Detecting selection in multiple populations by modelling ancestral admixture components

Jade Yu Cheng,[*,1,2] Aaron J. Stern,[3] Fernando Racimo,[1] Rasmus Nielsen,[1,2,4]

[1]Lundbeck GeoGenetics Centre, Globe Institute, University of Copenhagen, Oster Voldgade 5-7, Copenhagen 1350 Denmark

[2]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[3]Graduate Group in Computational Biology, University of California, Berkeley, Berkeley, CA 94720, USA

[4]Department of Statistics, University of California, Berkeley, Berkeley, CA 94720, USA

***Corresponding author:** E-mail: jade.cheng@cs.au.dk

Associate Editor: xxx

Abstract

One of the most powerful and commonly used approaches for detecting local adaptation in the genome is the identification of extreme allele frequency differences between populations. In this paper, we present a new maximum likelihood method for finding regions under positive selection. It is based on a Gaussian approximation to allele frequency changes and it incorporates admixture between populations. The method can analyze multiple populations simultaneously and retains power to detect selection signatures specific to ancestry components that are not representative of any extant populations. Using simulated data, we compare our method to related approaches, and show that it is orders of magnitude faster than the state-of-the-art, while retaining similar or higher power for most simulation scenarios. We also apply it to human genomic data and identify loci with extreme genetic differentiation between major geographic groups. Many of the genes identified are previously known selected loci relating to hair pigmentation and morphology, skin and eye pigmentation. We also identify new candidate regions, including various selected loci in the Native American component of admixed Mexican-Americans. These involve diverse biological functions, like immunity, fat distribution, food intake, vision and hair development.

Key words: Positive selection, admixture, population structure, human evolution, selective sweeps

Introduction

The emergence of population genomic data has facilitated fine-scale detection of regions under recent positive selection in humans and other species. There are multiple different methods for carrying out such selection scans. Some of these rely on patterns of long-range linkage-disequilibrium (Sabeti *et al.*, 2007; Voight *et al.*, 2006), one of the characteristic genomic footprints left by a selective sweep (Kim and Nielsen, 2004; Kim and Stephan, 2002; McVean, 2007). However, this pattern fades rapidly over time, and these methods are, consequently, best suited for detecting very recent selective sweeps from *de novo* mutations. Other techniques, based on

distortions in the allele frequency spectrum caused by positive selection, can allow for the detection of more ancient events, but are generally only applicable to one population at a time (DeGiorgio *et al.*, 2016; Fay and Wu, 2000; Fu and Li, 1993; Huber *et al.*, 2016; Nielsen, 2005; Tajima, 1989).

A different class of methods for detecting selection involves analyzing patterns of allele frequency differentiation between populations. The basic idea is that regions that have experienced episodes of positive selection will display frequency differences between populations that are stronger than what would be expected under pure genetic drift. For example, one can compute Wright's fixation index (F_{ST}) locally across different regions of a genome, and look for extreme outliers (Akey *et al.*, 2002; Beaumont and Balding, 2004; Beaumont and Nichols, 1996). Population differentiation methods can detect more ancient selective events than linkage disequilibrium-based methods (Sabeti *et al.*, 2006), and are sensitive to different types of positive selection events, including sweeps from a *de novo* mutation, sweeps from standing variation, incomplete sweeps, and adaptive introgression (Bonhomme *et al.*, 2010; Fumagalli *et al.*, 2015; Racimo *et al.*, 2017; Yi *et al.*, 2010). Recent methods have allowed researchers to detect excess local differentiation on particular branches of a 3-population tree (Racimo, 2016; Yi *et al.*, 2010), a 4-population tree (Cheng *et al.*, 2017b) or an abitrarily large tree (Librado and Orlando, 2018), albeit without modeling post-split admixture events.

A generalization of these approaches was developed by Coop *et al.* (2010), Günther and Coop (2013), and Gautier (2015). It involves the detection of genomically local distortions from a genome-wide covariance matrix, which is used as a neutral baseline. An advantage of this approach is that one can apply it to an arbitrary number of populations. Other researchers have used hierarchical Bayesian models (Foll and Gaggiotti, 2008; Foll *et al.*, 2014) or principal component analysis (Duforet-Frebourg *et al.*, 2016; Luu *et al.*, 2017) to model patterns of population differentiation to identify local distortions across the genome. Another method extended single-locus differentiation-based methods to the analysis of haplotype differentiation (Fariello *et al.*, 2013). More recently, Mathieson *et al.* (2015) developed an admixture-aware selection test based on a linear model and applied it to human data. The analysis took advantage of the fact that present-day European populations could be modeled as a mixture of three highly differentiated ancestral components. Regions of the genome that exhibited strong deviations from the genome-wide mixture proportions were therefore strong candidates for positive selection. Finally, Refoyo-Martínez *et al.* (2019) developed a method to test for selection on an admixture graph, which represents the history of divergence and admixture events

among populations. Although useful for detecting selection in the presence of admixture, it still requires the user to specify which individuals belong to which populations, and to infer the graph in advance.

Here, we introduce a new selection detection framework that can explicitly model admixture and detect selection from populations of admixed ancestries. It can simultaneously compare arbitrarily many populations and ancestry components and is encoded in a flexible framework for testing selection on a specific lineage or set of lineages. The method allows the user to identify signals of positive selection via population differentiation, without relying on self-reported ancestry or admixture correction to group individuals into populations. The method can also determine if a selective event is specific to a particular population or shared among different populations.

Unlike previous methods, we fully take advantage of admixed populations, and we do not require the user to *a priori* categorize samples into populations, or to correct allele frequencies to account for recent admixture. Thus, the selection scan does not rely on user-supplied sample labels or ancestry compositions. The methods identifies positive selection by searching for loci showing distortions in the population covariance matrix, relative to the genome-wide baseline. It provides a flexible framework to specifically test for selection on individual components or sets of components. This functionality allows researchers to accommodate specific evolutionary scenarios into the range of testable hypotheses, including local adaptation, adaptive introgression, and convergent selection. The method first co-estimates the population structure of the input panel and the allele frequencies of the ancestral admixture compomnents through an unsupervised learning process (Cheng *et al.*, 2017a), before testing for selection on the ancestral components themselves. Researchers can also use the method to examine estimated population structure and visualize trees connecting the ancestral components using plotting functionalities provided by our software package, Ohana, as part of the analysis pipeline.

Methods

Basic model

The new method is based on the Ohana inference framework (Cheng *et al.*, 2017a), which works with both genotype calls and genotype likelihoods. In brief, the classical Structure model (Pritchard *et al.*, 2000) is used to infer allele frequencies, ancestry components, and admixture proportions using maximum likelihood (ML). Then a covariance matrix among components is inferred using a multivariate Gaussian distribution while enforcing constraints imposed by the assumption of a tree structure. The covariance between leaf nodes is proportional to the amount of shared phylogentic history between the nodes. Consider, for example, the

example of the matrix and corresponding tree in the left side of Fig. 1. In this tree all branches have length 0.1 and the tree is rooted in node A. The covariance between node E and node B is then 0.1, because B and E share one edge in the path from A. However, the covariance between node C and E is 0.2 because they share two edges in common in the path from A. The covariance matrix, $\Omega = \{\Omega_{ij}\}$, can be converted into a distance matrix, $d = \{d_{ij}\}$, using the rule $d_{ij} = \Omega_{ii} + \Omega_{jj} - 2\Omega_{ij}$. Treeness can then be tested using the four point condition applied to d.

This system is underdetermined because the tree can be rooted in any node (see e.g., Felsenstein (1985)), and the same joint probability distribution is obtained no matter which rooting is chosen. We root the tree in one of the ancestry components and condition on the allele frequencies in this component when calculating the joint distribution of allele frequencies in the other components. This idea is similar to Felsenstein's restricted maximum likelihood approach (Felsenstein, 1985). We emphasize that the rooting is arbitrary but that it does not imply any assumptions about this component actually being ancestral

We estimate the covariance matrix Ω via ML. This matrix has size $(K-1)\times(K-1)$, where K is the number of populations assuming a joint density of allele frequencies given by

$$P(f_j \,|\, \Omega,\, \mu_j, f_{aj}) \sim \mathcal{N}\left(f_{aj},\, \mu_j(1-\mu_j) \begin{bmatrix} \Omega_{1,1} & \cdots \Omega_{1,k-1} \\ \vdots & \vdots \\ \Omega_{k-1,1} & \cdots \Omega_{1,k-1} \end{bmatrix} \right) \tag{1}$$

where f_{aj} is the allele frequency in the ancestry component arbitrarily assigned as ancestral and f_j is a vector of the allele frequencies in the other $K-1$ components, at SNP j. μ_j is the mean allele frequency for SNP j (averaged over all components). Notice that this model of joint allele frequencies is similar to the model implemented in TreeMix (Pickrell and Pritchard, 2012) which also uses a Gaussian approximation to allele frequency change. The full likelihood function is obtained by taking the product of Eq. (1) over all SNPs in the genome. The method for optimizing this function is described in a subsequent section.

Selection model

Following the genome-wide estimation of Ω, a natural extension of this framework is to detect SNPs that deviate strongly from the globally estimated covariance structure. The idea of testing for deviations from a Gaussian distribution follows Günther and Coop (2013), but differs in the use of an enforced tree-structure, an ML inference framework and fast optimization algorithms, thereby avoiding some of the computational challenges associated with Markov Chain Monte Carlo (MCMC). We also notice that admixture is incorporated into the inference framework, thereby enabling the possibility to test for positive

selection that acted on the ancestral components of a panel, before interbreeding occurred between the ancestors of the sampled individuals

Ohana uses a likelihood ratio test that identifies SNPs with allele frequency patterns that are poorly described by the genome-wide pattern. After estimating Ω jointly for all SNPs, each SNP is then independently tested for deviations from this model, using a scalar factor introduced to certain elements of the covariance matrix. This scalar factor can be introduced in different ways depending on which selection hypotheses are tested. In our analyses, we chose to scale the covariance matrix such that one of its diagonal values is multiplied by a scalar, α, corresponding to differences in allele frequency in one of the ancestry components relative to the rest, e.g.:

$$\Omega_\alpha = \begin{bmatrix} \Omega_{1,1} & \cdots & \Omega_{1,k-1} \\ \vdots & & \vdots \\ \Omega_{k-1,1} & \cdots & \alpha \cdot \Omega_{1,k-1} \end{bmatrix} \quad (2)$$

The value of α is then estimated via ML using Eq. (1) (assuming all other values in Ω_α is fixed at the genomic ML estimates) and a likelihood ratio is formed by testing the hypothesis of $\alpha = 1$ against the alternative of $\alpha > 1$. A significantly high likelihood ratio indicates a larger deviation in allele frequency in a focal component than expected under the globally estimated null-model. Fig. 1 shows an example. This test can also be implemented to test selection on ancestral non-terminal lineages by multiplying the corresponding values in the covariance matrix by a scaling factor.

Under the null-hypothesis, the likelihood ratio test statistic is expected to approximately follow a 50:50 mixture between a χ^2_1-distribution and a point mass at zero (Self and Liang, 1987) because α is bounded at 1, and we use this asymptotic distribution to calculate p-values.

In summary, we estimate a scaling factor for one or more components of the covariance matrix in a multivariate normal model of allele frequency distribution among populations. For each candidate SNP, we then compare the estimated covariance matrix to that obtained genome-wide, using a likelihood ratio test.

Optimization

To estimate allele frequencies, we assume a classical stucture/admixture model (Pritchard *et al.*, 2000) and first estimate Q, a matrix of admixture proportions for each individual, and F, the matrix of allele frequencies for all loci, using a quadratic programming algorithm described in full detail in Cheng *et al.* (2017a) and we refer the reader to the description in this paper. This method can also incorporate genotype likelihoods.

Conditional on these estimates of values of f_j and f_{aj} for all j, we then maximize the likelihood in Eq. (1) for Ω. This optimization is done using the Nelder-Mead simplex method (Nelder and Mead, 1965). It uses Cholesky decomposition (Cholesky, 1910) to determine the positive semi-definiteness of a matrix and to compute matrix

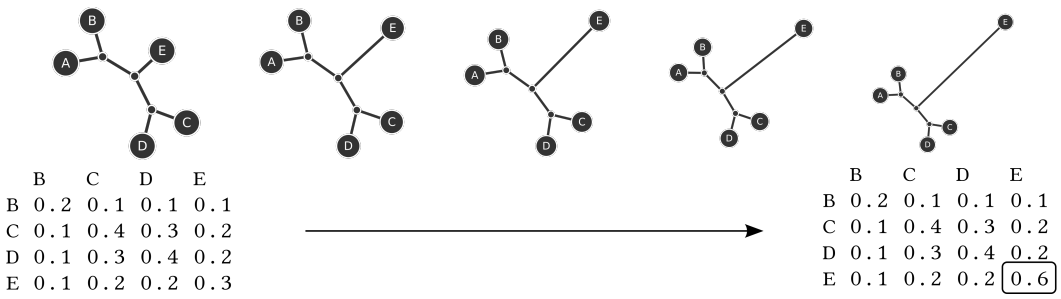

FIG. 1. Selection hypotheses and their encodings as covariance matrices. In this example, the ancestry component E is assumed to be the potential target of selection. The entry E:E in the covariance matrix is therefore allowed to deviate from the globally estimated value.

inverses and determinants. For the initial starting point, we use sample covariances:

$$S_c = \frac{1}{J}\sum_j^J (x_j - \bar{x}_j)(x_j - \bar{x}_j)^T$$

$$x_j = \begin{bmatrix} f_1 \\ \vdots \\ f_{K-1} \end{bmatrix} \quad \bar{x}_j = \begin{bmatrix} f_A \\ \vdots \\ f_A \end{bmatrix} \qquad (3)$$

To enforce treeness, instead of using a costly constrained optimization, we convert the covariance matrix into a distance matrix, $d = \{d_{ij}\}$, which is converted into a tree using the Neighbor-Joining algorithm (Saitou and Nei, 1987). We then use the covariance matrix induced by this procedure. For estimating α during a selection scan for a single SNP, conditionally on the globally estimated value of Ω, we use a simple Golden-section search algorithm (Kiefer, 1953).

Simulations

We conducted population genetic simulations using the forward simulator SLiM 3 (Haller and Messer, 2019). We consider 3 distinct demographic models (Fig. 2):

- A basic 4-population tree with no admixture (Fig. 2a): an ancestral population splits into 4 subpopulations at times 4000, 2000, and 800 generations before present, following the topology in Fig. 2a. Selection is simulated on the yellow branch in Fig. 2a. Tests for selection are conducted for yellow ancestry (i.e. the main ancestry component in the third branch).
- A 4-population tree with admixture (Fig. 2b): The same model as in (1), but split times are shifted backwards in time by 100 generations; at 100 generations before present, selection is turned off, and each population is supplanted by a (1/3,1/3,1/3) mixture of the other three populations. Tests for selection are conducted for yellow ancestry (i.e. the most depleted ancestry component in the third branch).
- A model based on human demography of Mexican (MXL), Northwestern European (CEU), CHB (EAS), and African Yoruba (YRI) populations (Fig. 2c): The model is based on parameter estimates from Gravel *et al.* (2011); Gutenkunst *et al.* (2009). MXL is modeled as a (1/2,1/2) mixture of CEU and

Native American (NA) ancestry. We simulate selection only in the ancestral NA population (i.e. no ongoing selection in MXL). We use Ohana to test for selection in this NA ancestry component, which is only observed in the admixed MXL individuals.

In simulations (1) and (2) we assume all populations are constant in size with $N_e = 10{,}000$. For all simulations, we simulate a locus of 2Mbp with mutation and recombination rates $\mu = r = 10^{-8}$ per bp per generation. In all cases, we sample 20 diploid individuals from each extant population (i.e. 160 chromosomes sampled). We simulate a single selected site occurring within a ±10kbp window of the center of the simulated locus. In order to simulate selection during particular time periods, we simulate sweeps from standing variation (an initial frequency f), although we consider such low frequencies (down to $f = 0.0001$) that these should produce indistinguishable patterns from those produced by hard sweeps (Przeworski *et al.*, 2005). For each demographic scenario, we consider 4 different selection coefficients ($s = 0$, 0.01, 0.02 and 0.05) and 3 different ranges of starting frequencies for the selected allele (f in [0.0001,0.001), [0.001,0.01), and [0.01,0.1)). (Simulations under model (3) exclude sweeps with $f < 0.001$ because the ancestral NA population size is too small for any such variation at that low frequency.) We use a neutral burn-in phase of 100,000 generations. For all simulations, as is typical in forward simulations, we scale times down by a factor of 10, and scale up the selection coefficients and mutation and recombination rates by a factor of 10, in order to ease computational burden. In all simulation scenarios we use 1000 independent replicates. Open-source implementations of each model are provided at `https://github.com/35ajstern/ohana_simulation_models`.

We compared Ohana's performance to that of two other state-of-the-art methods: pcadapt and BayPass (Duforet-Frebourg *et al.*, 2016; Gautier, 2015). Like Ohana, both methods depend on some sort of empirical null model. To this end, we simulated 3 20Mb-long neutral regions under otherwise the same settings as previously described, with $s = 0$, in order to generate a null dataset for calibrating each method. In the case of Ohana and BayPass, this null dataset is used to estimate the covariance matrix for each population; in pcadapt, we append this null dataset to each region we test for selection (we do this because the pcadapt package does not have an equivalent two-step process for calculating PCs in one region and testing for deviation from these PCs in a separate region). In all cases, we filter out SNPs with MAF < 0.05 prior to any analysis. In Ohana, we test for selection in specific ancestry groups; by contrast, BayPass and pcadapt test for any significant deviation from the empirical covariance matrix (BayPass models population-level covariance, whereas pcadapt

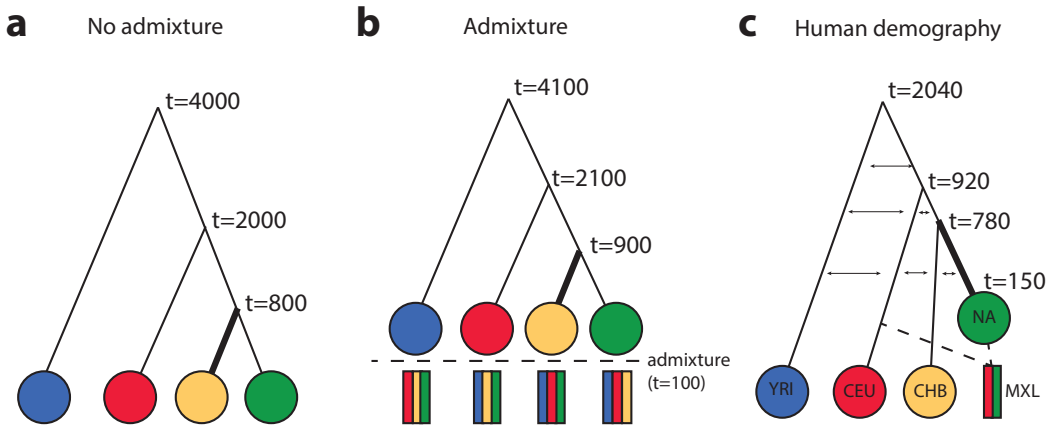

FIG. 2. Illustration of simulation models. (a) Model 1, a basic model of four-population split with no admixture. (b) Model 2, a four-population split with subsequent admixture. (c) Model 3, a four-population model mimicking human demographic models. Population size changes in Model 3 are omitted from the visualization for simplicity. Selection is simulated to operate on the branch that has a larger width.

models individual-level covariance). In this sense it is important to keep in mind that Ohana is performing a more specific test for selection, and can be used to methodologically attribute selection to a particular ancestral component / branch.

Results

Simulations

Power to detect selection

We first evaluated the performance of our method against comparable methods in detecting whether a locus has evolved under positive selection (Fig. 3a-c). For all tests we use the empirical null distribution to find the threshold associated with 5% false positive rate (FPR). We compute power as the proportion of simulations with test statistic exceeding this threshold. For all three methods, the test statistic was the site statistic with the maximum value across the whole 2Mbp locus; with Ohana, the test statistic used was the log-likelihood ratio (testing for selection in the specified ancestry group); with BayPass, the test statistic was the 'XtX' statistic; and with pcadapt, the test statistic was the selection test P-value, using $K = 3$ PCs (since there are 4 ancestral populations in each case). For all three methods, power increased uniformly with the value of the selection coefficient (Fig. 3a-c). However, different demographic scenarios result in different power levels; e.g., all methods were better-powered under the simple demographic scenario (Model 1, Fig. 2a,3a) compared to selection pre-admixture (Model 1, Fig. 2b,3b) or in the human demographic model (Model 3, Fig. 2c,3c).

pcadapt was significantly less well-powered than Ohana in most scenarios e.g., 72% vs 88% and 90% vs 98% power under moderate and strong selection, respectively (see Fig. 3a). In most scenarios, we found Ohana to have power equivalent to or greater than that of BayPass; e.g., in Models 1 and 2, we fail to reject that Ohana and BayPass have different power curves with 95%

confidence (Fig. 3a,b). However, in one simulation scenario, we found Ohana had significantly higher power than Baypass, under a model of human demography (Fig. 3c). However, we note that although BayPass and Ohana have similar power, Ohana's test is by design more specific, as it is testing for selection in a specified ancestry group; hence power calculations are inherently more lenient for BayPass than Ohana. In Fig. 3a-c we show results assuming $0.001 \leq f < 0.01$, and present full results illustrating the entire ROC curve (i.e., not conditioned on FPR=0.05) and for other values of f, in Fig. 4 and Figs. S1, S2. Also notice the overall low power in Fig. 4. The main reason for this low power is that selection is acting in a relatively short period in the past, and that the population has experienced 50% admixture after selection. The strong admixture after selection tends to obscure much of the selection signal.

Efficacy for fine-mapping the causal site

We also considered the performance of Ohana for fine mapping the position of the causal site (Fig. 3d-f). We considered the distribution of the distance between the site of the test statistics described above (i.e. the locus-wide max statistic) and the causal site (in the center of the 2Mbp locus). We plot the empirical cumulative distribution of these distances for different values of the selection coefficient under each demographic model. We found that in cases where Ohana is well-powered to detect selection, there is considerable power to narrow the position of the causal site down to ±10kbp of the max test statistic (e.g. $>40\%$ power for $s \geq 0.01$ under Model 1, and 25% power for $s = 0.05$ under Models 2 3; see Fig. 3d-f). Interestingly, under Models 1 and 2, there is similar power to fine-map sites with moderate and strong selection ($s = 0.02$ vs 0.05, see Fig. 3d,e); by contrast, under Model 3, there is significantly higher power to fine-map sites with strong selection (Fig. 3f); this dramatic difference may be due to the effects of demography on the pattern of hitchhiking surrounding the causal site. In Fig. 3d-f we show results assuming $0.001 \leq f < 0.01$, and present full results illustrating other values of f, in Figs. S3-S5.

Computational efficiency

In addition to comparing power to detect selection, we compared computational efficiency of Ohana and BayPass, which we showed in previous sections was the most competitive method in terms of statistical power (Fig. 3g). We found that Ohana was $>250X$ faster than BayPass (mean selection scan runtimes: Ohana, 0.626 secs. (±0.008 secs.); BayPass, 168 secs. (±2 secs.); $N =$ 1,000 replicates). We reiterate that our power comparison revealed Ohana to generally have comparable power to that of BayPass, despite multiple orders of magnitude improvement in computational efficiency.

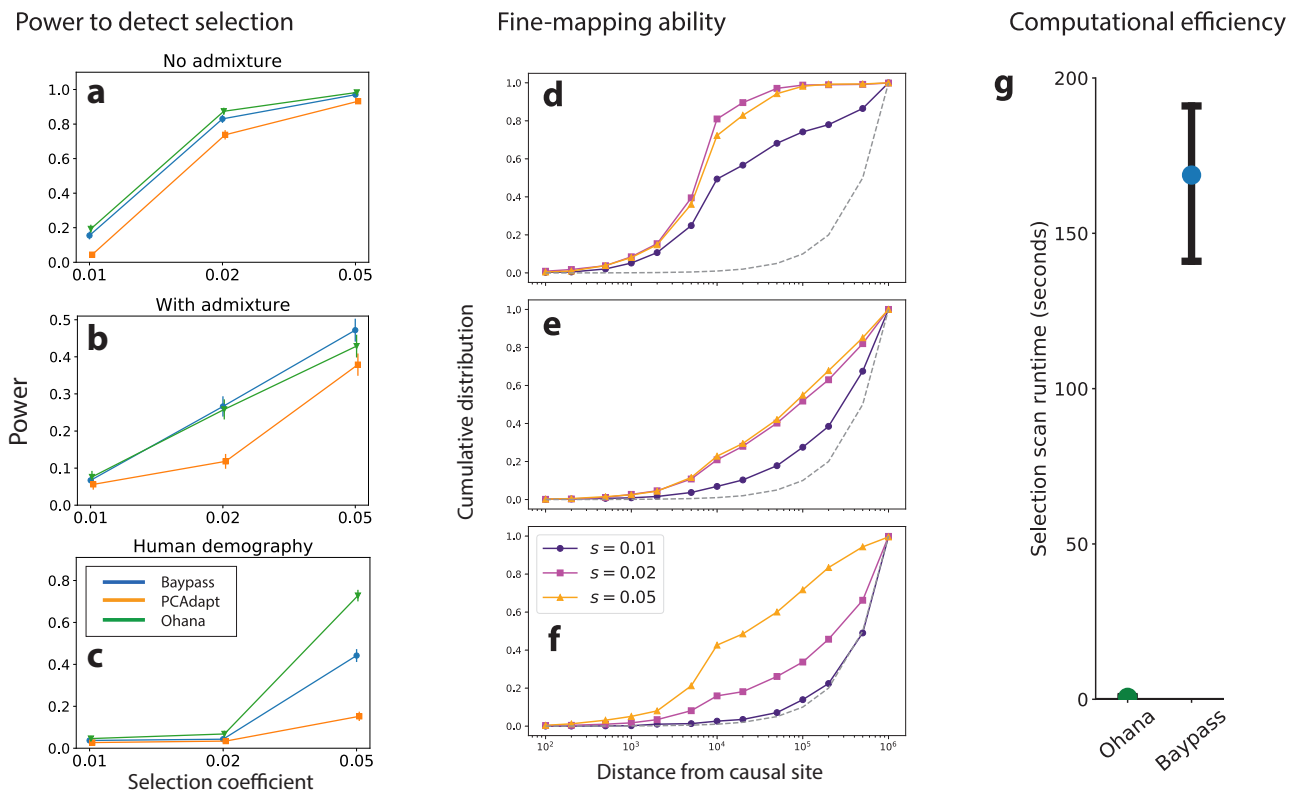

FIG. 3. Simulation tests of Ohana performance and efficiency in detecting and mapping selected sites. (a-c) Power to detect selection relative to two comparable methods, BayPass and pcadapt. Error bars are 95% CIs; (d-f) efficacy of Ohana to fine-map the causal site; (g) computational efficiency compared to that of BayPass. Error bars are 5-95th percentiles.

Analysis of real data

We identified regions in the genome that are likely to have been under the influence of positive selection using a merged dataset containing several population panels from phase 3 of the 1000 Genomes Project (Consortium *et al.*, 2015). We randomly selected 64 genomes from each of 4 populations from the 1000 Genomes project: the British from Great Britain (GBR), the Han Chinese from Beijing (CHB), the Yoruba Africans (YRI) and the admixed Mexican-Americans from Los Angeles (MXL) (the number 64 was chosen because it was the size of the smallest panel). We only included variable sites with no missing data and a minimum allele frequency of 0.05 across the entire merged panel. In total, we analyzed 5,601,710 variable sites across the autosomal genome. We inferred genome-wide allele frequencies and covariances for the latent ancestry components as described in the Methods section, using $K=4$. To scan for covariance outliers, we performed four hypothesis-driven scans, in which we specifically searched for selection separately in each of the four inferred ancestry components in our dataset (Fig. 5, Table 1).

After running these scans, we queried the CADD server (Rentzsch *et al.*, 2019) to obtain functional, conservation and regulatory annotations for the top candidate SNPs, including SIFT (Sim *et al.*, 2012), PolyPhen (Adzhubei *et al.*, 2013), GERP (Davydov *et al.*, 2010),

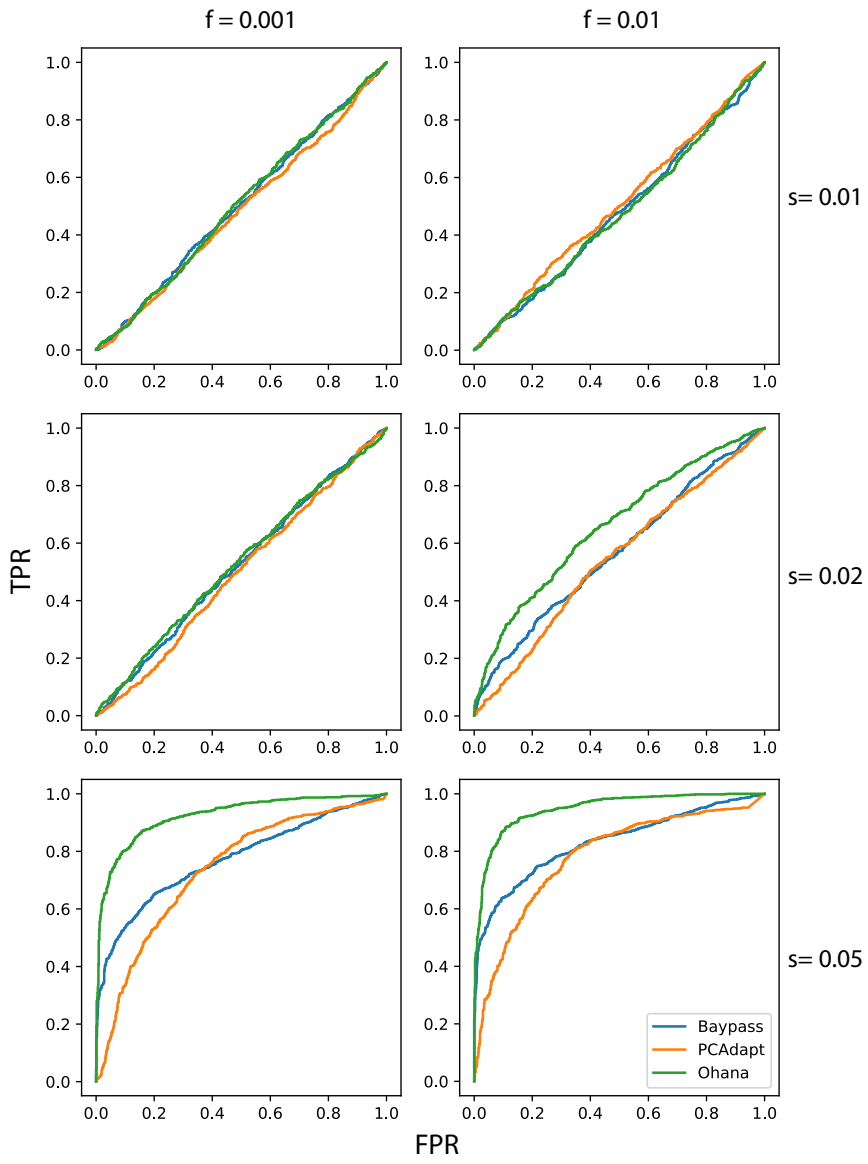

FIG. 4. ROC curves for Ohana vs. state-of-the-art methods, assessed using simulations with various values of the initial allele frequency at the beginning of selection (f) and different selection coefficients (s). Here the demographic model used was our human model (with selection in the Native American lineage), vs. other demographic models considered (i.e. basic tree without and with admixture, Figs. S1 and S2, respectively.

PhastCons (Siepel *et al.*, 2005), PhyloP (Pollard *et al.*, 2010) and Segway (Hoffman *et al.*, 2012) annotations, so as to find the changes most likely to be disruptive. We discuss some of these below. We also queried the GTEx cis-eQTL database (Lonsdale *et al.*, 2013), the UK Biobank GeneATLAS (Canela-Xandri *et al.*, 2018), and the GWAS catalog (MacArthur *et al.*, 2017), to look for trait-associated SNPs. We particularly focus on SNPs that have both high log-likelihood ratios in favor of positive selection ($LLRS > 15$) and high CADD scores in favor of functional disruption (> 10).

Below, we describe some of the top SNPs with high LLRS and their surrounding regions, for those cases in which available genic, expression

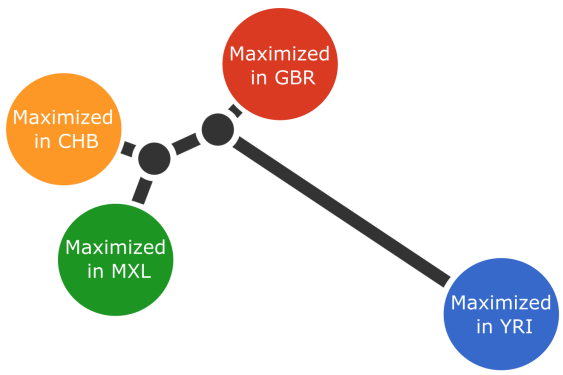

FIG. 5. Inferred unrooted tree of latent ancestry components for the analysis including the CHB, YRI, MXL and GBR genomic panels. We label each component by the population in which it is maximized, but emphasize that the components and the populations are not equivalent entities.

or regulatory information can provide us some clue as to the possible organismal function that may have been affected by the selective event. We particularly focus on the Native American ancestry scan (Figure 6), as few selection scans have been performed in this population, but also briefly summarize the results from the other scans.

European ancestry scan

Results for the top 30 loci in the European ancestry scan are presented in Table S1. Most loci have been previously shown to be under selection in Europeans populations, including SLC45A2, SLC24A5, BNC2, the OCA2/HERC2 region, the LCT/MCM6 region and the TLR region (Barreiro *et al.*, 2009; Bersaglieri *et al.*, 2004; Mathieson *et al.*, 2015; Vernot and Akey, 2014; Voight *et al.*, 2006). We notice that, in several cases, the presumed causal SNP previously identified in the literature coincides with the SNP with the strongest selection signal. This is the case, for example, for rs1426654 (SLC24A5) (Kimura *et al.*, 2009; Lamason *et al.*, 2005) and for rs16891982 (SLC45A2) (Branicki *et al.*, 2008). This suggest that the top SNPs for other loci, for which the causal SNPs are not yet known, may be good candidates for further tests of functional effects.

East Asian ancestry scan

We also performed a scan where we sought to recover SNPs that were candidates for selection in the ancestry component that is prevalent among our CHB samples. Results for the top 30 loci in this scan are in Table S2. Here, we also recover several candidate regions that have been previously reported in East Asian selection scans, including ABCC11, POU2F3, ADH1B, FADS1 and TARBP1 (Liu *et al.*, 2018; Ohashi *et al.*, 2011; Peng *et al.*, 2010; Refoyo-Martínez *et al.*, 2019; Vernot and Akey, 2014). Here, as in the previous scan, the top-scoring SNPs also tend to have the strongest phenotypic associations. For example, the highest scoring SNP (rs17822931) is the well-known missense variant in ABCC11, which is involved in sweat and earwax production (Yoshiura *et al.*, 2006).

Yoruba / ancestral non-African ancestry scan

Because our algorithm relies on an unrooted tree of the ancestry components (Fig. 5), we cannot distinguish between SNPs under positive selection in the terminal branch leading to the Yoruba / Sub-Saharan Africans and the ancestral non-African branch (Table S3). Nevertheless, more careful study of the allele frequencies of these SNPs in other populations may serve to distinguish among these scenarios in the future.

As in the the other ancestry scans, we also retrieve several genes that have been previously reported in positive selection studies. For example, the highest-scoring SNP is a missense variant in SLC39A4 (rs1871534) that has been reported to be under selection in Sub-Saharan Africa and to be causal for zinc deficiency (Engelken *et al.*, 2014).

Native American ancestry scan

The Native American ancestry scan yielded several novel candidates for positive selection (Table S4). As this ancestry has been less studied than the other aforementioned populations in the selection scan literature, we highlight some of the more interesting regions here.

The top SNP (rs140736443) is located in an intron of LINC00871. This SNP does not have a high CADD score (= 1.125), but is very close to a SNP (rs10133371) with a very high LLRS (= 16.54) and CADD score (= 15.99). This SNP is also intronic but is highly conserved in primates (PhastCons = 0.972) and is located in a GERP conserved element (P = 1.92e-21). LINC00871 is a long non-coding RNA gene that has been associated with number of children born (Barban *et al.*, 2016), although the specific trait-associated SNP in that study does not have a high LLRS. This gene also contains a suggestive association to longevity in females (Zeng *et al.*, 2018), although this study was under-powered to retrieve genome-wide significant associations.

The third top SNP (rs2316155) has a low CADD score (= 0.633) but is located near two SNPs with high LLRS (rs1466182, rs1466183) that overlap a regulatory region (ENSR00000088366) and have high CADD scores (= 16.8 and 19.5, respectively). Both of these SNPs have high PhastCons conservation scores across primates, mammals and vertebrates, and both overlap a GERP conserved element.

The sixth top SNP (rs10508343) has a low CADD score but lies very close to another SNP (rs17143255) with a high LLRS and a very high CADD score (= 14.16). The latter is an intergenic SNP overlapping a GERP conserved element between LINC00708 and GATA3, which has been shown to lead to abnormal hair shape and growth in mice when mutated (Kaufman *et al.*, 2003). Interestingly, SNPs overlapping LINC00708 have been recently associated with hair shape in a GWAS of admixed Latin Americans (Adhikari et al. 2016). There is also a high-LLRS SNP in this region that is significantly associated with the response to treatment for acute lymphoblastic leukemia (rs10508343) (Yang *et al.*, 2009).

The seventh top SNP (rs16959274) is a GTEx eQTL for GOLGA8A for tibial artery and skeletal muscle, and for GOLGA8B in pancreas. These two genes are members of the same gene family, and code for an auto-antigen localized in the surface of the Golgi complex (Eystathioy *et al.*, 2000).

The tenth top SNP (rs12580697) is a GTEx eQTL for TMTC1 in whole blood and has

a moderately high CADD score (= 8.676). TMTC1 codes for an endoplasmic reticulum transmembrane protein that is involved in calcium homeostasis (Sunryd *et al.*, 2014).

The eleventh top SNP (rs75607199) has a low CADD score but lies near three other SNPs (rs41325445, rs4901738 and rs59250732) with almost equally high LLRS and high CADD scores (= 13.49, 19.7 and 12.67, respectively). All of these SNPs are intronic and overlap OTX2-AS1, a long non-coding RNA gene. The SNP with the highest CADD score (rs4901738) is located in a GERP conserved element and has high PhastCons conservation scores across primates and mammals (>0.98). They all lie upstream of OTX2, coding for a developmental transcription factor implicated in microphtalmia (Ragge *et al.*, 2005), retinal dystrophy (Vincent *et al.*, 2014) and pituitary hormone deficiency (Diaczok *et al.*, 2008). In mice, this gene has been found to be involved in the embryonic development of the brain (Boncinelli *et al.*, 1993), photoreceptor development (Nishida *et al.*, 2003) and susceptibility to stress (Peña *et al.*, 2017).

The fourteenth top SNP (rs78441257) has a fairly high CADD score (= 12.72) and lies in a GERP conserved element of the 3' UTR of LRAT. This gene is implicated in retinal dystrophy (Thompson *et al.*, 2001) and retinitis pigmentosa (Sénéchal *et al.*, 2006).

The fifteenth top SNP (rs1919550) is a GTEx eQTL for FBXO40 in whole blood, but does not have a high CADD score. However, it lies near a SNP (rs9813391) with a high LLRS that leads to a nonsynonymous change (R145Q) in ARGFX - a homeobox gene - and another SNP (rs4676737) with both a high LLRS and high CADD score (= 14.07) overlapping a repressor region in an intron of FBXO40. The latter SNP is a GTEx eQTL for IQCB1 in fibroblasts, muscular esophagus and thyroid. IQCB1 is associated with Senor-Loken syndrome (Otto *et al.*, 2005), a ciliopathic eye disorder.

The twenty-second top SNP (rs4946567) is an eQTL of TBC1D32 in cerebellar brain. This SNP has a high CADD score (= 11.02) and is conserved across vertebrates (vertebrate PhyloP = 0.916, vertebrate PhastCons = 0.747). Interestingly, the region in which it is located also harbors signature of selection in Yucatan miniature pigs (Kim *et al.*, 2015; Kwon *et al.*, 2019). TBC1D32 plays a role in cilia assembly (Ko *et al.*, 2010) and may be involved in ciliopathic congenital abnormalities, including midline cleft, microcephaly, and microphthalmia (Adly *et al.*, 2014).

The twenty-third and twenty-fourth top SNPs (rs5758430, rs4822061) are close to each other and lie in a large region with several high-LLRS SNPs. They are both linked GTEx eQTLs to several genes in a variety of different tissues. They are also both significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank GeneATLAS

($P < 10^{-8}$). Although these SNPs do not have particularly high CADD scores, there are several neighboring linked high-LLRS, high-CADD SNPs with significant associations to the same traits, including splice site and missense mutations. We also find two significantly-associated SNPs in the GWAS catalog in this region ($P < 10^{-8}$): rs4822024 is associated with Vitiligo (Jin *et al.*, 2012) and rs13054099 is associated with neuroticism (Nagel *et al.*, 2018).

We also repurposed our aforementioned neutral simulations under human demography to estimate the false discovery rate (FDR) of these selected variants in aggregate. We estimate the expected number of SNPs to exceed a threshhold logLR T, assuming a genome length of 3×10^9 bp, a simple LD structure of 2Mbp blocks, and ascertaining the SNP with the top logLR within each block. Under this approach, we find that at the cutoffs of top 1, 5, 10, 20, and 30 SNPs, the FDR is approximately 0.0% (i.e., up to simulation precision), 15.1%, 22.6%, 30.1%, and 42.6%, respectively. We encourage users of the program to do similar simulations for estimating false discovery rates for inferences made on their specific data sets.

Signals of selection in Mexican ancestry (MXL)

We wanted to verify that our method was picking up signals of selection that were supported by alternative methods not explicitly relying on single-SNP patterns of population differentiation. For this, we used the program CLUES (Stern *et al.*, 2019), which relies on a likelihood approach based on reconstructed approximation to the ancestral recombination graph along the genome (Table S5). We applied CLUES using parameters corresponding to the demographic history of Mexican-ancestry (MXL) individuals in the 1000 Genomes Project (i.e., effective population size inferred by the method Relate (Speidel *et al.*, 2019)) to the set of hits identified using Ohana with selection acting on the Native American branch. We found that 9 out of the 10 tested SNPs showed significant ($p < 0.05$) signals of positive selection in MXL, under the asymptotic interpretation of the log-likelihood ratio statistics, supporting the evidence that these top hits in Native American ancestry have been targets of selection.

To learn more about the mode and time-frame of selection in these loci, we also used CLUES (Stern *et al.*, 2019) to estimate the trajectory of allele frequency changes for the 10 loci in the Native American component mentioned in Table 1 (Figure S6). In all cases, the estimated allele frequency trajectory was compatible with relative old selection leading to alleles with current day intermediate frequencies, typically between 0.4 and 0.6, i.e incomplete sweeps. The fact that we only detect incomplete sweep might be related to the filtering procedure we have used to eliminate SNPs with small MAF . The fastest change in allele frequency is found for the SNP in CSMD1

(s71523639) which currently is at frequency close to 0.5 but was at a frequency of approx. zero 700 generation ago, suggesting relative strong selection on a *de novo* mutation.

Discussion

We describe a new modeling framework that can detect signals of positive selection on ancestry components, using allele frequency patterns across admixed populations. It models admixture explicitly and works with an arbitrary number of populations with or without admixed ancestries. It also does not rely on labeling of samples into particular populations, and allows for testing of different positive selection models reflecting different historical adaptive hypotheses. It is in many ways similar to the Bayesian methods by Coop *et al.* (2010) and Günther and Coop (2013) in the structure of the likelihood function. The major differences being the use of optimization of the likelihood function in Ohana instead of MCMC used by Coop *et al.* (2010) and Günther and Coop (2013), which provides some computational advantages. The methods also differ in other ways, including the enforcement of a tree-structure in Ohana, the use of ancestry components to model selection in hypothesized ancestral populations in Ohana, and the functionality to perform branch-specific detection of selection, or detection of selection in multiple branches if one has an a priori selection hypothesis one wants to test.

The run-time complexity of our method is linear in the number of markers, but we still recommend a high-performance cluster to be used in a typical genomic analysis. With parallelization, a selection scan takes <10 minutes to analyze a 6 Mbp genome for <10 ancestry components using 100 cores. An example of how to perform this parallelization can be found on the project's wiki page on GitHub: `https://github.com/jade-cheng/ohana`

Our method works by testing for selection in specific components of the ancestry covariance matrix. We also explored what would occur if we used a likelihood model in which the ancestry covariance matrix was multiplied by a scalar, so as to find "global" candidates for selection rather than testing for selection in particular ancestries. We found however, that this was not an optimal way to detect candidates for selection, as it is biased towards finding many variants in highly drifted populations, likely because the excess variance in the Wright-Fisher process is not well modelled by the multivariate Gaussian assumption, especially at the boundaries of fixation and extinction.

We note, however, that the latent ancestry components inferred by Ohana and other similar programs cannot be strictly interpreted as corresponding to existing populations (now or in the past) and that the labels we assign to them ("European", "Asian", "African", etc.) are largely for convenience. This is especially true when the

studied individuals are not descended from recent admixture events among highly differentiated populations, so care should be taken in the interpretation of the identity of these components. We refer the reader to Lawson *et al.* (2018); Mathieson and Scally (2020) for more in-depth studies and discussions on the assumptions and limitations of latent ancestry inference methods.

We note that there is currently some debate in the field on the possibility that F_{ST} outliers could be caused by negative selection in various forms (see e.g, Johri *et al.* (2020), Schrider (2020), Matthey-Doret and Whitlock (2019)). Although it has been argued that such an effect is unlikely to explain F_{ST} outliers in real data (Schrider (2020), Matthey-Doret and Whitlock (2019)), our method will be similarly challenged by this effect, as the information used is very similar to that of F_{ST} outlier scans.

When specifically testing for candidates for selection in the "European", "East Asian" and "Sub-Saharan African" components, we identified several well-known candidates under positive selection, including OCA2, SLC24A5, SLC45A2, ABCC1 and SLC39A4. Many of our top scoring SNPs were also previously known to be causal for particular traits, as in the case of rs17822931 in ABCC11 in East Asians, rs16891982 in SLC45A2 in Europeans, rs1426654 in SLC24A5 in Europeans and rs1871534 in SLC39A4 in Sub-Saharan Africans.

Our scan for positive selection in the Native American ancestry component of Latin Americans yielded several novel candidates for adaptation in the human past. We found signatures of selection near genes involved in fertility (LINC00871), hair shape and growth (LINC00708), immunity (GOLGA8A / GOLGA8B and IRAK4), vision (OTX2 and LRAT), the nervous system (MDGA2) and various ciliopathies (IQCB1 and TBC1D32). Several of the highest-scoring SNPs in the candidate regions are known to be cis-eQTLs to their nearby genes, as is the case for rs12580697 / TMTC1 (involved in calcium homestasis) and rs4676737 / IQCB1 (involved in ciliopathies). We also found individual SNPs with high likelihood ratio scores in favor of selection that are associated with a variety of phenotypes, including rs12426688 (fat percentage), rs10508343 (response to leukemia treatment), rs34670506 (insomnia), and the cluster of high-scoring SNPs that include rs5758430 and rs4822061, among other SNPs. This particular cluster is especially interesting, as the SNPs in the region are associated with a variety of traits related to body fat distribution, food intake and white blood cells, suggesting a possible underlying phenotype related to these traits that may have driven an adaptive event. Estimates of the FDR suggest that the lion's share of these SNPs are selected, especially towards the higher end (e.g., the top 8 SNPs have an FDR of $<10\%$).

We provide a list of functional annotations for all the SNPs with high LLRS (>15) within a 2Mb region surrounding each of the top genome-wide SNPs, including CADD, conservation, regulatory and protein deleteriousness scores, which we hope will guide future functional validation studies in these regions of the genome (Table S6).

In conclusion, Ohana provides a fast and flexible selection-detection and hypothesis-testing framework. It is easy to use and has in-built visualization functionalities to explore patterns on a genome-wide and locus-specific scale. We believe it will be a useful tool for biologists aiming to study positive selection and understanding the genomic basis of adaptation, particularly in cases where demographic histories are complex or not well characterized.

Supplementary Material

Supplementary tables S1-S6 and figures S1-S5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors are grateful to Thomas Mailund, Mikkel Schierup, Christian Storm Pedersen, and the GenomeDK staff for their support during the course of this research. We also thank Leo Speidel for providing coalescence time and effective population size estimates for the 1000 Genomes Project. FR was funded by a Villum Fonden Young Investigator award (project no. 00025300). This research was supported by NIH grant R01GM138634.

Data Availability

All data analyzed in this manuscript are previously published publicly available data. The software Ohana described in the paper is open source and available at https://github.com/jade-cheng/ohana.

References

Adly, N., Alhashem, A., Ammari, A., and Alkuraya, F. S. 2014. Ciliary genes tbc 1 d 32/c6orf170 and sclt 1 are mutated in patients with ofd type ix. *Human mutation*, 35(1): 36–40.

Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. 2013. Predicting functional effect of human missense mutations using polyphen-2. *Current protocols in human genetics*, 76(1): 7–20.

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. 2002. Interrogating a high-density snp map for signatures of natural selection. *Genome research*, 12(12): 1805–1814.

Barban, N., Jansen, R., De Vlaming, R., Vaez, A., Mandemakers, J. J., Tropf, F. C., Shen, X., Wilson, J. F., Chasman, D. I., Nolte, I. M., *et al.* 2016. Genome-wide analysis identifies 12 loci influencing human reproductive behavior. *Nature genetics*, 48(12): 1462–1472.

Barreiro, L. B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J. K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., *et al.* 2009. Evolutionary dynamics of human toll-like receptors and their different contributions to host defense. *PLoS genetics*, 5(7).

Beaumont, M. A. and Balding, D. J. 2004. Identifying adaptive genetic divergence among populations from genome scans. *Molecular ecology*, 13(4): 969–980.

Beaumont, M. A. and Nichols, R. A. 1996. Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 263(1377): 1619–1626.

Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E., and Hirschhorn, J. N. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *The American Journal of Human Genetics*, 74(6): 1111–1120.

Boncinelli, E., Gulisano, M., and Broccoli, V. 1993. Emx and otx homeobox genes in the developing mouse brain. *Journal of neurobiology*, 24(10): 1356–1366.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., and SanCristobal, M. 2010. Detecting selection in population trees: the lewontin and krakauer test extended. *Genetics*, 186(1): 241–262.

Branicki, W., Brudnik, U., Draus-Barini, J., Kupiec, T., and Wojas-Pelc, A. 2008. Association of the slc45a2 gene with physiological human hair colour variation. *Journal of human genetics*, 53(11-12): 966–971.

Canela-Xandri, O., Rawlik, K., and Tenesa, A. 2018. An atlas of genetic associations in uk biobank. *Nature genetics*, 50(11): 1593–1599.

Cheng, J. Y., Mailund, T., and Nielsen, R. 2017a. Fast admixture analysis and population tree estimation for snp and ngs data. *Bioinformatics*, 33(14): 2148–2155.

Cheng, X., Xu, C., and DeGiorgio, M. 2017b. Fast and robust detection of ancestral selective sweeps. *Molecular ecology*, 26(24): 6871–6891.

Cholesky, A.-L. 1910. Sur la résolution numérique des systèmes d'équations linéaires. *Bulletin de la Sabix. Société des amis de la Bibliothèque et de l'Histoire de l'École polytechnique*, (39): 81–95.

Consortium, . G. P. *et al.* 2015. A global reference for human genetic variation. *Nature*, 526(7571): 68–74.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4): 1411–1423.

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., and Batzoglou, S. 2010. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS computational biology*, 6(12).

DeGiorgio, M., Huber, C. D., Hubisz, M. J., Hellmann, I., and Nielsen, R. 2016. Sweepfinder2: increased sensitivity, robustness and flexibility. *Bioinformatics*, 32(12): 1895–1897.

Diaczok, D., Romero, C., Zunich, J., Marshall, I., and Radovick, S. 2008. A novel dominant negative mutation of otx2 associated with combined pituitary hormone deficiency. *The Journal of Clinical Endocrinology & Metabolism*, 93(11): 4351–4359.

Duforet-Frebourg, N., Luu, K., Laval, G., Bazin, E., and Blum, M. G. 2016. Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular biology and evolution*, 33(4): 1082–1093.

Engelken, J., Carnero-Montoro, E., Pybus, M., Andrews, G. K., Lalueza-Fox, C., Comas, D., Sekler, I., de la Rasilla, M., Rosas, A., Stoneking, M., *et al.* 2014. Extreme population differences in the human zinc transporter zip4 (slc39a4) are explained by positive selection in sub-saharan africa. *PLoS genetics*, 10(2).

Eystathioy, T., Jakymiw, A., Fujita, D. J., Fritzler, M. J., and Chan, E. K. 2000. Human autoantibodies to a novel golgi protein golgin-67: high similarity with golgin-95/gm 130 autoantigen. *Journal of autoimmunity*, 14(2): 179–187.

Fariello, M. I., Boitard, S., Naya, H., SanCristobal, M., and Servin, B. 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. *Genetics*, 193(3): 929–941.

Fay, J. C. and Wu, C.-I. 2000. Hitchhiking under positive darwinian selection. *Genetics*, 155(3): 1405–1413.

Felsenstein, J. 1985. Phylogenies and the comparative method. *The American Naturalist*, 125(1): 1–15.

Foll, M. and Gaggiotti, O. 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a bayesian perspective. *Genetics*, 180(2): 977–993.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A., and Excoffier, L. 2014. Widespread signals of convergent adaptation to high altitude in asia and america. *The American Journal of Human Genetics*, 95(4): 394–407.

Fu, Y.-X. and Li, W.-H. 1993. Statistical tests of neutrality of mutations. *Genetics*, 133(3): 693–709.

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M. E., Korneliussen, T. S., Gerbault, P., Skotte, L., Linneberg, A., *et al.* 2015. Greenlandic inuit show genetic signatures of diet and climate adaptation. *Science*, 349(6254): 1343–1347.

Gautier, M. 2015. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, 201(4): 1555–1579.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., Bustamante, C. D., Project, . G., *et al.* 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29): 11983–11988.

Günther, T. and Coop, G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics*, 195(1): 205–220.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. 2009. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics*, 5(10).

Haller, B. C. and Messer, P. W. 2019. Slim 3: forward genetic simulations beyond the wright–fisher model. *Molecular biology and evolution*, 36(3): 632–637.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., and Noble, W. S. 2012. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5): 473.

Huber, C. D., DeGiorgio, M., Hellmann, I., and Nielsen, R. 2016. Detecting recent selective sweeps while controlling for mutation rate and background selection. *Molecular ecology*, 25(1): 142–156.

Jin, Y., Birlea, S. A., Fain, P. R., Ferrara, T. M., Ben, S., Riccardi, S. L., Cole, J. B., Gowan, K., Holland, P. J., Bennett, D. C., *et al.* 2012. Genome-wide association analyses identify 13 new susceptibility loci for generalized vitiligo. *Nature genetics*, 44(6): 676–680.

Johri, P., Charlesworth, B., Howell, E. K., Lynch, M., and Jensen, J. D. 2020. Revisiting the notion of deleterious sweeps. *bioRxiv*.

Kaufman, C. K., Zhou, P., Pasolli, H. A., Rendl, M., Bolotin, D., Lim, K.-C., Dai, X., Alegre, M.-L., and Fuchs, E. 2003. Gata-3: an unexpected regulator of cell lineage determination in skin. *Genes & development*, 17(17): 2108–2122.

Kiefer, J. 1953. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, 4(3): 502–506.

Kim, H., Song, K. D., Kim, H. J., Park, W., Kim, J., Lee, T., Shin, D.-H., Kwak, W., Kwon, Y.-j., Sung, S., *et al.* 2015. Exploring the genetic signature of body size in yucatan miniature pig. *PloS one*, 10(4).

Kim, Y. and Nielsen, R. 2004. Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167(3): 1513–1524.

Kim, Y. and Stephan, W. 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2): 765–777.

Kimura, R., Yamaguchi, T., Takeda, M., Kondo, O., Toma, T., Haneji, K., Hanihara, T., Matsukusa, H., Kawamura, S., Maki, K., *et al.* 2009. A common variation in edar is a genetic determinant of shovel-shaped incisors. *The American Journal of Human Genetics*, 85(4): 528–535.

Ko, H. W., Norman, R. X., Tran, J., Fuller, K. P., Fukuda, M., and Eggenschwiler, J. T. 2010. Broad-minded links cell cycle-related kinase to cilia assembly and hedgehog signal transduction. *Developmental cell*, 18(2): 237–247.

Kwon, D.-J., Lee, Y.-S., Shin, D., Won, K.-H., and Song, K.-D. 2019. Genome analysis of yucatan miniature pigs to assess their potential as biomedical model animals. *Asian-Australasian journal of animal sciences*, 32(2): 290.

Lamason, R. L., Mohideen, M.-A. P., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., *et al.* 2005. Slc24a5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science*, 310(5755): 1782–1786.

Lawson, D. J., Van Dorp, L., and Falush, D. 2018. A tutorial on how not to over-interpret structure and admixture bar plots. *Nature Communications*, 9(1): 1–11.

Librado, P. and Orlando, L. 2018. Detecting signatures of positive selection along defined branches of a population tree using lsd. *Molecular biology and evolution*, 35(6): 1520–1535.

Liu, S., Huang, S., Chen, F., Zhao, L., Yuan, Y., Francis, S. S., Fang, L., Li, Z., Lin, L., Liu, R., *et al.* 2018. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and chinese population history. *Cell*, 175(2): 347–359.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., *et al.* 2013. The genotype-tissue expression (gtex) project. *Nature genetics*, 45(6): 580.

Luu, K., Bazin, E., and Blum, M. G. 2017. pcadapt: an r package to perform genome scans for selection based on principal component analysis. *Molecular ecology resources*, 17(1): 67–77.

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., Junkins, H., McMahon, A., Milano, A., Morales, J., *et al.* 2017. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic acids research*, 45(D1): D896–D901.

Mathieson, I. and Scally, A. 2020. What is ancestry? *PLoS Genetics*, 16(3): e1008624.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., Novak, M., *et al.* 2015. Genome-wide patterns of selection in 230 ancient eurasians. *Nature*, 528(7583): 499–503.

Matthey-Doret, R. and Whitlock, M. C. 2019. Background selection and fst: Consequences for detecting local adaptation. *Molecular Ecology*, 28(17): 3902–3914.

McVean, G. 2007. The structure of linkage disequilibrium around a selective sweep. *Genetics*, 175(3): 1395–1406.

Nagel, M., Jansen, P. R., Stringer, S., Watanabe, K., de Leeuw, C. A., Bryois, J., Savage, J. E., Hammerschlag, A. R., Skene, N. G., Muñoz-Manchado, A. B., *et al.* 2018. Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nature genetics*, 50(7): 920–927.

Nelder, J. A. and Mead, R. 1965. A simplex method for function minimization. *The computer journal*, 7(4): 308–313.

Nielsen, R. 2005. Molecular signatures of natural selection. *Annu. Rev. Genet.*, 39: 197–218.

Nishida, A., Furukawa, A., Koike, C., Tano, Y., Aizawa, S., Matsuo, I., and Furukawa, T. 2003. Otx2 homeobox gene controls retinal photoreceptor cell fate and pineal gland development. *Nature neuroscience*, 6(12): 1255–1263.

Ohashi, J., Naka, I., and Tsuchiya, N. 2011. The impact of natural selection on an abcc11 snp determining earwax type. *Molecular biology and evolution*, 28(1): 849–857.

Otto, E. A., Loeys, B., Khanna, H., Hellemans, J., Sudbrak, R., Fan, S., Muerb, U., O'Toole, J. F., Helou, J., Attanasio, M., *et al.* 2005. Nephrocystin-5, a ciliary iq domain protein, is mutated in senior-loken syndrome and interacts with rpgr and calmodulin. *Nature*

genetics, 37(3): 282–288.

Peña, C. J., Kronman, H. G., Walker, D. M., Cates, H. M., Bagot, R. C., Purushothaman, I., Issler, O., Loh, Y.-H. E., Leong, T., Kiraly, D. D., *et al.* 2017. Early life stress confers lifelong stress susceptibility in mice via ventral tegmental area otx2. *Science*, 356(6343): 1185–1188.

Peng, Y., Shi, H., Qi, X.-b., Xiao, C.-j., Zhong, H., Runlin, Z. M., and Su, B. 2010. The adh1b arg47his polymorphism in east asian populations and expansion of rice domestication in history. *BMC evolutionary biology*, 10(1): 15.

Pickrell, J. and Pritchard, J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *Nature Precedings*, pages 1–1.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*, 20(1): 110–121.

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155(2): 945–959.

Przeworski, M., Coop, G., and Wall, J. D. 2005. The signature of positive selection on standing genetic variation. *Evolution*, 59(11): 2312–2323.

Racimo, F. 2016. Testing for ancient selection using cross-population allele frequency differentiation. *Genetics*, 202(2): 733–750.

Racimo, F., Marnetto, D., and Huerta-Sánchez, E. 2017. Signatures of archaic adaptive introgression in present-day human populations. *Molecular biology and evolution*, 34(2): 296–317.

Ragge, N. K., Brown, A. G., Poloschek, C. M., Lorenz, B., Henderson, R. A., Clarke, M. P., Russell-Eggitt, I., Fielder, A., Gerrelli, D., Martinez-Barbera, J. P., *et al.* 2005. Heterozygous mutations of otx2 cause severe ocular malformations. *The American Journal of Human Genetics*, 76(6): 1008–1022.

Refoyo-Martínez, A., da Fonseca, R. R., Halldórsdóttir, K., Árnason, E., Mailund, T., and Racimo, F. 2019. Identifying loci under positive selection in complex population histories. *Genome research*, 29(9): 1506–1520.

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., and Kircher, M. 2019. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic acids research*, 47(D1): D886–D894.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T., Altshuler, D., and Lander, E. 2006. Positive natural selection in the human lineage. *science*, 312(5780): 1614–1620.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., *et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164): 913–918.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4): 406–425.

Schrider, D. R. 2020. Background Selection Does Not Mimic the Patterns of Genetic Diversity Produced by Selective Sweeps. *Genetics*, 216(2): 499–519.

Self, S. G. and Liang, K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398): 605–610.

Sénéchal, A., Humbert, G., Surget, M.-O., Bazalgette, C., Bazalgette, C., Arnaud, B., Arndt, C., Laurent, E., Brabet, P., and Hamel, C. P. 2006. Screening genes of the retinoid metabolism: novel lrat mutation in leber congenital amaurosis. *American journal of ophthalmology*, 142(4): 702–704.

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., *et al.* 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and

yeast genomes. *Genome research*, 15(8): 1034–1050.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P. C. 2012. Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(W1): W452–W457.

Speidel, L., Forest, M., Shi, S., and Myers, S. R. 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, 51(9): 1321–1329.

Stern, A. J., Wilton, P. R., and Nielsen, R. 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from dna sequence data. *PLoS genetics*, 15(9): e1008384.

Sunryd, J. C., Cheon, B., Graham, J. B., Giorda, K. M., Fissore, R. A., and Hebert, D. N. 2014. Tmtc1 and tmtc2 are novel endoplasmic reticulum tetratricopeptide repeat-containing adapter proteins involved in calcium homeostasis. *Journal of Biological Chemistry*, 289(23): 16085–16099.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics*, 123(3): 585–595.

Thompson, D. A., Li, Y., McHenry, C. L., Carlson, T. J., Ding, X., Sieving, P. A., Apfelstedt-Sylla, E., and Gal, A. 2001. Mutations in the gene encoding lecithin retinol acyltransferase are associated with early-onset severe retinal dystrophy. *Nature genetics*, 28(2): 123–124.

Vernot, B. and Akey, J. M. 2014. Resurrecting surviving neandertal lineages from modern human genomes. *Science*, 343(6174): 1017–1021.

Vincent, A., Forster, N., Maynes, J. T., Paton, T. A., Billingsley, G., Roslin, N. M., Ali, A., Sutherland, J., Wright, T., Westall, C. A., *et al.* 2014. Otx2 mutations cause autosomal dominant pattern dystrophy of the retinal pigment epithelium. *Journal of medical genetics*, 51(12): 797–805.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. 2006. A map of recent positive selection in the human genome. *PLoS biology*, 4(3).

Yang, J. J., Cheng, C., Yang, W., Pei, D., Cao, X., Fan, Y., Pounds, S. B., Neale, G., Treviño, L. R., French, D., *et al.* 2009. Genome-wide interrogation of germline genetic variation associated with treatment response in childhood acute lymphoblastic leukemia. *Jama*, 301(4): 393–403.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., *et al.* 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. *science*, 329(5987): 75–78.

Yoshiura, K.-i., Kinoshita, A., Ishida, T., Ninokata, A., Ishikawa, T., Kaname, T., Bannai, M., Tokunaga, K., Sonoda, S., Komaki, R., *et al.* 2006. A snp in the abcc11 gene is the determinant of human earwax type. *Nature genetics*, 38(3): 324–330.

Zeng, Y., Nie, C., Min, J., Chen, H., Liu, X., Ye, R., Chen, Z., Bai, C., Xie, E., Yin, Z., *et al.* 2018. Sex differences in genetic associations with longevity. *JAMA network open*, 1(4): e181670–e181670.

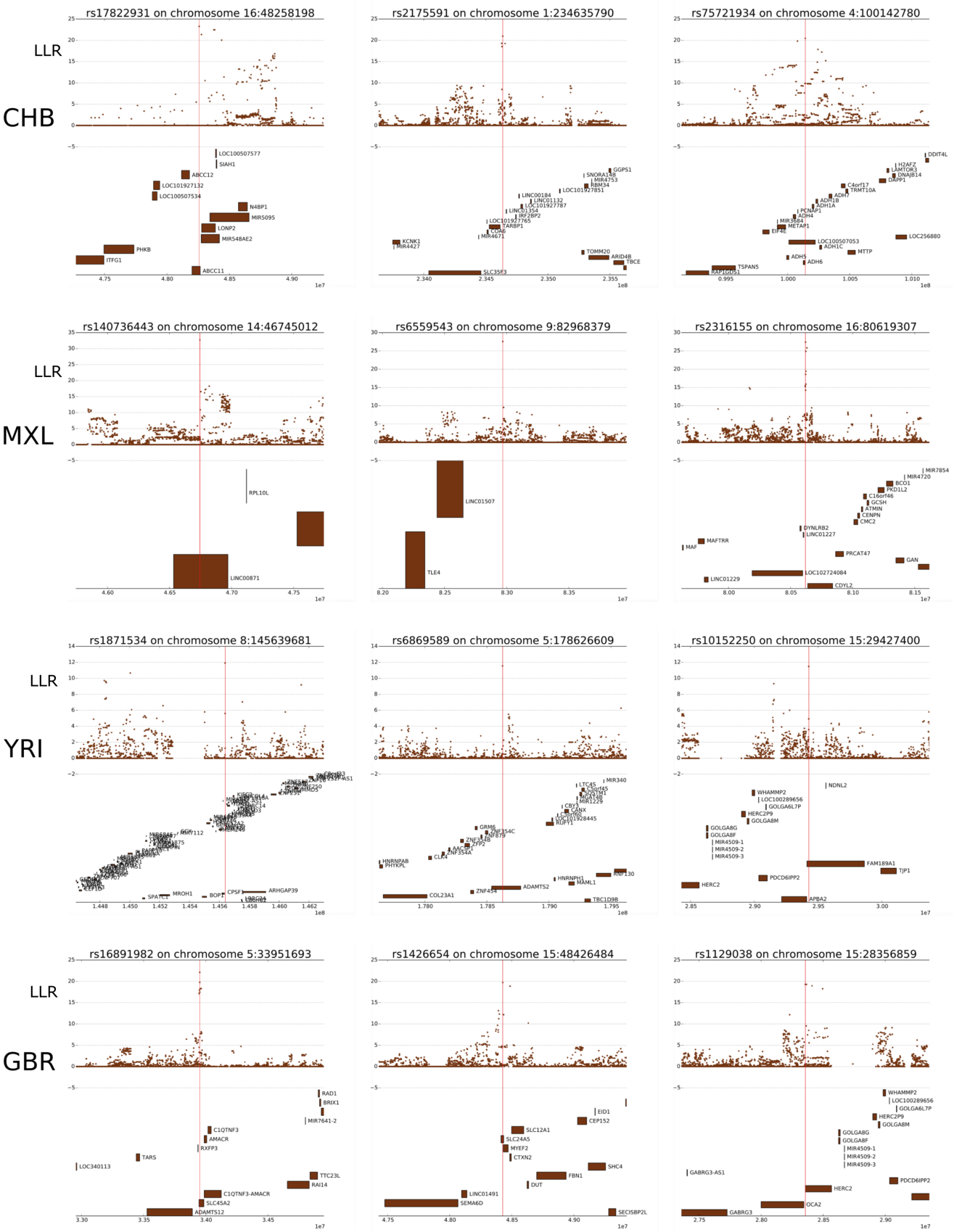

FIG. 6. Top 5 annotated peaks in each of the ancestry-specific selection studies. MXL-specific = scan for selection in Native American ancestry of MXL. GBR-specific = scan for selection in European ancestry of GBR. CHB-specific = scan for selection in CHB ancestry of CHB. YRI-specific = scan for selection in Yoruba African ancestry or ancestral non-African ancestry. We analyzed 5,601,710 variable sites across the autosomal genomes. We inferred genome-wide allele frequencies and covariances as described in the Methods section. We applied a likelihood model for each SNP by rescaling all variances and covariances by a scalar multiplier α. Descriptions of each candidate region are in Table 1. LLR = Log-likelihood ratio score.

Table 1. Top 10 most differentiated SNPs from each of the ancestry-specific scans. LLRS = log-likelihood ratio score for positive selection.

chr	pos	rsid	LLRS	target ancestry	nearest gene
5	33951693	rs16891982	22.085902	European	SLC45A2
15	48426484	rs1426654	19.707464	European	SLC24A5
15	28356859	rs1129038	19.290553	European	HERC2
15	28495956	rs12912427	18.270213	European	HERC2
9	16792200	rs10962596	15.819739	European	BNC2
1	1385211	rs1312568	15.066101	European	ATAD3C
2	136407479	rs1446585	14.957582	European	R3HDM1
2	136616754	rs182549	14.629386	European	MCM6
1	204784969	rs3940119	14.393216	European	NFASC
4	38798648	rs5743618	14.38681	European	TLR1
16	48258198	rs17822931	23.271759	CHB	ABCC11
16	48375777	rs6500380	22.474103	CHB	LONP2
1	234635790	rs2175591	20.95541	CHB	TARBP1
4	100142780	rs75721934	20.453247	CHB	LOC100507053
11	61579427	rs72643557	20.114033	CHB	FADS1
11	120154631	rs12224052	19.696284	CHB	POU2F3
21	43974948	rs228088	19.518001	CHB	SLC37A1
11	133043841	rs79802711	19.157192	CHB	OPCML
5	128016573	rs79478220	18.476104	CHB	FBN2
19	51441759	rs11084040	18.158963	CHB	KLK5
14	46745012	rs140736443	32.730697	Native American	LINC00871
9	82968379	rs6559543	27.584847	Native American	LINC01507
16	80619307	rs2316155	27.399123	Native American	LINC01227
14	21647765	rs77549780	27.355769	Native American	LINC00641
12	14189549	rs12425115	25.867367	Native American	GRIN2B
10	8150713	rs10508343	25.609772	Native American	GATA3
15	34936250	rs16959274	25.424824	Native American	GOLGA8B
8	4490837	rs71523639	24.59957	Native American	CSMD1
1	14301862	rs72640512	24.455822	Native American	PRDM2
12	29817716	rs12580697	23.967094	Native American	TMTC1
8	145639681	rs1871534	11.906794	Yoruba / Ancestral Non-African	SLC39A4
5	178626609	rs6869589	11.541667	Yoruba / Ancestral Non-African	ADAMTS2
15	29427400	rs10152250	11.48232	Yoruba / Ancestral Non-African	FAM189A1
1	1106112	rs6670693	11.447873	Yoruba / Ancestral Non-African	TTLL10
4	3666494	rs58827274	11.341367	Yoruba / Ancestral Non-African	LOC100133461
17	2631985	rs4790359	11.118134	Yoruba / Ancestral Non-African	PAFAH1B1
9	136769888	rs2789823	11.031687	Yoruba / Ancestral Non-African	VAV2
6	169656029	rs6930377	10.824098	Yoruba / Ancestral Non-African	THBS2
17	29350769	rs8073072	10.794224	Yoruba / Ancestral Non-African	RNF135
5	173642871	rs10067518	10.787147	Yoruba / Ancestral Non-African	HMP19