Detecting Selection in Multiple Populations by Modeling Ancestral Admixture Components

Cheng, Jade Yu; Stern, Aaron J.; Racimo, Fernando; Nielsen, Rasmus

Published in:
Molecular Biology and Evolution

DOI:
10.1093/molbev/msab294

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY-NG

Citation for published version (APA):
Detecting Selection in Multiple Populations by Modeling Ancestral Admixture Components

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

1Lundbeck GeoGenetics Centre, Globe Institute, University of Copenhagen, Copenhagen, Denmark
2Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA
3Graduate Group in Computational Biology, University of California, Berkeley, Berkeley, CA, USA
4Department of Statistics, University of California, Berkeley, Berkeley, CA, USA

*Corresponding author: E-mail: jade.yu.cheng@gmail.com.
Associate Editor: Yuseob Kim

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 article, 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, such as 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 (Voight et al. 2006; Sabeti et al. 2007), one of the characteristic genomic footprints left by a selective sweep (Kim and Stephan 2002; Kim and Nielsen 2004; 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 (Tajima 1989; Fu and Li 1993; Fay and Wu 2000; Nielsen 2005; DeGiorgio et al. 2016; Huber et al. 2016).

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 (Beaumont and Nichols 1996; Akey et al. 2002; Beaumont and Balding 2004). 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; Yi et al. 2010; Fumagalli et al. 2015; Racimo et al. 2017). Recent methods have allowed researchers to detect excess local differentiation on particular branches of a three-population tree (Yi et al. 2010; Racimo 2016), a four-population tree (Cheng, Xu, et al. 2017), or an arbitrarily large tree (Librado and Orlando 2018), albeit without modeling postsplit 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 (Dufour-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-Martinez 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. These methods identify 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 coestimates the population structure of the input panel and the allele frequencies of the ancestral admixture components through an unsupervised learning process (Cheng, Mailund, et al. 2017), 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.

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. 1a–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 2 Mbp 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 “X×X” statistic; and with pcadapt, the test statistic was the selection test P value, using K = 3 PCs (because there are four ancestral populations in each case). For all three methods, power increased uniformly with the value of the selection coefficient (fig. 1a–c). However, different demographic scenarios result in different power levels; for example, all methods were better-powered under the simple demographic scenario (model 1, figs. 1a and 2a) compared with selection preadmixture (model 1, figs. 1b and 2b) or in the human demographic model (model 3, figs. 1c and 2c).

pcadapt was significantly less well-powered than Ohana in most scenarios, for example, 72% versus 88% and 90% versus 98% power under moderate and strong selection, respectively (fig. 1a). In most scenarios, we found Ohana to have power equivalent to or greater than that of BayPass; for example, in models 1 and 2, we fail to reject that Ohana and BayPass have different power curves with 95% confidence (fig. 1a and b). However, in one simulation scenario, we found Ohana had significantly higher power than Baypass, under a model of human demography (fig. 1c). 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 figure 1a–c, we show results assuming 0.001 ≤ 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 figure 3 and supplementary figures S1 and S2, Supplementary Material online. Also note the overall low power in figure 3. 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. 1d–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 2 Mbp locus). We plot the empirical cumulative distribution of these distances for different values of the selection coefficient under each
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 6 kb of the max test statistic (e.g., >40% power for $s=0.01$ under model 1 and 25% power for $s=0.05$ under models 2 and 3; see fig. 1d–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. 1d and e); in contrast, under model 3, there is significantly higher power to fine-map sites with strong selection (fig. 1f); this dramatic difference may be due to the effects of demography on the pattern of hitchhiking surrounding the causal site. In figure 1d–f, we show results assuming $0.001 \leq f < 0.01$ and present full results illustrating other values of $f$ in supplementary figures S3–S5, Supplementary Material online.

**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. 1g). We found that Ohana was $>250\times$ faster than BayPass (mean selection scan runtimes: Ohana, 0.626 s [±0.008 s]; BayPass, 168.8 s [±2 s]; $N=1,000$ 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.

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 data set containing several population panels from phase 3 of the 1000 Genomes Project (1000 Genomes Project Consortium et al. 2015). We randomly selected 64 genomes from each of four 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 Materials and 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 data set (fig. 4 and 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), 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

---

**Fig. 3.** ROC curves for Ohana versus 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), versus other demographic models considered (i.e., basic tree without and with admixture, supplementary figs. S1 and S2, Supplementary Material online, respectively).
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, generic, expression, 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 (fig. 5), 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 supplementary table S1, Supplementary Material online. 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 (Bersaglieri et al. 2004; Voight et al. 2006; Barreiro et al. 2009; Vernot and Akey 2014; Mathieson et al. 2015). 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) (Lamason et al. 2005; Kimura et al. 2009) 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 supplementary table S2, Supplementary Material online. Here, we also recover several candidate regions that have been previously reported in East Asian selection scans, including ABCC11, POU2F3, ADH1B, FADS1, and TARBP1 (Peng et al. 2010; Ohashi et al. 2011; Vernot and Akey 2014; Liu et al. 2018; Refoyo-Martinez et al. 2019). 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. 4), 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 (supplementary table S3, Supplementary Material online). 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 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 (supplementary table S4, Supplementary Material online). 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 an SNP (rs101333731) 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 noncoding 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 (ENSR0000008366) 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 (rs105083434) 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 autoantigen 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 11th 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, 19.70, and 12.67, respectively). All of these SNPs are intronic and overlap OTX2-AS1, a long noncoding 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 microphthalmia (Ragge et al. 2005), retinal dystrophy (Vincent et al. 2014), and pituitary hormone deficiency (Diaz 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 (Pená et al. 2017).

### Table 1. Top Ten Most Differentiated SNPs from Each of the Ancestry-Specific Scans.

<table>
<thead>
<tr>
<th>Chr</th>
<th>Pos</th>
<th>Rsid</th>
<th>LLRS</th>
<th>Target Ancestry</th>
<th>Nearest Gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>33951693</td>
<td>rs16891982</td>
<td>22.085902</td>
<td>European</td>
<td>SLC45A2</td>
</tr>
<tr>
<td>15</td>
<td>48426484</td>
<td>rs1426654</td>
<td>19.707464</td>
<td>European</td>
<td>SLC24A5</td>
</tr>
<tr>
<td>15</td>
<td>28356859</td>
<td>rs1129038</td>
<td>19.290553</td>
<td>European</td>
<td>HERC2</td>
</tr>
<tr>
<td>15</td>
<td>28495956</td>
<td>rs12912427</td>
<td>18.270213</td>
<td>European</td>
<td>HERC2</td>
</tr>
<tr>
<td>9</td>
<td>16792200</td>
<td>rs10962596</td>
<td>15.819739</td>
<td>European</td>
<td>BNC2</td>
</tr>
<tr>
<td>1</td>
<td>1385211</td>
<td>rs1312568</td>
<td>15.067101</td>
<td>European</td>
<td>ATAD3C</td>
</tr>
<tr>
<td>2</td>
<td>136407479</td>
<td>rs1446585</td>
<td>14.957852</td>
<td>European</td>
<td>R3MD1</td>
</tr>
<tr>
<td>2</td>
<td>13661754</td>
<td>rs182549</td>
<td>14.629386</td>
<td>European</td>
<td>MCM6</td>
</tr>
<tr>
<td>1</td>
<td>204784969</td>
<td>rs3940119</td>
<td>14.393216</td>
<td>European</td>
<td>NFASC</td>
</tr>
<tr>
<td>4</td>
<td>38798646</td>
<td>rs5743618</td>
<td>14.386819</td>
<td>European</td>
<td>TLR1</td>
</tr>
<tr>
<td>16</td>
<td>48258198</td>
<td>rs17822931</td>
<td>23.271759</td>
<td>CHB</td>
<td>ABCC11</td>
</tr>
<tr>
<td>16</td>
<td>48375777</td>
<td>rs6500380</td>
<td>22.474103</td>
<td>CHB</td>
<td>LONP2</td>
</tr>
<tr>
<td>1</td>
<td>234635790</td>
<td>rs2175591</td>
<td>20.95541</td>
<td>CHB</td>
<td>TARBP1</td>
</tr>
<tr>
<td>4</td>
<td>100142780</td>
<td>rs75721934</td>
<td>20.453247</td>
<td>CHB</td>
<td>LOC100507053</td>
</tr>
<tr>
<td>11</td>
<td>61579427</td>
<td>rs72643557</td>
<td>20.114033</td>
<td>CHB</td>
<td>FADS1</td>
</tr>
<tr>
<td>11</td>
<td>120154631</td>
<td>rs12224052</td>
<td>19.696284</td>
<td>CHB</td>
<td>POU2F3</td>
</tr>
<tr>
<td>21</td>
<td>43794984</td>
<td>rs228088</td>
<td>19.518001</td>
<td>CHB</td>
<td>SLC37A1</td>
</tr>
<tr>
<td>11</td>
<td>133043841</td>
<td>rs79802711</td>
<td>19.157192</td>
<td>CHB</td>
<td>OPMC1</td>
</tr>
<tr>
<td>12</td>
<td>128016573</td>
<td>rs79478220</td>
<td>18.476104</td>
<td>CHB</td>
<td>FB2</td>
</tr>
<tr>
<td>19</td>
<td>51441759</td>
<td>rs11084804</td>
<td>18.158963</td>
<td>CHB</td>
<td>KLK5</td>
</tr>
<tr>
<td>14</td>
<td>46745012</td>
<td>rs140736443</td>
<td>32.730697</td>
<td>Native American</td>
<td>LINCO0871</td>
</tr>
<tr>
<td>9</td>
<td>82968379</td>
<td>rs6559543</td>
<td>27.584847</td>
<td>Native American</td>
<td>LINCO1507</td>
</tr>
<tr>
<td>10</td>
<td>80619307</td>
<td>rs2316155</td>
<td>27.399123</td>
<td>Native American</td>
<td>LINCO1227</td>
</tr>
<tr>
<td>14</td>
<td>21647765</td>
<td>rs77549780</td>
<td>27.355769</td>
<td>Native American</td>
<td>LINCO0641</td>
</tr>
<tr>
<td>12</td>
<td>14189549</td>
<td>rs12425115</td>
<td>25.867367</td>
<td>Native American</td>
<td>GRIN2B</td>
</tr>
<tr>
<td>10</td>
<td>8150713</td>
<td>rs10508343</td>
<td>25.609772</td>
<td>Native American</td>
<td>GATA3</td>
</tr>
<tr>
<td>15</td>
<td>34936250</td>
<td>rs16959274</td>
<td>25.424824</td>
<td>Native American</td>
<td>GOLGA8B</td>
</tr>
<tr>
<td>8</td>
<td>4490837</td>
<td>rs71523639</td>
<td>24.595927</td>
<td>Native American</td>
<td>CSMD1</td>
</tr>
<tr>
<td>1</td>
<td>14301862</td>
<td>rs72640452</td>
<td>24.455822</td>
<td>Native American</td>
<td>PRDM2</td>
</tr>
<tr>
<td>12</td>
<td>29817716</td>
<td>rs12580697</td>
<td>23.967094</td>
<td>Native American</td>
<td>TMTC1</td>
</tr>
<tr>
<td>8</td>
<td>145639681</td>
<td>rs1871534</td>
<td>11.906794</td>
<td>Yoruba/Ancestral Non-African</td>
<td>SLCP9A6</td>
</tr>
<tr>
<td>5</td>
<td>178626609</td>
<td>rs6869885</td>
<td>11.541667</td>
<td>Yoruba/Ancestral Non-African</td>
<td>ADAMTS2</td>
</tr>
<tr>
<td>15</td>
<td>29427400</td>
<td>rs10152250</td>
<td>11.48232</td>
<td>Yoruba/Ancestral Non-African</td>
<td>FAM189A1</td>
</tr>
<tr>
<td>1</td>
<td>1106112</td>
<td>rs6670963</td>
<td>11.447873</td>
<td>Yoruba/Ancestral Non-African</td>
<td>TTLL10</td>
</tr>
<tr>
<td>4</td>
<td>3666494</td>
<td>rs58827274</td>
<td>11.341367</td>
<td>Yoruba/Ancestral Non-African</td>
<td>LOC100133461</td>
</tr>
<tr>
<td>17</td>
<td>2631988</td>
<td>rs4790359</td>
<td>11.18134</td>
<td>Yoruba/Ancestral Non-African</td>
<td>PAFAH1B1</td>
</tr>
<tr>
<td>9</td>
<td>136769888</td>
<td>rs2789823</td>
<td>11.031687</td>
<td>Yoruba/Ancestral Non-African</td>
<td>VAV2</td>
</tr>
<tr>
<td>16</td>
<td>169656029</td>
<td>rs6930377</td>
<td>10.824098</td>
<td>Yoruba/Ancestral Non-African</td>
<td>THBS2</td>
</tr>
<tr>
<td>23</td>
<td>29503769</td>
<td>rs8073072</td>
<td>10.794224</td>
<td>Yoruba/Ancestral Non-African</td>
<td>RNF13</td>
</tr>
<tr>
<td>7</td>
<td>173642871</td>
<td>rs10067518</td>
<td>10.787147</td>
<td>Yoruba/Ancestral Non-African</td>
<td>HMP19</td>
</tr>
</tbody>
</table>

LLRS = log-likelihood ratio score for positive selection.
SNP by rescaling all variances and covariances by a scalar multiplier α. We inferred genome-wide allele frequencies and covariances as described in Materials and Methods section. We applied a likelihood model for each 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 Materials and Methods section. We applied a likelihood model for each 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 Materials and Methods section. We applied a likelihood model for each selection in Yoruba African ancestry or ancestral non-African ancestry. We analyzed 5,601,710 variable sites across the autosomal genomes.

Fig. 5. 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 Materials and 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.
The 14th 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 15th top SNP (rs1919550) is a GTEx eQTL for FBXO40 in whole blood, but does not have a high CADD score. However, it lies near an SNP (rs9813391) with a high LLRS that leads to a nonsynonymous change (R145Q) in ARFX—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 Senior-Loken syndrome (Otto et al. 2005), a ciliopathic eye disorder.

The 22nd 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 and 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 23rd and 24th top SNPs (rs5758430 and 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 significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank significantly associated with several traits related to body fat, food intake and white blood cells in the UK Biobank.
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 toward finding many variants in highly drifted populations, likely because the excess variance in the Wright–Fisher process is not well modeled 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., Matthey-Doret and Whitlock 2019; Johri et al. 2020; Schrider 2020). Although it has been argued that such an effect is unlikely to explain $F_{ST}$ outliers in real data (Matthey-Doret and Whitlock 2019; Schrider 2020), 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 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 LRP4), 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 rs1258069/TMT1C (involved in calcium homeostasis) 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 toward the higher end (e.g., the top eight SNPs have an FDR of < 10%).

We provide a list of functional annotations for all the SNPs with high LLRS (> 15) within a 2 Mb 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 (supplementary table S6, Supplementary Material online).

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 that 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.

### Materials and Methods

#### Basic Model

The new method is based on the Ohana inference framework (Cheng, Mailund, et al. 2017), 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 phylogenetic history between the nodes. Consider, for example, the example of the matrix and corresponding tree in the left side of figure 6. 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 $d_{ij} = \Omega_{ij} + \Omega_{ji} - 2\Omega_{0}$. Treelessness 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 ML approach (Felsenstein 1985). We emphasize that the root is arbitrary but that it does not imply any assumptions about this component actually being ancestral.

We estimate the covariance matrix $\Omega$ 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

$$P(f|\Omega, \mu, f_0) \sim \mathcal{N}(f_0, \mu(1 - \mu)),$$

where $f_0$ is the allele frequency in the ancestry component arbitrarily assigned as ancestral and $f$ is a vector of the allele frequencies in the other $K - 1$ components, at SNP $j$. $\mu_j$ is the mean allele frequency for SNP $j$ (averaged over all components). Note 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

$$Pritchard 2012) which also uses a Gaussian approximation to

similar to the model implemented in TreeMix (Pickrell and

nents). Note that this model of joint allele frequencies is

mean allele frequency for SNP

is described in a subsequent section.

SNPs in the genome. The method for optimizing this function

Selection Model

Following the genome-wide estimation of $\Omega$, 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 MCMC. We also note 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 $\Omega$ 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 elements is multiplied by a scalar, $\alpha$, corresponding to differences in allele frequency in one of the ancestry components relative to the rest, for example:

$$\Omega = \begin{bmatrix}
\Omega_{1,1} & \cdots & \Omega_{1,K-1} \\
\vdots & \ddots & \vdots \\
\Omega_{K-1,1} & \cdots & \alpha \cdot \Omega_{1,K-1}
\end{bmatrix}.$$

The value of $\alpha$ is then estimated via ML using equation (1) (assuming all other values in $\Omega_{x}$ 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. Figure 6 shows an example. This test can also be implemented to test selection on ancestral nonterminal 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 $\chi^2_1$-distribution and a point mass at zero (Self and Liang 1987) because $\alpha$ 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 structure/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, Mailund, et al. (2017) 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 \( f_{a0} \) for all \( j \), we then maximize the likelihood in equation (1) for \( \Omega \). This optimization is done using the Nelder–Mead simplex method (Nelder and Mead 1965). It uses Cholesky decomposition (Cholesky 1910) to determine the positive semidefiniteness of a matrix and to compute matrix inverses and determinants. For the initial starting point, we use sample covariances:

\[
S_c = \frac{1}{J} \sum_{j=1}^{J} (x_j - \bar{x}_j)(x_j - \bar{x}_j)^T x_j = \begin{bmatrix} f_1 \\ \vdots \\ f_{J-1} \\ f_J \end{bmatrix} \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_{J-1} \\ \bar{x}_J \end{bmatrix} \] (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 \( z \) during a selection scan for a single SNP, conditionally on the globally estimated value of \( \Omega \), 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 three distinct demographic models (fig. 2):

- A basic four-population tree with no admixture (fig. 2a): An ancestral population splits into four subpopulations at times 4,000, 2,000, and 800 generations before present, following the topology in figure 2a. Selection is simulated on the yellow branch in figure 2a. Tests for selection are conducted for yellow ancestry (i.e., the main ancestry component in the third branch).

- A four-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) and 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 2 Mbp 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 \( \pm 10 \) kb 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 four different selection coefficients \( s = 0, 0.01, 0.02, \) and \( 0.05 \) and three 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 \( f < 0.0001 \) 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 1,000 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 (Gautier 2015; Duforet-Frebourg et al. 2016). Like Ohana, both methods depend on some sort of empirical null model. To this end, we simulated three 20 Mb-long neutral regions under otherwise the same settings as previously described, with \( s = 0 \), in order to generate a null data set for calibrating each method. In the case of Ohana and BayPass, this null data set is used to estimate the covariance matrix for each population; in pcadapt, we append this null data set 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; in contrast, BayPass and pcadapt test for any significant deviation from the empirical covariance matrix (BayPass models population-level covariance, whereas pcadapt 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.

**Supplementary Material**

Supplementary data are available at Molecular Biology and Evolution online.
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 number R01GM138634.

Data Availability

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

References


