Robust Identification of Local Adaptation from Allele Frequencies

Torsten Günther*,[1] and Graham Coop†,[1]
*Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany, and †Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, California 95616

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 detect nonparametric correlations with environmental variables; these correlations 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} and is shown to be more powerful, as we account for population history. We also extend the model to next-generation sequencing of population pools—a cost-efficient way to estimate population allele frequencies, but one that introduces an additional level of sampling noise. The utility of these methods is demonstrated in simulations and by reanalyzing human SNP data from the Human Genome Diversity Panel populations and pooled next-generation sequencing data from Atlantic herring. An implementation of our method is available from http://gcbias.org.

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 (Mayr 1942), with diverse examples including latitudinal clines in 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). 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; Endler 1986) 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 to understand 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 in 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, population allele frequencies among multiple populations are not statistically independent, as populations vary in their relationship to one another due to varying amounts of shared genetic drift and migration over time. Failure to account for differences in sample size and the shared history of populations can lead to a high rate of

doi: 10.1534/genetics.113.152462
Manuscript received April 23, 2013; accepted for publication June 17, 2013
Supporting information is available online at http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152462/-/DC1.
[1]Corresponding authors: Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. E-mail: tguenther@zoho.com; and Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, California 95616. E-mail: gmcoop@ucdavis.edu

false positives and negatives when searching for the signal of local adaptation due to the unaccounted sources of variance and nonindependence among populations (Robertson 1975; Excoffier *et al.* 2009; Bonhomme *et al.* 2010). 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 (Hancock *et al.* 2008; Coop *et al.* 2010); Bayenv 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, an arbitrary 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 that 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. 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, 2011b,c; 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).

A range of other model-based methods have been developed in parallel to detect environmental correlations. One of the earliest of these was the method of Joost *et al.* (2007), which accounted for differences in sample size but did not account for population structure. Building on this, Poncet *et al.* (2010) developed a fast method based on generalized estimation equations, to allow population structure for correlated sample frequencies within a set of population clusters (but no relatedness between these clusters). A recent simulation study (De Mita *et al.* 2013) has shown that environmental correlation methods that explicitly account for population structure (*e.g.*, Coop *et al.* 2010; Poncet *et al.* 2010) outperform those that do not account for structure, with Bayenv having somewhat higher power and greater robustness to population structure, likely due to it allowing arbitrary relatedness between populations (at the expense of computational speed).

Recently, Guillot (2012) developed an inference method very similar to Bayenv, offering large gains in computational efficiency. These gains are at the expense of constraining the covariance matrix in an explicit isolation-by-distance parametric form. In addition, Frichot *et al.* (2013) presented a latent factor mixed model that estimates the effect of population history and environmental correlations simultaneously. The Frichot *et al.* (2013) method resulted in a slightly higher power than that of 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.

These methods (Coop *et al.* 2010; Poncet *et al.* 2010; Guillot 2012; Frichot *et al.* 2013) all seem to have similar performance in terms of power, suggesting that the detection of linear correlations in the presence of population structure has reached a reasonable level of development. However, further work is needed to extend the utility and robustness of these methods to enhance their application to population genomic data.

One concern about applications of these 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 nonparametric statistics, such as Spearman's ρ, which may also offer increased power to robustly detect linear and nonlinear relationships. The difficulty is that such tests do not acknowledge the differences in sample size or the covariance in allele frequencies across populations. To address this some authors (Fumagalli *et al.* 2011; Hancock *et al.* 2011a) have used a nonparametric partial Mantel test, which can partially account for this covariance and makes fewer model assumptions. However, the partial Mantel approach is known to perform very poorly when both genotypes and environmental variables are spatially autocorrelated (see Guillot and Rousset 2013, for discussion). As such, we currently lack a framework to perform robust, nonparametric inference of environmental correlations in population genomic data.

To overcome these difficulties we use the framework laid out in Bayenv to provide the user with a set of "standardized" allele frequencies at each SNP. In these standardized allele frequencies 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 application of this we show how these standardized allele frequencies can be used to develop a powerful test that robustly infers environmental correlations in the face of outlier populations. As a further example of how these "standardized allele frequencies" can be used, we construct a global F_{ST}-like statistic, which accounts for shared population history and sampling noise, to identify loci with unusually high allele frequency variance among populations (a test that is closely related to that of Bonhomme *et al.* 2010). We demonstrate the utility of these approaches through simulation and reanalyzing SNP genotyping data from the Centre d'Etude du Polymorphisme Hmain (CEPH) Human Genome Diversity Panel (HGDP) (Conrad *et al.* 2006; Li *et al.* 2008).

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; Lamichhaney *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. We extend the model of Bayenv to account for the sampling of reads in pooled NGS experiments. We show that this improves the power to detect environmental correlations in pooled resequencing data. We also show the utility of our approach by the reanalysis of a pooled next-generation sequence data set from Atlantic herring (*Clupea harengus*) populations along a salinity gradient (Lamichhaney *et al.* 2012).

The extensions to Bayenv detailed here are schematically depicted in Figure 1, and all of these extensions are implemented in Bayenv2.0, available from http://gcbias.org.

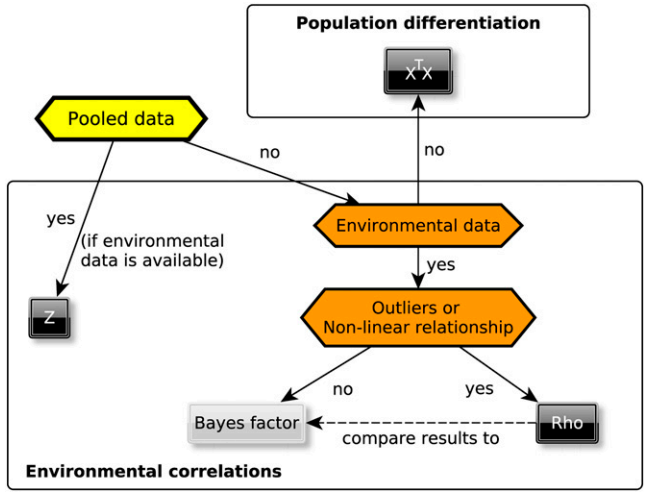

Figure 1 Overview of all statistics implemented in Bayenv2.0 and a decision-making support on which to calculate in what situations the statistics for environmental correlations require information about environmental variables. X^TX can be calculated for data with and without environmental information to detect differentiation patterns caused by different environmental factors. Black boxes represent statistics introduced with Bayenv2.0 while gray shows statistics already calculated by Bayenv1.0.

Methods

General model of Bayenv

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

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

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

$$p_{jl}=g\left(\theta_{jl}\right)=\begin{cases}0 & \text{if } \theta_{jl}<0\\ \theta_{jl} & 0\le\theta_{jl}\le 1\\ 1 & \theta_{jl}>1,\end{cases} \tag{2}$$

i.e., that the masses <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

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

where Ω is the variance–covariance matrix of allele frequencies among populations. We can write the joint probability of our counts at a locus and the θ_l as

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

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 $\hat{\Omega}$, 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) use 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:

$$P\left(\theta_l|\hat{\Omega},\epsilon_l,\beta,Y\right)\sim \mathrm{MVN}\left(\epsilon_l+\beta Y,\epsilon_l(1-\epsilon_l)\hat{\Omega}\right). \quad (6)$$

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 $\hat{\Omega}$, integrating out the uncertainty in θ_l, ϵ_l, and β (under a uniform prior on β between -0.2 and 0.2).

Tests based on standardized allele frequencies

The linear relationship between the transformed allele frequencies (Equation 6) may not be the best fit in all situations, as other monotonic relationships 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 nonparametric 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 nonindependent 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

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

If $\theta_l \sim \mathrm{MVN}(\epsilon_l, \epsilon_l(1-\epsilon_l)\Omega)$, then $X_l \sim \mathrm{MNV}(0, \mathbb{I})$ 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, as a range of other factorizations of the covariance matrix are possible.

If we wish to test the correlation of our transformed allele frequencies with an environmental variable, we 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 0, variance 1), then our transformed environmental variable is

$$Y' = C^{-1}Y. \quad (8)$$

The fact that we need to do this follows from the fact that we are interested in predicting θ_l by Y, and so what we are doing is conceptually equivalent to multiplying the left- and right-hand sides of the equation relating the two by C^{-1} to remove the effect of the covariance in allele frequencies. This rotation by C^{-1} means that X and Y' are linear superpositions of θ and Y, respectively, and so the population labels no longer apply to particular entries in these vectors. The transform will tend to exaggerate the environmental variable differences between very closely related populations. Furthermore, if part of the variation in the environmental variable precisely matches the major 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 Supporting Information, Figure S1, Figure S2, Figure S3, and Figure S4 for examples).

We do not get to observe θ_l, so we obtain a representative sample of M draws from the posterior $(X_l^{(1)}, \ldots, X_l^{(M)})$. 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

$$\rho_l\left(X_l^{(1)},\ldots,X_l^{(M)}\right)=\frac{1}{M}\sum_{i=1}^{M}\rho\left(X_l^{(i)},Y'\right); \quad (9)$$

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

In this article, 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 computationally 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 $\hat{\Omega}$, 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 timescales. Thus, while we model drift at a locus as being multivariate normal (*i.e.*, θ_l has a prior given by Equation 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.

A robust statistic to detect extreme population differentiation: This transformation to provide a set of X_l can also be used to propose an F_{ST}-like statistic. Specifically we can write the variance of X_l at a locus as

$$\mathrm{Var}(X_l) = X_l^T X_l = \frac{(\theta_l - \epsilon)^T \Omega^{-1} (\theta_l - \epsilon)}{\epsilon_l (1 - \epsilon_l)}. \tag{10}$$

Note that while the multiplication by C^{-1} is not a unique transform, $X_l^T X_l$ would be the same no matter which matrix factorization was used. This statistic is closely related to F_{ST}, which can be expressed as

$$\frac{\mathrm{Var}(p_{lj})}{(\epsilon_l (1 - \epsilon_l))}, \tag{11}$$

where $\mathrm{Var}(p_{lj})$ is the variance of our allele frequency across populations (see Nicholson *et al.* 2002; Balding 2003, for discussion). Thus $X^T X$ is directly analogous to F_{ST}, but it accounts for the variance–covariance structure of the populations in question. If θ_l is truly multivariate normal, then $X_l^T X_l$ is distributed $\sim \chi_J^2$. 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). As we do not observe X_l and rather obtained a posterior distribution of X_l, we take the average $X_l^T X_l$ across the sample of X_l from our MCMC (we term this $X^T X$).

Our $X^T X$ statistic is very closely related to the work of Bonhomme *et al.* (2010), who independently developed a similar test statistic (extended Lewontin and Krakauer test, FLK) for the case of a known population tree (see also Nei and Maruyama 1975; Robertson 1975, for earlier discussion of the effect of a population tree on the Lewontin and Krakauer 1973 test). Specifically, Bonhomme *et al.* (2010) use an equation analogous to Equation 10, but where Ω is a covariance matrix specified by a neighbor-joining population tree estimated from the data [using Reynolds' genetic distance between population pairs (Reynolds *et al.* 1983)]. Thus we expect similar performance of $X^T X$ to that of FLK in situations where the covariance matrix is well approximated by a tree and $X^T X$ to outperform FLK under models such as isolation by distance and more complex models where population history is poorly represented by a tree.

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 in population j is m_{jl}, and we observe r_{jl} reads of allele 1. Assuming that each individual contributed the same number of chromosomes to the pool, the sequenced reads are the result of binomial sampling

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

where i/n_j is the unknown sample allele frequency in the pooled sample. Summing over this unknown frequency

$$P\left(r_{jl} \middle| m_{jl}, p_{jl}, n_j\right) = \binom{m_{jl}}{r_{jl}} \sum_i \left(\frac{i}{n_j}\right)^{r_j} \left(1 - \frac{i}{n_j}\right)^{m_{jl} - r_{jl}} \binom{n_j}{i} p_{jl}^i \left(1 - p_{jl}\right)^{n_j - i} \tag{13}$$

(again assuming Hardy–Weinberg equilibrium) gives us the probability of our sampled reads given the population frequency. This replaces the binomial probability (Equation 1) in the joint probability given by Equation 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 (between -0.1 and 0.1) and update β along with ϵ_l and θ_l. For our update on β we use a random-walk sampler with a small normal deviate ($\sigma = 0.01$) and accept this move with the ratio of the joint posterior of our proposed parameters to that of our current 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 $\beta \neq 0$.

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; Li *et al.* 2008) and the environmental variables measured at these sampling locations (also used in Hancock *et al.* 2008; Coop *et al.* 2010). These simulations were implemented in R (R Development Core Team 2011). We employed a single Bayenv2.0 estimate of the covariance matrix $\hat{\Omega}$ 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 $\alpha = 0.5$, $\beta = 3$) if not otherwise stated. Then population allele frequencies were drawn from the multivariate normal $\mathrm{MVN}(\epsilon_i, \epsilon_i(1 - \epsilon_i)\hat{\Omega})$, with values >1 or <0 replaced with 1 and 0, respectively. Allele counts were

then drawn binomially from these frequencies, with a sample size of 20 chromosomes per population (if not otherwise stated).

To construct a null distribution we calculated our test statistics 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$(\epsilon_i + \beta Y, \epsilon_i(1-\epsilon_i)\hat{\Omega})$, using a range of β's. Again our test statistics were calculated over 100,000 MCMC iterations. 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 the upper α tail of the null distribution. Note that by taking this empirical approach we are setting our false-positive rate to 5%, and as such we do not need to investigate our false-positive rate.

Testing differentiation statistics: To compare the power of differentiation statistics, we simulated data sets based on the HGDP populations, using the minimum winter temperature and latitude as environmental variables (latitude is obviously not an environmental variable, but is taken as a proxy for many environmental variables that are correlated with latitude). To highlight the effects of different sample sizes among populations, we simulated populations using the original sample sizes of the HGDP data (Li *et al.* 2008). Sets of neutral and selected SNPs were simulated as described above, and the differentiation statistics F_{ST} [using the pegas R library (Paradis 2010)], X^TX, and the related FLK [using an R script provided by the authors (Bonhomme *et al.* 2010)] were calculated. The ancestral frequency was used as an outgroup population for the calculation of FLK. The ancestral allele frequency ϵ was set to 0.5 for all SNPs simulated for these tests as we discovered noisy power estimates for FLK when the ancestral frequency was drawn from a beta distribution (not shown). All power estimates again used empirical cutoffs based on the neutral simulations.

Simulation of pooled samples: 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 overdispersion of read depths compared to the Poisson. Coverages for each population and SNP were independently drawn from a negative binomial distribution NB(r, p), for which we set $r = 5$ and set p to obtain the respective coverage mean [*i.e.*, NB(r, p) has a mean of $pr/(1-p)$ and a variance of $pr/(1-p)^2$]. This represents a case where the variance–mean ratio increases for higher average coverages. This pattern is generally consistent with observations from pooled next-generation data in *Arabidopsis thaliana* (T. Günther, C. Lampei, I. Barilar, and K. J. Schmid, unpublished results). 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. The sign of β was assigned to be positive or negative at random.

Data analysis

Bayenv2.0 was employed to reanalyze a genome-wide data set of 640,698 SNPs from 52 HGDP-CEPH populations (Li *et al.* 2008; Hancock *et al.* 2010), using both Bayes factors and our nonparametric test statistic (ρ_l). We restricted our analysis to minimum winter temperature, as most winter climate variables show outliers. All environmental variables were normalized to a mean of zero and a standard deviation of one. The covariance matrix for this reanalysis was estimated from a random subset of 5000 SNPs after 100,000 MCMC iterations. Bayes factors and correlation coefficients for each SNP were estimated using 100,000 MCMC iterations. In addition to these test statistics, we sampled X_l every 500 MCMC generations and averaged X^TX over these values for each SNP. SNP positions and gene annotations were obtained from Ensembl (Flicek *et al.* 2012) and Entrez Gene (Maglott *et al.* 2011).

To explore the use of Bayenv on pooled data we reanalyzed a pooled sequence data set from eight Atlantic herring (*C. harengus*) populations that were sampled along a salinity gradient (Lamichhaney *et al.* 2012). Lamichhaney *et al.* (2012) sequenced eight population pools of 50 individuals and aligned their reads back to a *de novo* assembly of the *C. harengus* exome, with each pool being sequenced to on average 30-fold depth. They identified 440,817 SNPs in these data and concentrated their analyses on SNPs with at least 40 reads in each population to reduce sampling noise. We reanalyzed the data of Lamichhaney *et al.* (2012), using allele count data provided by the authors. We first estimated the covariance matrix, using a random subset of 1000 SNPs, chosen to have ≥40-fold coverage in all populations. To explore the performance of our pooled sequencing extension of Bayenv we calculated the statistic Z for the correlation of allele frequencies and salinity across the eight populations. The environmental correlation test was calculated for all SNPs with at least 40 reads in all populations (36,794 SNPs) and for 100,000 SNPs chosen at random irrespective of sequencing depth. All statistics were calculated during 100,000 MCMC iterations.

Results

Using tests based on standardized allele frequencies

We explored the performance of tests based on our standardized transformed population frequencies (X_l). Before we calculated test statistics on our standardized allele frequencies, we examined whether the multivariate standardization (as in Equation 7) had removed the covariance among populations from our standardized X_l. We first calculated the sample covariance matrix, using the sample frequencies for all 2333 autosomal SNPs of Conrad *et al.* (2006) shown in Figure 2A. 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

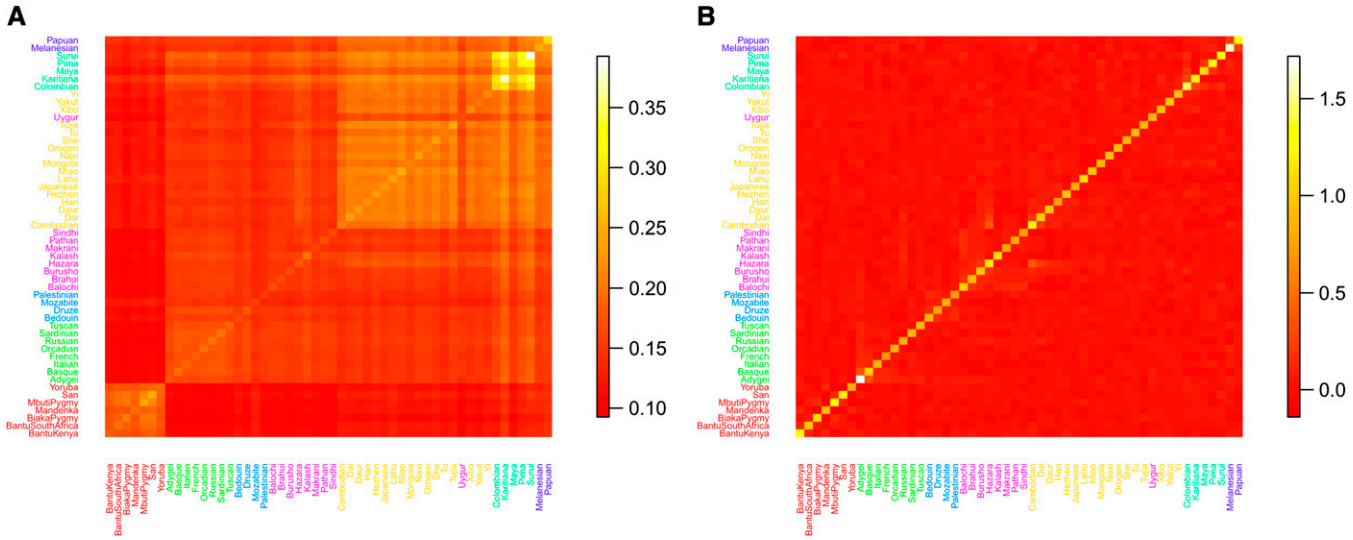

Figure 2 (A and B) Covariance among HGDP populations estimated by Bayenv2.0 (A) and the covariance calculated on the X's for the same SNPs (B) (all autosomal SNPs of Conrad *et al.* 2006). Populations are colored according to broad geographic regions used in Rosenberg *et al.* (2002).

structure in humans (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 burn-in) for each of these 2333 SNPs and calculated $\frac{1}{L}\sum_{l=1}^{L} X_l X_l^T$. The resulting sample covariance matrix (shown in Figure 2B) 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 whether the X^TX statistic follows a χ^2-distribution with 52 d.f. (*i.e.*, the number of populations). A QQ plot of the X^TX and the expected χ^2_{52}-distribution shows that the mean of the distribution approximately matches that of the χ^2_{52}, whereas the variances do not (Figure S5). Therefore, while X^TX provides a potentially suitable summary statistic for identifying empirical outliers, we cannot assume a distributional form to those outliers under a null neutral model.

Detecting extreme population differentiation: To test the utility of X^TX as a statistic to detect overly differentiated loci, we simulated adaptation of the HGDP populations to minimum winter temperature and latitude and explored different ways to detect such environmental effects with an empirical cutoff score determined by simulations of a neutral model. In addition to Bayenv's Bayes factors, we calculated the differentiation statistics F_{ST}, X^TX, and FLK (Bonhomme *et al.* 2010). The environmental correlation Bayes factors clearly outperformed the differentiation statistics for both environments (Figure 3), which is consistent with the intuition that knowing the responsible environmental factor should improve the detection power (De Mita *et al.* 2013). Our simulations show that FLK and X^TX clearly outperform the traditional F_{ST}-based tests, suggesting that accounting for the relationship among populations increases the power of differentiation statistics. Notably, FLK and X^TX have similar power with slightly higher estimates for X^TX. The similar level of power of the two tests is perhaps not unsurprising, as the history of the worldwide HGDP populations can be well represented by a tree-like relationship [with a small number of migration events (Pickrell and Pritchard 2012)].

Correlations using standardized allele frequencies: 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 calculated both Spearman's ρ and Pearson's r between Y' and our transformed allele frequencies averaged across the posterior on these transformed frequencies. Statistics based on our Bayesian model clearly outperform correlation tests calculated for point estimates from sample allele frequencies (Figure 4, A and B). 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 4B) and a slightly lower power of Spearman's ρ, which is not surprising, as all simulated effects are linear. 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', which led 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 (Equation 6) implies a linear relationship between the transformed allele frequencies and the environmental variable. However, the fitting and

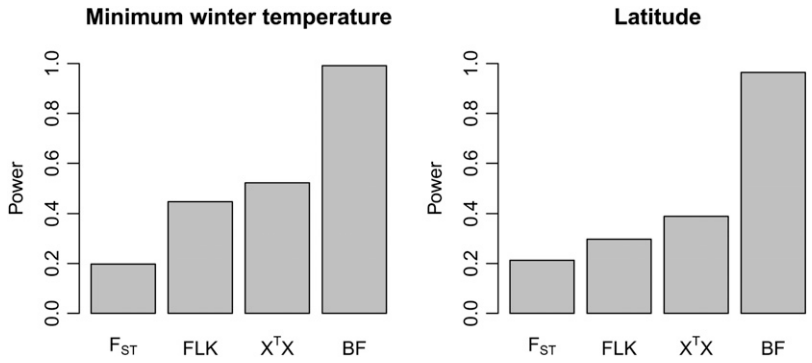

Figure 3 Power of differentiation statistics and Bayenv2.0 to detect different simulated environmental effects.

significance of this linear model may be misled by populations that are statistical outliers. For instance, linear models might mistakenly identify a SNP as a strong candidate, when a 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 of correlation with an environmental variable.

To test such a case, we used winter minimum temperature as a climate variable since one population, the Yakuts from northeast Russia, is characterized by a very low minimum temperature (Figure 4C). We simulated allele frequencies for the HGDP populations as described above but to create outliers, we set the allele frequencies of the Yakuts to 0. Bayes factors and Pearson's correlation coefficient r, which both assume a linear relationship, showed an excess of false positives (Figure 4D), while a nonparametric 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 4D).

Robust candidates in the HGDP data

We next explored the use of our standardized X_l for identifying robust putative candidates for adaptive evolution in a genome-wide data set of 640,698 SNPs from a global sample of 52 human populations [HGDP-CEPH (Li *et al.* 2008; Hancock *et al.* 2010)].

Environmental correlation statistics in the HGDP data: 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 reanalyzed all SNPs of the HGDP data, calculating both Bayes factors and $\rho_l(X_l, Y')$ using Spearman's rank correlation coefficient. 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% of 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 Table S1 and Table S2. Among our top hits for both statistics are multiple SNPs that fall in the gene *MKL1* (megakaryoblastic leukemia 1), which is a myocardin-related transcription factor that is involved in smooth muscle cell differentiation, mammary gland function, and cytoskeletal signaling (Parmacek 2007; Maglott *et al.* 2011) and has been associated with various disease phenotypes (Ma *et al.* 2001; Hinohara *et al.* 2009; Scharenberg *et al.* 2010).

To exemplify the effect of an outlier in Figure S6, we compare two SNPs that fall in our top 20 Bayes factors, but that have very different Spearman's ρ. While in general, as seen above, there is good agreement between the Bayes factors and $|\rho_l(X_l, Y')|$, 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.

Population differentiation statistics in the HGDP data: We also explored our test statistic X^TX, which is designed to highlight loci that deviate strongly from the expected pattern of population structure, by calculating it for each of the 640,698 HGDP SNPs. These X^TX statistics have been uploaded as a genome browser track to http://hgdp.uchicago.edu/. The empirical distribution is shown in Figure 5. The empirical distribution clearly differs from the expected χ^2_{52}-distribution, with 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 examine the relationship between X^TX and global F_{ST} we took per SNP values of global F_{ST} previously calculated among the population groups as colored in Figure 2 (values

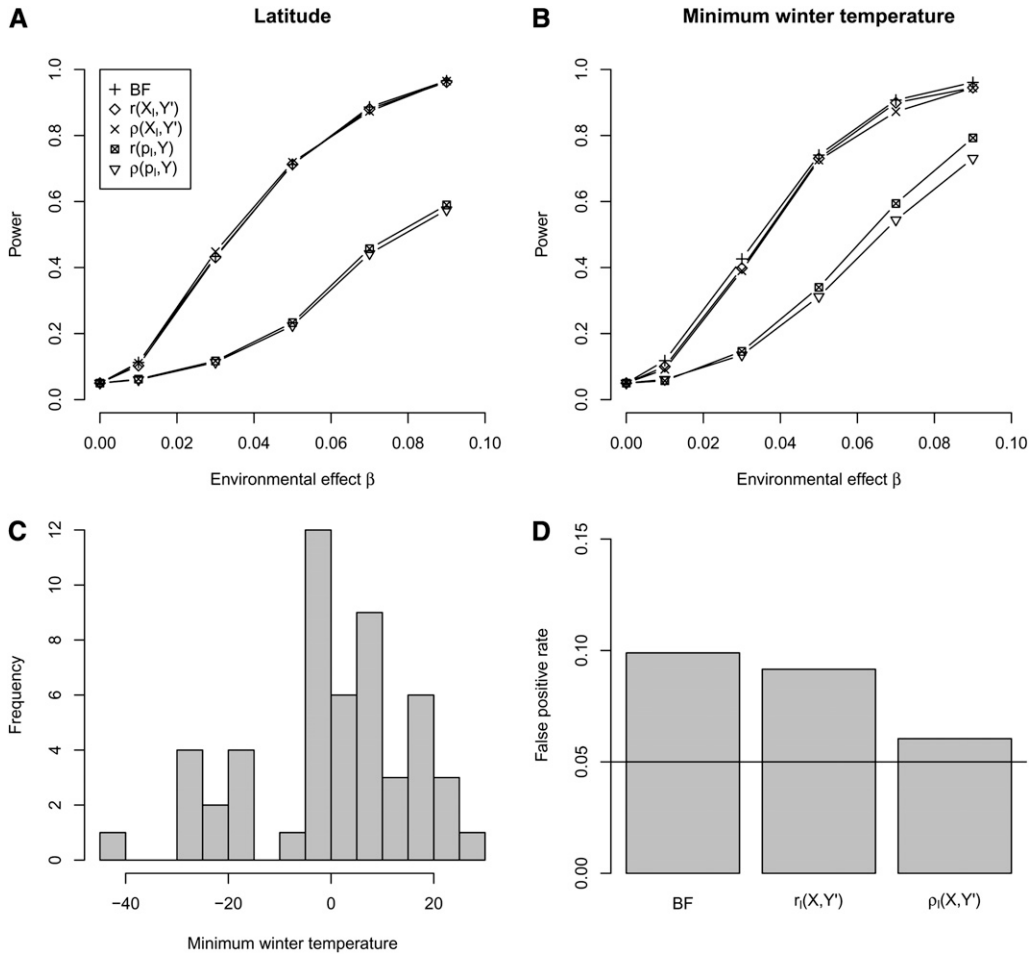

Figure 4 (A and B) Performance of Bayes factors compared to power of correlation coefficients on *X* based on simulations for all 52 HGDP populations for the environmental variables latitude (A) and minimum winter temperature (B). (C) Histogram of minimum winter temperature for the 52 HGDP populations. (D) False-positive rate of different statistics if one allele is fixed in the coldest population.

from Coop *et al.* 2009; Pickrell *et al.* 2009). The Spearman's ρ between X^TX and F_{ST} was 0.48. Looking at the extremes of both distributions, X^TX 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% of signals.

In Table S3 we present the top 100 X^TX SNPs in the genome, along with their nearest genes and global F_{ST} values. To briefly explore where known signals fall in our empirical distribution in Figure 5, we also plot as arrows the maximum X^TX for SNPs that fall within 50 kbp up- and downstream of 10 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 X^TX in the genome falls close to *SLC24A5* (Nakayama *et al.* 2002; Lamason *et al.* 2005), while a SNP close to *SLC45A2* is the second-largest signal in the genome. More generally, 5 of these 10 pigmentation genes fall in the top 1% and 9 genes fall in the top 5% of the empirical distribution. A SNP close to the gene *EDAR*, one of the highest pairwise F_{ST}'s between East Asia and Western Eurasia HGDP populations, is also in the top 10 SNPs (Sabeti *et al.* 2007).

The weak overlap in the tails of the genome-wide X^TX and F_{ST} means that they are finding different sets of candidate SNPs, presumably due to the reweighting of allele frequencies in X^TX. For example, our 12th highest SNP for X^TX falls close to *MCHR1*, with our 21st highest gene being a nonsynonymous variant (rs133072) in the same gene. *MCHR1* (melanin-concentrating hormone receptor 1) is known to play a 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 human obesity has been debated (Wermter *et al.* 2005; Rutanen *et al.* 2007; Kring *et al.* 2008; Speliotes *et al.* 2010). Both of these SNPs are nearly fixed differences between East Asia and the American HGDP populations (Figure S7 and Figure S8). This strong difference between regions that share a recent history and, thus, covariance among allele frequencies (Figure 2) makes these SNPs an interesting pattern for X^TX. However, neither of these two SNPs 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.

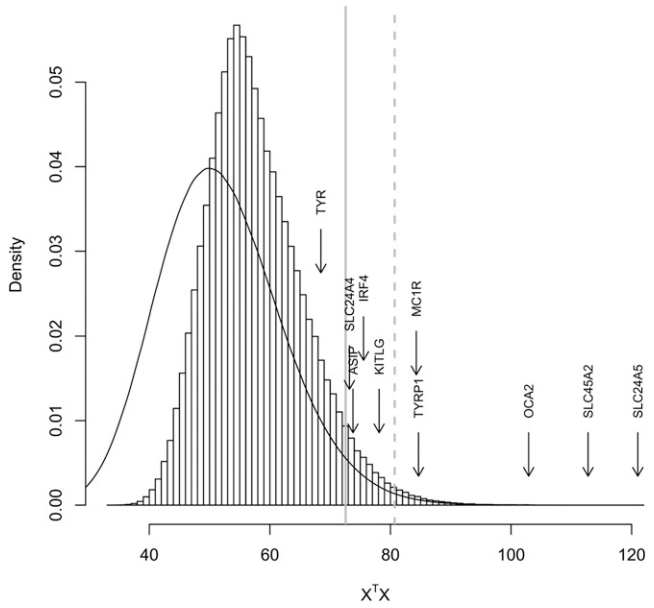

Figure 5 Histogram of X^TX calculated for all HGDP SNPs. The labels of candidate genes are shown at the maximum X^TX of any SNP within 50 kbp up- and downstream of the particular gene. The solid shaded line shows the position beyond which 5% of all X^TXs fall, and the dashed shaded line denotes the top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

Analysis 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, for a discussion of drawbacks). 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 of how Bayenv behaves for different coverages as increasing sequencing coverage is not the same as increased numbers of sampled individuals.

Power simulations: We simulated data that resemble 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. First, we 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 by failing to account for the second level of sampling.

We then ran power simulations of Bayenv matched to the HGDP data, with Z as a test statistic, using the true sample frequencies (black squares in Figure 6) and the read counts (incorrectly) as the input data for the previous version of Bayenv (Bayenv1.0, black circles in Figure 6). 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 circles in Figure 6). The true sample frequencies naturally resulted in the best power as there is no additional sampling noise (Figure 6A). 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 6A). As most applications may consist of a smaller number of populations, we additionally sampled two subsets (Figure 6, B and C). On all of these, as expected, Bayenv using the true sample frequencies outperformed 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. Including the sampling of reads into the model clearly had a positive effect on power in our population subsets, while incorrectly using the read counts as input did not reach similar powers, even with high coverages (Figure 6, B and C). However, the 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. Somewhat surprisingly, we did not observe any advantages of the extended model in detection power if all 52 HGDP populations are simulated (Figure 6A), perhaps because differences in coverage over populations are averaged over so many populations.

We note that both our original implementation and Bayenv2.0, which acknowledges the pooled nature of the data, will outperform tests of association that use the sample frequency computed from the read count data. This follows from the fact that both of these tests more fully acknowledge sampling noise. This is an important issue for next-generation data, as most technologies currently lead to high variation in coverage along the genome, which will be a substantial source of additional noise. Failure to acknowledge this noise across loci will greatly reduce the power of tests of environmental correlations.

Analysis of empirical data from Atlantic herring: Finally, we applied Bayenv2.0 to the data of Lamichhaney *et al.* (2012), which consist of pooled sequence for eight Atlantic herring populations sampled along a salinity gradient, and identified outliers for differentiation using F_{ST} and P-values of a χ^2-test on the read counts. For their F_{ST} analysis, Lamichhaney *et al.* (2012) selected a subset of their data set consisting of the 36,794 SNPs (of 440,817 SNPs in total) with at least 40 reads in each population; this sampling cutoff was implemented to reduce the impact of sampling noise on F_{ST}. While this is a sensible approach, it does discard data. Approaches, like those in Bayenv2.0, that acknowledge sampling should make better use of these data, as they can identify candidate loci in regions with lower coverage while avoiding false positives due to sampling noise.

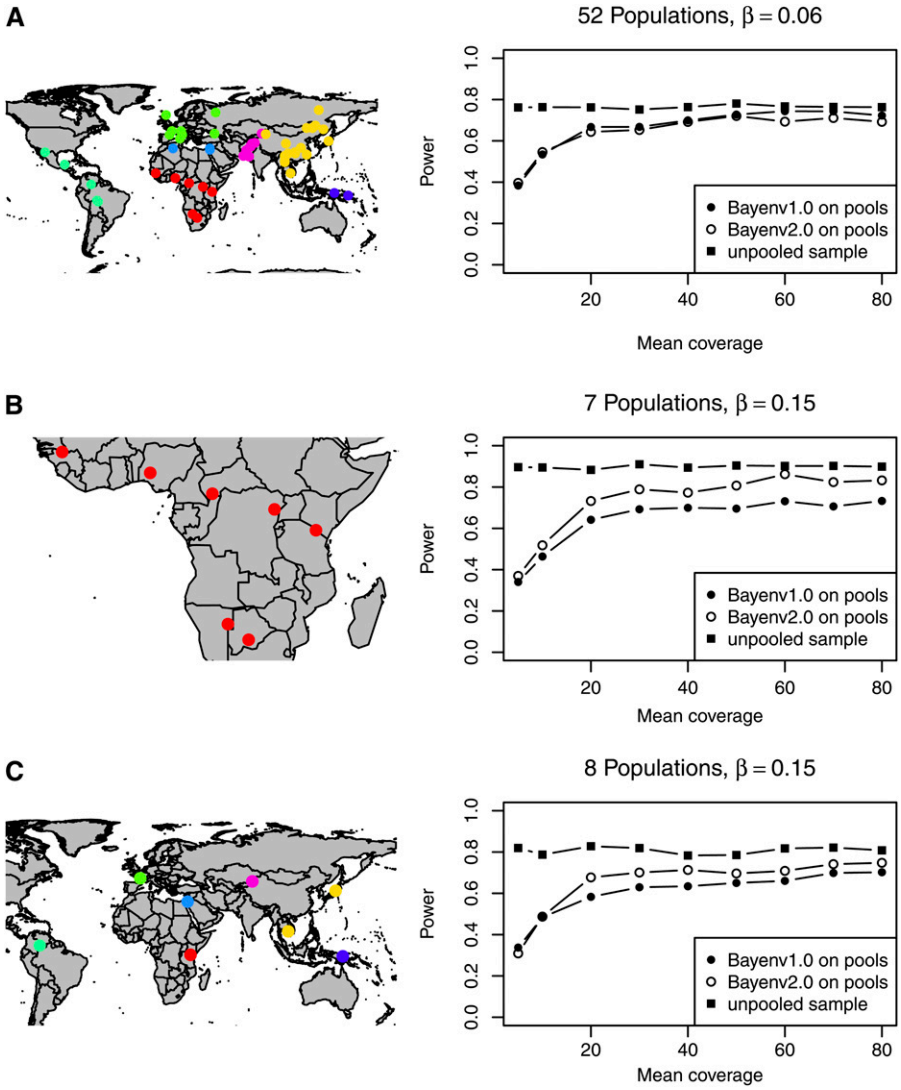

Figure 6 Power to detect environmental correlations with latitude in pooled samples: (A–C) in all 52 HGDP populations (A), in the seven sub-Saharan populations (Yoruba, San, Mbuti Pygmy, Mandenka, Biaka Pygmy, Bantu South Africa, and Bantu Kenya) (B), and in one population per broader geographic region (Bantu Kenya, French, Bedouin, Cambodian, Japanese, Uygur, Colombian, and Papuan) (C). Populations are colored as in Figure 2.

We calculated Z for the salinity gradient on the 36,794 SNPs (with coverage >40×) and for a random subset of 100,000 SNPs chosen regardless of their coverage. In addition, we calculated F_{ST} and P-values of a χ^2-test on the read counts for these data [to compare to the original study (Lamichhaney *et al.* 2012)]. To demonstrate how these three tests treat sites with different coverages, we show the mean coverage of sites in the upper 10–0.5% tail in Figure 7A. SNPs in the tail of F_{ST} show a strong bias toward loci with lower coverages, which in turn implies that many of these loci are likely false positives due to sampling error being mistaken for extreme population differentiation. On the other hand, SNPs in the tails of the χ^2-test's P-value are enriched for high-coverage positions, reflecting the fact that this significance of the χ^2-test will be greater for sites with higher coverage, even if those sites do not have particularly high levels of differentiation. Using our Z statistic, which accounts for the sampling error while looking for differentiation along an environmental gradient, we find that our outliers seem much less biased toward either extremely high or extremely low coverage. This observation, along with our higher power to detect environmental correlations than differentiation-based approaches (Figure 3 and De Mita *et al.* 2013), suggests that our Z statistic is potentially detecting more true signals of loci that are strongly differentiated along the salinity gradient while the number of false positives is reduced.

Among the contigs in which the top 200 SNPs for our Z statistic were found, only 8 were not named as candidates by Lamichhaney *et al.* (2012), suggesting good agreement with their extreme outliers. More generally, we find >70% of our top 1% Z SNPs among the top 5% of F_{ST} and χ^2 (Figure 7B). This enrichment increases with coverage, reassuringly suggesting that the approaches are in stronger agreement about outliers in regions with high coverage. The overlap is not perfect; the tails of the differentiation statistics should also include SNPs involved in adaptation to other environmental variables and outliers caused by selection in single populations. Due to the current poor annotation of herring data available in public databases, we refrain from further speculation about the biological relevance of our top hits.

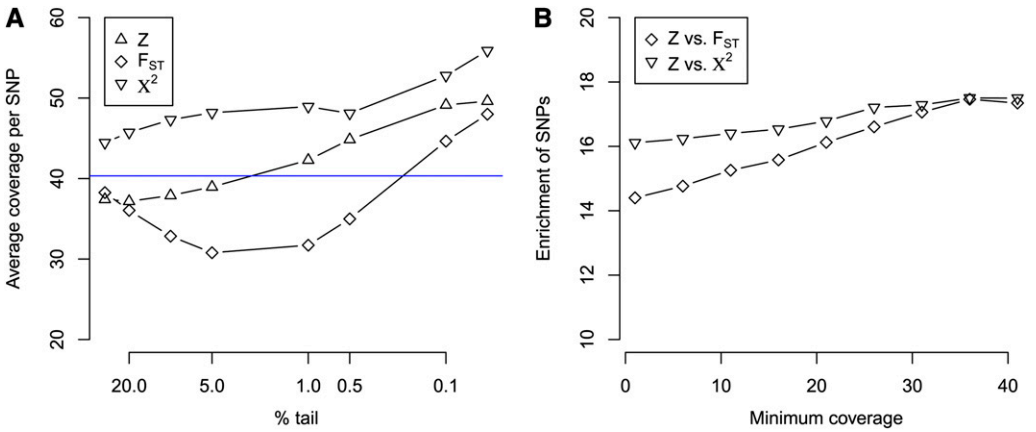

Figure 7 (A) Average sequencing coverage (per SNP across all eight populations) for different tails of the statistics' empirical distributions when using salinity as an environmental variable. The blue horizontal line denotes the mean across all SNPs. (B) Enrichment compared to expectations of the top 1% Z SNPs among the top 5% F_{ST} and χ^2 SNPs for different minimum coverages per SNP.

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 of 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, a development that will allow us to connect environmental variables to 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 an 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 confounding due to heritable variation in the phenotype elsewhere in the genome. Our model could be used to detect loci that 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 that covaried with environmental gradients, but they may be underpowered, because 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. Since the X_l are estimated using a single sample of Ω from the posterior, they are dependent on the quality of this $\hat{\Omega}$. With a large number of SNPs used to estimate the covariance matrix, MCMC runs converge to a small set of Ω's with little variance, so in practice this is not a major concern (for smaller numbers of loci, investigators could average the covariance matrix across the MCMC output).

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 that 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 $\mathrm{Var}(p_{lj})/(\epsilon_l(1-\epsilon_l))$, where $\mathrm{Var}(p_{lj})$ 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 Equation 10, and so X^TX can be seen as closely related to F_{ST} calculated 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 strongly increases our power to detect unusual allele frequencies compared to that of global F_{ST}. The fact that we remove the covariance between closely related populations also means that, unlike in F_{ST}-based methods, we do not have to arbitrarily clump populations to identify globally differentiated SNPs. While in this article 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 such cases perhaps the sample frequencies and the Cholesky decomposition of the sample covariance matrix could instead be used to mitigate the computational burden. This could also be done for environmental correlations for large numbers of samples.

Ideally our X^TX statistic would have a parametric distribution under a general null model in which 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 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 resequencing of populations is a powerful and cost-efficient way to estimate allele frequencies (Zhu *et al.* 2012; but see Cutler and Jensen 2010). The downside 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, it 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 (Cutler and Jensen 2010; Futschik and Schlötterer 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, as we do not consider quality scores or sequencing errors. Researchers 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 acknowledgment of sequencing errors and a probabilistic approach to 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 (Daub *et al.* 2013). The development of such techniques poses an important challenge for future method development.

Acknowledgments

We thank Jeremy Berg, Gideon Bradburd, Yaniv Brandvain, Fabian Freund, Chuck Langley, Joe Pickrell, Jonathan Pritchard, Peter Ralph, Jeffrey Ross-Ibarra, Karl Schmid, and Alisa Sedghifar for helpful discussions and comments on earlier versions of the manuscript. We acknowledge the various comments we received from readers of the preprint of the manuscript available on the arXiv. The X^TX estimates for the HGDP data were kindly made available through the HGDP selection browser (hgdp.uchicago.edu/) by Joe Pickrell. We also thank Alvaro Martinez Barrio and Leif Andersson for sharing the herring data. T.G. was supported by the German Federal Ministry for Education and Research (Synbreed, 0315528D) and by a Volkswagen Foundation scholarship (I/84225), affording him a visit to the University of California, Davis. G.C. was supported by a Sloan Foundation fellowship.

Literature Cited

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. Proc. Natl. Acad. Sci. USA 107: 1160–1165.

Balding, D. J., 2003 Likelihood-based inference for genetic correlation coefficients. Theor. Popul. Biol. 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. Ecol. Lett. 15: 769–777.

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. Proc. Natl. Acad. Sci. USA 107: 11459–11464.

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. Am. J. Hum. Genet. 89: 368–381.

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 Genet. 6: e1001116.

Boitard, S., C. Schlötterer, V. Nolte, R. V. Pandey, and A. Futschik, 2012 Detecting selective sweeps from pooled next-generation sequencing samples. Mol. Biol. Evol. 29: 2177–2186.

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 Biol. 8: e1000451.

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

Cavalli-Sforza, L. L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. Ser. B 164:362–379.

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. Nat. Genet. 38: 1251–1260.

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li *et al.*, 2009 The role of geography in human adaptation. PLoS Genet. 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: 1411–1423.

Cutler, D. J., and J. D. Jensen, 2010 To pool, or not to pool? Genetics 186: 41–43.

Daub, J. T., T. Hofer, E. Cutivet, I. Dupanloup, L. Quintana-Murci *et al.*, 2013 Evidence for polygenic adaptation to pathogens in the human genome. Mol. Biol. Evol. 30: 1544–1558.

De Mita, S., A.-C. Thuillet, L. Gay, N. Ahmadi, S. Manel *et al.*, 2013 Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol. Ecol. 22: 1383–1399.

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). Mol. Ecol. 19: 3789–3805.

Endler, J. A., 1986 *Natural Selection in the Wild*. Princeton University Press, Princeton, NJ.

Excoffier, L., T. Hofer, and M. Foll, 2009 Detecting loci under selection in a hierarchically structured population. Heredity 103: 285–298.

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. Mol. Ecol. 21: 4748–4769.

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 Res. 40: D84–D90.
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: 86–89.
Frichot, E., S. D. Schoville, G. Bouchard, and O. François, 2013 Testing for associations between loci and environmental gradients using latent factor mixed models. Mol. Biol. Evol. 30: 1687–1699.
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 Genet. 7: e1002355.
Futschik, A., and C. Schlötterer, 2010 The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186: 207–218.
Guillot, G., 2012 Detection of correlation between genotypes and environmental variables. A fast computational approach for genomewide studies. arXiv: 1206.0889
Guillot, G., and F. Rousset, 2013 Dismantling the Mantel tests. Methods Ecol. Evol. 4: 336–344.
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 Genet. 4: e32.
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. Proc. Natl. Acad. Sci. USA 107(Suppl): 8924–8930.
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. Mol. Biol. Evol. 28: 601–614.
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 Genet. 7: e1001375.
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 Genet. 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: 920–924.
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. Hum. Genet. 126: 539–547.
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: 133–147.
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. Annu. Rev. Anthropol. 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. Curr. Biol. 22: 83–90.
Joost, S., A. Bonin, M. W. Bruford, L. Després, C. Conord *et al.*, 2007 A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol. Ecol. 16: 3955–3969.
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: 1709–1723.
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. Nat. Genet. 42: 348–354.
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. Mol. Biol. Evol. 29: 3143–3152.
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 Genet. 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: 245–260.
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: 138–145.
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: 1782–1786.
Lamichhaney, S., A. Martinez Barrio, N. Rafati, G. Sundström, C.-J. Rubin *et al.*, 2012 Population-scale sequencing reveals genetic differentiation due to local adaptation in Atlantic herring. Proc. Natl. Acad. Sci. USA 109: 19345–19350.
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: 175–195.
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: 1100–1104.
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. Ecol. Evol. 2: 1–18.
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. Nat. Genet. 28: 220–221.
Maglott, D., J. Ostell, K. D. Pruitt, and T. Tatusova, 2011 Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res. 39: D52–D57.
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. Proc. Natl. Acad. Sci. USA 99: 3240–3245.
Mayr, E., 1942 *Systematics and the Origin of Species*, *from the Viewpoint of a Zoologist*. Harvard University Press, Cambridge, MA.
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. J. Hum. Genet. 47: 92–94.
Nei, M., and T. Maruyama, 1975 Letters to the editors: Lewontin-Krakauer test for neutral genes. Genetics 80: 395.
Nicholson, G., A. V. Smith, F. Jonsson, O. Gustafsson, K. Stefansson *et al.*, 2002 Assessing population differentiation and isolation from single-nucleotide polymorphism data. J. R. Stat. Soc. 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. Nat. Rev. Genet. 12: 443–451.
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. Mol. Ecol. 21: 4931–4941.
Paradis, E., 2010 pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419–420.
Parmacek, M. S., 2007 Myocardin-related transcription factors: critical coactivators regulating cardiovascular development and adaptation. Circ. Res. 100: 633–644.
Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8: e1002967.
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 Res. 19: 826–837.
Poncet, B. N., D. Herrmann, F. Gugerli, P. Taberlet, R. Holderegger *et al.*, 2010 Tracking genes of ecological relevance using a genome scan in two independent regional population samples of Arabis alpina. Mol. Ecol. 19: 2896–2907.
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 Genet. 5: e1000505.
Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. 20: R208–R215.
Pyhäjärvi, T., M. B. Hufford, S. Mezmouk, and J. Ross-Ibarra, 2012 Complex patterns of local adaptation in teosinte. arXiv:1208.0634.
R Development Core Team, 2011 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.
Ralph, P., and G. Coop, 2010 Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics 186: 647–668.
Reynolds, J., B. S. Weir, and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics 105: 767–779.
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: 2381–2385.
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: 2902–2907.
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. Theor. Popul. Biol. 75: 312–319.
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. Int. J. Biochem. Cell Biol. 42: 1911–1914.
Scheinfeldt, L. B., S. Soi, S. Thompson, A. Ranciaro, D. Woldemeskel *et al.*, 2012 Genetic adaptation to high altitude in the Ethiopian highlands. Genome Biol. 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: 72–75.
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. Nat. Genet. 42: 937–948.
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. Proc. Natl. Acad. Sci. USA 101: 4712–4717.
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. Nat. Genet. 42: 260–263.
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 Genet. 7: e1001336.
Weir, B. S., and W. G. Hill, 2002 Estimating F-statistics. Annu. Rev. Genet. 36: 721–750.
Wermter, A.-K., K. Reichwald, T. Büch, F. Geller, C. Platzer *et al.*, 2005 Mutation analysis of the MCHR1 gene in human obesity. Eur. J. Endocrinol. 152: 851–862.
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: 75–78.
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. Nat. Genet. 38: 203–208.
Zhou, X., and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44: 821–824.
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.

Communicating editor: M. A. Beaumont

GENETICS

Supporting Information
http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.113.152462/-/DC1

Robust Identification of Local Adaptation from Allele Frequencies

Torsten Günther and Graham Coop

DOI: 10.1534/genetics.113.152462

Table S1: Top 100 Bayes factors for minimum winter temperature

Rank	SNP ID	BF	ρ_l	Genes within ± 50kbp
1	rs17002034	16734000.00	-0.48****	MKL1, COX6B1P3, CTA-229A8.5
2	rs3827382	777660.00	-0.49****	RP4-591N18.2, AL031594.1, MKL1
3	rs2075106	585830.00	-0.29**	RP5-1091E12.1, EGFR
4	rs915624	264470.00	-0.43****	
5	rs6500380	257850.00	-0.36***	UBA52P8, AC141846.4-001, LONP2, SIAH1
6	rs10503918	255690.00	0.34***	NRG1-IT3, NRG1
7	rs7870404	252090.00	0.43****	ATP5J2P3, RP11-490H9.1
8	rs103294	241970.00	-0.34***	RPS9, LILRB5, LILRB2, LILRA3, LILRA5, AC010492.4, AC098789.1, AC010492.5, AC010518.3, AC008984.5, AC008984.7, VN1R104P, AC008984.6, MIR4752, U6, LILRA4, AC008984.2
9	rs3816186	221650.00	-0.39****	MTA3, AC074375.1, U6
10	rs9306345	194310.00	-0.36***	SGSM3, RP5-1042K10.13, RP5-1042K10.10, RP5-1042K10.12, MKL1
11	rs2104012	178340.00	-0.31**	RP11-199O14.1
12	rs7789059	177590.00	0.26**	CNTNAP2
13	rs6942733	128120.00	-0.39****	POP7, EPO, ZAN, EPHB4
14	rs2326794	113470.00	-0.40****	RP11-73O6.4, LAMA2
15	rs8035855	110760.00	0.34***	RP11-107F6.3, RP11-23P13.4, MGA, MAPKBP1, JMJD7-PLA2G4B, JMJD7
16	rs10965206	110500.00	0.30**	RP11-59O6.3, CBWD1, C9orf66, DOCK8
17	rs7974925	70478.00	-0.24*	KSR2
18	rs6001912	67447.00	-0.41****	SGSM3, RP5-1042K10.13, RP5-1042K10.10, RP5-1042K10.12, MKL1
19	rs4923183	67016.00	-0.29**	NELL1
20	rs1568765	63642.00	0.37***	
21	rs1783391	61123.00	-0.44****	
22	rs10869887	58948.00	0.44****	ATP5J2P3, RP11-490H9.1
23	rs900102	53950.00	0.27**	NRG1-IT2, RN5S263, NRG1
24	rs9819616	43242.00	-0.27**	U3
25	rs13001721	36716.00	-0.38****	
26	rs893188	26801.00	-0.36***	AC012313.1, ZNF584, ZNF132, ZNF324B, ZNF324, ZNF446, Metazoa_SRP, U6, SLC27A5
27	rs954106	23982.00	-0.31**	7SK, U6
28	rs6463224	23585.00	-0.27**	C7orf50, GPR146, GPER, RP11-449P15.1, AC091729.7, AC091729.8
29	rs9984991	23452.00	-0.35***	AP000475.2
30	rs989465	23082.00	-0.37***	NRG1-IT2, RN5S263, NRG1
31	rs6001913	22636.00	-0.42****	SGSM3, RP5-1042K10.12, MKL1
32	rs144173	22257.00	-0.34***	RP11-126L15.4, EPHB4, SLC12A9, ZAN, Metazoa_SRP, TRIP6
33	rs6001932	22113.00	-0.39****	RP5-1042K10.12, RP4-591N18.2, AL031594.1, MKL1
34	rs301624	22078.00	0.23*	RP11-463O9.5, RP11-463O9.6, FOXC2, FOXL1, RP11-58A18.1
35	rs4910295	19999.00	-0.40****	MTND5P21, CTD-3224I3.3, GALNTL4
36	rs10273967	19889.00	0.24*	CNTNAP2
37	rs774890	19383.00	-0.26**	SLC26A10, B4GALNT1, OS9, RP11-571M6.1, RP11-571M6.7, snoU13, RP11-571M6.8
38	rs2791708	19352.00	0.33***	PIGR, FCAMR, C1orf116, snoU13
39	rs344934	18974.00	0.28**	RP5-817C23.1, GADD45A, Metazoa_SRP, GNG12

40	rs2294352	18861.00	-0.39****	SGSM3, RP5-1042K10.13, RP5-1042K10.10, RP5-1042K10.12, MKL1
41	rs13116347	17784.00	-0.30**	RP11-76G10.1, NR3C2
42	rs1584586	17677.00	-0.38****	RP11-145F16.2, TSC22D2
43	rs10787530	16321.00	-0.34***	AFAP1L2, Metazoa_SRP, ABLIM1
44	rs957224	14887.00	0.33***	RP11-3L8.3, LURAP1L, Metazoa_SRP
45	rs473818	14572.00	0.31**	DLG2
46	rs12446160	14547.00	0.36***	UBA52P8, AC141846.4-001, LONP2, SIAH1
47	rs4675161	14505.00	0.35***	AC097662.2, COL4A3
48	rs6127972	13737.00	-0.28**	RP4-813D12.2, BMP7, RP4-813D12.3
49	rs10953303	13193.00	-0.36***	EPO, ZAN, EPHB4
50	rs3731739	13047.00	-0.31**	snoU13, SESTD1
51	rs2208363	12784.00	-0.35***	RP5-1185H19.2, MAB21L3
52	rs1344779	12396.00	-0.33***	C12orf75
53	rs2145419	11682.00	0.36***	
54	rs11712748	11428.00	-0.30**	RP11-454H13.6, ZBTB11, RP11-454H13.5, RPL24, U6, Y_RNA
55	rs3785461	10981.00	0.30**	RPS20P35, RP1-56K13.2, RP1-56K13.3, RP1-56K13.1, C17orf98, RPL23, LASP1, SNORA21
56	rs4489283	10892.00	0.29**	NRG1
57	rs722158	10884.00	0.32**	RP11-317J19.1, ME3
58	rs6449438	10618.00	-0.34***	AC006499.8, AC006499.5, AC006499.4, AC006499.3, AC006499.2, AC006499.1
59	rs658624	10568.00	0.34***	TMPRSS4, SCN4B, SCN2B, AMICA1
60	rs6478215	9938.60	-0.30**	LINC00474
61	rs6763648	9201.80	0.32**	RP11-430J3.1, HMGB1P36, U6
62	rs1760907	8743.60	0.37***	RPPH1, RP11-203M5.2, CCNB1IP1, PARP2, TEP1, RN5S382
63	rs840966	8655.20	0.32**	AC012370.3, AC007389.1, AC007389.2, SPRED2, AC074391.1
64	rs3743024	8572.00	0.21*	RP11-23P13.4, MAPKBP1, JMJD7-PLA2G4B, JMJD7, PLA2G4B, MIR4310, RN5S393, SPTBN5
65	rs2496737	8569.40	-0.38****	RP11-8L18.3, RP11-8L18.2, PARD3
66	rs12828	8259.50	-0.37***	RP11-679B19.2, WWOX
67	rs2868355	7891.40	0.16	HNRNPD, HNRPDL, IGBP1P4, SNORD42, ENOPH1
68	rs4918844	7577.00	0.26**	HABP2, NRAP
69	rs3807496	7564.60	-0.37***	TSPAN13, AGR2
70	rs3804778	7326.60	-0.32**	RP11-234A1.1, RP11-454H13.1, RG9MTD1, PCNP, U6
71	rs3778112	7122.20	-0.38****	RP11-73O6.4, LAMA2
72	rs13072301	6711.00	0.22*	FAM172BP, RP11-234A1.1, RP11-454H13.1, SENP7, RG9MTD1, PCNP
73	rs12297142	6708.30	0.27**	RP11-13G14.4, FICD
74	rs622572	6671.90	-0.30**	AC009294.1
75	rs10935800	6594.20	-0.31**	RP11-145F16.2, TSC22D2
76	rs194741	6449.10	-0.32**	RP11-723P16.2, ZFP36L1, Metazoa_SRP
77	rs314320	6431.50	-0.31**	RP11-126L15.4, EPHB4, SLC12A9, TRIP6, ZAN, Metazoa_SRP, SRRT
78	rs2355097	6404.00	0.29**	AC073109.2, THSD7A
79	rs602618	6361.00	-0.40****	RP11-479A21.1, ADRA2A
80	rs134810	6214.20	0.37***	CTA-929C8.5, CTA-929C8.7, CTA-929C8.6
81	rs639847	6107.10	0.36***	EDARADD
82	rs4737872	6016.30	-0.32**	PREX2
83	rs1917950	5972.40	-0.32**	CNTNAP2
84	rs7997069	5831.30	-0.39****	
85	rs7845867	5698.20	0.29**	CSMD1

86	rs6001954	5572.60	-0.41****	RP4-591N18.2, COX6B1P3, AL031594.1, MKL1
87	rs633715	5512.30	-0.24*	RP4-798P15.2, SEC16B
88	rs9985057	5406.60	0.34***	DSCAM-AS1, DSCAM
89	rs845561	5322.00	-0.30**	RP5-1091E12.1, EGFR
90	rs10781417	5318.20	-0.36***	ATP5J2P3, RP11-490H9.1
91	rs10063964	5240.30	-0.25**	CTD-2227I18.1, ANKRD55
92	rs10737764	5227.80	0.30**	RP5-1185H19.2, MAB21L3
93	rs6762432	5215.10	0.30**	TSC22D2
94	rs1990606	5176.70	0.33***	HMGB1P4, MYO3B, AC007277.3
95	rs7678436	5105.50	0.31**	FAM184B, DCAF16, NCAPG, LCORL
96	rs11819326	5096.80	-0.41****	
97	rs2425445	5055.80	0.23*	RP4-644L1.2, MAFB
98	rs2811659	4803.30	-0.17	
99	rs11589250	4751.30	-0.41****	RLF, TMCO2, U6
100	rs2496742	4708.10	-0.35***	RP11-8L18.3, RP11-8L18.2, PARD3

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

Table S2: Top 100 ρ_l for minimum winter temperature

Rank	SNP ID	ρ_l	BF	Genes within ± 50kbp
1	rs3827382	-0.49	777660.00****	RP4-591N18.2, AL031594.1, MKL1
2	rs17002034	-0.48	16734000.00****	MKL1, COX6B1P3, CTA-229A8.5
3	rs10869887	0.44	58948.00****	ATP5J2P3, RP11-490H9.1
4	rs1783391	-0.44	61123.00****	
5	rs7870404	0.43	252090.00****	ATP5J2P3, RP11-490H9.1
6	rs915624	-0.43	264470.00****	
7	rs11698307	-0.42	382.95***	RP11-560A15.4, RP11-560A15.3, BMP7
8	rs6001913	-0.42	22636.00****	SGSM3, RP5-1042K10.12, MKL1
9	rs6792339	-0.42	151.32**	MIR3921, C3orf26, FILIP1L
10	rs1871847	-0.42	667.70***	COL8A1, C3orf26, FILIP1L
11	rs12550228	0.42	272.63**	
12	rs10056699	-0.42	212.92**	CCDC11P1
13	rs6001912	-0.41	67447.00****	SGSM3, RP5-1042K10.13, RP5-1042K10.10, RP5-1042K10.12, MKL1
14	rs9632856	0.41	546.86***	NKX3-1, NKX2-6, RP11-213G6.2, RP11-175E9.1
15	rs2835190	-0.41	50.32**	AP000688.8, RPL23AP3, RUNX1
16	rs4837745	0.41	13.66*	MIR147A
17	rs10510925	-0.41	785.22***	ADAMTS9-AS2
18	rs11589250	-0.41	4751.30***	RLF, TMCO2, U6
19	rs1979635	-0.41	102.21**	SORCS3
20	rs10739560	0.41	25.61**	MIR147A
21	rs6001954	-0.41	5572.60***	RP4-591N18.2, COX6B1P3, AL031594.1, MKL1
22	rs11819326	-0.41	5096.80***	
23	rs7848719	-0.41	2465.70***	PALM2, PALM2-AKAP2, AKAP2
24	rs602618	-0.40	6361.00***	RP11-479A21.1, ADRA2A
25	rs12536200	-0.40	3535.30***	CNTNAP2
26	rs2326794	-0.40	113470.00****	RP11-73O6.4, LAMA2
27	rs4910295	-0.40	19999.00****	MTND5P21, CTD-3224I3.3, GALNTL4
28	rs2282537	-0.39	1029.50***	POU2F3, TMEM136, ARHGEF12
29	rs7997069	-0.39	5831.30***	
30	rs12364607	-0.39	28.71**	ST14, RP11-567O18.1, ZBTB44, DDX18P5
31	rs2294352	-0.39	18861.00****	SGSM3, RP5-1042K10.13, RP5-1042K10.10, RP5-1042K10.12, MKL1
32	rs3816186	-0.39	221650.00****	MTA3, AC074375.1, U6
33	rs10008679	0.39	2512.30***	
34	rs2691269	-0.39	616.05***	KLK13, KLK14, CTU1, SIGLEC9
35	rs17715017	-0.39	1950.70***	
36	rs6001932	-0.39	22113.00****	RP5-1042K10.12, RP4-591N18.2, AL031594.1, MKL1
37	rs6942733	-0.39	128120.00****	POP7, EPO, ZAN, EPHB4
38	rs10506959	0.39	306.37**	
39	rs11582490	-0.39	66.66**	TSEN15P2, NRD1, Y_RNA, MIR761
40	rs1946086	0.38	542.69***	KIRREL3
41	rs420302	0.38	374.55***	
42	rs3778112	-0.38	7122.20***	RP11-73O6.4, LAMA2
43	rs10869873	0.38	668.43***	ATP5J2P3, RP11-490H9.1, FOXB2
44	rs17591848	-0.38	3309.20***	LINC00332
45	rs13001721	-0.38	36716.00****	
46	rs1122127	-0.38	29.63**	RP11-90D11.1, KB-1047C11.2
47	rs4501094	-0.38	1815.40***	ATP2B2-IT1, ATP2B2
48	rs12932768	-0.38	290.64**	RP11-509E10.1, RBFOX1, RP11-420N3.2
49	rs1030089	0.38	13.94*	AC074391.1
50	rs6563798	-0.38	1359.00***	LINC00332

51	rs3859379	0.38	621.97***	
52	rs2626636	0.38	297.82**	RP11-525K10.3
53	rs9288052	-0.38	523.11***	AC009478.2, AC009478.1
54	rs4849387	0.38	38.88**	DPP10
55	rs12130516	-0.38	634.77***	HHAT
56	rs17617922	0.38	21.83**	TRPM3
57	rs2809936	-0.38	42.40**	NRD1, AL589663.1, RAB3B
58	rs6439989	0.38	360.90**	RP11-764I5.1, RP11-438D8.2, ACPL2
59	rs1584586	-0.38	17677.00****	RP11-145F16.2, TSC22D2
60	rs6446629	-0.38	436.26***	AFAP1
61	rs17480463	-0.38	794.99***	FGGY
62	rs17583067	-0.38	26.99**	RP11-661C8.2, RP11-43F13.3, AC116351.1, RP11-661C8.3, AC116351.2, NKD2
63	rs856615	-0.38	69.12**	TSEN15P2, OSBPL9, Y_RNA, MIR761, NRD1
64	rs2496737	-0.38	8569.40***	RP11-8L18.3, RP11-8L18.2, PARD3
65	rs920300	0.38	127.61**	NKX3-1, NKX2-6, RP11-213G6.2, RP11-175E9.1
66	rs12743478	-0.38	41.82**	NTNG1
67	rs10901091	-0.38	14.79*	EIF4A1P3, RAPGEF1, SNORA67, Metazoa_SRP, RP11-323H21.3
68	rs3737002	-0.38	1388.80***	RP11-78B10.2, CR1
69	rs4566993	-0.38	183.73**	RP11-981G7.4, PRSS55
70	rs2846063	-0.38	70.98**	AP000783.2, AP000783.1, MIR4493
71	rs4361385	-0.38	89.02**	AFAP1
72	rs895710	0.38	1305.60***	
73	rs16910142	0.38	12.13*	RP11-180I4.2, RP11-180I4.1
74	rs17084654	0.38	22.17**	
75	rs7911274	-0.37	2357.00***	
76	rs2245822	-0.37	27.73**	XXbac-BPG299F13.15, XXbac-BPG299F13.16, USP8P1, RPL3P2, WASF5P, XXbac-BPG248L24.13, XXbac-BPG248L24.10, HLA-C
77	rs7858354	-0.37	210.12**	PALM2, PALM2-AKAP2, AKAP2
78	rs903919	0.37	32.58**	SKI, MORN1
79	rs1760907	0.37	8743.60****	RPPH1, RP11-203M5.2, CCNB1IP1, PARP2, TEP1, RN5S382
80	rs134810	0.37	6214.20***	CTA-929C8.5, CTA-929C8.7, CTA-929C8.6
81	rs879502	0.37	2205.90***	AP000470.2
82	rs1029225	-0.37	2537.30***	LINC00160, AP000330.8, CLIC6
83	rs1492343	-0.37	153.31**	COL8A1
84	rs17294590	-0.37	32.44**	CPNE4
85	rs1323923	-0.37	50.91**	PCDH9-AS3, PCDH9-AS4, PCDH9
86	rs11145093	0.37	410.48***	PCA3, PRUNE2
87	rs12828	-0.37	8259.50***	RP11-679B19.2, WWOX
88	rs2477009	-0.37	1682.80***	RP11-8L18.3, RP11-8L18.2, PARD3
89	rs9515201	0.37	1440.70***	snoU13, COL4A2
90	rs989465	-0.37	23082.00****	NRG1-IT2, RN5S263, NRG1
91	rs12092568	-0.37	202.31**	SLC25A39P1, RP11-480I12.10, KDM5B-AS1, RP11-480I12.5, RP11-480I12.7, RP11-480I12.9, KDM5B, RP11-480I12.4, RABIF
92	rs12918809	-0.37	247.48**	RP11-509E10.1, RBFOX1, RP11-420N3.2
93	rs17186	-0.37	60.33**	AC087859.1
94	rs13093976	-0.37	428.40***	IFT57
95	rs3807496	-0.37	7564.60***	TSPAN13, AGR2
96	rs617938	0.37	102.57**	CTD-2127H9.1, OSMR
97	rs7764258	-0.37	11.62*	RP3-522P13.2, Y_RNA, LRRC16A
98	rs6657710	-0.37	57.41**	NTNG1
99	rs2041704	0.37	74.90**	

100	rs1568765	0.37	63642.00****

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

Table S3: Top 100 X^TX signals in the HGDP data

Rank	SNP ID	X^TX	continental F_{ST}	Genes within ± 50kbp
1	rs1834640	121.02	0.76****	SLC24A5, MYEF2
2	rs28777	112.77	0.43**	RP11-1084J3.1, RP11-1084J3.3, RXFP3, SLC45A2, AMACR, C1QTNF3
3	rs2250072	112.50	0.66****	RP11-208K4.1, SLC24A5, MYEF2
4	rs1419138	111.65	0.42**	RP11-354M20.3, RAB11FIP2
5	rs260690	109.69	0.60***	EDAR
6	rs1257016	109.15	0.26*	AC079395.1, FAM178B, snoU13, RN5S101
7	rs10886189	108.27	0.51***	RP11-354M20.3, RAB11FIP2
8	rs6500380	108.03	0.59***	UBA52P8, AC141846.4-001, LONP2, SIAH1
9	rs12477830	107.31	0.51***	AC019100.7, RP11-443K8.1, SULT1C2P1, SULT1C2
10	rs6497573	107.06	0.26*	RP11-101E7.2, C16orf52, VWA3A, SDR42E2
11	rs260714	106.93	0.58***	EDAR
12	rs12172281	106.76	0.27*	MKL1, MCHR1, GAPDHP37
13	rs13054099	106.76	0.31*	SLC25A17, ST13, DNAJB7, RP3-408N23.4, JTBP1, MIR4766, XPNPEP3
14	rs12476238	106.56	0.50***	AC019100.7, RP11-443K8.1, SULT1C2P1, SULT1C2
15	rs3922756	106.50	0.32*	INS, TH, IGF2, INS-IGF2, MIR4686
16	rs2424641	106.44	0.31*	
17	rs1257029	106.38	0.27*	AC079395.1, TRIM43CP, AC018892.8, AC018892.3, FAM178B, snoU13, RN5S101
18	rs5996039	105.76	0.32*	POLR3H, PMM1, DESI1, CSDC2, XRCC6
19	rs5996092	105.64	0.31*	SREBF2, SHISA8, TNFRSF13C, CENPM, SEPT3, CTA-250D10.23, CTA-250D10.15, CTA-250D10.19, MIR378I, MIR33A
20	rs3750997	105.50	0.08	AP002387.1, RP11-660L16.2, DHCR7, NADSYN1
21	rs133072	105.47	0.28*	MKL1, MCHR1, GAPDHP37
22	rs749767	105.45	0.28*	AC135050.5, RP11-196G11.4, RP11-388M20.2, ZNF668, ZNF646, PRSS53, RP11-196G11.1, VKORC1, BCKDK, KAT8, PRSS8, PRSS36, AC135050.2
23	rs11055962	105.41	0.28*	RP11-515B12.1, ATF7IP
24	rs8139993	105.30	0.32*	PMM1, DESI1, Y_RNA, CSDC2, XRCC6
25	rs11230851	105.24	0.37**	RP11-810P12.1, BEST1, FTH1, U6
26	rs8137373	105.19	0.25*	RANGAP1, ZC3H7B, U6, TEF
27	rs2156208	105.14	0.31*	RP11-640A1.1
28	rs3927	105.12	0.27*	RANGAP1, ZC3H7B, TEF
29	rs8135759	104.86	0.21	RP5-1042K10.12, MKL1
30	rs6761501	104.77	0.47**	EDAR
31	rs9837708	104.63	0.61****	FOXP1
32	rs1866694	104.39	0.36**	RP11-787D18.2, RP11-787D18.1, CLVS1
33	rs6542787	104.39	0.46**	EDAR
34	rs509360	104.36	0.23	RP11-467L20.10, RP11-467L20.9, RP11-467L20.7, DAGLA, C11orf9, AP002380.1, C11orf10, FEN1, FADS1, MIR611, MIR1908, FADS2
35	rs1136348	104.24	0.25*	RP11-430L17.1, FBXO42, C1orf144
36	rs6990312	103.93	0.38**	EBAG9, SNORD112, SYBU
37	rs10202644	103.89	0.27*	snoU13, RN5S101, FAM178B
38	rs4820437	103.85	0.26*	ZC3H7B, TEF, U6
39	rs10871454	103.82	0.37**	RP11-196G11.2, AC135050.5, HSD3B7, STX1B, STX4, ZNF668, AC135050.1, ZNF646, PRSS53, RP11-196G11.1
40	rs202654	103.72	0.24*	TEF, TOB2, PHF5A, ACO2
41	rs260698	103.67	0.54***	EDAR

42	rs1701930	103.67	0.28*	KLKP1, AC011483.1, KLK2, KLK4, KLK5, KLK6
43	rs1256991	103.58	0.29*	AC079395.1, TRIM43CP, AC018892.8, AC018892.3, AC018892.5, AC018892.6, FAM178B, snoU13, RN5S101
44	rs17740937	103.54	0.22	RP11-1A16.1
45	rs11020772	103.35	0.33*	RNU6-16, GRM5
46	rs7483764	103.30	0.32*	RNU6-16, GRM5
47	rs10483393	103.30	0.38**	RP11-187E13.1, RP11-187E13.2
48	rs3742000	103.29	0.37**	RPS2P41, ADAM1, AC003029.4, MAPKAPK5, TMEM116
49	rs2305884	103.13	0.22	CTF2P, FBXL19-AS1, AC135048.13, FBXL19, ORAI3, SETD1A, HSD3B7, STX1B
50	rs12913832	102.91	0.36**	OCA2, HERC2
51	rs1257017	102.90	0.24*	AC079395.1, FAM178B, snoU13, RN5S101
52	rs2441727	102.85	0.43**	snoU40, CTNNA3
53	rs7238925	102.85	0.17	ZCCHC2
54	rs12440301	102.84	0.61****	RP11-208K4.1, SLC24A5, MYEF2
55	rs1561277	102.79	0.28*	ZRANB3
56	rs139533	102.78	0.30*	L3MBTL2, CHADL, RANGAP1, ZC3H7B
57	rs1873933	102.75	0.19	GULOP, EPHX2, CLU
58	rs7498665	102.71	0.16	RP11-1348G14.5, RP11-24N18.1, RP11-22P6.2, RP11-22P6.3, ATXN2L, TUFM, SH2B1, ATP2A1, SNORA43, MIR4721, RABEP2
59	rs139553	102.66	0.33*	MEI1, CCDC134, RP5-821D11.7, Y_RNA, SREBF2
60	rs1468253	102.65	0.39**	RPL7AP60, RP3-521E19.3, C12orf51
61	rs139528	102.62	0.30*	L3MBTL2, CHADL, RANGAP1, ZC3H7B
62	rs11066322	102.40	0.34*	PTPN11
63	rs897986	102.37	0.21	FBXL19-AS1, AC135048.13, FBXL19, ORAI3, SETD1A, HSD3B7, STX1B
64	rs12049408	102.35	0.23	RP1-224A6.3, LINC00339, HSPG2, CELA3B, CELA3A, U6, Metazoa_SRP
65	rs4788102	102.31	0.16	RP11-1348G14.4, RP11-1348G14.1, RP11-1348G14.5, RP11-24N18.1, RP11-22P6.2, RP11-22P6.3, ATXN2L, TUFM, SH2B1, ATP2A1, SNORA43, MIR4721, RABEP2
66	rs889548	102.26	0.35**	RP11-196G11.4, RP11-388M20.2, RP11-388M20.7, RP11-388M20.8, ZNF646, PRSS53, RP11-196G11.1, VKORC1, BCKDK, KAT8, PRSS8, PRSS36, AC135050.2
67	rs8104441	102.23	0.43**	AC011483.1, KLK4, KLK5, KLK6, KLK7
68	rs1796045	102.11	0.23	snoU13, RN5S101, FAM178B
69	rs1348587	102.09	0.37**	GALNT13
70	rs5751080	102.05	0.22	RANGAP1, ZC3H7B, U6, TEF
71	rs692804	102.01	0.37**	RP11-778O17.4, TRIM29, OAF, POU2F3
72	rs6731972	102.01	0.46**	AC064847.4, ASXL2, KIF3C
73	rs126092	101.94	0.31*	MEI1, CCDC134, Y_RNA, RP5-821D11.7
74	rs2339941	101.87	0.37**	RP1-267L14.3, TMEM116, ERP29, Y_RNA, MIR3657, NAA25
75	rs4299060	101.83	0.32*	RP11-326E7.1
76	rs6583859	101.76	0.56***	RP11-280G19.1
77	rs6730157	101.59	0.28*	RAB3GAP1, SNORA40, ZRANB3
78	rs7584385	101.48	0.29*	GKN1, ANTXR1
79	rs876251	101.46	0.20	AC093690.1, RP11-731I19.1, BRE
80	rs9611613	101.32	0.24	ACO2, POLR3H, PMM1, CSDC2, DESI1
81	rs9323160	101.26	0.25*	
82	rs6802472	101.25	0.44**	FOXP1

83	rs12992554	101.23	0.33*	EDAR
84	rs4833103	101.19	0.34*	TLR10, TLR1, TLR6
85	rs4820425	101.19	0.28*	RP11-12M9.3, RP11-12M9.4, AL080243.1, Y_RNA
86	rs1999618	101.17	0.33*	RP11-1A16.1
87	rs886205	101.13	0.37**	RP3-462E2.3, ACAD10, RP11-162P23.2, ALDH2
88	rs1572018	101.12	0.43**	CALM2P3, KBTBD6, snoU13, Metazoa_SRP, MIR3168, KBTBD7
89	rs1796028	101.11	0.23	snoU13, RN5S101, FAM178B
90	rs7974383	101.02	0.35**	RP3-521E19.3, C12orf51, RPL6
91	rs133070	100.93	0.26*	MKL1, MCHR1, GAPDHP37
92	rs12891534	100.92	0.26*	CEP128
93	rs932206	100.91	0.24*	AC068492.1, CXCR4
94	rs853577	100.91	0.44**	RP11-354M20.3, RAB11FIP2, CASC2
95	rs4757108	100.87	0.31*	CTC-497E21.5, RP11-413N13.1
96	rs12889337	100.86	0.34*	RP11-1A16.1
97	rs7090105	100.73	0.53***	RP11-537A6.9, TTC18, ANXA7, Y_RNA
98	rs4757894	100.73	0.43**	NAV2-AS1, NAV2, DBX1
99	rs5758314	100.69	0.22	ZC3H7B, TEF, U6
100	rs7304572	100.67	0.35**	RP1-267L14.3, NAA25, Y_RNA, MIR3657, TRAFD1

*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$, ****: $p < 0.0001$

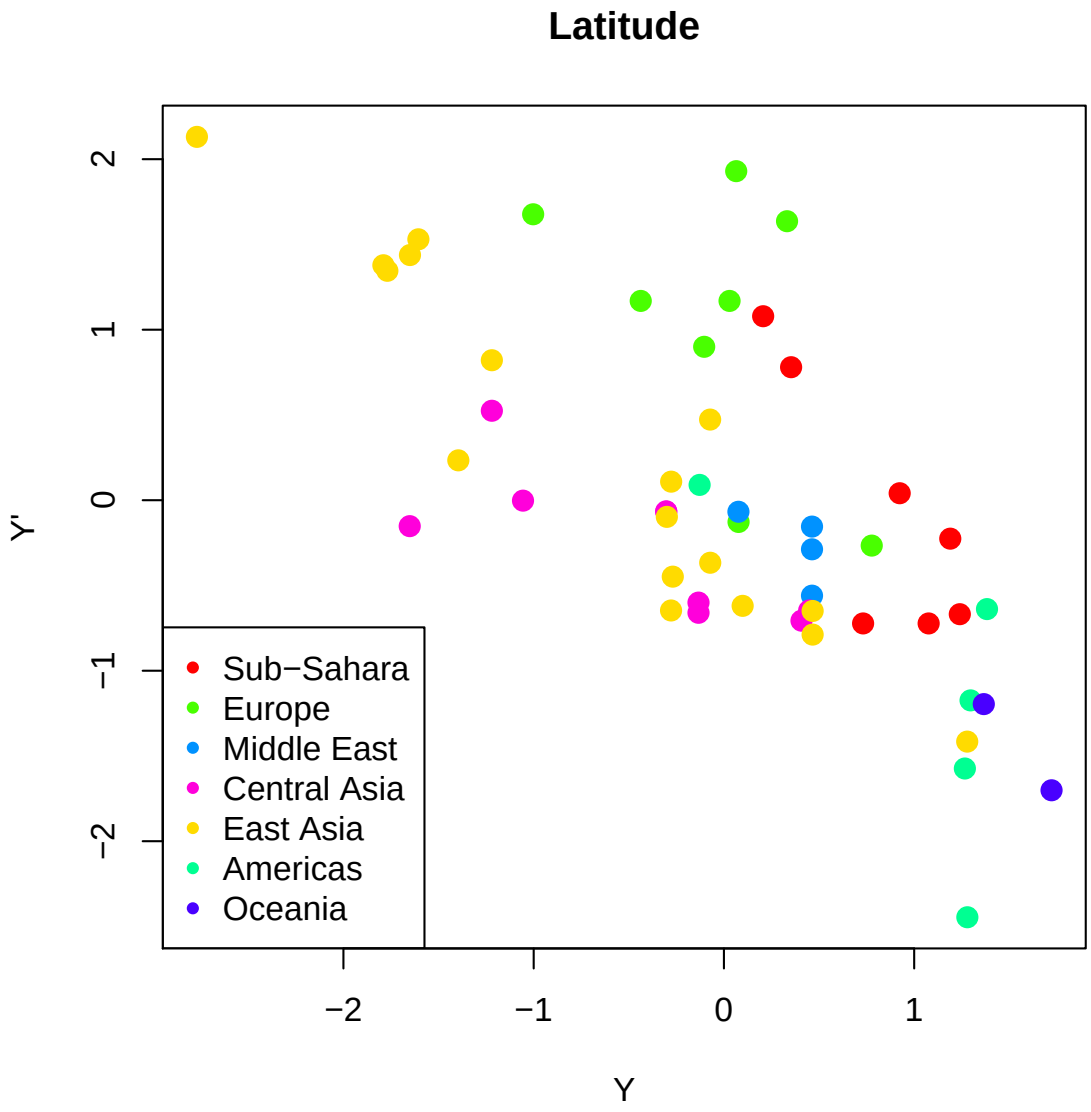

Figure S1: Transformation of normalized latitude Y to Y'.

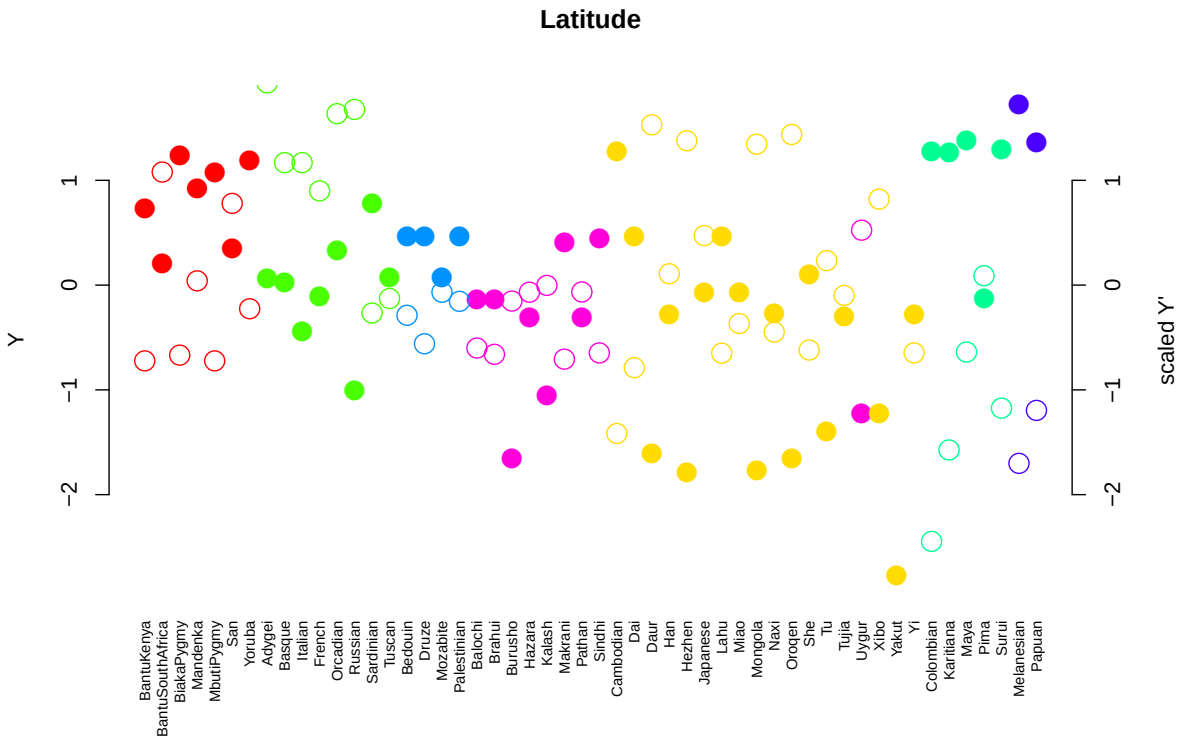

Figure S2: Transformation of normalized latitude Y (full symbols) to Y' (open symbols).

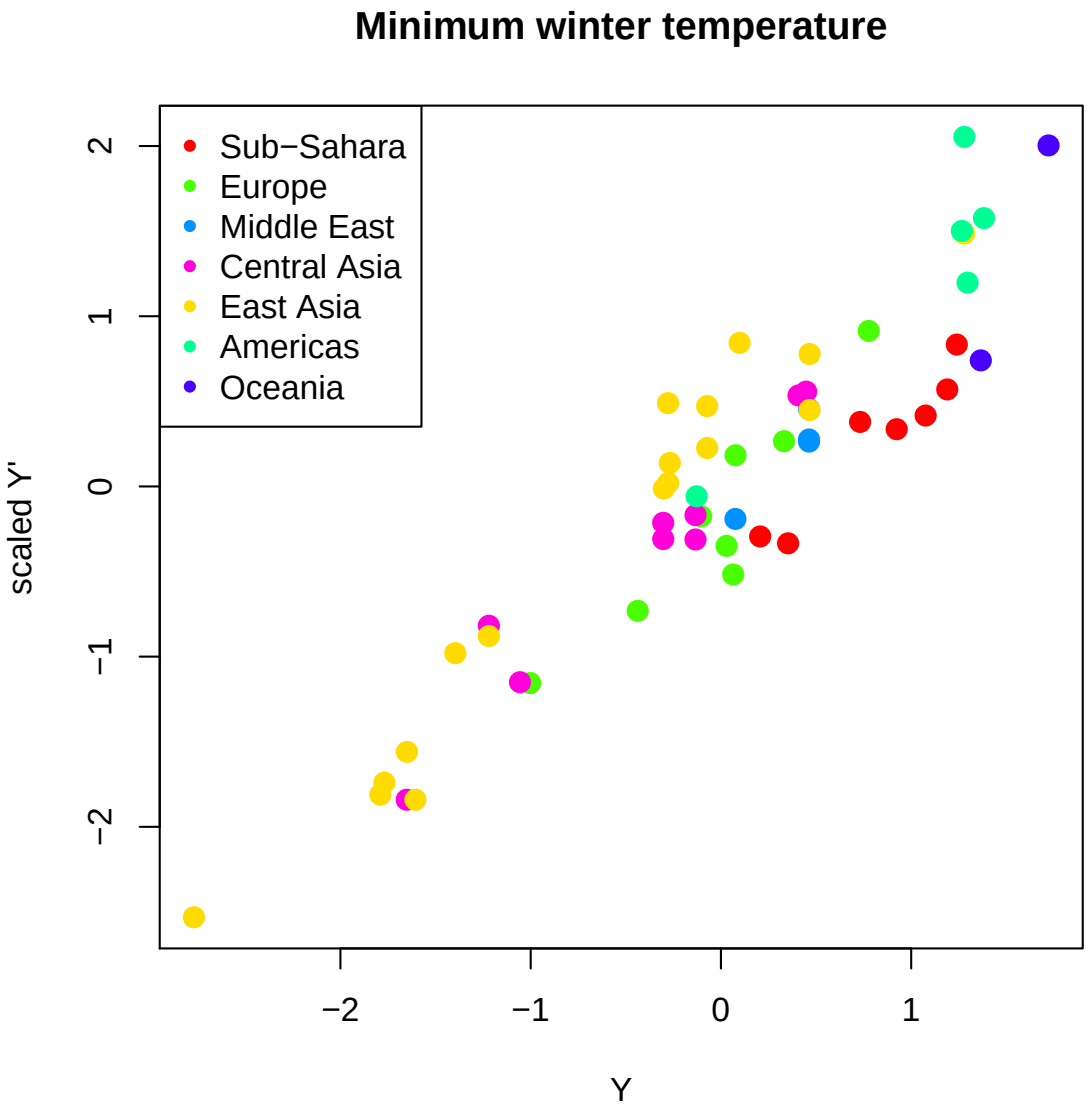

Figure S3: Transformation of normalized minimum winter temperature Y to Y'.

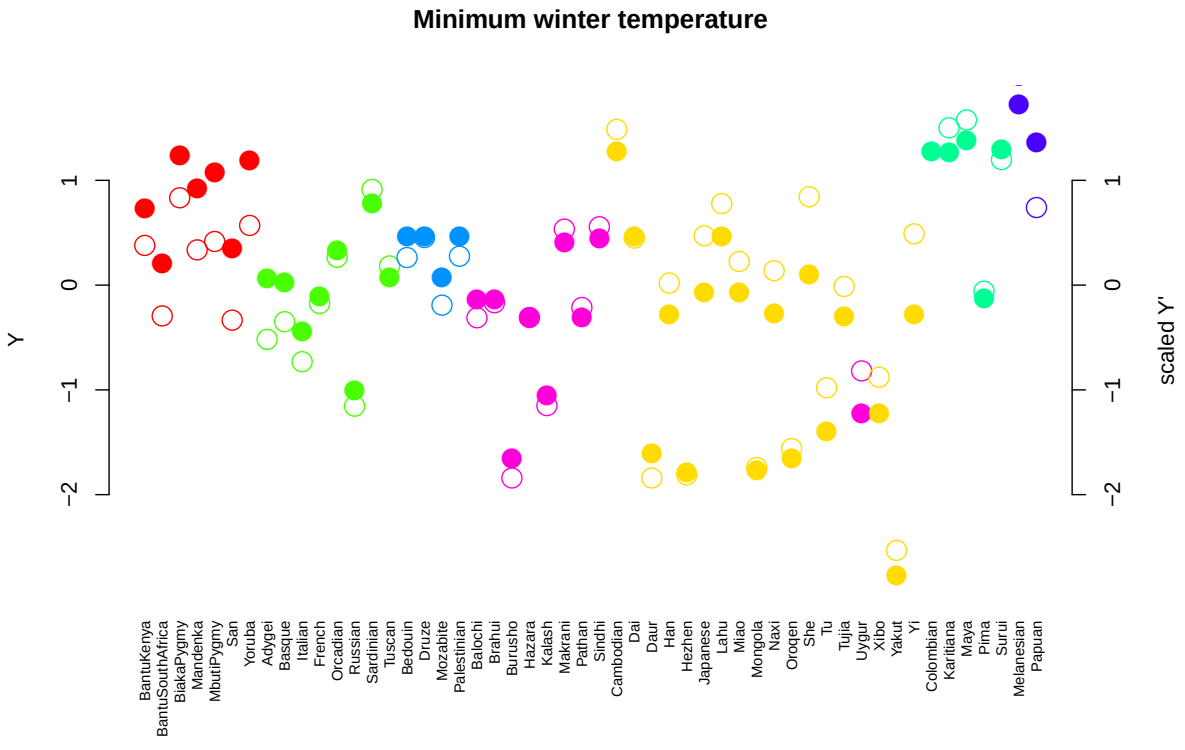

Figure S4: Transformation of normalized minimum winter temperature Y (full symbols) to Y' (open symbols).

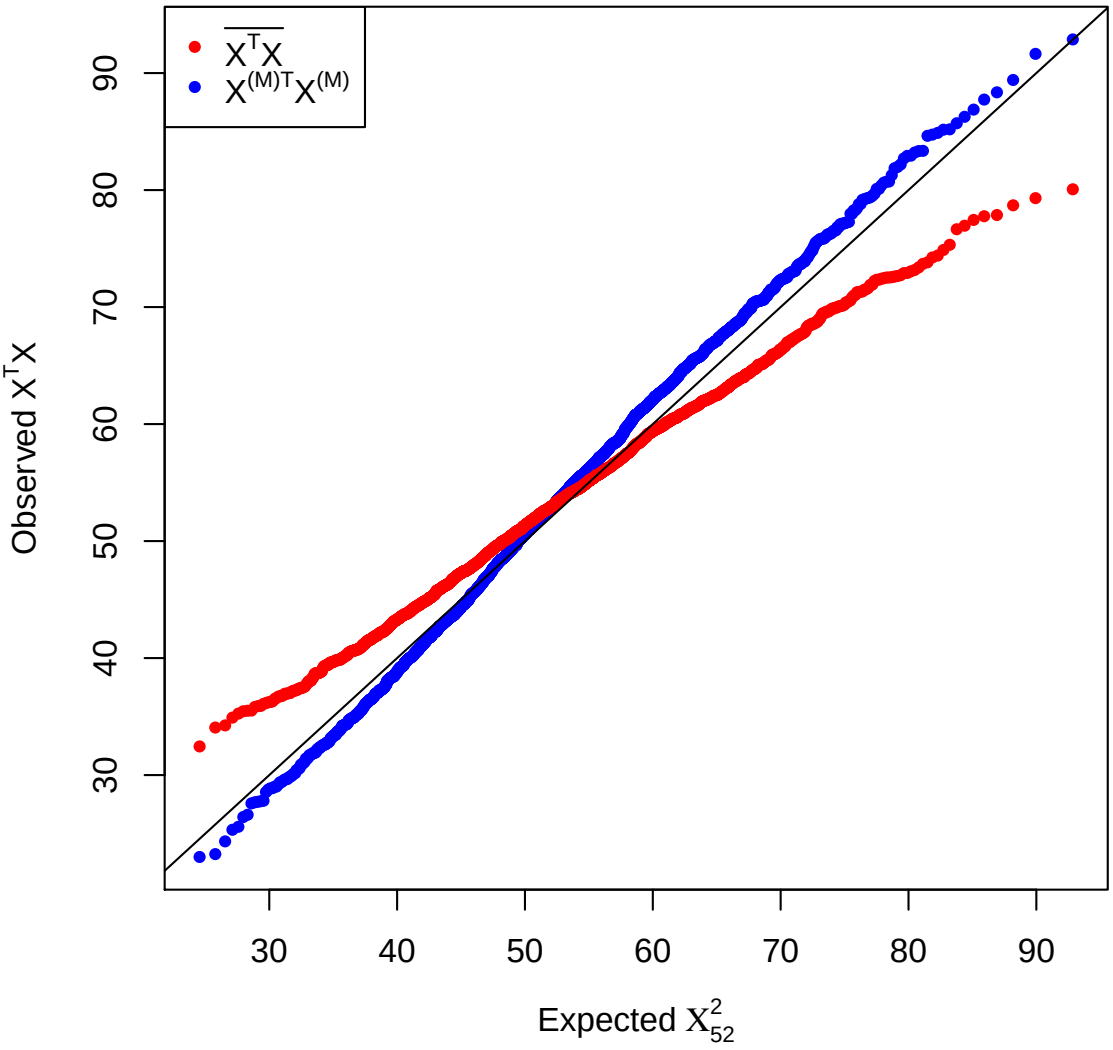

Figure S5: QQ-plot of $X_l^T X_l$ calculated in two different ways and the χ^2_{52} distribution, which is expected if X_l would follow MVN$(0, I)$. $X^T X$ was calculated in two different ways, first using the final generation of the MCMC ($X_l^{(M)T} X_l^{(M)}$) and the second using the average $X_l^T X_l$ across all M samples for each locus l ($\overline{X_l^T X_l}$). 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 $< 10^{-6}$). 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.

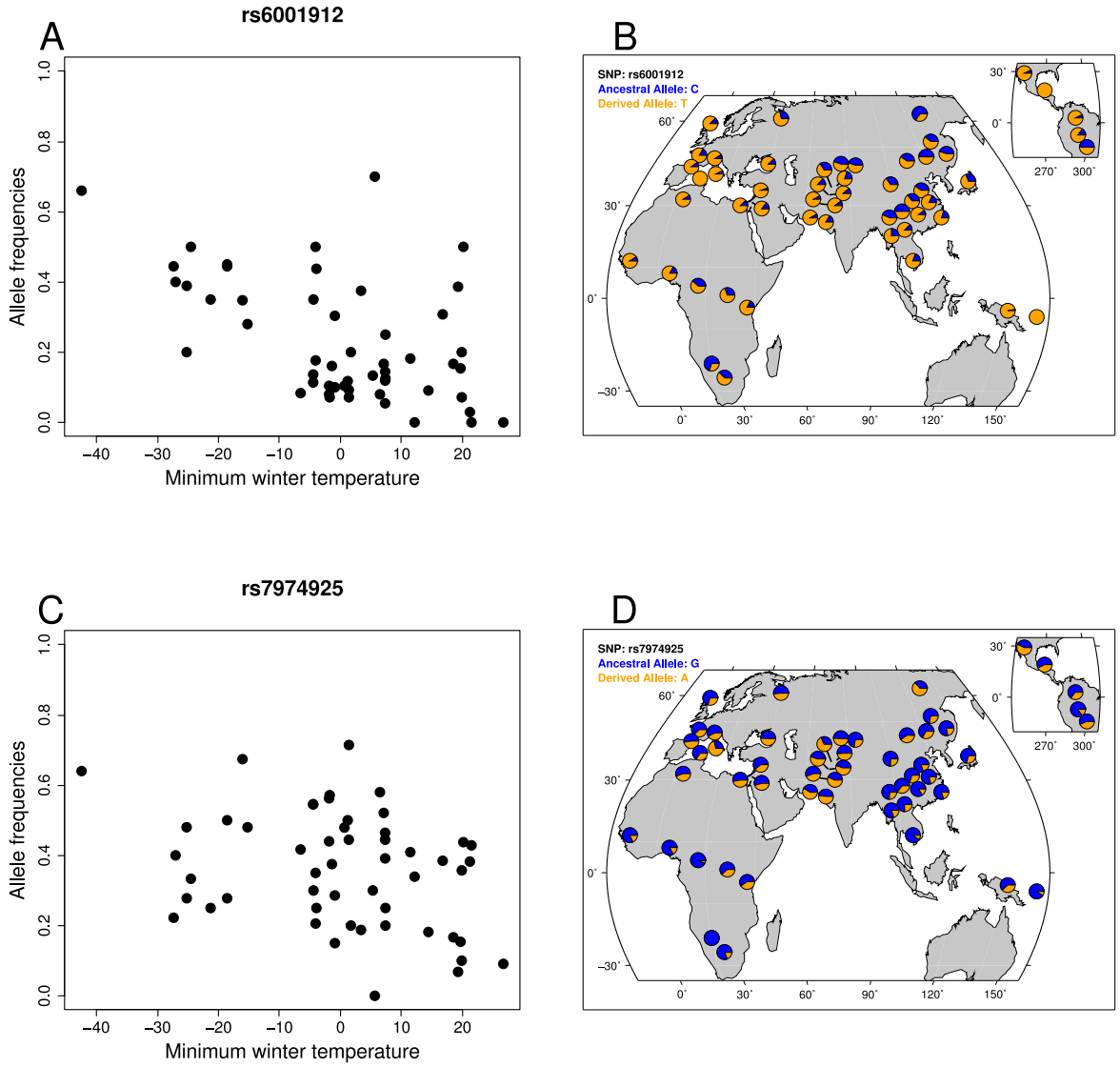

Figure S6: Two expemplarily chosen SNP from the top 20 Bayes factors. Both are characterized by similarly high Bayes factors (Supplementary Table 1) and extreme allele frequencies in the Yakuts. (A) Allele frequencies and standardized minimum winter temperatures of rs6001912 which is among the top 25 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 (hgdp.uchicago.edu).

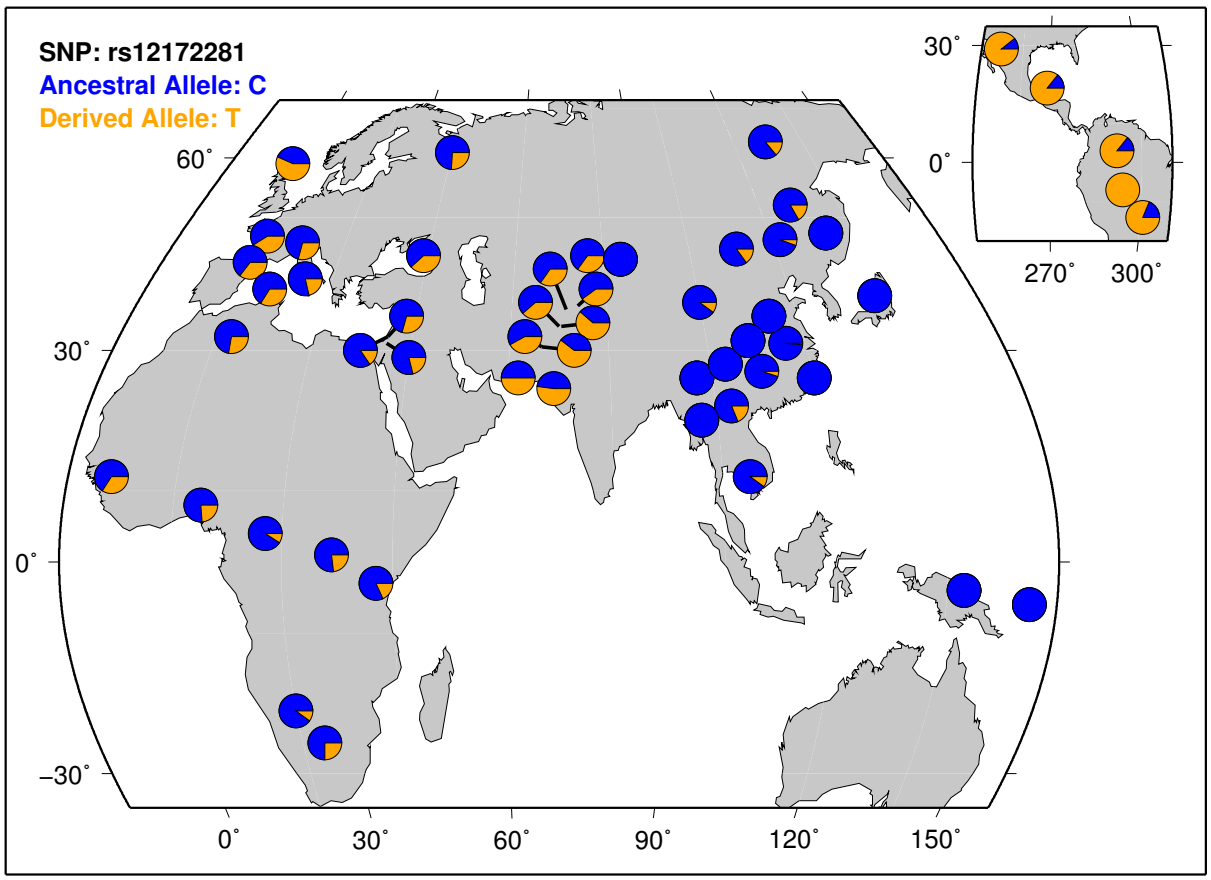

Figure S7: Geographical distribution of our 12th highest SNP for $X^T X$, rs12172281. This plot was downloaded from the HGDP selection browser (hgdp.uchicago.edu).

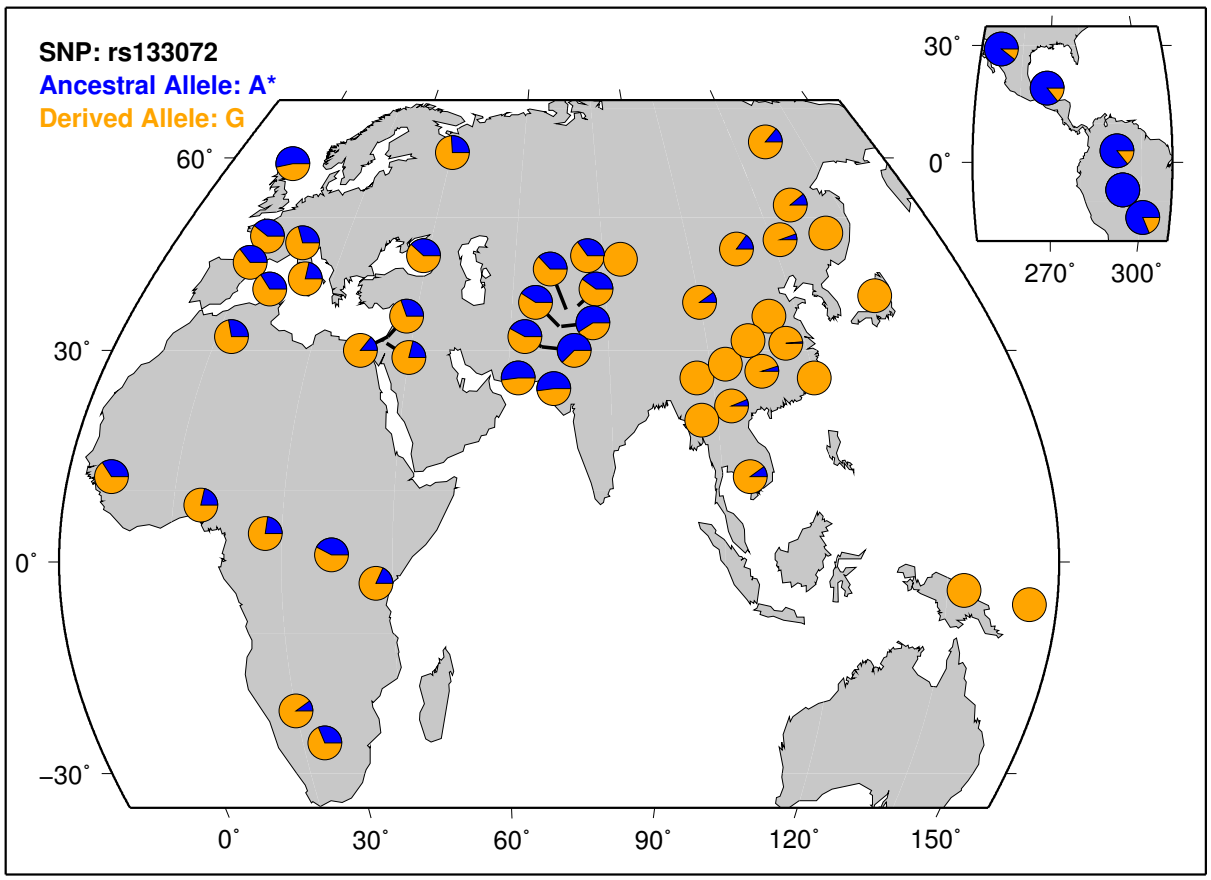

Figure S8: Geographical distribution of our 21st highest SNP for $X^T X$, rs133072. This plot was downloaded from the HGDP selection browser (hgdp.uchicago.edu).