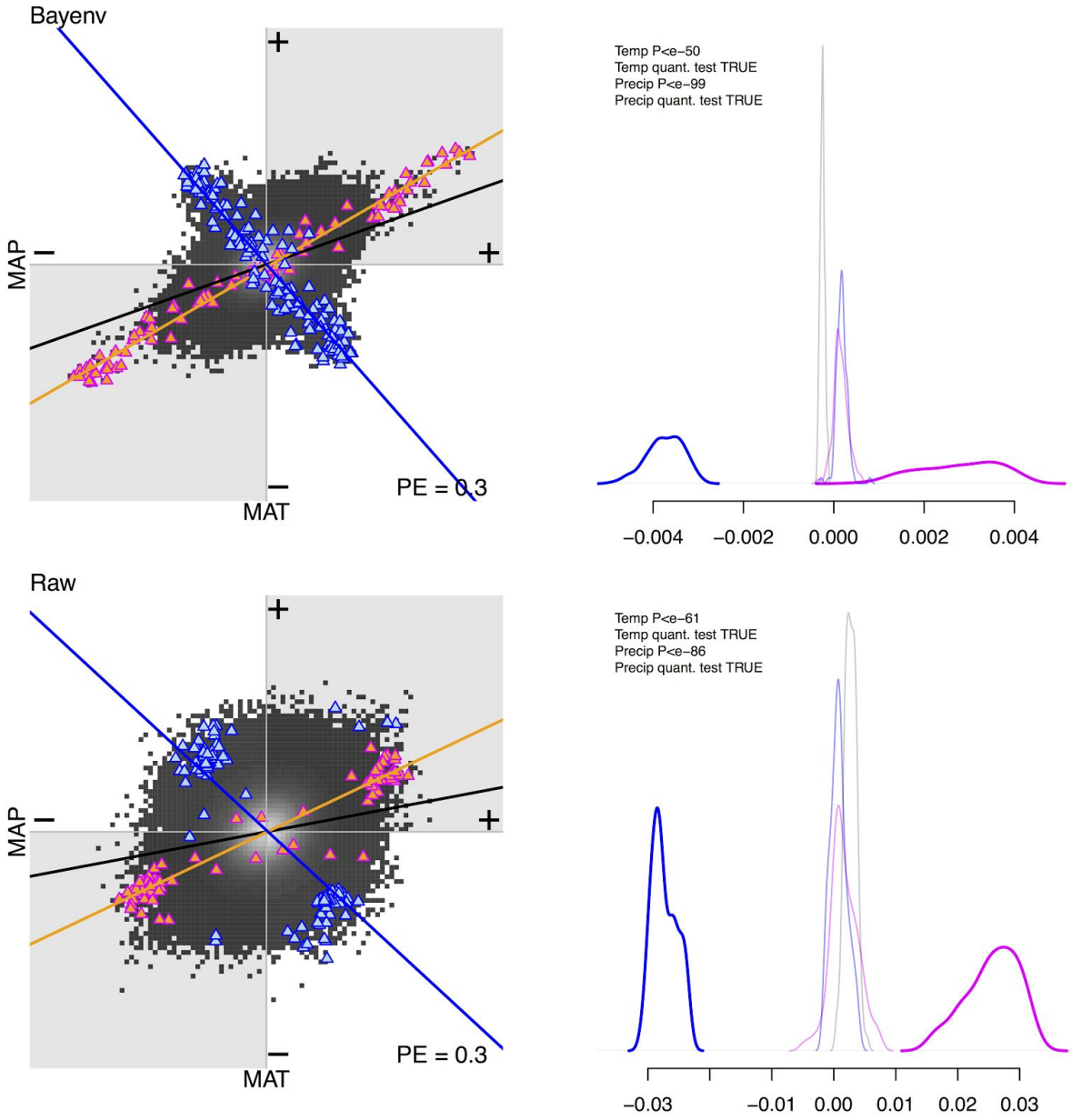


Figure MAIN RESULT. The left column shows the Galaxy plot (the Sperman's ρ association between allele frequency and the x or y variable) for the genome-wide pattern (greyscale), the temperature group (magenta) and the precipitation group (blue). The top row shows allelic patterns based on Bayenv2-corrected allele frequencies and the bottom row is the results based on uncorrected (raw) associations. The right column shows the distribution of covariance in effect sizes from bootstrapping the result of the significance testing for 5 groups: (i) the temperature group (solid magenta line), (ii) genome-wide SNPs matched to the temperature group (light magenta line), (iii) the precipitation group (solid blue line), (iv) genome-wide SNPs matched to the precipitation group, and (v) other candidate SNPs not in the temperature or preciptation groups. Significance was assessed by determining if the 95% confidence intervals do not overlap (quant. test TRUE if if they do not overlap).

Multivariate environments: STOPPED EDITING HERE

We next expanded the galaxy plots to interpret allelic effects across multiple environmental variables, which we visualized as a function of latitude. Broad-scale geographic patterns reflect species range, regional topography and associated macroclimatic variation: the environments inhabited by the populations collected at lower latitudes tended to be warmer and wetter, while the environment for populations collected at higher latitudes tended to be colder and drier. Despite this overall large-scale macroclimatic variation, the Temperature and Precipitation Groups appear to be adapting to mesoclimatic variation across temperature and precipitation gradients. For environments, we focused on five temperature-related variables: degree days below 0C(DD0); mean annual temperature C(MAT), Mean coldest month temperature C (MCMT); Extreme minimum temperature over 30 years (C)(EMT); and Frost-free period (days)(FFP), and two precipitation-related variables: mean annual precipitation (mm)(MAP) and Hargreaves climatic moisture deficit (mm)(CMD).

Taken together, at southern latitudes the Temperature Group is adapting to wetter and warmer conditions at low latitudes, and at northern latitudes adapting to drier and colder conditions (Figure GALAXY ENVIRONMENT). On the other hand, at southern latitudes the Precipitation Group is adapting to drier and warmer conditions at high elevations, and at northern latitudes adapting to wetter and colder conditions at lower elevations (Figure GALAXY ENVIRONMENT). These patterns are summarized in Table SUMMARY.

Next paragraph to do: compare Bayenv corrected and uncorrected patterns. Basically they are always the same, although the significance changes sometimes.

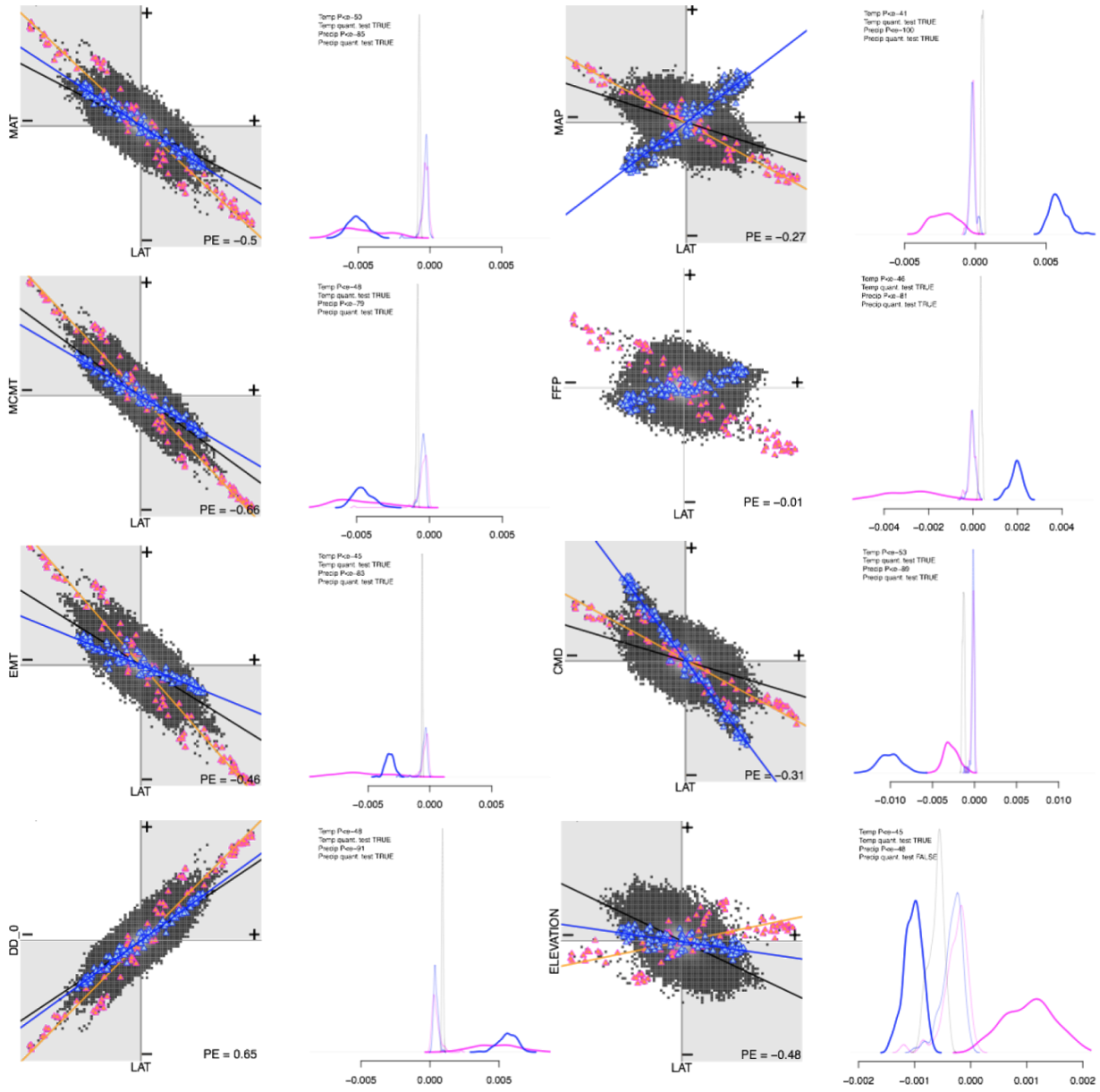


Figure GALAXY ENVIRONMENT. Patterns of allelic correlations (Bayenv2-corrected) with latitude versus allelic correlations with environmental variables. (I know the resolution is crappy, but hopefully you get the picture). See table X for a summary.

Table summary of above 2 figures:

	Temp Group- Allele 1 (left)	Precipitation Group- Allele 1 (left)	Temp Group - Allele 2 (right)	Precipitation Group- Allele 2 (right)
Latitude (LAT)	**lower**	**lower**	**higher**	**higher**
Longitude (LONG)	**TO DO**			
Degree-days < 0C (DD0)	**fewer**	**fewer**	**more**	**more**
Mean annual precip (MAP)	**wetter**	**drier**	**drier**	**wetter**
Mean annual temp (MAT)	**warmer**	**warmer**	**colder**	**colder**
Mean coldest month temp (MCMT)	**warmer**	**warmer**	**colder**	**colder**
Extreme minimum temp (EMT)	**warmer**	**warmer**	**colder**	**colder**
Frost free period (FFP)	**longer**	**shorter**	**shorter**	**longer**
Climate moisture deficit (CMD)	**higher**	**higher**	**lower**	**lower**

Table SUMMARY. Each row represents a trait or environment, and table entry represents the expected value of that trait or environment for an individual that possesses the allele designated in columns.

TO DO: Add Figure of correlations among environments

Annotation

No GO terms were over-represented in the temperature or precipitation groups (FDR 5%) nor over all 107 top candidate genes, a result previously found in this dataset for a larger group of candidate genes than presented here (Yeaman et al 2016). This suggests that the genes underlying adaptation to these environmental gradients had diverse functions. Several ...

Other landscape visualization (mapping?)

It would help to visualize the allele frequencies in hot/dry cold/wet etc environments on the landsape, but not sure how to do this. One question in my mind is what is the frequencies of the alleles in the Temperature group doing in the environmental combinations that they don't show adaptation to - are they just at intermediate frequency?

Discussion

This picture about sums it up:

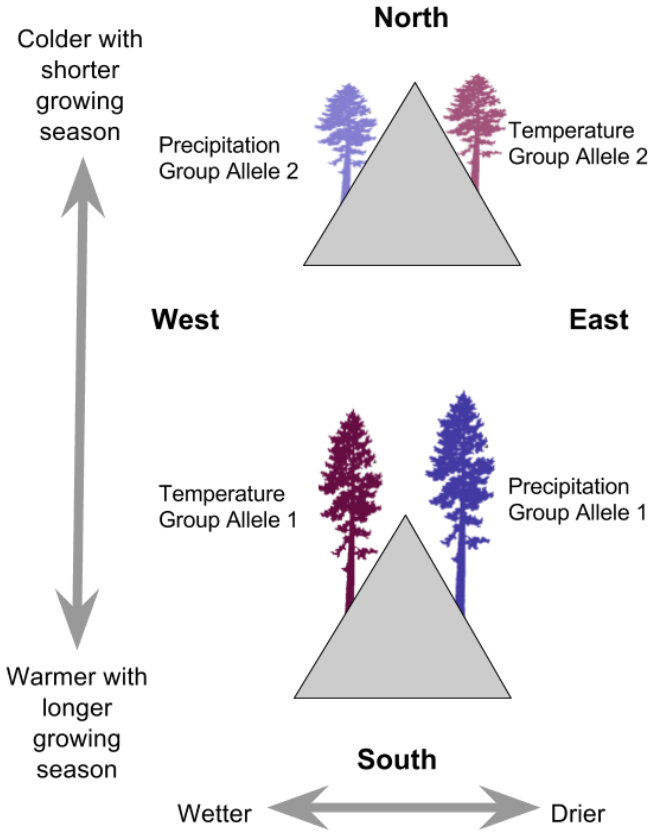


Figure UNDERSTAND: Conceptual understanding of how the Temperature and Precipitation Groups are adapting to the landscape. The top of the graph represents North and the left side of the graph is West. The triangles represent mountain ranges. One allele of the Precipitation Group is adapting to warmer, drier conditions at higher elevations at southern latitudes: conditions that will be more likely on the east side of mountains (dark blue tree in lower right corner). The other allele in the Precipitation Group is adapting to colder, wetter conditions at lower elevations at northern latitudes: conditions that will be more likely on the west side of mountains (light blue tree in upper left corner). One allele of the Temperature Group is adapting to warmer, wetter conditions at lower elevations at southern latitudes: conditions that will be more likely on the west side of mountains (dark red tree in lower left corner). The other allele of the Temperature Group is adapting to colder, drier conditions at higher elevations at northern latitudes: conditions that will be more likely on the east side of mountains (light red tree in upper right corner). From common garden experiments in this species, we know that trees tend to grow at overall higher elevations at southern latitudes and lower elevations at northern latitudes. To be clear, every tree will have alleles from both groups, but the picture just represents the climatic conditions that each group shows allelic patterns of local adaptation to.

Some other points for discussion:

- Data visualization was necessary to reveal these patterns
- If we had treated candidate SNPs as one group, they would have “canceled” each other out in certain kinds of analyses and the pattern would not have been revealed.

- Importance of a priori determining a candidate set for visualization, but recognition that our set is probably missing some important candidates.
- It's important (and simple) to interpret the magnitude and direction of the genetic effect in landscape genomic patterns, not just the significance of any particular outlier
- What is the putative function of the candidates
- Temp Group
    - This seems very extreme: warm-wet-low vs. cold-dry-high. It's like local adaptation to super-harsh and super-unharsh conditions
- Precip Group
    - In contrast, these could be seen as "less harsh" local adaptations, and map more directly to water stress (hot dry would be stressed and cold-wet would be unstressed).

Data Accessibility

Author Contributions

Acknowledgements

References

Blanquart F, Kaltz O, Nuismer SL, Gandon S (2013) A practical guide to measuring local adaptation. *Ecology letters*, **16**, 1195–205.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23: 2633-2635

Chuine, I., G.E. Rehfeldt and S.N. Aitken. 2001. Height growth determinants in pines: a case study of *Pinus contorta* and *Pinus monticola*. Can. J. For. Res. 36: 1059-1066.

Dabney A, Storey JD qvalue: Q-value estimation for false discovery rate control. *In*, Ed R package version 1.40.0.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

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

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

Hannerz, M., S.N. Aitken, J.N. King and S. Budge. 1999. Effects of genetic selection for growth on frost hardiness in western hemlock. Can. J. For. Res. 29(4): 509-516.

Kawecki, T. J. & Ebert, D. Conceptual issues in local adaptation. *Ecol Lett.* **7,** 1225–1241 (2004).

Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology and Evolution*, **17**, 183–189.

Lotteps KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575

Savolainen, O., Lascoux, M. & Merila, J. Ecological genomics of local adaptation. *Nat Rev Genet* **14,** 807–820 (2013).

Stinchcombe, J. R. & Hoekstra, H. E. Combining population genomics and quantitative genetics: finding the genes underlying ecologically important traits. *Heredity (Edinb).* **100,** 158–70 (2008).

Wang, T., Hamann, A., Spittlehouse, D., and Murdock, T. N. 2012. ClimateWNA – High-resolution spatial climate data for western North America. Journal of Applied Meteorology and Climatology 61: 16-29.

Williams, E.R. and M. Talbot. 1993. *ALPHA+: Experimental design for variety trials. Design User Manual*: CSIRO, Canberra and SASS, Edinburgh.

Yang J, Benyamin B, McEvoy BP *et al.* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, **42**, 565–569.

Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics*, **88**, 76–82.