Robust identification of local adaptation from allele frequencies

Torsten Günther[1] and Graham Coop[2]
[1] Institute of Plant Breeding, Seed Science and Population Genetics,
University of Hohenheim, Stuttgart, Germany.
[2] Department of Evolution and Ecology & Center for Population Biology,
University of California, Davis, USA.
To whom correspondence should be addressed: torsten.guenther@uni-hohenheim.de, gmcoop@ucdavis.edu

Abstract

Comparing allele frequencies among populations that differ in environment has long been a tool for detecting loci involved in local adaptation. However, such analyses are complicated by an imperfect knowledge of population allele frequencies and neutral correlations of allele frequencies among populations due to shared population history and gene flow. Here we develop a set of methods to robustly test for unusual allele frequency patterns, and correlations between environmental variables and allele frequencies while accounting for these complications based on a Bayesian model previously implemented in the software Bayenv. Using this model, we calculate a set of 'standardized allele frequencies' that allows investigators to apply tests of their choice to multiple populations, while accounting for sampling and covariance due to population history. We illustrate this first by showing that these standardized frequencies can be used to calculate powerful tests to detect non-parametric correlations with environmental variables, which are also less prone to spurious results due to outlier populations. We then demonstrate how these standardized allele frequencies can be used to construct a test to detect SNPs that deviate strongly from neutral population structure. This test is conceptually related to F_{ST} but should be more powerful as we account for population history. We also extend the model to next-generation sequencing of population pools, which is a cost-efficient way to estimate population allele frequencies, but it implies an additional level of sampling noise. The utility of these methods is demonstrated in simulations and by re-analyzing human SNP data from the HGDP populations. An implementation of our method is available from http://gcbias.org

1 Introduction

The phenotypes of individuals within a species often vary clinally along environmental gradients (Huxley, 1939). Such phenotypic clines have long been central to adaptive arguments in evolutionary biology, with many diverse examples including skin pigmentation in humans (Jablonski, 2004), body size and temperature tolerance in *Drosophila* (Hoffmann and Weeks, 2007), and flowering time in plants (Stinchcombe *et al.*, 2004), which all vary clinally with latitude. Unsurprisingly, comparisons of allele frequencies between populations that differ in environment were among the earliest population genetic tests for selection (Cavalli-Sforza, 1966; Lewontin and Krakauer, 1973), and have continued to be central to population genetics to this day (e.g. Coop *et al.*, 2009; Akey *et al.*, 2010).

The falling cost of sequencing and genotyping means that such comparisons can now be made on a genome-wide scale, allowing us to start understanding the genetic basis of local adaptation

across a broad range of organisms. However, such studies need to acknowledge the sampling issues inherent in population genetic studies of natural populations. In assessing correlations between allele frequencies and environmental variables or looking for loci with unusually high levels of differentiation, two broad technical issues need to be addressed. First, sample allele frequencies are noisy estimates of the population allele frequency, and this issue is exacerbated when sample sizes differ across populations. Second, when multiple populations are compared they are not statistically independent. These populations have experienced varying amounts of shared genetic drift and migration over time and they will consequently vary in their relationship to each other (Robertson, 1975; Nicholson *et al.*, 2002; Excoffier *et al.*, 2009; Bonhomme *et al.*, 2010). Failure to account for differences in sample size and the shared history of populations could lead to a high rate of false-positive and negatives due to the unaccounted sources of variance and non-independence among populations. Therefore, accounting for these potential biases should provide additional precision in the identification of loci responsible for adaptation. To accommodate these sources of noise the Bayesian method Bayenv was developed (Coop *et al.*, 2010) that attempts to account for these two factors while testing for a correlation between allele frequencies and an environmental variable. To control for a general relationship between populations, a covariance matrix of allele frequencies is estimated from a set of control markers. This model of covariance is then used as a null model against an alternative model which allows for a linear relationship between the (transformed) allele frequencies at a particular locus and an environmental variable of interest. Inference under these models is performed using Markov chain Monte Carlo (MCMC) to integrate over the posterior of the parameters.

Recently, methods closely related to Bayenv have been developed and applied to detect environmental correlations while accounting for population structure. The most similar approach is by Guillot (2012) who offered large gains in computational efficiency for a model very similar to Bayenv, but where the covariance matrix has an explicit isolation by distance parametric form, by making use of approximations to perform inference in an MCMC-free framework (Rue *et al.*, 2009; Lindgren *et al.*, 2011). Frichot *et al.* (2012) presented a Latent Factor Mixed Model that estimates the effect of population history and environmental correlations simultaneously. The Frichot *et al.* (2012) method resulted in a slightly higher power than Bayenv to detect environmental correlations in simulations, perhaps in part as a result of the simultaneous inference of fixed and random effects reducing the effect of selected loci inflating the covariance matrix. Finally, Fumagalli *et al.* (2011) and Hancock *et al.* (2011a) used a non-parametric partial mantel test, which makes fewer model assumptions and so should be less sensitive to non-normality. However, the partial mantel test is not well calibrated when both genotypes and environmental variables are spatially autocorrelated (see Guillot and Rousset, 2011, for discussion), and so the p-values should be interpreted with caution.

Bayenv has been successfully applied to identify loci putatively involved in local adaptation to environmental variables across a range of different species (e.g. Hancock *et al.*, 2008, 2010, 2011c,b; Eckert *et al.*, 2010; Fumagalli *et al.*, 2011; Jones *et al.*, 2011; Cheng *et al.*, 2012; Fang *et al.*, 2012; Keller *et al.*, 2012; Limborg *et al.*, 2012; Pyhäjärvi *et al.*, 2012). However, further work is needed to make Bayenv robust to outliers and and to extend it to next-generation data applications. One concern about applications of such methods is that linear models are not robust to outliers, which can lead to spurious correlations. For example, if a single population has both an extreme allele frequency and an extreme environmental variable, while all other populations show no correlation, then the linear model may be misled (see Hancock *et al.*, 2011b; Pyhäjärvi *et al.*,

2012, for examples). This sensitivity can be overcome by using rank-based non-parametric statistics, such as Spearman's ρ, which may also offer increased power to detect non-linear relationships. The difficulty is that such tests do not acknowledge the differences in sample size or the covariance in allele frequencies across populations. To overcome these difficulties we provide the user with a set of 'standardized' allele frequencies at each SNP, where the effect of unequal sampling variance and covariance among populations has been approximately removed. This affords users a general framework to utilize statistics of their choosing to investigate environmental correlations or other sources of allele frequency variation. As an example of how these 'standardized' allele frequencies' can be used we construct a global F_{ST}-like statistic that accounts for shared population history and sampling noise.

We also extend Bayenv to deal with some of the statistical challenges posed by next-generation sequencing. Recently, pooled next-generation sequencing (NGS) of multiple individuals from a population has gained in popularity (e.g. Turner *et al.*, 2010, 2011; He *et al.*, 2011; Kolaczkowski *et al.*, 2011; Boitard *et al.*, 2012; Fabian *et al.*, 2012; Kofler *et al.*, 2012; Orozco-Terwengel *et al.*, 2012), as it offers a cost efficient alternative to sequencing of single individuals. However, estimating allele frequencies from read counts sequenced from a pool implies a second level of sampling variance (Futschik and Schlötterer, 2010; Zhu *et al.*, 2012), which needs to be considered in population genetic analyses such as Bayenv. We include the sampling of reads in pooled NGS experiments into the model to account for the additional sampling noise incurred.

These extensions to Bayenv are implemented in Bayenv2.0 available from `http://gcbias.org`. We demonstrate the utility of these approaches through simulation and re-analyzing SNP genotyping data from the CEPH Human Genome Diversity Panel (HGDP, Conrad *et al.*, 2006; Li *et al.*, 2008).

2 Methods

2.1 General model of Bayenv

First, we briefly explain the underlying model of Bayenv for the sake of completeness. Further details about the model and inference method can be found in Coop *et al.* (2010). Consider a biallelic locus l with a population allele frequency p_{jl} in population j where n_j alleles have been sampled from this population in total. We assume that the observed count of allele 1, k_{jl}, in this population is the result of binomial sampling from this population frequency:

$$P(k_{jl}|p_j, n_j) = \binom{n_j}{k_{jl}} {p_{jl}}^{k_{jl}} (1 - p_{jl})^{n_j - k_{jl}}. \tag{1}$$

We follow the model of Nicholson *et al.* (2002) by assuming that a simple transform of the population allele frequency p_{jl} in subpopulation j at locus l represents a normally distributed deviate around an "ancestral" frequency ϵ_l. Specifically we assume that

$$p_{jl} = g(\theta_{jl}) = \begin{cases} 0 & \text{if } \theta_{jl} < 0 \\ \theta_{jl} & 0 \leq \theta_{jl} \leq 1 \\ 1 & \theta_{jl} > 1. \end{cases} \tag{2}$$

i.e. that the mass < 1 and > 1 are placed as point masses at 0 and 1, representing the loss or fixation of the allele in population j respectively. We then assume that that the marginal distribution of

θ_{jl} is normally distributed, around an 'ancestral' mean frequency ϵ_l with variance proportional to $\epsilon_l(1-\epsilon_l)$ (inspired by the model of Nicholson *et al.*, 2002). We denote the vector of transformed population allele frequencies at a locus by θ_l where $\theta_l = (\theta_{1l}, \ \ldots, \ \theta_{Jl})$ when J is the number of populations. As we do not expect that the populations are independent from each other, we assume that θ_l follows a multivariate normal distribution

$$P(\theta_l|\Omega, \epsilon_l) \sim MVN(\epsilon_l, \epsilon_l(1-\epsilon_l)\Omega). \tag{3}$$

We can write the joint probability of our counts at a locus and the θ_l as

$$P((k_{1l}, \ldots, k_{Jl}), \theta_l|\Omega, \epsilon_l, (n_{1l}, \ldots, n_{Jl})) \sim MVN(\epsilon_l, \epsilon_l(1-\epsilon_l)\Omega) \prod_{j=1}^{J} P(k_{jl}|p_{jl} = g(\theta_{jl}), n_{jl}). \tag{4}$$

We place priors on Ω (inverse Wishart) and the ϵ_l at each SNP (symmetric Beta). Assuming that our SNPs are independent, we write the joint probability of all of our loci and parameters as

$$P(\Omega) \prod_{l=1}^{L} P((k_{1l}, \ldots, k_{Jl}), \theta_l|\Omega, \epsilon_l, (n_{1l}, \ldots, n_{Jl}))P(\epsilon_l). \tag{5}$$

Our posterior is this joint probability normalized by the integral over Ω and the ϵ_l and θ_l at all of the loci.

We then use MCMC to sample posterior draws of the covariance matrix (Ω) from a set of unlinked, putatively neutral control SNPs. Our observations showed that the MCMC converges quickly to a small set of covariance matrices for each data set given a sufficient number of independent SNPs (Coop *et al.*, 2010). Given this tight distribution, we use a single draw of Ω, denoted by $\widehat{\Omega}$, after a sufficient burn in. The entries of the matrix Ω are closely related to the matrix of pairwise F_{ST} (Weir and Hill, 2002; Samanta *et al.*, 2009), and so this model provides a flexible model of population history; for example Pickrell and Pritchard (2012) used a similar model to infer a tree-like graph of population history and Guillot (2012) uses a related model as a model of isolation by distance.

Next, we formulate an alternative model where an environmental variable Y, standardized to have mean 0 and variance 1, has a linear effect β on the allele frequencies:

$$P(\theta_l|\widehat{\Omega}, \epsilon_l, \beta, Y) \sim MVN(\epsilon_l + \beta Y, \epsilon_l(1-\epsilon_l)\widehat{\Omega}). \tag{6}$$

To express the support for the alternative model at a locus l, Coop *et al.* (2010) calculated a Bayes factor (BF) by taking the ratio of probability of the alternative and the null model given the data and $\widehat{\Omega}$, integrating out the uncertainty in θ_l, ϵ_l, and β (under a uniform prior on β between -0.2 and 0.2).

2.2 Tests based on standardized allele frequencies

The linear relationship between the transformed allele frequencies (eqn. (6)) may not be the best fit in all situations, as other monotonic relationships (e.g. exponential, logarithmic, saturating) could be viewed as biologically realistic in some cases. Additionally, there may be situations in which a linear model is not robust to outliers and so will spuriously identify loci as strong correlations. Therefore, we provide a general framework to allow investigators to apply statistics of their choice,

such as rank-based non-parametric statistics, to detect environmental correlations, while taking advantage of the Bayenv framework. These statistics could in theory be applied to the raw sample frequencies; in practice, however, that can lead to high false positive and false negative rates as sample allele frequencies are naturally noisy because of the process of sampling and non-independent due to the covariance among populations. The multivariate normal framework employed by Bayenv offers a natural way to attempt to standardize θ_l to be variates with mean zero, variance one, and no covariance. These allele frequencies allow standard statistics that rely on these assumptions to be applied more directly. We denote the Cholesky decomposition of the covariance matrix C ($\Omega = CC^T$, where C is an upper-triangular matrix), which can be thought of as being equivalent to the square root of the matrix, and so analogous to the standard deviation of θ_l. To standardize the θ_l for effects of unequal sampling variance, and covariance among populations we write

$$X_l = C^{-1}\frac{(\theta_l - \epsilon_l)}{\sqrt{\epsilon_l(1-\epsilon_l)}}. \tag{7}$$

If $\theta_l \sim \text{MVN}(\epsilon_l, \epsilon_l(1-\epsilon_l)\Omega)$ then $X_l \sim \text{MVN}(0, \mathbb{I})$ where $\mathbb{I}$ is the identity matrix (i.e. $\mathbb{I}_{i,j} = 1$ if $i = j$ and $\mathbb{I}_{i,j} = 0$ otherwise). Note that this transform is not unique, but

$$X_l^T X_l = \frac{\theta_l^T \Omega^{-1} \theta_l}{\epsilon_l(1-\epsilon_l)} \tag{8}$$

is. Furthermore, if θ_l is truly multivariate normal then $X_l^T X_l$ is distributed $\sim \chi^2_J$. This suggests that $X_l^T X_l$ is a natural test statistic to identify loci that deviate away strongly from the multivariate normal distribution, e.g. due to selection. Furthermore, this form naturally accounts for hierarchical population structure, or other models of population structure, that can confound F_{ST}-style outlier analyses (Excoffier *et al.*, 2009). Our $X_l^T X_l$ statistic extends the ideas of Bonhomme *et al.* (2010), who developed a similar test statistic for the case of a known population tree (see also Robertson (1975), for earlier discussion of the effect of a population tree on the Lewontin and Krakauer (1973) test).

If we wish to test the correlation of our transformed allele frequencies with an environmental variable, we will also need to similarly transform our environmental variable, to ensure that our frequencies and environmental variable are in the same frame of reference. Specifically if our environmental variable is Y (standardized to be mean zero, variance 1) then our transformed environmental variable is

$$Y' = C^{-1}Y. \tag{9}$$

Note that this transform will exaggerate the environmental variable difference between very closely related populations. Furthermore, if part of the variation in the environmental variable precisely matches the major of axis of variation in the genetic data, then applying the transform may remove much of this variation. Both of these effects seem desirable properties, as we are interested in identifying correlations discordant with the patterns expected from drift. However, users should visually inspect Y and Y' to understand how the transform has altered the environmental variable (see Supplementary Figures 1-4 for examples).

We do not get to observe θ_l so we obtain a representative sample of M draws from the posterior $(X_l^{(1)}, \cdots, X_l^{(M)})$. Given these draws there is an enormous variety of ways that we could choose to summarize the support for the correlation with our environmental variable Y'. Here we choose

to write

$$\rho_l(X_l^{(1)}, \cdots, X_l^{(M)}) = \frac{1}{M}\sum_{i=1}^{M}\rho(X_l^{(i)}, Y'), \tag{10}$$

i.e. ρ_l is the mean of the function $\rho()$ over our posterior draws of X_l.

In the present paper, we calculated Pearson's and Spearman's correlation coefficients (as our $\rho()$) as alternative tests to the Bayes factors. To obtain an appropriate sample from the posterior in a computational efficient manner, these statistics were calculated between X_l and Y' every 500 MCMC generations and then averaged over the complete MCMC run. Our draws of X_l will therefore be weakly autocorrelated, but as ρ_l is a mean this does not affect its expectation.

While this standardization, for a known $\widehat{\Omega}$, would work perfectly if our θ_l were really multivariate normal, in reality this is only an approximation, as even under the null model deviations due to drift are only approximately normal over short time-scales. Thus, while we model drift at a locus as being multivariate normal (i.e. θ_l has a prior given by eqn. (3)), if the true model is more complex the joint probability of this along with our count data (and our uncertainty in Ω) may force θ_l to not be MVN$(0, \mathbb{I})$. While, under these circumstances, X_l will conform to those assumptions better than θ_l, we still choose to use the empirical distribution of ρ_l across SNPs rather than rely on asymptotic results, which may not hold.

2.3 Sequencing of pooled samples

If genotyping is conducted as sequencing of population pools, an additional step of sampling is included. At a site l the total coverage of in population j is m_{jl} and we observe r_{jl} reads supporting allele 1. Assuming that each individual contributed the same number of chromosomes to the pool, we can conclude that the sequenced reads are the result of binomial sampling

$$P\left(r_{jl}|\frac{i}{n_j}, m_{jl}\right) = \binom{m_{jl}}{r_{jl}}\frac{i}{n_j}^{r_{jl}}\left(1-\frac{i}{n_j}\right)^{m_{jl}-r_{jl}}, \tag{11}$$

where $\frac{i}{n_j}$ is the unknown sample allele frequency in the pooled sample. Summing over this unknown frequency

$$P\left(r_{jl}|m_{jl}, p_{jl}, n_j\right) = \binom{m_{jl}}{r_{jl}}\sum_i\left(\frac{i}{n_j}\right)^{r_j}\left(1-\frac{i}{n_j}\right)^{m_{jl}-r_{jl}}\binom{n_j}{i}{p_{jl}}^i\left(1-p_{jl}\right)^{n_j-i} \tag{12}$$

gives us the probability of our sampled reads given the population frequency. This replaces the binomial probability (eqn. (1)) in the joint probability given by eq. (4). In Coop *et al.* (2010) the Bayes factors were approximated by an importance sampling technique while performing MCMC under the null model, i.e. $\beta = 0$. This allowed the rapid calculation of the Bayes factor for many environmental variables with little extra computational cost. However, Bayes factors calculated by this technique are noisy, and so here we also implement an MCMC to estimate the posterior on β. We place a uniform prior on β and update β along with ϵ_l and θ_l. For our update on β we use a small normal deviate ($\sigma = 0.01$) and accept this move with the ratio of the joint posterior of our current parameters to that of our old parameters. As a simple summary of the posterior support for $\beta \neq 0$, we look at the skew of the posterior away from zero. Specifically we estimate the proportion (f) of the marginal posterior on β that is above 0, and then take $Z = |0.5 - f|$ as a test statistic, with values of Z close to 0.5 showing strong support for $\beta \neq 0$.

2.4 Power simulations

The extended model was implemented in Bayenv2.0. Simulations were conducted to evaluate the power of these extensions. To use both a realistic covariance among populations and realistic environmental values, we based these simulations on SNP data from the HGDP populations (Conrad *et al.*, 2006) and the environmental variables measured at these sampling locations (also used in Hancock *et al.*, 2008; Coop *et al.*, 2010). We employed a single Bayenv2.0 estimate of the covariance matrix $\widehat{\Omega}$ from the original SNP data (sampled after 100,000 MCMC iterations) to simulate population allele frequencies. For each SNP, an ancestral frequency ϵ_l was drawn from a beta distribution (with parameters $\alpha = 0.5, \beta = 3$). Then population allele frequencies were drawn from the multivariate normal $MVN(\epsilon_i, \epsilon_i(1-\epsilon_i)\widehat{\Omega})$ using the MASS package for GNU R (R Development Core Team, 2011). In contrast to the more empirical approach in Coop *et al.* (2010), who used observed SNP frequencies, these simulated population frequencies allow us to vary sample size and sequencing coverage for a population. For the simulation of pooled NGS data, we assume that the depth of coverage of a pool follows a negative binomial distribution, which allows for the over-dispersion of read depths compared to the Poisson. Coverages for each population and SNP were independently drawn from a negative binomial distribution $NB(r, p)$ where we set $r = 5$ and set p to obtain the respective coverage mean ($NB(r, p)$ has a mean of $pr/(1-p)$ and a variance of $pr/(1-p)^2$). This represents an extreme case where the variance-mean-ratio increases for higher average coverages. Such pattern is generally consistent with observations from pooled next-generation data generated along an altitudinal gradient in *Arabidopsis thaliana* (T.G., C. Lampei, O. Simon and K.J. Schmid, unpublished results).

To construct a null distribution we calculated Bayes factors or our test statistic Z for these simulated SNPs and an environmental variable Y during 100,000 MCMC iterations. For a second set of SNPs, an environmental effect was simulated by drawing their population allele frequencies from a multivariate normal $MVN(\epsilon_i + \beta Y, \epsilon_i(1-\epsilon_i)\widehat{\Omega})$. Again Bayes factors or our test statistic Z were calculated over 100,000 MCMC iterations. An environmental effect of $|\beta| = 0.06$ was simulated when all 52 HGDP populations were used and $|\beta| = 0.15$ was used for simulations of smaller population subsets; positive and negative βs were simulated in identical proportions for all simulations where Z was calculated. To estimate power for our ρ_l statistics samples of 20 chromosomes from each of the 52 HGDP populations were simulated and β was varied between 0.01 and 0.09. Power estimates were based on the proportion of these SNPs that were detected at a certain significance level α (5% here), i.e. the fraction of our simulations (with a β) in upper α tail of the null distribution.

2.5 Data application

Bayenv2.0 was used to re-analyze a genome-wide data set of 640,698 SNPs from 52 HGDP-CEPH populations (Li *et al.*, 2008; Hancock *et al.*, 2010) using Bayes factors and our non-parametric test statistic (ρ_l). We restricted our analysis to winter conditions, as most winter climate variables show outliers and a non-normal distribution. All environmental variables were normalized to a mean of zero and a standard deviation of one. The covariance matrix was estimated from a random subset of 5,000 SNPs after 100,000 MCMC iterations. Bayes factors and correlation coefficients for each SNP were estimated during 100,000 MCMC iterations. In addition to these test statistics, we sampled X_l every 500 MCMC generations and calculated $X_l^T X_l$. These values were averaged per SNP to calculate $\overline{X_l^T X_l}$ and to check for deviations from the multivariate normal distribution for

each SNP individually. SNP positions and gene annotations were obtained from Ensembl (Flicek *et al.*, 2012) and Entrez Gene (Maglott *et al.*, 2011).

3 Results

3.1 Using tests based on standardized allele frequencies

We explored the performance of tests based on our standardized transformed population frequencies (X_l). Before we calculate test statistics on our standardized allele frequencies, we first examined whether the multivariate standardization (as in eqn. (7)) had removed the covariance among populations from our standardized X_l. We first calculated the sample covariance matrix using the sample frequencies for 2,333 HGDP SNPs (the dataset of Conrad *et al.*, 2006) shown in Figure 1A. Specifically, denoting the vector of sample frequencies by k_l/n_l we calculated $\frac{1}{L}\sum_{l=1}^{L}(k_l/n_l)(k_l/n_l)^T$. As expected, there is substantial structure in this sample covariance matrix between regions, which corresponds to known population structure (Coop *et al.*, 2010). Then we calculated the sample covariance matrix of the X_l across these SNPs using Bayenv2.0; specifically we took a single draw of X_l (after a burnin) for each of these 2,333 SNPs, and calculated $\frac{1}{L}\sum_{l=1}^{L} X_l X_l^T$. The resulting sample covariance matrix (shown in Figure 1B) is close to the identity matrix in form, demonstrating that the majority of the covariance between populations has been removed. This suggests that our X_l are appropriately standardized for the application of correlation tests averaging across our uncertainty in X_l at each locus.

To further test the normality of X_l, we checked if $X_l^T X_l$ follows a χ^2 distribution with 52 degrees of freedom (i.e. the number of populations, see Methods). To test this, $X^T X$ was calculated in two different ways, first using the final generation of the MCMC ($X_l^{(M)T} X_l^{(M)}$) and the second using the average $X_l^T X_l$ across all M samples for each locus l ($\overline{X_l^T X_l}$). Figure 2 shows a QQ-plot of the $X_l^T X_l$ and the expected χ^2_{52} distribution. The mean of each distribution approximately matches that of the χ^2_{52}, whereas the variances do not. The estimates based on single samples from the MCMC show a somewhat higher variance. The averaging, on the other hand, led to a smaller variance, indicating that this approach is slightly over-conservative. Both observed distributions are not consistent with the expected χ^2_{52} distribution (Kolmogorov-Smirnov tests, both p-values $< 10^{-6}$). Therefore, while $X_l^T X_l$ provides a potentially suitable summary statistic for identifying empirical outliers, we cannot assume a distributional form to those outliers under a null neutral model. We chose to use $\overline{X_l^T X_l}$, as it averages over our uncertainty in the sample frequencies, and so should be more robust to outliers due to small sample sizes.

To explore the power of standard correlation tests applied to our standardized X_l, in comparison to the Bayes factors, we again conducted power simulations based on the HGDP data. We also calculated both Spearman's ρ and and Pearson's r between Y' and our transformed allele frequencies averaged across the posterior on these transformed frequencies. We transformed our latitude and minimum winter temperature value, our Y's, to give us Y' as in eqn. (9) (see Supplementary Figures 1-4). Statistics based on our Bayesian model clearly outperform correlation tests calculated for point estimates from sample allele frequencies (Figure 3). This improvement in power is due to the fact that the methods based on the sample frequencies fail to incorporate the sampling noise and the relationship among populations. All three tests based on Bayenv performed effectively identically with marginal advantages of the Bayes factors for minimum winter temperature (Figure 3B) and a slightly lower power of Spearman's ρ, which is not surprising, as all simulated effects are linear.

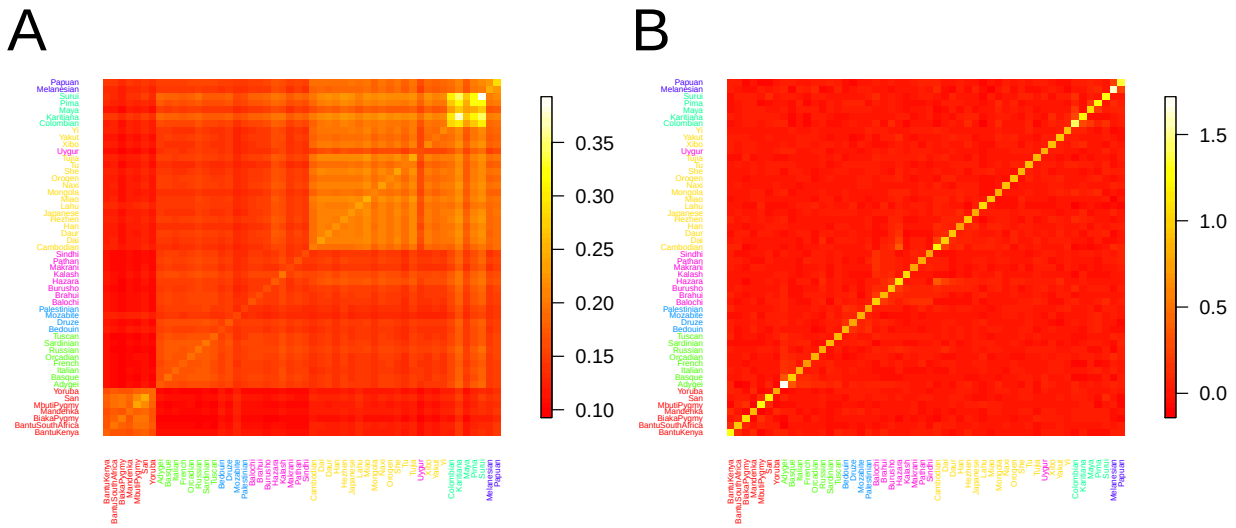

Figure 1: Covariance among HGDP populations estimated by Bayenv2.0 and the covariance calculated on the Xs for the same SNPs. Populations are colored according to broad geographic regions used in Rosenberg *et al.* (2002).

We expect that the relative performance of the rank-based test, i.e. Spearman's ρ, may be reduced as the number of populations is decreased. We also tested the power of the X_l tests incorrectly using Y in place of Y'; this gave rise to power curves intermediate between the two sets (data not shown). Overall these results show that correlation tests based on X_l perform well.

The alternative model of Bayenv (eqn. (6)) implies a linear relationship between the transformed allele frequencies and the environmental variable. However, the fitting and significance of this linear model may be misled by populations that are statistical outliers. For instance, linear models might mistakenly identify cases as strong candidates, when allele frequencies and environment for all but one population are consistent with our null model and this single outlier population features both an extreme environment and an extreme allele frequency. We note that the extreme allele frequency may be due to a component of drift not well modeled by our MVN framework, or due to a selection pressure (or response) poorly correlated with our environmental variable of interest. While loci of the latter form are of interest as *genomic* outliers, we believe researchers interested in particular environmental variables would consider such loci spurious, and would prefer a set of candidates where many populations support a consistent pattern.

To test such a case, we simulated allele frequencies for the HGDP populations based on $MVN(\epsilon_l, \epsilon_l(1-\epsilon_l)\widehat{\Omega})$ as described above. Winter minimum temperature was used as climate variable since one population, the Yakuts from north-east Russia, is characterized by a very low minimum temperature (Figure 4A). To create outliers, we set the allele frequencies of the Yakuts to 0. Both statistics based on linear models, Bayes factors and Pearson's correlation coefficient r showed an excess of false positives (Figure 4B), while a non-parametric statistic, in this case Spearman's rank correlation coefficient ρ, was much less sensitive to these outliers, with a false-positive rate very close to the expected value of 5% (Figure 4B).

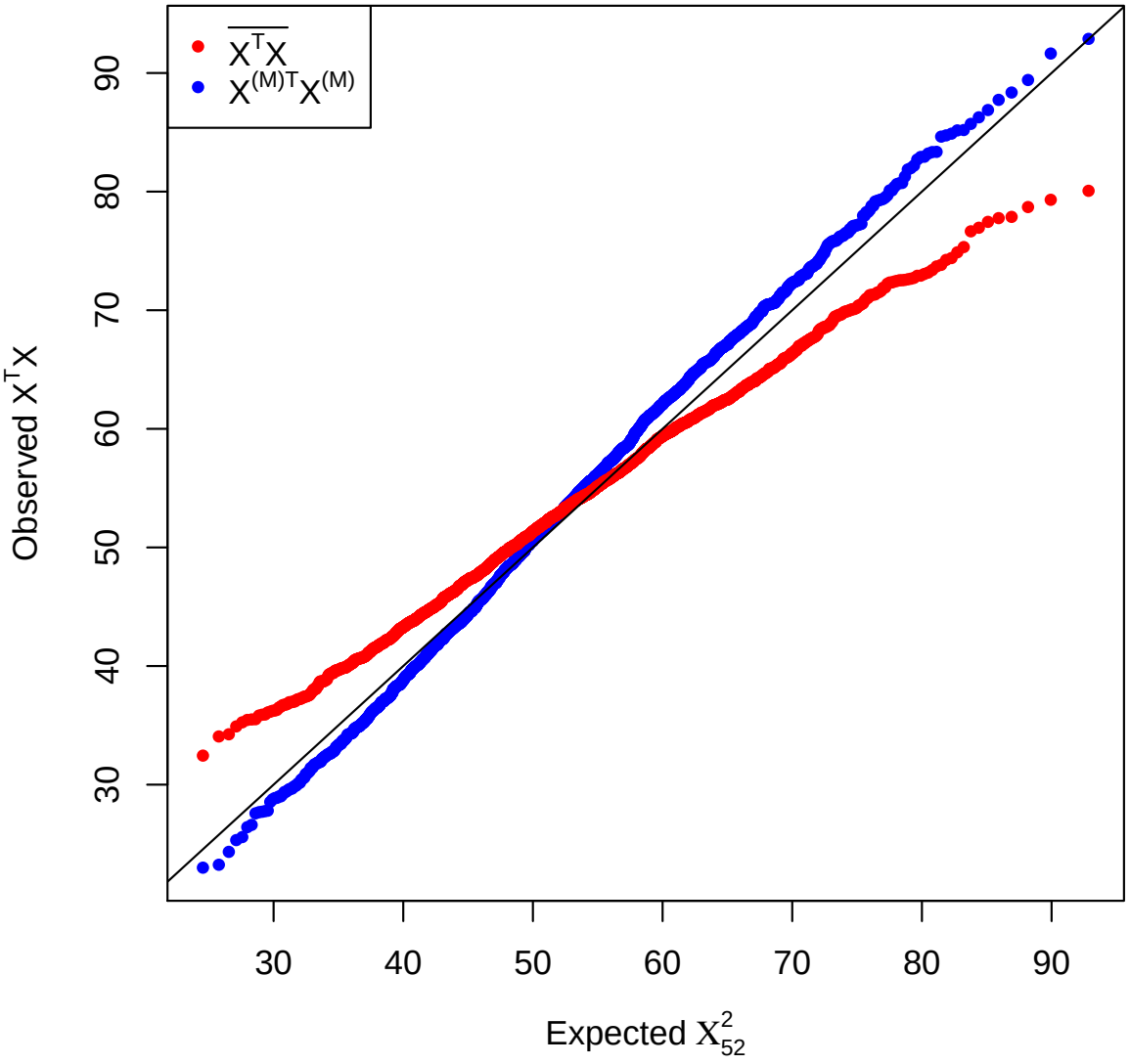

Figure 2: QQ-plot of $X_l^T X_l$ calculated in two different ways and the χ^2_{52} distribution, which is expected if X_l would follow MVN$(0, \mathbb{I})$.

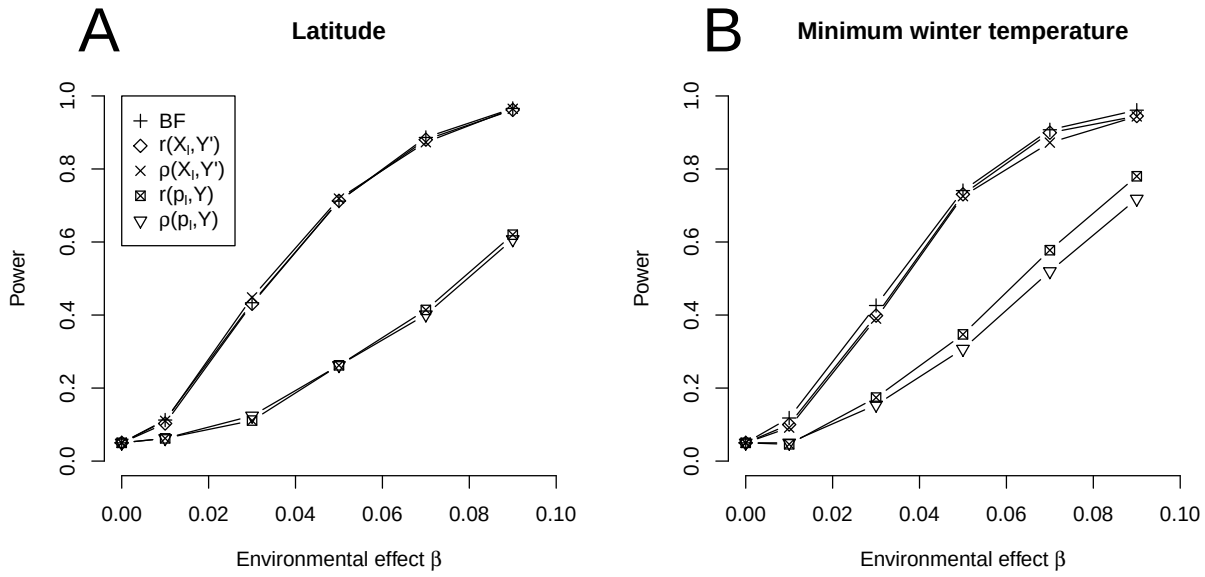

Figure 3: Power of Bayes factors compared to power of correlation coefficients on X based on simulations for all 52 HGDP populations for the environmental variables latitude (A) and minimum winter temperature (B).

3.2 Simulation of pooled data

Pooled sequencing of multiple individuals has increased in popularity, as it is considerably cheaper than barcoding all individuals and sequencing them separately (but see Cutler and Jensen, 2010). The use of allele frequencies estimated from the resulting read counts seems to be a reasonable application of our method. However, it raises the question how Bayenv behaves for different coverages as increasing sequencing coverage is not the same as increased numbers of sampled individuals. Therefore, we simulated data that resembles the HGDP populations and then pooled 10 diploid individuals (i.e. 20 chromosomes) from each population and used the populations' respective latitudes as our environmental variable.

We first experimented with incorrectly using read counts in place of the chromosome counts (i.e. assuming r_{jl} and m_{jl} were k_{jl} and n_{jl}, respectively), and found that this resulted in an excess of extreme Bayes factors for high coverages under the null (data not shown). We found this inflation to be most pronounced when read depths are greater than the actual sample size, and this is likely due to false certainty about the population frequencies. We then ran power simulations of Bayenv matched to the HGDP data, using Z as a test statistic, with the true sample frequencies (black squared in Figure 5), and incorrectly using the read counts as the input data for the previous version of Bayenv (Bayenv1.0, black circles in Figure 5). Bayenv2.0, which accounts for both stages of binomial sampling in pooled data (as described above), was also applied to the same read counts (white dots in Figure 5). The true sample frequencies naturally resulted in the best power as there is no additional sampling noise (Figure 5A). For higher mean coverages the power of Bayenv1.0 using the read counts as sample allele frequencies was almost as good as the power using true sample allele frequencies (Figure 5A). As most applications may consist of a smaller number of

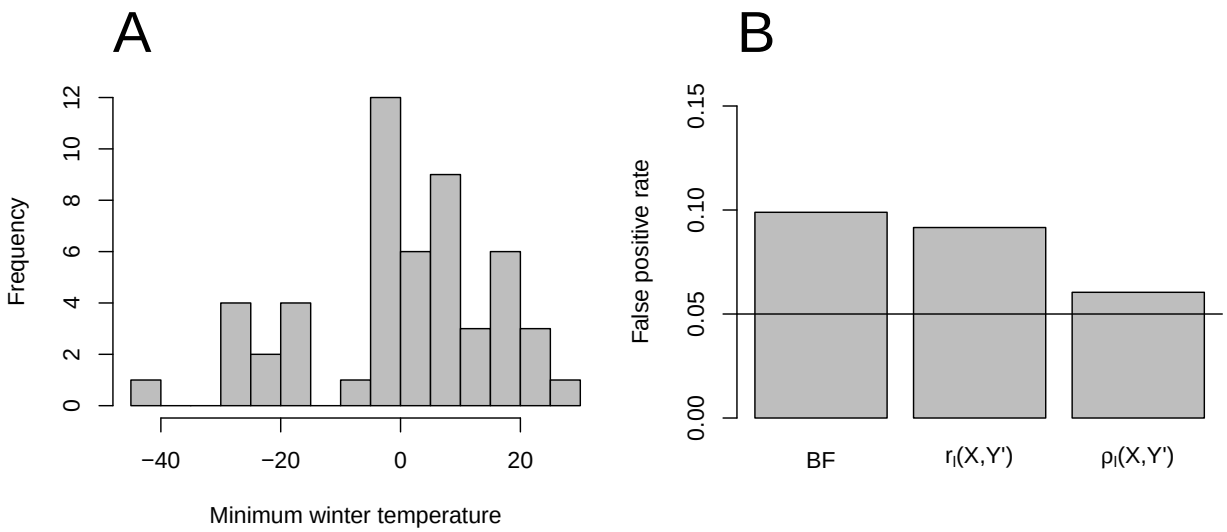

Figure 4: False-positives induced by populations at extreme conditions. (A) Histogram of minimum winter temperature for the 52 HGDP populations. (B) False-positive rate of different statistics if one allele is fixed in the coldest population.

populations, we additionally sampled two subsets consisting of all HGDP sub-Saharan African populations (7 populations; Yoruba, San, Mbuti Pygmy, Mandenka, Biaka Pygmy, Bantu South Africa, Bantu Kenya; Figure 5B) and eight populations spread over the entire globe (Bantu Kenya, French, Bedouin, Cambodian, Japanese, Uygur, Colombian, Papuan; Figure 5C). On all of these, Bayenv using the true sample frequencies out-performed Bayenv1.0 using the read counts.

In part the poor power in pooled studies is unavoidable due to the additional sampling noise. However, the loss of power is likely boosted by failing to properly account for this second stage of sampling, which leads to poor performance due to variation in depth across populations and SNPs. The extended model of Bayenv2.0 should compensate the loss of power to some extent. Somewhat surprisingly, we did not observe any advantages of the extended model in detection power if all 52 HGDP populations are simulated (Figure 5A). The small differences between the extended model and incorrectly using the read counts as the input are mainly due to convergence of the MCMC, which is somewhat slower incorporating both levels of sampling. Including the sampling of reads into the model had a clearly positive effect on power in our population subsets and incorrectly using the read counts as input did not reach similar powers for high coverages (Figure 5B, C). However, power of Bayenv2.0 was still considerably low for mean coverages $< 20\times$, suggesting that such low read depths do not provide enough certainty for reliable frequency estimation.

Notably, a simulated effect of identical magnitude was detected with a higher power in the seven sub-Saharan populations than in the eight worldwide populations. Additionally, the power difference between the extended model and incorrectly using the read counts as the input was higher in the sub-Saharan populations. This demonstrates the effect of different covariance structures among populations (see Figure 1A) and their relation to the environmental variable on the performance of Bayenv. Presumably this lower power in the world-wide samples is due to the fact that the levels of drift are higher between world-wide populations than within Africa.

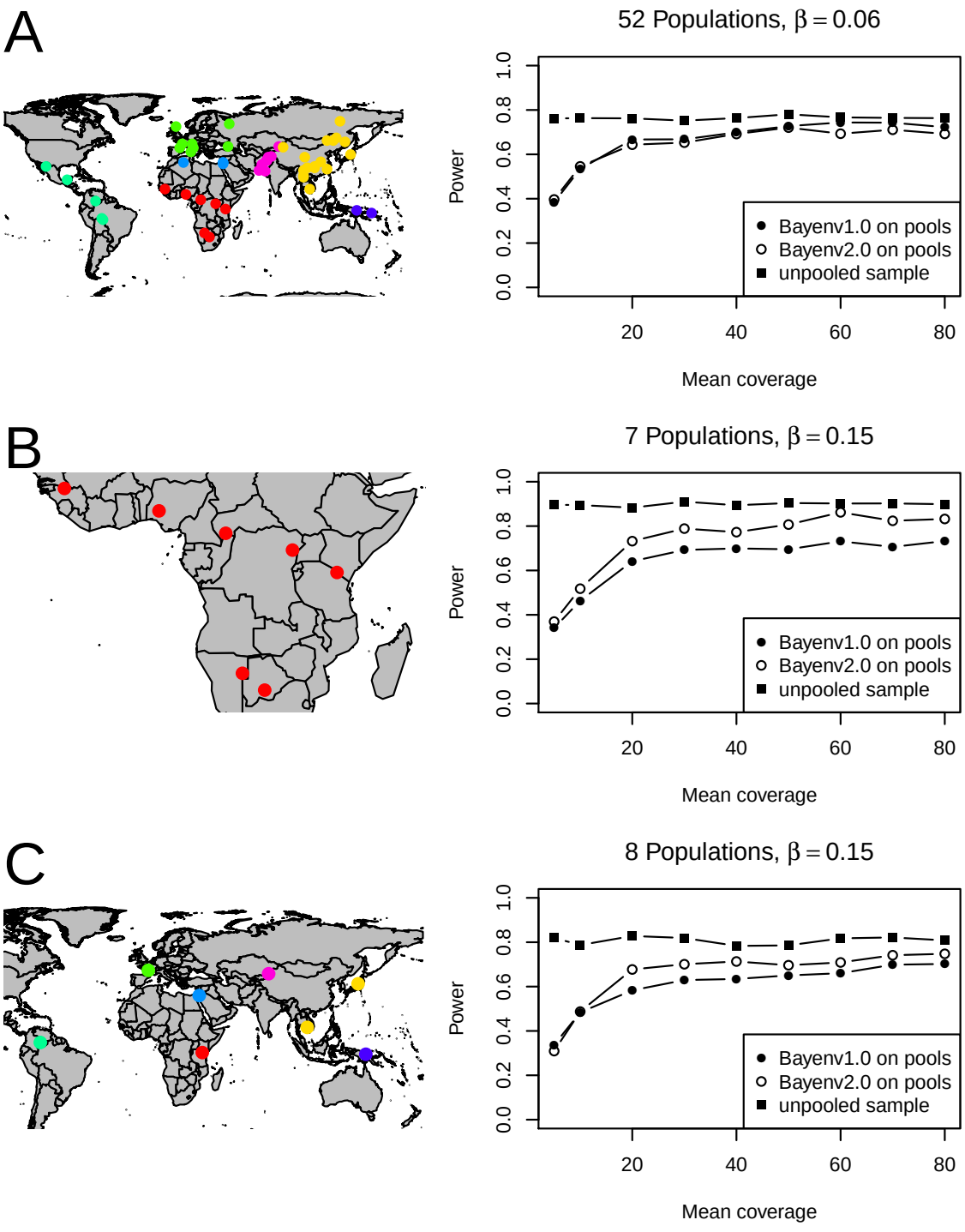

Figure 5: Power to detect environmental correlations with latitude in pooled samples. (A) in all 52 HGDP populations, (B) in the seven sub-Saharan populations and (C) in one population per broader geographic region. Populations are colored as in Figure 1.

3.3 Robust candidates in the HGDP data

Finally we explored the use of our standardized X_l for identifying robust putative candidates for adaptive evolution in the HGDP data of Li *et al.* (2008).

As described above, populations with outliers in terms of allele frequencies and/or environments can potentially lead to spurious correlations. For example, the use of minimum winter temperature as an environmental variable could generate false positive correlations in analyses of the HGDP data because of the extremely low temperature for the Yakut population. To explore this, we used minimum winter temperature and re-analyzed all 640,698 SNPs of the HGDP data, calculating both Bayes factors and $\rho_l(X_l, Y')$ using Spearman's rank correlation coefficient ρ. Our Bayes factors and $|\rho_l(X_l, Y')|$ are correlated across SNPs (Spearman's $\rho = 0.72$) and show an overlap of 29 SNPs in their top 100 most extreme SNPs, 142 SNPs in the top 500 and 2.8 % in the top 5 % signals. These overlaps are substantial but suggest that our two tests are detecting somewhat different signals, which likely reflects in part the influence of outlier populations.

The 100 strongest signals of the Bayes factor analysis and $|\rho_l(X_l, Y')|$ are shown in Supplementary Tables 1 and 2. The top 5 Bayes factors include SNPs that fall in potential candidate genes, such as epidermal growth factor receptor (*EGFR*, Hancock *et al.*, 2008) and a non-synonymous SNP in zonadhesin (*ZAN*, Gasper and Swanson, 2006), both of which were previously identified in small scale selection scans. We also find a SNP (rs6500380) located in a region associated with earwax type (i.e. wet or dry) which has been subject to a selective sweep in East Asian populations (Ohashi *et al.*, 2011). Further signals fall in genes involved in fat metabolism, which is a plausible trait for the adaptation to low temperatures. Among our top hits multiple SNPs fall in the gene *MKL1* (megakaryoblastic leukemia 1), which is a myocardin-related transcription factor that has been associated with various disease phenotypes (Ma *et al.*, 2001; Hinohara *et al.*, 2009; Scharenberg *et al.*, 2010), but is also involved in smooth muscle cell differentiation, mammary gland function, and cytoskeletal signaling (Parmacek, 2007; Maglott *et al.*, 2011).

To exemplify the effect of an outlier, we compare two SNPs that fall in our top 20 Bayes factors. Both SNPs, rs6001912 and rs7974925 (Figure 6A, C), are characterized by similarly high Bayes factors (Supplementary Table 1) and extreme allele frequencies in the Yakuts (Figure 6B, D). However, only rs6001912 is among the top 25 signals for both statistics, whereas rs7974925 is only among the top 5 % of Spearman's ρ (Supplementary Tables 1, 2). This suggests that the Bayes factor signal at rs7974925 is strongly driven by the low allele frequency in the Yakuts, and the signal at rs6001912 is more robust even without this outlying data point (Figure 6A, C). We suggest that the Bayes factors, or other linear model test statistics, should be used in conjunction with robust test statistics such as those described here to avoid spurious signals due to outliers. As these both can be calculated from the same MCMC run, this should be reasonably computationally efficient.

We also explored our test statistic $\overline{X_l^T X_l}$, designed to highlight loci that deviate strongly from the expected pattern of population structure, calculated for each of the 640,698 HGDP SNPs. These have been uploaded as a genome browser track to `http://hgdp.uchicago.edu/`. The empirical distribution is shown in Figure 7. The empirical distribution clearly differs from the expected χ^2_{52} distribution, having a higher mean and a lower variance than expected. This again highlights that we do not have a good theoretical expectation for the distribution and so must use the empirical ranks to judge how interesting a signal is. To briefly explore where known signals fall in our empirical distribution in Figure 7, we also plot as arrows the maximum $\overline{X_l^T X_l}$ for SNPs that fall within 50 kbp up- and downstream of ten well known pigmentation genes (list taken from Pickrell *et al.*,

2009). As these arrows represent maximums across a number of SNPs around the gene, they will necessarily be more extreme than an average draw from this distribution. However, the extreme signals at a number of these genes demonstrate that the method is detecting loci with extreme allele frequency patterns. The SNP with the most extreme value of $\overline{X_l^T X_l}$ in the genome falls close to *SLC24A5* (Lamason *et al.*, 2005), with a SNP close to *SLC45A2* being the second largest signal in the genome (Nakayama *et al.*, 2002). More generally, five of these ten pigmentation genes fall in the top 1% and nine genes fall in the top 5% of the empirical distribution. A SNP close to the gene *EDAR*, one of the highest pairwise F_{ST} between East Asia and Western Eurasia HGDP populations, is also in the top ten SNPs (Sabeti *et al.*, 2007).

To examine the relationship between $\overline{X_l^T X_l}$ and global F_{ST} we took per SNP values of global F_{ST} previously calculated among the colored groupings depicted in Figure 1 (values from Pickrell *et al.*, 2009; Coop *et al.*, 2009). The Spearman's ρ between $\overline{X_l^T X_l}$ and F_{ST} was 0.48. Looking at the extremes of both distributions, $\overline{X_l^T X_l}$ and F_{ST} show an overlap of 6 SNPs in their top 100 most extreme SNPs, 37 SNPs in the top 500 and 1.4 % in the top 5 % signals. In Supplementary table 3 we present the top 100 $\overline{X_l^T X_l}$ SNPs in the genome, along with their nearest genes and global F_{ST} values.

The weak overlap in the tails of the genome-wide $\overline{X_l^T X_l}$ and F_{ST} means that they are finding different sets of candidate SNPs, presumably due to the reweighting of allele frequencies in $\overline{X_l^T X_l}$. For example, our 12^{th} highest SNP for $\overline{X_l^T X_l}$ falls close to *MCHR1*, with our 21^{st} highest gene being a non-synonymous variant (rs133072) in this gene. *MCHR1* (Melanin-concentrating hormone receptor 1) is known to play a role in role in the intake of food, body weight, and energy balance in mice (Marsh *et al.*, 2002), and the effect of the nonsynonymous variant on obesity has been debated (Wermter *et al.*, 2005; Rutanen *et al.*, 2007; Kring *et al.*, 2008) but the variant did not achieve genome-wide significance in a large genome-wide association meta-analysis of BMI (Speliotes *et al.*, 2010). Both of these SNPs are nearly fixed differences between East Asia and the American HGDP populations (Supplementary Figures 5 and 6). This strong difference between regions that share a recent history and, thus, covariance among allele frequencies (Figure 1), makes these SNPs an interesting pattern for $\overline{X_l^T X_l}$. However, neither of the two SNP has an extremely impressive global F_{ST} (falling in only the 5% tail), presumably because East Asia and the American HGDP populations are only two of seven groups in the global F_{ST} calculation and the other five groups do not show an interesting pattern.

4 Discussion

In this article we have presented a method to more robustly identify loci where allele frequencies correlate with environmental variables. We have also described a method to detect loci that are outliers with respect to genome-wide population structure, while accounting for the differential relatedness across populations.

Many available tests for selection are designed to detect rapid complete sweeps from new mutations; however, such events are likely just a small percentage adaptive genetic change (Coop *et al.*, 2009; Pritchard *et al.*, 2010; Cao *et al.*, 2011; Hernandez *et al.*, 2011). Analyzing allele frequencies across multiple populations offers the opportunity to detect selection acting on standing variation and polygenic phenotypes. The falling cost of genotyping means that typing individuals from many populations is now in reach, which will allow us to connect environmental variables to

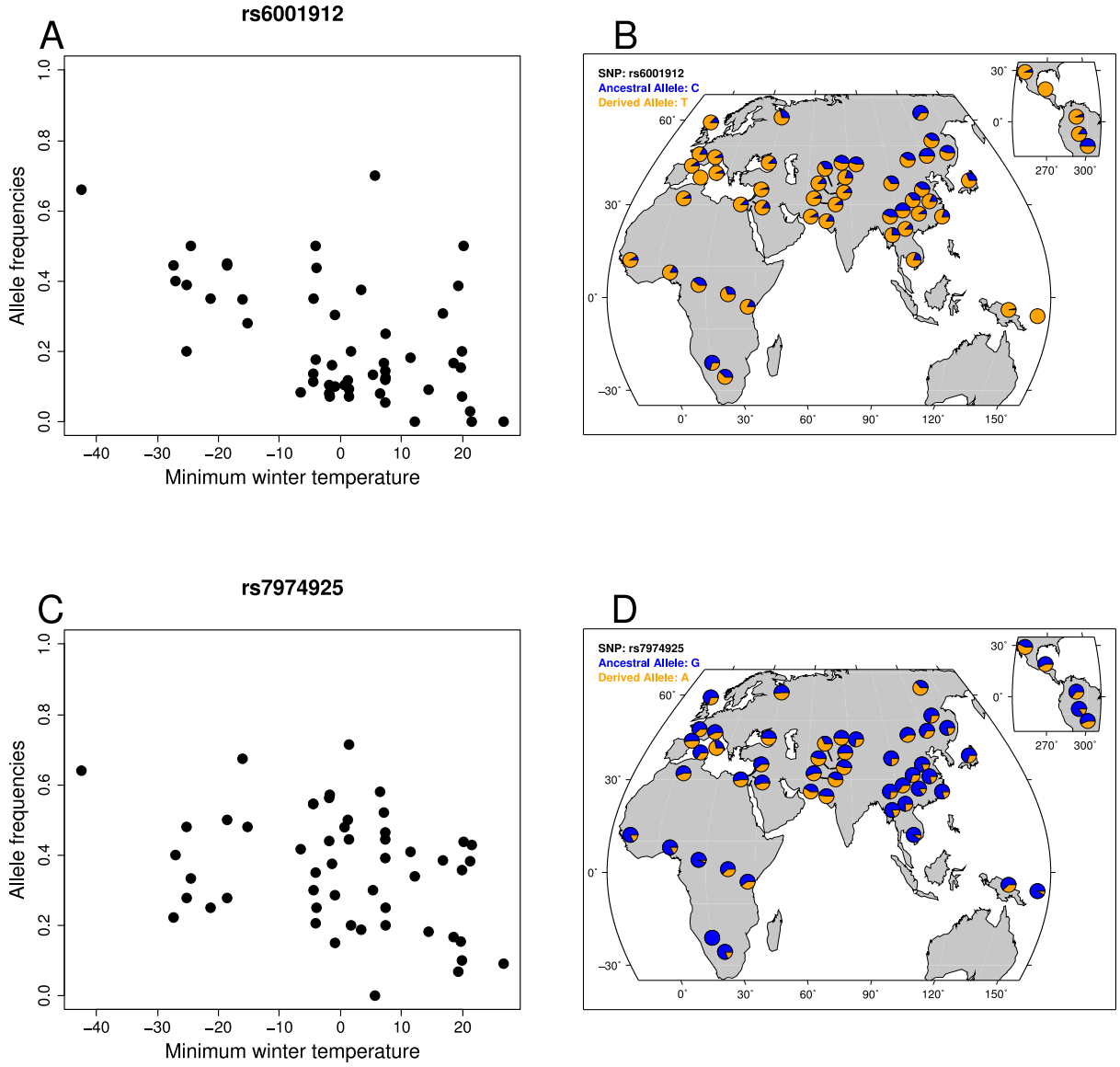

Figure 6: Two exemplarily chosen SNP from the top 20 Bayes factors. (A) Allele frequencies and standardized minimum winter temperatures of rs6001912 which is among the top 25 SNPs of both statistics BF and ρ, (B) shows the geographical distribution of rs6001912. (C) rs7974925 is among the top 20 BFs but only the top 7,000 ρ signals which is mainly caused by the two outlier populations, (D) shows the geographical distribution of rs7974925. Plots of geographic distributions were downloaded from the HGDP selection browser (Pickrell *et al.*, 2009).

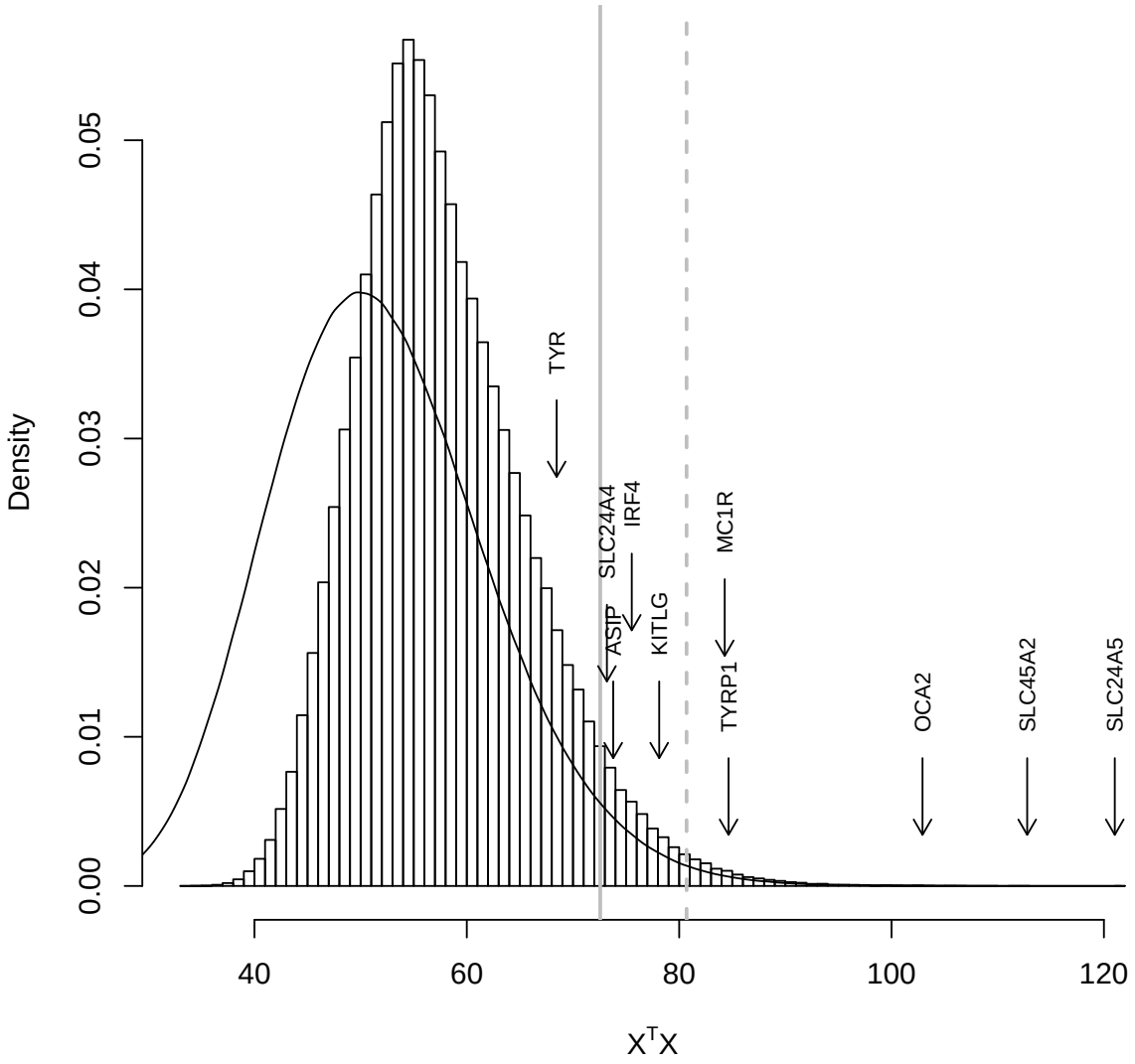

Figure 7: Histogram of $X_l^T X_l$ calculated for all HGDP SNPs. The labels of candidate genes are shown at the maximum $X_l^T X_l$ of any SNP within 50kbp up- and downstream of the particular gene. The solid line shows the position beyond which 5% of all $X_l^T X_l$ fall, the dashed line denotes the top 1%. The solid black line shows the density of the expected χ^2_{52} distribution.

more subtle adaptive genetic variation. However, we stress that loci detected by the approaches discussed above are obviously at best just candidates for being involved in adaptation to a particular climate variable, or set of climate variables, and so additional evidence is needed to build the adaptive case at any locus.

Our use of the covariance matrix of population allele frequencies when looking for environmental correlations is conceptually similar to linear mixed model (LMM) approaches that account for kinship structure in genome-wide association studies (GWAS) (e.g. Yu *et al.*, 2006; Kang *et al.*, 2008, 2010; Zhou and Stephens, 2012, who use a observed relatedness matrix as the covariance matrix of the random effect). One important difference is that we seek to predict allele frequencies at a locus using the environmental variable, whereas these LMM methods are predicting a phenotype as a function of genotypes at a locus. In our approach the equivalent of the random effect matrix is a proxy for a neutral model of allele frequency variation, while in the application to GWAS the kinship matrix accounts for the cofounding due to heritable variation in the phenotype elsewhere in the genome. Our model could be used to detect loci that were strongly covaried with population mean phenotypes (e.g. phenotypes measured at the breed level in dogs Boyko *et al.*, 2010). However, the method used this way would have a high rate of false positives if there are large environmental effects on the phenotype that coincide with the principal axes of the covariance matrix. Similarly, LMM approaches could be used to identify loci which covaried with environmental gradients, but they may be underpowered as their random effects model does not attempt to reflect a model of genetic drift.

Standardized allele frequencies We introduced a set of tests based on using our model of the covariance of allele frequencies to produce a set of standardized allele frequencies (X_l). The calculation of standardized allele frequencies allows us to calculate a variety of statistics while taking advantage of the other features of Bayenv2.0's approach to account for covariance among populations and sampling noise. The removal of covariance is often a standard step in multivariate analysis; here we remove this covariance structure in a way that acknowledges the approximate form of genetic drift and the bounded nature of allele frequencies. By integrating our statistic across the posterior for X_l, we are averaging across our uncertainty in allele frequencies, which should further increase our power.

As an example of the usefulness of the X_l, we explored their application in identifying robust correlations with environmental variables. While the use of Spearman's ρ on these transformed allele frequencies results in a small loss of power, it is much less sensitive to outliers and able to detect any monotonic relationship. Therefore, a combined approach which takes a set of SNPs in the intersection of the tail of Bayes factors and in the tail of Spearman's ρ on our transformed allele frequencies should provide best results.

Our transformed allele frequencies could also be used to detect and distinguish between the effects of multiple environmental variables shaping variation at a locus. This could be accomplished by including the multiple transformed environmental variables (Y') into a linear model to predict the X_l at a locus or by applying appropriate transformed ecological niche models (ENM) to the X_l to understand the predictors of allele frequencies at a locus (see Fournier-Level *et al.*, 2011; Banta *et al.*, 2012, for applications of ENMs to allele frequencies). However, there is limited information about the effects of even a single environmental variable from contemporary allele frequencies if neutral allele frequencies are autocorrelated on the same scale as environmental variation (as is the case in humans). Therefore, we caution that in many situations there will be very limited power

to learn about the effect of multiple environmental variables.

Using our X_l statistics, we also introduced a method to identify loci that are outliers from the general pattern of population structure (our X^TX statistic). This statistic is closely related to F_{ST}, which can be expressed as $Var(p_{lj})/(\epsilon_l(1-\epsilon_l))$, where $Var(p_{lj})$ is the variance of our allele frequency across populations (see Nicholson *et al.*, 2002; Balding, 2003; Bonhomme *et al.*, 2010, for discussion). Our statistic, which is the variance of X_l, can be written as eqn. (8), and so X^TX can be seen as closely related to calculating F_{ST} on the standardized allele frequencies. Importantly, by removing the covariance, we reweight populations so that a small change shared across many closely related populations is downweighted. This reweighting therefore should increase our power to detect unusual allele frequencies compared to global F_{ST}. The fact that we remove the covariance between closely related populations also means that, unlike F_{ST}-based methods, we do not have to arbitrarily clump populations in order to identify globally differentiated SNPs. While in this paper we use the 52 HGDP population labels, in principle Bayenv2.0 could be run treating each individual as a population, allowing X^TX to be calculated without regard to any population label. However, this would be computationally time-consuming with thousands of individuals. In that case perhaps the sample frequencies and the sample covariance matrix, could instead be used to mitigate the computational burden.

Ideally our X^TX statistic would have a parametric distribution under a general null model where only drift and migration shaped our frequencies. That might allow us to make statements about what fraction of allele frequency change was due to selection. Indeed, as noted above, if our population frequencies were truly multivariate normal, our X^TX statistic would be χ^2 distributed if our sample sizes were sufficiently large. This assumption would be approximately met if our levels of drift were sufficiently small, such that the transition density of allele frequencies was well approximated by a normal (see Price *et al.*, 2009; Bhatia *et al.*, 2011, for recent empirical applications along these lines). However, when levels of drift are higher, our normal approximation will be break down, as demonstrated by the poor fit of the χ^2 to the transformed HGDP frequencies. The distribution of our statistic could be obtained by simulation if the population history were known. In practice, we are skeptical that our knowledge of population genetic history will be sufficiently accurate to make this feasible, but simulations may be useful in guiding the setting of approximate significance levels.

Pooled Next-generation sequencing Recent empirical validations have shown that pooled re-sequencing of populations is a powerful and cost-efficient way to estimate allele frequencies (Zhu *et al.*, 2012), but see Cutler and Jensen (2010). The down-side of the saving of costs in library preparation and sequencing is the potential for increased sampling noise in the allele frequency estimates (Futschik and Schlötterer, 2010; Zhu *et al.*, 2012) and the loss of haplotype information (although some haplotypic information can be recovered, Long *et al.*, 2011). We account for the sampling of sequencing reads as an additional level of binomial sampling in the model of Bayenv2.0. Our power simulations show that accommodating the extra level of sampling in pooled designs can help to improve the power. However, they also highlight the large unavoidable loss in power due to increased sampling noise when the depth of coverage is low. The only way that this can be circumvented is through increasing sequencing coverage to provide sufficient certainty in the estimated allele frequencies and, thus, sufficient power to detect environmental correlations. Although low fold sequencing of many populations may help to increase power in some situations, is likely that for some species (notably humans) sampling, and not sequencing, will be the limiting

resource in the future.

Our model of pooled resequencing in Bayenv2.0 implies uniform sampling of reads from each individual. Therefore, we do not account for the possibility of an unequal number of chromosomes per individual due to measurement errors, different DNA content per individual, or differences caused during DNA extraction, all of which might cause additional noise in the allele frequency estimation (Futschik and Schlötterer, 2010; Cutler and Jensen, 2010). This additional noise, if it is constant across loci, should be absorbed into the covariance matrix in Bayenv2.0, which will result in a reduction in power. However, including a sufficient number of individuals in each pool should mitigate this effect (Zhu *et al.*, 2012). Furthermore, our model assumes perfectly called bases since we do not consider quality scores or sequencing errors. Rearchers dealing with NGS data should exercise caution with these issues. However, examining multiple population pools simultaneously provides some straightforward approaches to minimize error rates in SNP calling, such as calling only SNPs supported by a minimum number of reads in at least one population (Futschik and Schlötterer, 2010). Such strategies are already good practice in studies of pooled samples and should be used in combination with the Bayenv model. For the application to individual based NGS data, further possible extensions of our model include sequencing errors and probabilistic genotype calling (see Nielsen *et al.*, 2011, for a discussion on SNP calling from NGS data).

Outlook The population genomic comparison of closely related populations that differ strongly in environmental variables has already yielded many great candidate loci (see for example, altitude adaptation in Tibetans, Beall *et al.*, 2010; Simonson *et al.*, 2010; Yi *et al.*, 2010). The methods developed here and elsewhere are part of realizing the power of these population comparisons. Such empirical studies also highlight the current deficiencies of such methods, as some of the best signals in these studies are not shared across populations with broadly similar environments, and instead indicate that adaptation has occurred through independent mutations in the same gene or pathway. For example, high altitude adaptation seems to have a different genetic basis in highland Ethiopian and Andean populations (Bigham *et al.*, 2010; Scheinfeldt *et al.*, 2012). Methods based on environmental correlations will fail to detect such cases, unless the data are split into the appropriate geographic subsets (e.g. Hancock *et al.*, 2011c) on an appropriate geographic scale (Ralph and Coop, 2010). While shared standing variation will surely be part of the adaptive response across geographically separated instances of similar environments, ideally we would have methods that could cluster signals at the level of the gene or pathway to allow putative cases of parallel adaptation to be identified. The development of such techniques poses an important challenge for future method development.

5 Acknowledgements

We thank Gideon Bradburd, Yaniv Brandvain, Fabian Freund, Chuck Langley, Jonathan Pritchard, Peter Ralph, Jeffrey Ross-Ibarra, Karl Schmid, and Alisa Sedghifar for helpful discussions and comments on earlier versions of the manuscript. We also thank Joseph Pickrell for making X^TX available through the HGDP selection browser. TG was supported by the German Federal Ministery for Education and Research (Synbreed, 0315528D), and by a VolkswagenFoundation scholarship (I/84225) affording him to visit UC Davis. GC was supported by a Sloan Foundation fellowship.

References

Akey, J. M., A. L. Ruhe, D. T. Akey, A. K. Wong, C. F. Connelly, *et al.*, 2010 Tracking footprints of artificial selection in the dog genome. Proceedings of the National Academy of Sciences of the United States of America **107**: 1160–5.

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theoretical Population Biology **63**: 221–230.

Banta, J. A., I. M. Ehrenreich, S. Gerard, L. Chou, A. Wilczek, *et al.*, 2012 Climate envelope modelling reveals intraspecific relationships among flowering phenology, niche breadth and potential range size in *Arabidopsis thaliana*. Ecology Letters **15**: 769–77.

Beall, C. M., G. L. Cavalleri, L. Deng, R. C. Elston, Y. Gao, *et al.*, 2010 Natural selection on *EPAS1* (*HIF2alpha*) associated with low hemoglobin concentration in Tibetan highlanders. Proceedings of the National Academy of Sciences of the United States of America **107**: 11459–64.

Bhatia, G., N. Patterson, B. Pasaniuc, N. Zaitlen, G. Genovese, *et al.*, 2011 Genome-wide Comparison of African-Ancestry Populations from CARe and Other Cohorts Reveals Signals of Natural Selection. American Journal of Human Genetics **89**: 368–81.

Bigham, A., M. Bauchet, D. Pinto, X. Mao, J. M. Akey, *et al.*, 2010 Identifying Signatures of Natural Selection in Tibetan and Andean Populations Using Dense Genome Scan Data. PLoS Genetics **6**: e1001116.

Boitard, S., C. Schlötterer, V. Nolte, R. Pandey, and A. Futschik, 2012 Detecting selective sweeps from pooled next generation sequencing samples. Molecular Biology and Evolution .

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, *et al.*, 2010 Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics **186**: 241–262.

Boyko, A. R., P. Quignon, L. Li, J. J. Schoenebeck, J. D. Degenhardt, *et al.*, 2010 A simple genetic architecture underlies morphological variation in dogs. PLoS Biology **8**: e1000451.

Cao, J., K. Schneeberger, S. Ossowski, T. Günther, S. Bender, *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nature Genetics **43**: 956–963.

Cavalli-Sforza, L. L., 1966 Population structure and human evolution. Proceedings of the Royal Society of London. Series B, Containing papers of a Biological character. Royal Society (Great Britain) **164**: 362–79.

Cheng, C., B. J. White, C. Kamdem, K. Mockaitis, C. Costantini, *et al.*, 2012 Ecological Genomics of *Anopheles gambiae* Along a Latitudinal Cline in Cameroon: A Population Resequencing Approach. Genetics **190**: 1417–1432.

Conrad, D. F., M. Jakobsson, G. Coop, X. Wen, J. D. Wall, *et al.*, 2006 A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nature Genetics **38**: 1251–60.

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, *et al.*, 2009 The role of geography in human adaptation. PLoS Genetics **5**: e1000500.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. Genetics **185**: 1411–23.

Cutler, D. J., and J. D. Jensen, 2010 To pool, or not to pool? Genetics **186**: 41–3.

Eckert, A. J., A. D. Bower, S. C. González-Martínez, J. L. Wegrzyn, G. Coop, *et al.*, 2010 Back to nature: ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). Molecular Ecology **19**: 3789–805.

Excoffier, L., T. Hofer, and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity **103**: 285–98.

Fabian, D. K., M. Kapun, V. Nolte, R. Kofler, P. S. Schmidt, *et al.*, 2012 Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Molecular Ecology : n/a–n/a.

Fang, Z., T. Pyhäjärvi, A. L. Weber, R. K. Dawe, J. C. Glaubitz, *et al.*, 2012 Megabase-scale Inversion Polymorphism in the Wild Ancestor of Maize. Genetics **191**: 883–894.

Flicek, P., M. R. Amode, D. Barrell, K. Beal, S. Brent, *et al.*, 2012 Ensembl 2012. Nucleic Acids Research **40**: D84–90.

Fournier-Level, A., A. Korte, M. D. Cooper, M. Nordborg, J. Schmitt, *et al.*, 2011 A map of local adaptation in *Arabidopsis thaliana*. Science **334**: 86–9.

Frichot, E., S. Schoville, G. Bouchard, and O. François, 2012 Landscape genomic tests for associations between loci and environmental gradients .

Fumagalli, M., M. Sironi, U. Pozzoli, A. Ferrer-Admettla, L. Pattini, *et al.*, 2011 Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. PLoS Genetics **7**: e1002355.

Futschik, A., and C. Schlötterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics **186**: 207–18.

Gasper, J., and W. J. Swanson, 2006 Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. American Journal of Human Genetics **79**: 820–30.

Guillot, G., 2012 Detection of correlation between genotypes and environmental variables. A fast computational approach for genomewide studies .

Guillot, G., and F. Rousset, 2011 On the use of the simple and partial Mantel tests in presence of spatial auto-correlation .

Hancock, A. M., B. Brachi, N. Faure, M. W. Horton, L. B. Jarymowycz, *et al.*, 2011a Adaptation to Climate Across the *Arabidopsis thaliana* Genome. Science **334**: 83–86.

Hancock, A. M., V. J. Clark, Y. Qian, and A. Di Rienzo, 2011b Population genetic analysis of the uncoupling proteins supports a role for *UCP3* in human cold resistance. Molecular Biology and Evolution **28**: 601–14.

Hancock, A. M., D. B. Witonsky, G. Alkorta-Aranburu, C. M. Beall, A. Gebremedhin, *et al.*, 2011c Adaptations to Climate-Mediated Selective Pressures in Humans. PLoS Genetics **7**: e1001375.

Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, *et al.*, 2010 Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proceedings of the National Academy of Sciences of the United States of America **107 Suppl**: 8924–30.

Hancock, A. M., D. B. Witonsky, A. S. Gordon, G. Eshel, J. K. Pritchard, *et al.*, 2008 Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genetics **4**: e32.

He, Z., W. Zhai, H. Wen, T. Tang, Y. Wang, *et al.*, 2011 Two Evolutionary Histories in the Genome of Rice: the Roles of Domestication Genes. PLoS Genetics **7**: e1002100.

Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton, *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. Science **331**: 920–4.

Hinohara, K., T. Nakajima, M. Yasunami, S. Houda, T. Sasaoka, *et al.*, 2009 Megakaryoblastic leukemia factor-1 gene in the susceptibility to coronary artery disease. Human Genetics **126**: 539–47.

Hoffmann, A. A., and A. R. Weeks, 2007 Climatic selection on genes and traits after a 100 year-old invasion: a critical look at the temperate-tropical clines in *Drosophila melanogaster* from eastern Australia. Genetica **129**: 133–47.

Huxley, J. S., 1939 Clines: an auxiliary method in taxonomy. Bijdr. Dierk **27**: 491—-520.

Jablonski, N. G., 2004 the Evolution of Human Skin and Skin Color. Annual Review of Anthropology **33**: 585–623.

Jones, F. C., Y. F. Chan, J. Schmutz, J. Grimwood, S. D. Brady, *et al.*, 2011 A Genome-wide SNP Genotyping Array Reveals Patterns of Global and Repeated Species-Pair Divergence in Sticklebacks. Current Biology **22**: 83–90.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-Y. Kong, *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nature Genetics **42**: 348–54.

Kang, H. M., N. a. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, *et al.*, 2008 Efficient control of population structure in model organism association mapping. Genetics **178**: 1709–23.

Keller, S. R., N. Levsen, M. S. Olson, and P. Tiffin, 2012 Local Adaptation in the Flowering-Time Gene Network of Balsam Poplar, *Populus balsamifera L.* Molecular Biology and evolution .

Kofler, R., A. J. Betancourt, and C. Schlötterer, 2012 Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in *Drosophila melanogaster*. PLoS Genetics **8**: e1002487.

Kolaczkowski, B., A. D. Kern, A. K. Holloway, and D. J. Begun, 2011 Genomic differentiation between temperate and tropical Australian populations of *Drosophila melanogaster*. Genetics **187**: 245–60.

Kring, S. I. I., L. H. Larsen, C. Holst, S. r. Toubro, T. Hansen, *et al.*, 2008 Genotype-phenotype associations in obesity dependent on definition of the obesity phenotype. Obesity Facts **1**: 138–45.

Lamason, R. L., M.-A. P. K. Mohideen, J. R. Mest, A. C. Wong, H. L. Norton, *et al.*, 2005 *SLC24A5*, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science **310**: 1782–6.

Lewontin, R. C., and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics **74**: 175–95.

Li, J. Z., D. M. Absher, H. Tang, A. M. Southwick, A. M. Casto, *et al.*, 2008 Worldwide human relationships inferred from genome-wide patterns of variation. Science **319**: 1100–4.

Limborg, M. T., S. M. Blankenship, S. F. Young, F. M. Utter, L. W. Seeb, *et al.*, 2012 Signatures of natural selection among lineages and habitats in *Oncorhynchus mykiss*. Ecology and Evolution **2**: 1–18.

Lindgren, F., H. v. Rue, and J. Lindström, 2011 An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **73**: 423–498.

Long, Q., D. C. Jeffares, Q. Zhang, K. Ye, V. Nizhynska, *et al.*, 2011 PoolHap: Inferring Haplotype Frequencies from Pooled Samples by Next Generation Sequencing. PLoS ONE **6**: e15292.

Ma, Z., S. W. Morris, V. Valentine, M. Li, J. A. Herbrick, *et al.*, 2001 Fusion of two novel genes, *RBM15* and *MKL1*, in the t(1;22)(p13;q13) of acute megakaryoblastic leukemia. Nature Genetics **28**: 220–1.

Maglott, D., J. Ostell, K. D. Pruitt, and T. Tatusova, 2011 Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research **39**: D52–7.

Marsh, D. J., D. T. Weingarth, D. E. Novi, H. Y. Chen, M. E. Trumbauer, *et al.*, 2002 Melanin-concentrating hormone 1 receptor-deficient mice are lean, hyperactive, and hyperphagic and have altered metabolism. Proceedings of the National Academy of Sciences of the United States of America **99**: 3240–5.

Nakayama, K., S. Fukamachi, H. Kimura, Y. Koda, A. Soemantri, *et al.*, 2002 Distinctive distribution of AIM1 polymorphism among major human populations with different skin color. Journal of Human Genetics **47**: 92–4.

Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson, *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society: Series B **64**: 695–715.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song, 2011 Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics **12**: 443–451.

Ohashi, J., I. Naka, and N. Tsuchiya, 2011 The impact of natural selection on an *ABCC11* SNP determining earwax type. Molecular Biology and Evolution **28**: 849–57.

Orozco-Terwengel, P., M. Kapun, V. Nolte, R. Kofler, T. Flatt, *et al.*, 2012 Adaptation of *Drosophila* to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. Molecular Ecology .

Parmacek, M. S., 2007 Myocardin-related transcription factors: critical coactivators regulating cardiovascular development and adaptation. Circulation Research **100**: 633–44.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, *et al.*, 2009 Signals of recent positive selection in a worldwide sample of human populations. Genome Research **19**: 826–37.

Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data .

Price, A. L., A. Helgason, S. Palsson, H. Stefansson, D. St Clair, *et al.*, 2009 The impact of divergence time on the nature of population structure: an example from Iceland. PLoS Genetics **5**: e1000505.

Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology **20**: R208–15.

Pyhäjärvi, T., M. B. Hufford, S. Mezmouk, and J. Ross-Ibarra, 2012 Complex patterns of local adaptation in teosinte .

R Development Core Team, 2011 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ralph, P., and G. Coop, 2010 Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics **186**: 647–68.

Robertson, A., 1975 Gene frequency distributions as a test of selective neutrality. Genetics **81**: 775–785.

Rosenberg, N. A., J. K. Pritchard, J. L. Weber, H. M. Cann, K. K. Kidd, *et al.*, 2002 Genetic structure of human populations. Science **298**: 2381–5.

Rue, H. v., S. Martino, and N. Chopin, 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **71**: 319–392.

Rutanen, J., J. Pihlajamäki, M. Vänttinen, U. Salmenniemi, E. Ruotsalainen, *et al.*, 2007 Single nucleotide polymorphisms of the *MCHR1* gene do not affect metabolism in humans. Obesity **15**: 2902–7.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449**: 913–918.

Samanta, S., Y.-J. Li, and B. S. Weir, 2009 Drawing inferences about the coancestry coefficient. Theoretical Population Biology **75**: 312–9.

Scharenberg, M. A., R. Chiquet-Ehrismann, and M. B. Asparuhova, 2010 Megakaryoblastic leukemia protein-1 (*MKL1*): Increasing evidence for an involvement in cancer progression and metastasis. The International Journal of Biochemistry & Cell Biology **42**: 1911–4.

Scheinfeldt, L. B., S. Soi, S. Thompson, A. Ranciaro, D. Woldemeskel, *et al.*, 2012 Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biology **13**: R1.

Simonson, T. S., Y. Yang, C. D. Huff, H. Yun, G. Qin, *et al.*, 2010 Genetic evidence for high-altitude adaptation in Tibet. Science **329**: 72–5.

Speliotes, E. K., C. J. Willer, S. I. Berndt, K. L. Monda, G. Thorleifsson, *et al.*, 2010 Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nature Genetics **42**: 937–48.

Stinchcombe, J. R., C. Weinig, M. Ungerer, K. M. Olsen, C. Mays, *et al.*, 2004 A latitudinal cline in flowering time in *Arabidopsis thaliana* modulated by the flowering time gene *FRIGIDA*. Proceedings of the National Academy of Sciences of the United States of America **101**: 4712–7.

Turner, T. L., E. C. Bourne, E. J. Von Wettberg, T. T. Hu, and S. V. Nuzhdin, 2010 Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. Nature Genetics **42**: 260–3.

Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. PLoS Genetics **7**: e1001336.

Weir, B. S., and W. G. Hill, 2002 Estimating F-statistics. Annual Review of Genetics **36**: 721–50.

Wermter, A.-K., K. Reichwald, T. Büch, F. Geller, C. Platzer, *et al.*, 2005 Mutation analysis of the *MCHR1* gene in human obesity. European Journal of Endocrinology **152**: 851–62.

Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo, *et al.*, 2010 Sequencing of 50 human exomes reveals adaptation to high altitude. Science **329**: 75–8.

Yu, J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics **38**: 203–208.

Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nature Genetics **44**: 821–4.

Zhu, Y., A. O. Bergland, J. González, and D. A. Petrov, 2012 Empirical Validation of Pooled Whole Genome Population Re-Sequencing in *Drosophila melanogaster*. PLoS ONE **7**: e41901.