DISENTANGLING SIGNATURES OF SELECTION BEFORE AND AFTER EUROPEAN COLONIZATION IN LATIN AMERICANS

Javier Mendoza-Revilla[1,2,§*], Juan Camilo Chacón-Duque[3], Macarena Fuentes-Guajardo[4], Louise Ormond[1], Ke Wang[5], Malena Hurtado[2], Valeria Villegas[2], Vanessa Granja[2], Victor Acuña-Alonzo[6], Claudia Jaramillo[7], William Arias[7], Rodrigo Barquera Lozano[5,6], Jorge Gómez-Valdés[6], Hugo Villamil-Ramírez[8,9], Caio C. Silva de Cerqueira[10], Keyla M. Badillo Rivera[11], Maria A. Nieves-Colón[12], Christopher R. Gignoux[13], Genevieve L. Wojcik[14], Andrés Moreno-Estrada[15], Tábita Hunemeier[10], Virginia Ramallo[10,16], Lavinia Schuler-Faccini[10], Rolando Gonzalez-José[16], Maria-Cátira Bortolini[10], Samuel Canizales-Quinteros[8,9], Carla Gallo[2], Giovanni Poletti[2], Gabriel Bedoya[7], Francisco Rothhammer[4,17], David Balding[1,18], Matteo Fumagalli[19], Kaustubh Adhikari[20], Andrés Ruiz-Linares[1,21,22*¶] and Garrett Hellenthal[1*¶]

[1] Department of Genetics, Evolution and Environment, and UCL Genetics Institute, University College London, London, UK
[2] Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Lima, Perú
[3] Centre for Palaeogenetics & Department of Archaeology and Classical Studies, Stockholm University, Stockholm, Sweden
[4] Departamento de Tecnología Médica, Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, Chile.
[5] Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[6] National Institute of Anthropology and History, Mexico City, Mexico
[7] GENMOL (Genética Molecular), Universidad de Antioquia, Medellín, Colombia
[8] Unidad de Genómica de Poblaciones Aplicada a la Salud, Facultad de Química, UNAM-Instituto Nacional de Medicina Genómica, Mexico City, Mexico
[9] Universidad Nacional Autónoma de México e Instituto Nacional de Medicina Genómica, Mexico City, Mexico
[10] Departamento de Genética, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil
[11] Department of Genetics, Stanford School of Medicine, Stanford, California, United States
[12] Department of Anthropology, University of Minnesota Twin Cities, Minneapolis, Minnesota, United States
[13] University of Colorado Anschutz Medical Campus, Aurora, Colorado, United States
[14] Bloomberg School of Public Health, John Hopkins University, Baltimore, Maryland, United States
[15] Laboratorio Nacional de Genómica para la Biodiversidad (UGA-LANGEBIO), CINVESTAV, Irapuato, Guanajuato, Mexico
[16] Instituto Patagónico de Ciencias Sociales y Humanas-Centro Nacional Patagónico, CONICET, Puerto Madryn, Argentina
[17] Programa de Genetica Humana, ICBM, Facultad de Medicina, Universidad de Chile, Santiago, Chile
[18] Schools of BioSciences and Mathematics & Statistics, University of Melbourne, Melbourne, Australia
[19] Department of Life Sciences, Silwood Park campus, Imperial College London, Ascot, UK
[20] School of Mathematics and Statistics, Faculty of Science, Technology, Engineering and Mathematics, The Open University, Milton Keynes, UK
[21] Ministry of Education Key Laboratory of Contemporary Anthropology and Collaborative Innovation Center of Genetics and Development, Fudan University, Shanghai, China
[22] Aix-Marseille Université, CNRS, EFS, ADES, Marseille, France
[§] Current address: Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris, France.

¶These authors jointly supervised this work

*Correspondence to: javier.mendoza-revilla@pasteur.fr (J.M.R); andresruiz@fudan.edu.cn (A.R.L.); g.hellenthal@ucl.ac.uk (G.H.)

Abstract

Throughout human evolutionary history, large-scale migrations have led to intermixing (i.e., admixture) between previously separated human groups. While classical and recent work have shown that studying admixture can yield novel historical insights, the extent to which this process contributed to adaptation remains underexplored. Here, we introduce a novel statistical model, specific to admixed populations, that identifies loci under selection while determining whether the selection likely occurred post-admixture or prior to admixture in one of the ancestral source populations. Through extensive simulations we show that this method is able to detect selection, even in recently formed admixed populations, and to accurately differentiate between selection occurring in the ancestral or admixed population. We apply this method to genome-wide SNP data of ~4,000 individuals in five admixed Latin American cohorts from Brazil, Chile, Colombia, Mexico and Peru. Our approach replicates previous reports of selection in the HLA region that are consistent with selection post-admixture. We also report novel signals of selection in genomic regions spanning 47 genes, reinforcing many of these signals with an alternative, commonly-used local-ancestry-inference approach. These signals include several genes involved in immunity, which may reflect responses to endemic pathogens of the Americas and to the challenge of infectious disease brought by European contact. In addition, some of the strongest signals inferred to be under selection in the Native American ancestral groups of modern Latin Americans overlap with genes implicated in energy metabolism phenotypes, plausibly reflecting adaptations to novel dietary sources available in the Americas.

Introduction

Admixed populations offer a unique opportunity to detect recent selection. In the human lineage, genomic studies have demonstrated the pervasiveness of admixture events in the history of the vast majority of human populations (Patterson et al. 2012; Hellenthal et al. 2014; Lazaridis et al. 2014). By inferring the ancestral origins of particular genetic loci in the genomes of recently admixed individuals, recent studies have provided evidence that such admixture has facilitated the spread of adaptative genetic mutations in humans. Notable examples include the transfer of a protective allele in the Duffy blood group gene likely providing resistance to *Plasmodium vivax* malaria in Malagasy and Cape Verdeans from sub-Saharan Africans (Hodgson et al. 2014; Pierron et al. 2018; Hamid et al. 2021), and the transmission of the lactase persistence allele in the Fula pastoralists from Western Eurasians (Vicente et al. 2019).

An ideal setting in which to test whether and how admixture contributed to genetic adaptation is Latin America. The genetic make-up of present day Latin Americans stems mainly from three ancestral populations: indigenous Native Americans, Europeans (mainly from the Iberian Peninsula), and Sub-Saharan Africans (Wang et al. 2007; Moreno-Estrada et al. 2013; Moreno-Estrada et al. 2014; Homburger et al. 2015; Chacon-Duque et al. 2018; Luisi et al. 2020) that were brought together starting ~500 years ago. The admixed genomes of Latin Americans are thus the result of an intermixing process between human populations that had been evolving independently for tens-of-thousands of years and that were suddenly brought together in a new environment. In this new environment, the ancestral genomes were quickly subjected to novel pressures that were largely unfamiliar from where they firstly evolved. Therefore, the genomes of Latin Americans potentially harbor signals of both older adaptations present in each of the ancestral populations, and more recent adaptations attributable to beneficial variants, e.g. introduced from a particular ancestral population, increasing rapidly in frequency post-admixture. Motivated by this, several studies have explored the genomes of admixed Latin Americans for signatures of selection, for example focusing on events occurring since the admixture event (Tang et al. 2007; Basu et al. 2008; Ettinger et al. 2009; Guan 2014; Rishishwar et al. 2015; Deng et al. 2016; Zhou et al. 2016; Norris et al. 2020; Vicuna et al. 2020). These studies have relied on an approach similar to that of admixture mapping, where the ancestry of a genomic region in each admixed individual is assigned to a particular ancestral population, a technique known as local-ancestry-inference (LAI). Loci with significantly more inferred ancestry inherited from one ancestral population are assumed to have evolved under some form of selection (Tang et al. 2007).

In addition, the genetic make-up of Latin Americans offers the opportunity to detect selection in their ancestral populations, as large cohorts of Latin Americans can be leveraged to reconstruct genetic variation patterns in each source population. This is of particular use for exploring selection in Native Americans, since Native groups are currently underrepresented in genomic studies (Sirugo et al. 2019) and as a consequence only a few studies have centered on detecting adaptive signals of indigenous groups from the Americas. Such studies have identified strong selective signals at different genes, particularly at those related to immunity, highlighting the selective pressures that Native Americans were subjected to after they entered the continent (Lindo et al. 2018; Reynolds et al. 2019; Avila-Arcos et al. 2020).

With some exceptions (Cheng et al. 2021), these studies either limited their analyses to Latin Americans with high Native American ancestry or used LAI to infer loci in individuals that derive from a Native American source. However, such approaches may result in a reduction of statistical power due to removal of individuals with non-Native ancestry, inaccurate local ancestry estimation and/or through removing segments challenging to assign.

Here we present a novel statistical model that identifies loci that have undergone selection before or after an admixture event (which we refer to as pre- or post-admixture selection, respectively). In contrast to previous methods, this approach is based on allele frequencies and does not require assignments of local ancestry along the genome. We illustrate the utility of our new method by performing a selection scan in five Latin American cohorts collected as part from the CANDELA Consortium (Ruiz-Linares et al. 2014). Our results suggest that several loci have been subjected to natural selection in admixed Latin American populations, and in their ancestral populations, replicating many of these signals using LAI. Many of the putative selected SNPs are strongly associated to relevant phenotypes, or act as expression quantitative loci (eQTL) in relevant tissues, providing further evidence of their functional effect. Overall, our analyses highlight the usefulness of our method to detect signals of selection in admixed populations or their ancestral populations, and reveal novel candidate genes implicated in the adaptive history of groups from the American continent.

Results

Overview of AdaptMix

In part following Balding and Nichols (1995), and analogous to previous approaches (Long 1991; Mathieson et al. 2015; Cheng et al. 2021), our model AdaptMix assumes that, under neutrality, the allele frequencies of an admixed target population can be described using a beta-binomial model, with expected allele frequency equal to a mixture of sampled allele frequencies from a set of groups that act as surrogates to the admixing sources (fig. 1). In our case the admixed target population is a Latin American cohort, defined below, and we use three surrogate groups to represent Native American, European, and African admixing source populations. The mixture values are inferred a priori, e.g. using ADMIXTURE (Alexander et al. 2009) (fig. 1a) or SOURCEFIND (Chacon-Duque et al. 2018), as the average amount of ancestry that each admixed target individual matches to a set of reference populations. (The reference populations used by these programs may be the same as the surrogate populations, but they need not be as illustrated below.) We find the variance parameter that maximises the likelihood of this beta-binomial model across all SNPs. This variance term aims to limit the number of false-positives attributable to genetic drift in the target population following admixture and/or the use of inaccurate surrogates for the ancestral populations. Then, at each SNP, we calculate the probability of observing allele counts equal to or more extreme than those observed in the target population, hence providing a *P*-value testing the null hypothesis that the SNP is neutral (see Methods).

Assuming a pulse of admixture, this test is designed to detect selection occurring: (i) in the admixed population following the admixture event (e.g. along the purple line time period in fig. 1b), and/or (ii) in one (or more) of the source/surrogate pairings, i.e. following the split of the surrogate population from the admixing source it is representing (e.g. along the red and/or blue lines in fig. 1b). At SNPs with evidence of selection (i.e. low *P*-values), we distinguish between (i) and (ii) by exploring how genotype counts of admixed target individuals relate to their inferred admixture proportions contributed by each surrogate. Under scenario (i), we assume that selection affects all target individuals equally, regardless of their admixture proportions, which in turn assumes all ancestries were present when selection occurred. In contrast, under scenario (ii), we expect selection to more strongly affect one of the source/surrogate population pairings. Intuitively, if (ii) is true, individuals with nearly 100% ancestry from the source/surrogate pair experiencing selection will have genotype counts that deviate the most from expectations under the neutral model, while individuals with nearly 0% ancestry from this pair will have counts that closely follow the neutral model (fig. 1c). If instead (i) is true, this pattern is attenuated, though it can be challenging in practice to distinguish (ii) from (i) if allele frequencies strongly differ between surrogate groups (fig 1d). Assuming a multiplicative model of selection, we find the selection coefficients that maximize the fit of the data to model (i) and to model (ii) when separately treating each source/surrogate pair as the selected group. We report ratios of likelihoods, equivalent here to using differences in Akaike Information Criterion (AIC), to quantify our ability to distinguish among scenarios (i) and (ii).

In summary, for each tested SNP we infer (a) a *P*-value testing the null hypothesis of neutrality, (b) the relative evidence (i.e. likelihood ratios) for whether selection occurred post-admixture or in one of the admixing sources and (c) the selection strength summed across time.

Simulations

We tested our approach using simulations designed to resemble our Latin American cohort in terms of sample size, inferred admixture proportions, and the extent to which our surrogates match the true admixing sources (see Methods). At a false-positive rate of 5×10^{-5}, these simulations indicate we have ~50-90% power to detect selection for scenario (i) (i.e., post-admixture selection) with selection strength (*s*) of 1.15-1.20 per generation in homozygotes carrying two copies of the selected allele, and selection occurring over 12 generations under various modes of selection (additive, dominant, multiplicative, recessive) (fig. 2a, supplementary fig. S1). For scenario (ii), in the case of selection occurring in the Native American source, power depends on the overall amount of Native American ancestry (fig. 2a). As an example, Brazil-like simulations (<15% average Native American ancestry) show little power, Colombia-like simulations (~30% average Native American ancestry) typically exhibit >50% power, and other simulated populations (~50–70% average Native American ancestry) exhibit >75% power under scenario (ii) assuming *s*=1.1 per generation over 50 generations, with similar power if instead s~1.025 over 150 generations (supplementary fig S2). Detecting selection occurring in the European source depends on the overall amount of European ancestry in a similar manner (e.g., fig. 2a, supplementary fig. S3). For SNPs where we detect selection, we mis-classify the type of selection ≤2% of the time, e.g., concluding post-admixture selection when the truth is selection in the Native American source ~1% of the time across all selection coefficients (fig. 2b). However, our approach often fails to classify selection scenarios unless selection strengths are large (e.g., *s*>1.1).

Applying AdaptMix to the five Latin American cohorts of CANDELA

We divided Latin Americans into five cohorts based on country of origin: Brazil (n=190), Chile (n=896), Colombia (n=1125) Mexico (n=773), and Peru (n=834), using individuals sampled as part of the CANDELA Consortium (Ruiz-Linares et al. 2014), testing each cohort for selection separately (supplementary fig. S4). Analyzing each cohort by country of origin results in a higher number of individuals, and thus increases the statistical power to detect selection. As demonstrated in Chacon-Duque et al (2018), however, there is notable population sub-structure within each country. To test for robustness of our selection signals to this sub-structure, we supplemented each of these analyses by testing subsets of individuals within a country based on their inferred ancestry matching to Native American reference groups from Chacon-Duque et al. (2018). This gave six additional tested groups with sufficient ancestry represented: ‘Mapuche’ (n=434) in Chile, ‘Chibcha Paez’ (n=200) in Colombia, ‘Nahua’ (n=466) and ‘South Mexico’ (n=78) in Mexico, and ‘Andes Piedmont’ (n=195) and ‘Quechua’ (n=147) in Peru (supplementary fig. S5). To infer the proportion of African, European, and Native American ancestry in each Latin American, we applied unsupervised ADMIXTURE with *K*=3 clusters jointly to all CANDELA individuals and 553 Native American, Iberian, and West African reference individuals (fig. 1a).

Note that the choice of surrogate populations defines the selection test between each surrogate and its corresponding ancestral source in scenario (ii). In this way, our test acts as an analogue to F_{ST} comparing two populations, but while accounting for admixture in one of the populations. As an illustration, we tested the Brazilian cohort for selection using northwest Europeans from England and Scotland (GBR) from the 1000 Genomes Project (1KGP) (The 1000 Genomes Project Consortium 2015) as a surrogate for the Brazilian cohort's European ancestry source (supplementary fig. S6). Given the majority (~80%) of ancestry in the Brazilian cohort is related to Iberian Europeans, this test is most-powered to detect selection acting along the branch separating present-day northwest Europeans and descendants of Iberians who traveled to Brazil post-Columbus. In this analysis, we infer strongest signals of selection at the *HERC2*/*OCA2* and *LCT*/*MCM6* genes. This replicates previously reported selection signals when comparing northwest Europeans to present-day Iberians (Poulter et al. 2003; Bersaglieri et al. 2004), and likely indicates selection for lighter skin pigmentation and lactase persistence in northwest Europeans that is unrelated to any selection in the Americas. As another example, we also tested each Latin American cohort separately while using Han Chinese from Beijing (CHB) from the 1KGP as a surrogate for Native American ancestry (supplementary fig. S7). In this analysis, SNPs that follow model (ii) indicate selection along the branch separating present-day Han Chinese and Native American populations. For this test, we find the strongest signals of selection at previously reported selected genes in East Asians, including those related to alcohol metabolism such as *ADH7* and *ADH1B* (Galinsky et al. 2016; Gu et al. 2018) that both are classified as selection under model (ii). The strongest overall signal in this analysis overlapped the *POU2F3* gene, implicated in the regulation of viral transcription, keratinocyte differentiation and other cellular events, which has been reported to be under selection in Native American populations from throughout the Americas (Amorim et al. 2017).

For our main analyses, we use 205 Iberians (from 1KGP and Chacon-Duque et al. (2018)) to represent European ancestry surrogates. Therefore, given the likely short split time between present-day Iberians and Europeans that migrated to the Americas during the colonial era, we are underpowered to detect selection in the European source only (see simulations). We use 206 West

Africans from the 1KGP to represent the African ancestry source, which has been reported as a good proxy to the African genetic sources (from Chacon-Duque et al. (2018)). For this reason, we should similarly have low power to find selection occurring only in the African source/surrogate. At any rate we do not test for selection related to African ancestry, because the Latin American cohort here have ~6% African ancestry on average, limiting power further. We combined 142 individuals with <1% non-Native American inferred ancestry from 19 Native American groups (supplementary table S1) to represent the Native American surrogate. By using individuals sampled from geographically spread Native American groups as the Native American ancestry surrogate, we aim to identify regional selection signals experienced by some Native American groups but not others. We also expect to have the highest power when testing for selection type (ii) in Native Americans, as there is likely to be the most time separating this 'average' Native American surrogate and the admixing source of each regional Latin American cohort. To avoid confounding our inference, we excluded individuals with >1% inferred ancestry matching to surrogates other than Native Americans, Iberian Europeans, and West Africans using SOURCEFIND (Chacon-Duque et al. 2018). Also, since the time since admixture among these groups is relatively short in the CANDELA cohort (likely <15 generations ago), detecting selection post-admixture can only identify relatively strong selection signals (see simulations).

AdaptMix identifies 47 regions of putative selection

For each Latin American cohort, we considered SNPs under selection as those having *P*-values less than the 5×10^{-5} false-positive threshold in the population-matched neutral simulations, which corresponds to a model-based *P*-value of 6.75×10^{-6}–1.07×10^{-7} (supplementary table S2). For Chile, Colombia, Mexico and Peru, we report loci that pass these criteria both in the analysis of all individuals from that country and in at least one of three alternative analyses for that country that are designed to test for robustness to latent population structure (supplementary fig. S8). The first of these alternative analyses consisted of identifying signals of selection using AdaptMix on each of the six Native American subsets defined above (e.g., in either the 'Andes Piedmont' or 'Quechua' subset when testing for selection in Peruvians) (supplementary table S3). The other two alternative analyses were based on LAI. In particular we used ELAI (Guan 2014) to assign each genomic region of an admixed individual to a Native American, European, or African ancestral source. For the second alternative analysis, designed to test for post-admixture selection, we assessed whether the proportion of ancestry inferred from one of these three sources in a local region deviated substantially from the genome-wide average (supplementary table S4). For the third alternative analysis, designed to test for selection in the Native American source, we instead used the Population Branch Statistic (PBS) (Yi et al. 2010) to test for selection in one of the six Native

American subset groups defined above, using allele frequencies computed from LAI-inferred Native American segments from the subset of individuals representing that Native American group (see Methods) (supplementary fig. S5 and supplementary table S5).

Overall, we find 51 candidate regions to have evidence of positive or purifying selection passing the criteria above, 47 of which target protein-coding genes (supplementary table S6 and fig. 3). Four of these 47 candidate gene regions contain at least one SNP exhibiting strong evidence (likelihood ratio >1,000) of selection affecting all admixed individuals regardless of ancestry proportions, which we assume reflects post-admixture selection. Furthermore, 18 of these 47 regions exhibit strong evidence of selection containing at least one SNP (likelihood ratio >1,000) in the Native American source only. The 25 remaining candidate gene regions are unclassified into either type of selection (likelihood ratio ≤1,000).

To prioritize candidate casual genes, we annotated the protein-coding gene that had the highest overall Variant-to-Gene (V2G) scores (Ghoussaini et al. 2021) for the SNPs showing the strongest evidence of selection in each candidate gene region. The overall V2G score aggregates differentially weighted evidence of variant-gene association from several sources, including cis-QTL data, chromatin interaction experiments, *in silico* function predictions (e.g., Variant Effect Predictor from Ensembl), and distance between the variant and each gene's canonical transcription starting site. For each of these candidate genes we then annotated the phenotype with the highest overall association score based on the Open Targets Platform (Koscielny et al. 2017).

While most of these associated phenotypes represent genetic disorders, syndromes, or different types of measurements (medically or non-medically-related), many are also related to immune response and diet – two major selective forces that shape the human genome (Karlsson et al. 2014; Fan et al. 2016). We therefore organize the description of our candidate selection signals into two main sections below that cover only these two features, with signals of all other hits in supplementary table S6. For brevity, below we only highlight putatively selected regions where at least one significant SNP had an associated GWAS or eQTL signal. For our significant SNPs related to immune-response genes, GWAS signals included SNPs associated to white blood cell counts in a large multi-continental cohort (including Latin American individuals) (Chen et al. 2020), and eQTL signals included cis-associated SNPs to gene expression in 15 immune-related cell types from the DICE project (Schmiedel et al. 2018). For our significant SNPs related to diet, GWAS signals included metabolic, anthropometric, and lipid levels from the UK Biobank cohort

(Loh et al. 2018), and eQTL signals included cis-associated SNPs to gene expression in adipose, muscle, and liver tissue from the GTEx Project (Lonsdale et al. 2013).

Signals at immune-related genes

Fifteen of the 47 candidate gene regions contained at least one protein-coding gene either related to the development or regulation of the immune system or that has been previously associated to the quantification of immune cell types, susceptibility progression to infectious diseases, or autoimmune disorders. For example, we replicate a well-known signal encompassing several immune-related genes at 6p21 that are part of the human leukocyte antigen (HLA) system (fig. 4 and supplementary fig. S9-S11). These included SNPs (AdaptMix P-value<5.00×10^{-7}) near several MHC class I genes (*HLA-G*, *HLA-H*, *HLA-A*, and *HLA-J*) in each of the Chilean, Colombian, Mexican and Peruvian cohorts, with the Colombian cohort containing several SNPs classified as being selected post-admixture (likelihood ratio>1,000). Encouragingly, we inferred African ancestry enrichment (Z-score>2.5) in each cohort ~60kb downstream from our top AdaptMix signals using LAI, with maximum Z-score>9 (one-sided P-value<4.09×10^{-21}) in the Chilean cohort (fig. 4). In addition, other signals were inferred upstream in the Chilean cohort at a 5' UTR SNP of the *ZBTB12* gene (rs2844455, AdaptMix P-value=5.45×10^{-8}), the Mexican cohort at an intronic SNP of *HLA-DMA* (rs28724903, AdaptMix P-value=3.87×10^{-8}), and the Peruvian cohort at an intronic SNP of the MHC class III gene *STK19* (rs6941112, AdaptMix P-value=7.57×10^{-9}). Many of these HLA genes have been previously characterized as subject to be under selection post-admixture in different Latin American populations by showing an excess of African ancestry at the HLA locus (Tang et al. 2007; Basu et al. 2008; Ettinger et al. 2009; Guan 2014; Rishishwar et al. 2015; Deng et al. 2016; Zhou et al. 2016; Norris et al. 2020; Vicuna et al. 2020).

In addition to HLA, we infer previously unreported selection signals in four candidate gene regions that each harbor genes with well-established roles in the immune system, with each region containing at least one SNP significantly associated (P-value<5×10^{-8}) to white blood cell counts or the expression of an immune-related gene in immune cells (P-value<10^{-5}) (see Methods). Among these, one signal at 1p13 in the Chilean cohort encompasses the *CD101* gene (fig. 5a), which belongs to a family of cell-surface immunoglobulins superfamily proteins and plays a role as an inhibitor of T-cell proliferation (Soares et al. 1998; Bouloc et al. 2000). Within this region five SNPs are classified as being selected post-admixture and show also an increase of LAI-inferred European ancestry (maximum Z-score=3.40, one-sided P-value=3.36×10^{-4}). Strikingly, the region contains a synonymous SNP (Ile588, CADD score of 9.23) (rs3736907, AdaptMix P-value=1.05×10^{-9}) that strongly affects *CD101* expression in T cells (eQTL P-value < 2.42×10^{-10})

and is associated with neutrophil (GWAS P-value=2.08×10^{-10}) and total white cell count (GWAS P-value=3.61×10^{-9}) (fig. 5a).

The second signal, at 18p11 also in Chileans, encompasses the *PTPN2* gene, a tyrosine-specific phosphatase involved in the Janus kinase (JAK)-signal transducer and activator of transcription (STAT) signaling pathway (fig. 5b). The JAK-STAT pathway has an important role in the control of immune responses, and dysregulation of this pathway is associated with various immune disorders (Shuai and Liu 2003). Several SNPs with low AdaptMix P-values (P-value<1.69×10^{-7}) in the 18p11 region are also associated with eosinophil counts (GWAS P-value<1.13×10^{-10}) and the expression of *PTPN2* in natural killer (NK) cells (eQTL P-value<1.14×10^{-9}) (fig. 5b).

The other two novel signals, both in the Peruvian cohort, are consistent with selection in Native Americans only (likelihood-ratio>1,000). The first, at 17q25, contains the *CD300LF* gene that encodes for a membrane glycoprotein that contains an immunoglobulin domain, and which plays an important role in the maintenance of immune homeostasis by promoting macrophage-mediated efferocytosis (Borrego 2013). Notably, a 3'UTR SNP (rs9913698, AdaptMix P-value=3.11×10^{-9}) is strongly associated with monocyte count (GWAS P-value=1.00×10^{-33}), total white cell count (GWAS P-value=5.96×10^{-24}), lymphocyte count (GWAS P-value=2.50×10^{-19}), and neutrophil count (GWAS P-value=1.30×10^{-9}) (supplementary fig. S12). The second signal is at 22q11 adjacent to the *MIF* gene (fig. 5c), which is implicated in macrophage function in host defense through the suppression of anti-inflammatory effects of glucocorticoids (Calandra and Roger 2003). Variants within *MIF* have been recently associated to rheumatoid arthritis in southern Mexican patients (Santoscoy-Ascencio et al. 2020). The SNP rs2330635 (AdaptMix P-value=7.06×10^{-8}) is strongly associated to the expression of *MIF* in T-cells (eQTL P-value<8.63×10^{-5}) and NK cells (eQTL P-value=5.77×10^{-9}) and is also marginally associated to neutrophil counts (GWAS P-value=2.46×10^{-6}) (fig. 5c).

Overall, these findings suggest that some of the most robust signals of adaptation in the Americas can be ascribed to immune-related selective pressures. These plausibly resulted from both the introduction of novel pathogens after European colonization and the endemic pathogens encountered by the first Native Americans during the initial peopling of the continent.

Signals at genes related to diet

Among the 47 candidate regions, nine regions contained at least one protein-coding gene potentially related to dietary practices through their association with metabolism-related phenotypes or anthropometric-related measurements (supplementary table S6). Among these, we infer three previously unreported signals where at least one of the selected SNPs was associated to metabolic- or anthropometric-related phenotypes, or to the expression of the candidate gene in adipose, muscle, or liver tissue (see Methods). One of these three hits (rs4636058, AdaptMix *P*-value=5.70×10^{-10}), at 6p22 in the Chilean cohort, is classified as being selected post-admixture and shows an increase of LAI-inferred European ancestry (*Z*-score=3.78, one-sided *P*-value=7.82×10^{-4}). It is located at 6q22 and encompasses the *SLC35F1* gene, whose function is not known, though several studies have associated this gene with different measurements of cardiac function (Hoffmann et al. 2017; Warren et al. 2017; Giri et al. 2019). Notably, SNP rs4636058 is marginally associated to cholesterol levels (UKBB GWAS *P*-value=3.8×10^{-4}) and body fat percentage (UKBB GWAS *P*-value=4.29×10^{-4}). Another of these three hits, at 1q31 in the Mexican cohort, is consistent with selection in Native Americans (likelihood-ratio>1,000) (fig. 6a). The 1q31 signal includes an intronic SNP (rs1171148, AdaptMix *P*-value=2.31×10^{-8}) of *BRINP3*, a gene associated to body mass index in studies across different human groups (Pulit et al. 2019; Zhu et al. 2020). Within this region, various SNPs are associated to different metabolic-related phenotypes, including the SNP rs1171148 that is associated with hip circumference (UKBB GWAS *P*-value=4.96×10^{-8}) and marginally associated with body mass index (UKBB GWAS *P*-value=5.51×10^{-5}) (fig. 6a).

Finally, the third hit (rs5030938, AdaptMix *P*-value= 3.79×10^{-15}), which had the highest overall AdaptMix score, is inferred in the Peruvian cohort at 10q22 and indicates selection in Native Americans (likelihood-ratio>1,000) (fig. 6b). This SNP is associated with the expression of *HKDC1* in liver (eQTL *P*-value=2.19×10^{-5}), adipose visceral (eQTL *P*-value=1.46×10^{-5}), and adipose subcutaneous tissue (eQTL *P*-value=1.36×10^{-4}) (fig. 6b). *HKDC1* encodes and hexokinase that catalyzes the rate-limiting and first obligatory step of glucose metabolism (Ludvik et al. 2016), and several studies have associated variants within this gene with glucose levels in pregnant women (Hayes et al. 2013; Guo et al. 2015; Kanthimathi et al. 2016; Tan et al. 2019) and with weight at birth (Warrington et al. 2019).

Overall, these results support previous hypothesis that genes related to energy metabolism were probably critical in the establishment of stable human populations in distinct ecoregions (Hancock et al. 2010), including those of the Americas (Amorim et al. 2017; Reynolds et al. 2019).

Discussion

Analytical considerations

Here we present AdaptMix, a novel statistical model that identifies loci under selection in admixed populations. Our model is based on the principle that allele frequencies in an admixed population can be modeled as a linear combination of the allele frequencies in the ancestral populations proportional to their admixing contributions, and that deviations from the expectation can be a product of selection. This selection test is related to the work of Long (1991) and Mathieson et al. (2015). One difference is that our approach directly infers and models the variance of the predicted allele frequencies in the admixed population given the set of surrogates used for ancestral sources. This parameter can help control for large deviations in allele frequency arising solely from genetic drift experienced in the admixed population (Long 1991; Bhatia et al. 2014) and/or from using inaccurate proxies for one or more of the source populations. In some applications here, e.g. the Brazilian cohort, AdaptMix gives *P*-values with a median near 0.5 as expected under the null hypothesis of neutrality, indicating a correction approach such as genomic control may not be necessary as in Mathieson et al. (2015) (supplementary fig. S13). However, simulations under neutrality that follow a slightly different model than our inference approach (see Methods), shows AdaptMix gives both an excess of high and low *P*-values relative to the uniform distribution expected under neutrality (supplementary fig. S14). This suggests our *P*-values are not well-calibrated, perhaps reflecting deviations from the underlying model and necessitating caution when choosing thresholds for significance. We thus based our significance thresholds on neutral simulations tailored to each cohort, and focus only on the strongest association signals that resulted in low false-positive rates based on simulated neutral SNPs. However, we caution that necessarily simulations are over-simplifications of complex latent demographic processes, and more work is required to verify these signals.

Another important contribution of our test is that it can infer whether selection disproportionately affects one source/surrogate pairing or affects all ancestry backgrounds equally. We assume selection affecting all ancestry backgrounds indicates selection occurring post-admixture, which is more parsimonious than an alternative explanation of independent selection events differentiating allele frequencies between each admixing source and its surrogate. For inferred selection in a source/surrogate pairing, this can reflect selection occurring in that source and/or its surrogate, possibly even following the admixture event. Post-admixture selection affecting only one source may be possible in cases of selection only occurring in a particular environment that is correlated with admixture fractions. For example, selection we detect to occur in Native Americans may be attributable to Europeans introducing a new environmental pressure (e.g. infectious disease) that disproportionately affected fitness in indigenous Americans. However, the split time between the true Native American ancestral source and our Native American surrogate is likely much longer than the time since colonial era admixture, suggesting selection pre-admixture as a more plausible explanation given the longer time to act. Supporting this, our inferred selection coefficients (which are summed over time) in cases where we conclude selection in Native Americans are typically greater than 2 (supplementary table S6). If selection had occurred post-admixture continuously over the last 12 generations (corresponding to an admixture date of ~1650CE), this value approximately corresponds to a per generation selection coefficient ~0.16, which is strong relative to previous reports of recent selection in human populations (e.g. Hamid et al. (2021)). In contrast, our four signals concluding post-admixture selection infer a per generation selection coefficient <0.1, which falls more in line with previous inference of selection strengths.

For 18 genomic regions where we conclude selection in the Native American source (supplementary table S6), it is possible this is capturing selection in (some subset of) groups that comprise the Native American surrogate group we use here, rather than in the (more localized) Native American source of the admixed population. The lack of overlap in selection signals when analysing the five CANDELA cohorts, and lack of concordance of our signals with those from PBS testing for selection in this combined Native American surrogate (supplementary fig. S15), suggests our signals are not being driven by selection in this combined population in practice. Furthermore, when using PBS to test for selection in LAI-inferred Native American segments from individuals with high degrees of ancestry recently related to the tested Native American source, an analysis that does not use the combined Native American surrogate, PBS scores for SNPs in 6 of these 18 regions fall into the top 99.99th percentile (supplementary fig. S16-21), with the remaining 13 regions containing SNPs in the top 99th percentile. However, relative to our approach, LAI-based inference (e.g., Avila-Arcos et al. (2020)) may be more robust to using combined data from multiple populations to represent one surrogate, since it only requires matching to a subset of individual's haplotype patterns in the reference panel.

In general our approach has decreased power to distinguish whether selection occurred post-admixture versus in one of the ancestral sources, if reference population allele frequencies are very different and selection is weak (fig. 1c). Inferring excess ancestry matching using LAI would likely better capture post-admixture selection in such cases, e.g. a scenario where one population that is fixed (or nearly-fixed) for the protective allele intermixes with a population nearly-fixed for the non-protective allele, with the admixed population subsequently undergoing selection. An example of this is a recently reported excess of African ancestry, likely attributable to post-admixture selection, on the Duffy-null allele in inhabitants of Santiago Island in Cape Verde (Hamid et al. 2021). However, our test to detect whether *any* type of selection occurred should not be affected by these scenarios. In addition, our approach may identify post-admixture selection in scenarios that excess-ancestry LAI-based would miss by design, such as cases where the selected allele is at a similar frequency in all reference populations. Perhaps the most important contrast to LAI and other approaches detecting selection in admixed populations (Cheng et al. 2021), is that in principle our approach can be applied to populations that descend from the mixture of genetically similar groups, e.g. if using haplotype-based approaches (e.g. SOURCEFIND) to infer ancestry proportions. Future work should assess the power of this technique under such admixture settings.

While our method assumes a single pulse of admixture, theoretically our ability to diagnose and classify selection occurring in only one source should not be affected by multiple instances of (or continuous) admixture from that or any other source. This is because the signal of allele frequency deviation due to selection in such cases is entirely determined by the amount of ancestry inherited from that source, and not the number of admixture pulses. In contrast, if an admixed population experiences selection and then receives new migrants from one of the original admixing sources that are unaffected by this selection, e.g. later European migrants to the Americas, in theory this may attenuate our ability to determine that selection occurred post-admixture. However, in a simple scenario of one such additional admixture pulse, contributing 10-50% of DNA, the correct post-admixture selection theoretical model fits as well or better to the theoretical truth than does the incorrect model concluding selection in the source that did not contribute new migrants (supplementary fig. S22).

As noted above, and consistent with other tests comparing populations (Mathieson 2020), the choice of surrogate group can make a difference in the inferred selection signals. For example, our largest signal of Native American selection, at 10q22 and most strongly signalled in the ‘Andes Piedmont’ Peruvian subgroup, disappears if replacing the ‘combined Native American’ surrogate group with Han Chinese (CHB from the 1KGP) (supplementary fig. S7). In this case, the frequency of the putatively selected allele (rs5030938) is 67% in LAI-inferred Native haplotypes in the Peruvian ‘Andes Piedmont’ subgroup, which is notably higher than the 38-54% observed in LAI-inferred Native American haplotypes in four non-Peruvian sub-groups, and thus consistent with selection (supplementary table S7). However, it is lower than that of CHB (~76%,), which explains the lack of signal when using CHB as a surrogate. The frequency in Yakut, a Siberian group that perhaps better represents ancestral Native Americans than CHB does (Wang et al. 2007), is closer to that of frequency estimates across non-Peruvian Native American groups (0.46-0.5). In general, there is a trade-off between using surrogates more distantly related to the source, which may decrease power to find regional adaptation signals, versus choosing a more closely related surrogate, which may also decrease power by masking adaptation signatures that it shares with the target source (e.g. using Iberians as a surrogate for European ancestry of Latin Americans). Our inferred variance parameter can be used to investigate how well a given surrogate captures genetic variation in the target population, with for example the inferred variance using CHB as a surrogate ~5-10-fold higher relative to using the combined Native American surrogate.

Selection signals detected in the CANDELA cohort

The candidate genes we infer to be affected by selection in Latin Americans and their Native American ancestors are best viewed in the context of other previously reported signals. Reynolds et al. (2019) recently performed a selection scan in three Native North American populations and identified some of the strongest signals at immune-related genes including the interleukin 1 receptor Type 1 (*IL1R1*) gene in a sample from several closely related communities in the southeastern United States, and the mucin 19 (*MUC19*) gene in a central Mexican population. We do not replicate the MUC19 signal in the CANDELA Mexican cohort, which could indicate that the Native American component in this cohort is not closely related to that of the central Mexican Native American group. Nonetheless, we found some of our strongest signals of selection at several loci encompassing genes involved in the immune response, including *CD300LF* and *MIF,* detected as being selected in the Native American ancestors of Peruvians. Interestingly, *CD300LF* promotes macrophage-mediated efferocytosis, while *MIF* play a role regulating macrophage function through the suppression of glucocorticoids. These observations suggest that these two genes might have perhaps evolved in a coordinated manner, possibly due to their phagocytic-related role against the novel pathogens encountered in the Americas.

Regarding signals of selection post-admixture, several studies have consistently shown adaptive signals in different Latin American populations at HLA by showing an excess of matching to African reference haplotypes using LAI (Tang et al. 2007; Basu et al. 2008; Ettinger et al. 2009; Guan 2014; Rishishwar et al. 2015; Deng et al. 2016; Zhou et al. 2016; Norris et al. 2020; Vicuna et al. 2020). Given that African ancestry was enriched at this region, the authors suggested that certain African alleles could have conferred a selective advantage to certain infectious diseases most likely brought by Europeans. While AdaptMix is only able to classify selection in one cohort (Colombia) out of our four HLA signals, we also replicated this excess of African ancestry in each of the CANDELA cohorts (supplementary fig. S9). There is some debate as to whether these signals are genuine or attributable to confounders such as inaccurate LAI inference (Pasaniuc et al. 2013). To illustrate the validity of these concerns, people with entirely Northwest European ancestry from Britain infer excess ancestry related to Africa in HLA, which – though perhaps influenced by genuine selection at HLA in Northwest Europeans – presumably does not reflect genuine recent African ancestry (supplementary fig. S23). Instead this is more likely attributable to the relatively high degree of genetic diversity in HLA mimicking African genetic diversity, illustrating how these LAI-based tests can give false-positive signals when testing for post-admixture selection. This may explain why AdaptMix does not replicate the moderate amount of excess African ancestry inferred by LAI at HLA in the Brazilian cohort (supplementary fig. S9), which is predominantly of European ancestry. Indeed regions under selection in admixed populations may be particularly difficult to classify accurately using LAI, e.g. with the HLA region here having the lowest overall LAI classification probability (supplementary fig. S24), especially in cases where the reference population have not experienced similar selection and hence may have poorly matching genetic variation patterns. As our approach does not require LAI, it is robust to these issues. While our model is not able to classify selection as post-admixture at most of our HLA signals, allele frequency patterns in the admixed cohorts are consistent with post-admixture selection and often show allele frequencies drifting away from those expected under our neutral model and towards those of the African or European reference population (supplementary fig. S25). This is most evident in the Colombian cohort, consistent with Africans contributing protective alleles as previously suggested (Tang et al. 2007; Basu et al. 2008; Ettinger et al. 2009; Guan 2014; Rishishwar et al. 2015; Deng et al. 2016; Zhou et al. 2016; Norris et al. 2020; Vicuna et al. 2020). In addition to HLA, we also identified a novel post-admixture selection signal in the Chilean cohort that was accompanied by a significant increase of European ancestry at the *CD101* locus, again, suggesting that protective alleles from Europeans might have also been adaptive to counter Old World-borne diseases brought to the Americas.

The signals encompassing genes related to metabolic and anthropometric-related phenotypes are consistent with novel dietary practices in the Americas driving adaptation, with many signals with an effect on relevant phenotypes and/or tissues, classified as being selected in the Native American source. Previous studies have shown evidence of adaptation at genes related to metabolic-related phenotypes and attributed the adaptation to dietary pressures in Native Americans. Avila-Arcos et al. (2020) recently reported strong signals of selection in the Mexican Huichol at several genes associated to lipid metabolism, including *APOA5* and *ABCG5.* We do not replicate these signals in the CANDELA Mexican cohort, which could indicate that the Native American component in this cohort is not closely related to that of the Huichol. The signals at *APOA5* and *ABCG5* are in line with a previous finding of a strong selection signal at another ATP-binding cassette transporter A1 (*ABCA1*) gene that has been associated with low high-density lipoprotein cholesterol in Latin Americans (Villarreal-Molina et al. 2008; Acuña-Alonzo et al. 2010). As the ABCA1 protein carrying the putative selected allele shows a decrease cholesterol efflux, the authors suggest that this variant could have favored intracellular cholesterol and energy storage, which in turn might have beneficially influenced the ability to accommodate fluctuations in energy supply during severe famines and during the regulation of reproductive function (Acuña-Alonzo et al. 2010). Lindo et al. (2018) used a genomic transect of Andean highlanders from northern Peru, and found the strongest signals of selection at *MGAM*, a gene related to starch digestion. The authors attributed this finding to a dietary-related selective pressure perhaps brought by the transition to agriculture in this region. AdaptMix shows evidence in the CANDELA Peruvian cohort within *MGAM* (rs7810984, AdaptMix *P*-value=1.79×10^{-8}, above 99.9th percentile) only when using CHB as a surrogate for Native American ancestry. This again illustrates how the choice of surrogate populations defines the selection test between each surrogate and its corresponding ancestral source. It is possible that by including Andean Native Americans in our Native American source population (supplementary table S1) we are affecting the power to detect selection in the Andean Native American ancestors of the CANDELA Peruvian cohort, analogous to how Lindo et al. (2018) no longer detect selection at *MGAM* if using PBS to compare ancient and present-day (Aymara) Andean groups.

Studies have also reported signals of selection in Native Americans groups shared with Siberian populations, which the authors interpreted as an adaptation to polyunsaturated-rich diets prior or close to the peopling of the Americas, likely in the Arctic Beringia. These included a signal overlapping the *WARS2* and *TBX15* genes, previously associated to body fat distribution and adipose tissue differentiation (Racimo et al. 2017), and the fatty acid desaturase (FADS) gene cluster that modulates the manufacture of polyunsaturated fatty acids (Amorim et al. 2017; Harris et al. 2019) (but see Mathieson (2020) for an alternative explanation of the *FADS* signal). Again, we inferred moderate selection evidence at these regions in the CANDELA Peruvian cohort only when using CHB as surrogate for Native American ancestry (SNP rs2361028 near *TBX15*, AdaptMix *P*-value=1.8×10^{-7}, above 99.5th percentile; SNP rs174576 within *FADS2*, AdaptMix *P*-value=3.8×10^{-8}, above 99.5th percentile). It is thus tempting to suggest that the three novel signals of selection AdaptMix classifies as being under selection in Native Americans might be related to dietary pressures experienced prior or during the peopling of the Americas (e.g., the *BRINP3* signal detected in Mexicans), or as a product for a greater reliance of domesticated crops including potatoes (3400–1,600 CE) (Rumold and Aldenderfer 2016) (e.g., the *HKDC1* signal detected in Peruvians). However, it is important to note that other factors may also be attributable for some of these selection signals.

Of potential adaptive interest is the *STOX1* gene detected in the Peruvian cohort close to our highest overall selection signal within *HKDC1* at 10q22 (fig. 6b). Mutations within *STOX1* have been associated to preeclampsia (Van Dijk et al. 2005; van Dijk and Oudejans 2011), a pathology of pregnancy characterized by high blood pressure and signs of damage to other organ system that can be lethal for the mother and for the fetus (Sibai 2003). Interestingly, in the single linkage study on preeclampsia conducted in Andean Peruvian families to date, SNPs within *STOX1* show marginal association (*P*-value=0.004678) (supplementary fig. S26) (Badillo Rivera and Nieves Colón et al. 2021). Given that high altitude is linked to an increased incidence of preeclampsia (Zamudio 2007), it is possible that natural selection has acted on genes related to this condition. Furthermore, the fact that variants within *HKDC1* are associated with glucose levels in pregnant women (Hayes et al. 2013; Guo et al. 2015; Kanthimathi et al. 2016; Tan et al. 2019) and considering the relationship between abnormal glucose levels and preeclampsia (Joffe et al. 1998; Weissgerber and Mudd 2015), it is also possible that natural selection has targeted variants at *HKDC1* due to its role in glucose metabolism.

Lastly, other environmental factors may also be attributable for some of these selection signals, such as infectious diseases. There is growing evidence of a link between metabolic diseases and innate immunity or inflammation (Pickup and Crook 1998; Kominsky et al. 2010; Lumeng and Saltiel 2011; Robbins et al. 2014). For instance, it has been shown that cholesterol plays a key role in various infectious processes such as the entry and replication of flaviviral infection (Osuna-Ramos et al. 2018). Additional studies in indigenous American populations will be needed to disentangle the putative selective pressures at these loci.

Conclusion

We have presented a novel allele frequency-based method that identifies loci under selection in admixed populations, while determining whether the selection affected all ancestral sources equally, indicating selection following admixture, or in only one of the sources. The novel candidate genes under selection provide new insights into the adaptive traits necessary for the early habitation of the Americas and to respond to the challenge of infectious pathogens corresponding to European contact. Future functional investigations will allow a more detailed understanding of the consequences of selective pressures experienced in the American continent, including its effect on present-day health outcomes.

Materials and Methods

Genomic datasets

The Latin American individual samples analyzed here were part of CANDELA Consortium (Ruiz-Linares et al. 2014). The CANDELA Consortium samples (http://www.ucl.ac.uk/silva/candela) have been described in detail in previous publications (Ruiz Linares et al 2014; Chacon-Duque et al., 2018). These data include a total of 6,630 volunteers from five Latin American countries (Brazil, Chile, Colombia, Mexico and Peru). This dataset was genotyped on the Illumina HumanOmniExpress chip platform including 730,525 SNPs. We also collated reference populations from regions that have contributed to the admixture in Latin America. For Native American samples we used individuals previously genotyped by Chacon-Duque et al. (2018). This dataset compromises 19 Native American populations from throughout the Americas with genotype data (supplementary table S1). For all the analyses described, we have only retained Native American individuals that showed more than 99% Native American ancestry as estimated by ADMIXTURE (see below). For European samples, we used genotype data from Portuguese and Spanish, individuals previously genotyped by Chacon-Duque et al. (2018) and Spanish (IBS; Iberian Population in Spain) from the 1000 Genomes Project study (The 1000 Genomes Project Consortium 2015). For Sub-Saharan Africans, we used genotype data from Yoruba (YRI; Yoruba in Ibadan, Nigeria), and Luhya (LWK; Luhya in Webuye, Kenya) individuals from the 1KGP. The reference samples from Chacon-Duque et al. (2018) are described in more detail in the Supplementary Table 1 from the mentioned publication. For some of our analysis we also included the 103 Han Chinese from Beijing (CHB) and 85 Europeans from England and Scotland (GBR) from the 1KGP as a surrogate for the Native American and European source, respectively. Genotype data of the individuals from the 1KGP was downloaded from the 1000 Genomes Project FTP site available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/.

Data curation

We used PLINK v1.9 (Chang et al. 2015) to exclude SNPs and individuals with more than 5% missing data or that showed evidence of genetic relatedness as in Chacon-Duque et al. (2018). Due to the admixed nature of the Latin American samples, there is an inflation in Hardy-Weinberg *P*-values, and therefore we did not exclude SNPs based on Hardy-Weinberg deviation. After applying these filters, 625,787 autosomal SNPs and 7,986 individuals were retained for further analysis.

Selecting admixed Latin American and reference individuals

In order to select admixed Latin American individuals (i.e. individuals with varying degrees of Native American, European and African ancestry), we conducted an unsupervised ADMIXTURE analysis at *K=3* using a set of 103,426 LD-pruned SNPs including Native Americans, Iberian Europeans and West Africans. We then removed non-admixed Latin American individuals that we define as having less than 10% or more than 90% Native American genome-wide ancestry. To avoid confounding our selection inference due to underlying population structure, we further excluded individuals with >1% inferred ancestry matching to surrogates other than Native Americans, Iberian Europeans, and West Africans using SOURCEFIND estimates obtained for the same individuals in Chacon-Duque et al. (2018). After this filtering procedure, the five Latin American populations consisted of 190 Brazilians (BRA), 1125 Colombians (COL), 896 Chileans (CHL), 773 Mexicans (MEX) and 834 Peruvians (PER). From our Native American, European, and Sub-Saharan African reference populations, we also removed individuals that contained more than 1% of ancestry from another group based on the ADMIXTURE analysis described above. After this extra filter our final reference dataset was composed of 142 Native Americans, 205 Europeans, and 206 Sub-Saharan Africans.

Change in allele frequency under Wright-Fisher with multiplicative model of selection

Assuming a multiplicative model of selection and random mating, the frequency of the three genotypes in generation 1 at a biallelic locus with alleles A and a at frequencies p and $1 - p$, respectively, in the previous generation is:

AA	Aa	aa
$(1+s_1)^2 p^2/c_1$	$(1+s_1)2p(1-p)/c_1$	$(1-p)^2/c_1$

where $s_1 \in [-1, \infty]$ is the selection coefficient in generation 1 and $c_1 = (1+s_1)^2p^2 + (1+s_1)2p(1-p) + (1-p)^2$. Note that each copy of the A allele changes fitness by a factor of $(1+s_1)$.

More generally, the allele frequency p_g of allele A in generation g is:

$$p_g = \frac{(1+s)p}{1+sp}, \qquad (1)$$

where

$$s = \left[\sum_{i=1}^{g} s_i\right] + \left[\sum_{j=1}^{g-1}\left(s_j \sum_{i=j+1}^{g} s_i\right)\right] + \sum_{i=3}^{g} \prod_i \approx \sum_{i=1}^{g} s_i, \qquad (2)$$

with s_i the selection coefficient at generation i and Π_i the sum of the products of all $\binom{g}{i}$ combinations of $\{s_1, \ldots, s_i\}$ values. The approximation in equation (4) assumes the s_i are small, which should be a reasonable approximation based on e.g. estimated selection coefficients in humans.

Testing for evidence of selection at a SNP

To assess the evidence of selection at a SNP, we employ a model inspired by that used in Mathieson et al. (2015) and based on the Balding-Nichols formulation (Balding and Nichols 1995). In particular for the allele count X_j at SNP j in the target population, we assume:

$$Pr\left(X_j = x_j \middle| M, p_j, D\right) = Beta - Binomial\left(x_j; 2M, \frac{1-D}{D} p_j, \frac{1-D}{D}\left(1 - p_j\right)\right), \qquad (3)$$

where M is the number of target individuals. The above model implicitly assumes that the frequency of the allele in the target population follows a $Beta(mean = p_j, variance = Dp_j(1 - p_j))$. Under neutrality, we assume

$$p_j = \frac{1}{M} \sum_{k=1}^{K} \left(\left[\sum_{i=1}^{M} \alpha_k(i)\right] f_{jk}\right), \qquad (4)$$

where fjk is the sampled frequency of the allele in the surrogate population at SNP j for source k, and $\propto_k$ (i) is the inferred admixture proportion from population k in individual i. We first find $\widehat{D}$ as the value of D that maximizes $\prod_{j=1}^{J}\left[Pr\left(X_j \middle| M, p_j, D\right)\right]$, using the optim function in R with the 'Nelder-Mead' algorithm. Then, fixing $D = \widehat{D}$ in equation (3), for each SNP we find the two-sided P-value testing the null hypothesis that the observed allele counts follow this neutral model.

The variance under (3) is small for SNPs with very high or very low p_j, so such SNPs tend to reject this null model even in cases where the observed target population allele frequency does not deviate notably from its neutral expectation p_j in (4). Therefore, we used an alternative parameterisation where we assumed the frequency of the allele in the target population follows a $Beta\left(mean = p_j, variance = V\right)$. This was achieved by substituting D in equation (3) at SNP j with

$min\left[\frac{V}{p_j(1-p_j)}, 0.99999\right]$, necessary to ensure numerical stability, and finding $\hat{V}$. In practice this means that SNPs with minor allele frequency $< (1.00001 \times V)$ had variance $\left(0.99999p_j(1-p_j)\right)$ rather than V, though this approach gave sensible results in practice.

Determining whether selection occurred pre or post-admixture

Consider the scenario in supplementary fig. S27, where sampled population C descends from an admixture of unsampled populations A^* and B^*, who are represented by sampled surrogate populations A and B, respectively. Our test aims to distinguish whether selection occurred post-admixture along branch (e) versus along any of branches (a)-(d). Let f_c be the allele frequency of a sample from population C. At a neutral SNP:

$$E[\mathrm{f_c}] = \alpha f_{A^*} + (1-\alpha) f_{B^*}, \qquad (5)$$

where f_{A^*} and f_{B^*} are true allele frequencies of A^* and B^* at the SNP, respectively, and α is the admixture proportion from A^*. Letting f_k be the sampled allele frequency for population k serving as surrogate to the true admixing population k^*, it seems reasonable to assume:

$$E[\mathrm{f_c}] = \propto f_A + (1-\propto) f_B. \qquad (6)$$

Note that this also holds under selection along branch (f) in supplementary fig. S27, which we ignore here (but which can be tested by comparing allele frequencies in A and B). Equation (6) assumes that f_A and f_B are equally good proxied for the admixing populations' frequencies f_A^* and f_{B^*}, respectively, at the SNP, which may not be true. We test the effect of this using simulations, described below, in which surrogates vary in how well they reflect their respective true admixing sources.

In the case of a multiplicative model of selection along branch (e) in supplementary fig. S27 at this SNP, using equation (1) we assume:

$$E[\mathrm{f_c}] = \frac{(1+s)[\propto f_A + (1-\propto) f_B]}{1 + s[\propto f_A + (1-\propto) f_B]} \equiv E_c[f_c], \qquad (7)$$

where s is the selection strength (i.e. equation [2]) along branch (e).

Alternatively, under a multiplicative model for selection along branches (a) and/or (c) in supplementary fig. S27, with analogous results for selection along branches (d) and/or (b), instead we assume:

$$E[\mathrm{f_c}] = \propto \left[\frac{(1+s_A)f_A}{1+s_A f_A}\right] + (1-\propto)f_B = f_B + \propto \left[\frac{(1+s_A)f_A}{1+s_A f_A} - f_B\right] \equiv E_A[\mathrm{f_c}], \tag{8}$$

where s_A is the selection strength along branches (a) and/or (c). Importantly, $E_A[f_c]$ is linear in $\propto$, while $E_C[f_c]$, is not, which we aim to exploit to distinguish between these two scenarios.

Here, assuming CANDELA population T can be described as a mixture of K sources, we assume the genotype g_i of individual $i \in [1, \ldots, M]$ from T follows:

$$g_i \sim Binomial\big(2, f_T(i)\big). \tag{9}$$

Under neutrality, we set $f_T(i)$ in (9) to:

$$f_T^N(i) = \sum_{k=1}^{K} [\propto_k (i)\mathrm{f_k}], \tag{10}$$

where f_k is the sampled allele frequency at the given SNP for the surrogate population to the source contributing $\propto_k (i)$ admixture to individual i.

In the case of selection in T post-admixture, we generalize equation (7) and set $f_T(i)$ in (9) to:

$$f_T^P(i|s) = \frac{(1+s)[\sum_{k=1}^{K} \alpha_k(i) f_k]}{1+s[\sum_{k=1}^{K} \alpha_k(i) f_k]}. \tag{11}$$

For the alternative case of selection along the branches separating source A and its sampled surrogate A^*, we generalize equation (8) and replace $f_T(i)$ in (9) with:

$$f_T^A(i|s_A) = \left[\sum_{k!=A}^{K} \alpha_A(i) f_k\right] + \alpha_A(i)\left[\frac{(1+s_A)f_A}{1+s_\mathrm{A} f_A}\right]. \tag{12}$$

In practice, we fix $\propto_A(i)$ to be the proportion of DNA that each target individual i matches to surrogate k as inferred by ADMIXTURE. We define:

$$\mathrm{L}^{\mathrm{P}}(s) \equiv \prod_{i=1}^{M} \left[f_T^P(i|s)^{g_i} \left(1 - f_T^P(i|s)\right)^{2-g_i} \right], \qquad (13)$$

where g_i is the genotype for target individual i. We use the optim function in R with the 'Nelder-Mead' algorithm to find the maximum-likelihood estimate (MLE) $\hat{s}$, which is the value of s that maximizes equation (13).

Similarly we define:

$$L^A(s_A) \equiv \prod_{i=1}^{M} \left[f_T^A(i|s_A)^{g_i} \left(1 - f_T^A(i|s_A)\right)^{2-g_i} \right], \qquad (14)$$

again finding $\hat{s}_A$, as the MLE for s_A.

We note that $[2 - 2log(L^P(\hat{s})]$ and $[2 - 2log(L^A(\hat{s}_A))]$ are analogous to AIC values for these respective models. Following AIC theory, we calculate:

$$I = \frac{min[L^P(\hat{s}), L^A(\hat{s}_A)]}{max[L^P(\hat{s}), L^A(\hat{s}_A)]} \leq 1, \qquad (15)$$

where, relative to the model with higher likelihood out of (13) and (14), the model with smaller likelihood is I times as probable to minimise the loss of information when used to represent the unknown true model (Akaike 1974).

Note we could analogously calculate the likelihood under the neutral model, i.e., using equation (10). Then, as an alternative to the selection testing approach described in Section 'Testing for evidence of selection at a SNP', we could use a likelihood-ratio-statistic approach to test for selection using either (13) or (14) as the alternative model likelihood. We explored this alternative testing approach, but do not use it here because it gave lower *P*-values when simulating under neutrality. This observation may in part be alleviated if we estimated f_{k^*} under both the neutral and alternative models rather than fixing $f_{k^*} = f_k$. However, estimating f_{k^*} is confounded with estimating s or s_A under the alternative models.

Simulations

Estimating how well each surrogate reflects its corresponding true admixing source

We aimed to generate simulations that mimic our real data. To do so, we first generate a measure of how well a sampled surrogate population k reflects its corresponding true (unknown) source population. In particular, we estimate a drift parameter d_k in the following manner. First, at each SNP j we use nlminb in R to find the estimated values $\{\tilde{f}_1^j, \ldots, \tilde{f}_K^j\}$ for $\{f_{1^*}, \ldots, f_{K^*}\}$, respectively, that minimize:

$$\sum_{i=1}^{M} \left(x_i^j - \sum_{k=1}^{K} \alpha_k(i) f_{k^*} \right)^2, \qquad (16)$$

Where $x_i^j \in \{0,1,2\}$ is the allele count for the admixed target individual $i \in [1, \ldots, M]$ at the SNP and each $\tilde{f}_k^j \in [0,1]$. Then, for each source k, with observed allele counts G_k^j and total counts M_k^j at SNP j in the surrogate population, following Balding-Nichols (Balding and Nichols 1995) we assume:

$$G_k^j Beta \sim Binomial\left(M_k^j \frac{d_k}{1-d_k} \tilde{f}_k^j, \frac{d_k}{1-d_k} \left(1 - \tilde{f}_k^j\right) \right). \qquad (17)$$

We then used the 'Nelder-Mead' algorithm in the optim function in R to find the $d_k \in [0,1]$ that maximized the product of (17) across all SNPs. This gave the values reported in Table 1.

Large estimated d_k (>0.1) correspond to cases where there is little admixture from that source in our sampled individuals from that country, i.e. for African admixture in most countries and Native American admixture in Brazil. As values inferred using such little data are presumably unreliable, we cap them at 0.05 for the simulations below. While these values are a guide, in practice we adjusted these values by a multiple of 2-7 to generate neutral simulations that had the same inferred drift $\widehat{D}$, described in section 'Testing for evidence of selection at a SNP', as that observed in the real data.

Target	Native American	European	African
Brazil	0.173	0.007	0.102
Chile	0.02	0.011	0.226
Colombia	0.044	0.012	0.044
Mexico	0.024	0.007	0.223
Peru	0.015	0.009	0.119

Table 1. Inferred d_k measuring how well the sampled surrogate (column) reflect the true admixing sources for each target population (row).

Generating simulated allele frequencies

We simulated admixed individuals who had experienced selection, with genome-wide admixture proportions $\propto_k (i)$ from source populations $k \in [1, \ldots, K]$ for simulated individuals $i \in [1, \ldots, M]$ matching those inferred by ADMIXTURE in the real data. To do so, for each simulation we repeated the following procedure:

1. For each source k, at each SNP we sample starting allele frequencies f_{k^*} from a $Beta\left(\frac{d_k}{1-d_k} f_k, \frac{d_k}{1-d_k}(1 - f_k)\right)$, where f_k is the sampled frequency of the respective surrogate population and d_k are defined in Table 1 (but capped at 0.05).
2. We randomly select SNPs to undergo selection. If selection is occurring in source population k prior to admixture, we randomly sample from among SNPs for which $f_{k^*} < 0.5$. If selection is occurring post-admixture, we instead randomly sample from among SNPs for which $\sum_{i=1}^{M} (\sum_{k=1}^{K} f_{k^*} \propto_k (i))/M < 0.5$.
3. We randomly select neutral SNPs from among all remaining SNPs, i.e., those not among the SNPs chosen in (2), in the real data.
4. To simulate selection:
    - If selection is occurring prior to admixture, we simulate selection in the relevant source population for g generations under a specified model of selection (additive, dominant, multiplicative, recessive) using Wright-Fisher with a population size of N_e indiviuals.
    - If selection is occurring after admixture, we simulate selection separately in each of the source populations for g generations, under a specified model of selection using Wright-Fisher with a population size of N_e individuals per population.

5. At each SNP, we sample allele counts for each individual i from a $Binomial(2, p_i)$ with $p_i = \sum_{k=1}^{K}[f_k^g \propto_k (i)]$, where:
   - f_k^g = $\mathrm{f}_{\mathrm{k}^*}$ for neutral SNPs
   - $f_k^g = f_{k^*}$ at selected SNPs for source populations k not undergoing selection (i.e., in cases where selection is pre-admixture)
   - f_k^g is the sampled final frequencies in step (4) after g generations, at selected SNPs for source population k undergoing selection

We then analyse data from the simulated target population individuals using the real sampled data from the surrogate populations. For simulations here, we use $N_e = 10000$ for the African, European, and Native American source groups.

Our procedure in steps (4)-(5) to simulate selection and admixture ensures the admixed individuals have variable admixture proportions while remaining computationally tractable. An alternative to this would be to generate M admixed populations using observed f_k values, with the admixture proportions for population i equal to $\alpha_1(i), \dots, \alpha_K(i)$, and then simulate each admixed population for g generations using Wright-Fisher, either with or without selection. Such simulations would match the approach used by our model to classify selection as type (i) or type (ii) (Section 'Determining whether selection occurred pre- or post-admixture'). However, we chose the above for reasons of computational efficiency, as we have many individuals (i.e., $M > 1000$). Note also that our selection test (Section 'Determining whether selection occurred pre- or post-admixture') is different from this simulation procedure, in that our test models the combined allele frequency across all admixed individuals, using the mean admixture contributions across target individuals to calculate the expected frequency. This may explain why our model exhibits an excess of SNPs with small P-values even when simulating no selection. This is despite using all SNPs to infer our model's variance parameter, which is designed to make more SNPs fit the model (likely explaining the excess of high P-values we also see, e.g., in supplementary fig. S14). While including this variance parameter does somewhat control P-values by e.g., giving a median P-value near 0.5, as expected under neutrality, our no-selection simulations suggest caution in directly using our model's P-values for assessing selection evidence. This suggests some degree of plausible simulations would be helpful to calibrate the model's reported P-values.

Local ancestry analysis

Local ancestry assignment was conducted using the HMM approach implemented in ELAI (Guan

2014). The phased genotype data needed as input was obtained by using SHAPEIT2 (Delaneau et al. 2012) with default parameter settings. Genetic distances were obtained from the HapMap Phase II genetic map build GRCh37 (Gibbs et al. 2003). As reference continental panels, we used the same Native American, European, and African individuals as in our AdaptMix analysis. ELAI was run setting the admixture generation parameter to 20, and with 20 rounds of EM iterations. To obtain local ancestry assignment probabilities, we conducted 10 independent runs and averaged probabilities across all runs as recommended in the ELAI manual. To test for local ancestry deviations we estimated *Z*-scores for each ancestry across each locus, and obtained the corresponding one-sided *P*-values testing for a positive deviation.

Population Branch Statistic (PBS) analysis

We first selected Latin American individuals carrying a specific Native American ancestry component based on the inferred Native American ancestry proportions previously estimated by Chacon-Duque et al 2018 in the CANDELA sample. Specifically, for each Native American ancestry component, we selected CANDELA individuals with >10% inferred ancestry from that particular Native American ancestry component, and with <1% combined inferred ancestry combined across all other Native American components. Thus, each group of admixed Latin Americans was composed primarily of Native American ancestry from a particular Native American component, plus European and African ancestry. We then estimated allele frequencies for each Native American component by considering only alleles (i.e. haplotypes) that were considered of Native American origin with local-ancestry posterior probability >0.9. We only computed allele frequencies for a Native American component if all SNPs genome-wide had >100 alleles (haplotypes) assigned to Native American origin. This resulted in allele frequency estimates for six Native American components, including ‘Quechua’, ‘Andes Piedmont’, ‘Chibcha Paez’, ‘Nahua1’, ‘South Mexico’, and ‘Mapuche’ ancestral components (see Chacon-Duque et al. (2018) for a detail description of the inferred components). Pairwise F_{ST} were then estimated using Hudson’s estimator as in equation 9 of Bhatia et al. (2013). The branch length (*T*) between two populations was computed as $T = -log_{10}(1 - F_{ST})$ (Cavalli-Sforza 1969). The Population Branch Statistic (PBS) (Yi et al. 2010) combines the pairwise branch lengths between three populations, which was computed as:

$$PBS_{Target} = \frac{T^{Target,Control} + T^{Target,Outgroup} + T^{Control,Outgroup}}{2}.$$

PBS values were computed for each Native American component, using all possible pairwise combinations of the other Native components as the control and outgroup populations. The rationale of this analysis was to try to find signals of selection exclusive to a given Native American group (i.e. that likely occurred after the divergence between Native American lineages). For some of our analysis we also used the CHB population from the 1000 Genomes Project, the European reference population, or the African reference population, as control and outgroup populations.

Summary statistics for GWAS and eQTL data

To assess the biological consequence of selected variants, we queried summary statistics from GWASs of relevant phenotypes, and gene-expression data (i.e expression quantitative locus [eQTL] studies) from relevant cell or tissues. For our GWAS query, we retrieved data from immune and metabolic-related phenotypes, as these traits are known to have been subjected to strong selective pressures across several human groups (Fan et al. 2016). Immune-related phenotypes included (i) total white cell count, neutrophil count, lymphocyte count, monocyte count, basophil count, and eosinophil count from the Chen et al. (2020) GWAS study conducted across five continental ancestry groups. Metabolic-related phenotypes included body mass index (BMI), body fat percentage, type II diabetes status, hip circumference, waist circumference, HDL levels, LDL levels, cholesterol levels, and triglycerides levels (Loh et al. 2018). Summary statistics from these GWAS analyses were based on the UK BioBank cohort available at: http://www.nealelab.is/uk-biobank. For our eQTL query, we retrieved cis-associations summary statistics of 15 human immune cell types from the DICE (Database of Immune Cell Expression, Expression quantitative trait loci [eQTLs] and Epigenomics) project (Schmiedel et al. 2018), available at: https://dice-database.org/downloads. We also retrieved cis-association summary statistics from adipose (subcutaneous, and visceral omentum), muscle (skeletal), and liver tissue from the GTEx Project v7 (Lonsdale et al. 2013) available at: https://gtexportal.org/home/datasets.

Acknowledgements

We thank the volunteers for their enthusiastic support for this research. We also thank Alvaro Alvarado, Mónica Ballesteros Romero, Ricardo Cebrecos, Miguel Ángel Contreras Sieck, Francisco de Ávila Becerril, Joyce De la Piedra, María Teresa Del Solar, Paola Everardo Martínez, William Flores, Martha Granados Riveros, Rosilene Paim, Ricardo Gunski, Sergeant João Felisberto Menezes Cavalheiro, Major Eugênio Correa de Souza Junior, Wendy Hart, Ilich Jafet Moreno, Paola León-Mimila, Francisco Quispealaya, Diana Rogel Diaz, Ruth Rojas, and Vanessa Sarabia, for assistance with volunteer recruitment, sample processing and data entry. We also thank Francois Balloux, Aida Andres, Mark McCarthy, and Etienne Patin for helpful discussion and critical comments on earlier versions of the manuscript. We are very grateful to the institutions that allowed the use of their facilities for the assessment of volunteers, including: Escuela Nacional de

Antropología e Historia and Universidad Nacional Autónoma de México (México); Universidade Federal do Rio Grande do Sul (Brazil); 13° Companhia de Comunicações Mecanizada do Exército Brasileiro (Brazil); Pontificia Universidad Católica del Perú, Universidad de Lima and Universidad Nacional Mayor de San Marcos (Perú). Work leading to this publication received funded from: the National Natural Science Foundation of China (#31771393 to ARL), the Scientific and Technology Committee of Shanghai Municipality (18490750300 to ARL), Ministry of Science and Technology of China (2020YFE0201600 to ARL), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01 to ARL) and the 111 Project (B13016 to ARL), the Leverhulme Trust (F/07 134/DF to ARL), BBSRC (BB/I021213/1 to ARL), the Excellence Initiative of Aix-Marseille University - A*MIDEX (a French "Investissements d'Avenir" programme to ARL), Wellcome Trust/Royal Society (098386/Z/12/Z to GH), the National Institute for Health Research University College London Hospitals Biomedical Research Centre, BBSRC (BB/R01356X/1), Universidad de Antioquia (CODI sostenibilidad de grupos 2013- 2014 and MASO 2013-2014), Conselho Nacional de Desenvolvimento Científico e Tecnológico, Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (Apoio a Núcleos de Excelência Program) and Fundação de Aperfeiçoamento de Pessoal de Nível Superior. JM-R was supported by a doctoral scholarship from CONCYTEC-PERU (224-2014-FONDECYT).

Data availability

This project only analyses data that has been previously reported in other publications. Raw genotype data for reference populations can be accessed as described previously (The 1000 Genomes Project Consortium 2015; Chacon-Duque et al. 2018). Raw genotype data from CANDELA cannot be made available due to restrictions imposed by ethical approval. Summary statistics from the selection analysis will be deposited in a public repository upon publication.

Software availability

Scripts for selection analyses will be uploaded to a software developer public repository upon publication. The current version of AdaptMix presented in this study is available upon request from g.hellenthal@ucl.ac.uk.

Main Figure legends

\Fig. 1. Schematic and intuition of the AdaptMix model. (a) For each CANDELA individual (columns), ADMIXTURE-inferred proportions of ancestry related to Native American, European, and African reference individuals. **(b)** Assuming only two admixing sources in this illustration for simplicity, the model assumes ancestral populations (K_1' and K_2') contribute ancestry proportions $\alpha_{K_1}$and α_{K_2}, respectively, to an admixed population (X') that is ancestral to the tested population (X). Assuming neutrality, the expected allele frequency (p_0) of X' is estimated using these proportions and the allele frequencies surrogate populations K_1 and K_2 related to K_1' and K_2', respectively. The sampled allele frequency (p) of X is compared to p_0, with large deviations indicative of selection (shown with an asterisk in the distribution). **(c and d)** The relationship between p_0, the expected allele frequency in the admixed population under neutrality or selection, and α_{K_2}, the ancestry proportion contributed from ancestral population K_2'. If selection occurred prior to admixture during the split between populations K_2' and its surrogate K_2 (i.e. along the blue branch in **[a]**), this relationship increases linearly (blue lines), becoming more differentiated from neutrality (grey line) as the admixture from K_2' increases. In contrast, under selection post-admixture (i.e. along the purple branch in **a]**), the expected allele frequency (purple lines) can deviate from neutrality even when the admixture from K_2' is near 0. The difference between the post-admixture and pre-admixture lines is more clear when allele frequencies in populations K_1 and K_2 are similar (top plot). Solid blue and red lines indicate the allele frequencies in the surrogate populations K_1 and K_2, which are used to calculate p_0.

Fig 2. Performance of AdaptMix to detect and classify selection in simulated Latin American populations. (a) Power to detect selection post-admixture, selection in Native Americans, or selection in Europeans in simulated populations mimicking the Latin American cohorts. Power is based on a *P*-value cutoff that resulted in a false-positive rate of 5×10^{-5} in neutral simulations. The power estimated for a given selection coefficient is based on combining simulations using four different modes of selection (additive, dominant, multiplicative, recessive) over 12 generations for the post-admixture simulations, over 50 generations for the selection in Native American simulations, and over 25 generations for the selection in European simulations. Each simulation for a given combination of parameters consisted of 10,000 advantageous SNPs with a pre-selection minor allele frequency lower than 0.5. **(b)** The proportion of significant SNPs from (a) that were assigned to the correct simulated scenario of (left-to-right) post-admixture selection or selection in Native Americans or Europeans (using a likelihood ratio > 1,000 to make a call; otherwise 'Unclassified'). Rows give the true selection coefficient (legend at right), and the heatmap values

give the classification rate. Rows with N.A. shows instances with less than 50 selected SNPs for which the classification rate is not shown.

Fig. 3. Genome-wide selection scan in five Latin American cohorts. Manhattan plot showing the genomic regions identified as selected via AdaptMix in each Latin American cohort. The dashed horizontal lines indicate the *P*-values cutoffs corresponding to a false-positive rate of 5×10^{-5} based on neutral simulations. Different shapes represent the most likely selection model. Names of genes associated with significant SNPs are shown.

Fig 4. Regional selection plot at the HLA region in five Latin American cohorts. The top plot shows the $-\log_{10}$(*P*-values) of SNPs from AdaptMix, the middle plot shows *Z*-score values based on African local ancestry deviations, and the bottom plot shows genes in the region shaded in grey. Genomic coordinates are in Mb (build hg19 as reference) and genes shown include transcripts.

Fig. 5. Genetic loci with signals of selection at immune-related genes. (a), **(b)** and **(c)** Regional selection plot at three candidate regions of selection encompassing two immune-related genes in the Chilean and one immune-related gene in the Peruvian cohort, respectively. Each plot is composed of four panels (rows), consisting of $-\log_{10}$(*P*-values) of SNPs: (row 1) from AdaptMix; (row 2) associated with immune-related cell counts via GWAS (Chen et al 2020); (row 3) associated (as expression quantitative trait loci [eQTLs]) with expression of genes *CD101*, *PTPN2* and *MIF* for (a)-(c), respectively (Schmiedel et al. 2018); with (row 4) depicting genes in the region (in Mb, build hg19 as reference. Horizontal dashed lines give significance thresholds of (row 1) *P*-value $= 1\times10^{-5}$ based on neutral simulations (row 2) *P*-value $= 1\times10^{-5}$ (blue line) and *P*-value $= 5\times10^{-8}$ (red line), and (row 3) *P*-value $= 1\times10^{-4}$. **(d)**, **(e)** and **(f)** Derived allele frequency (DAF) in admixed Latin Americans (white circles) stratified by proportion of inferred Native American ancestry, for the SNPs highlighted (vertical dashed line) in top row panels. The sizes of the circles are proportional to the number of individuals in that particular bin. Lines give expected DAF under neutrality (grey), post-admixture selection (brown) or selection in the Native source (black). Horizontal dashed red, blue, and green lines depict DAF for surrogates to Native American, European, and African sources, respectively.

Fig. 6. Genetic loci with signals of selection at metabolic-related genes. (a) and **(b)** Regional selection plot at two candidate regions of selection encompassing metabolic-related genes in the Mexican and Peruvian cohorts, respectively. Each plot is composed of four panels consisting of $-\log_{10}$(*P*-values) of SNPs: (row 1) from AdaptMix; (row 2) from the UK Biobank GWAS; (row 3)

associated (as eQTLs) with expression of *BRINP3* and *HKDC1* for (a)-(b), respectively, (GTEx eQTL study); with (row 4) depicting genes in the region (in Mb, build hg19 as reference). Horizontal dashed lines give significance thresholds of (row 1) P-value = 1×10^{-5} based on neutral simulations (row 2) P-value = 1×10^{-5} (blue line) and P-value = 5×10^{-8} (red line), and (row 3) P-value = 1×10^{-4}. **(c)** and **(d)** Derived allele frequency (DAF) in admixed Latin Americans (white circles) stratified by proportion of inferred Native American ancestry, for the SNPs highlighted (vertical dashed line) in top row panels. The sizes of the circles are proportional to the number of individuals in that particular bin. Lines give expected DAF under neutrality (grey), post-admixture selection (brown) or selection in the Native American source (black). Horizontal dashed red, blue, and green lines depict DAF for surrogates to Native American, European, and African sources, respectively.

References

Acuña-Alonzo V, Flores-Dorantes T, Kruit JK, Villarreal-Molina T, Arellano-Campos O, Hünemeier T, Moreno-Estrada A, Ortiz-López MG, Villamil-Ramírez H, León-Mimila P. 2010. A functional ABCA1 gene variant is associated with low HDL-cholesterol levels and shows evidence of positive selection in Native Americans. Human molecular genetics 19:2877-2885.

Akaike H. 1974. A new look at the statistical model identification. IEEE transactions on automatic control 19:716-723.

Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19:1655-1664.

Amorim CE, Nunes K, Meyer D, Comas D, Bortolini MC, Salzano FM, Hunemeier T. 2017. Genetic signature of natural selection in first Americans. Proc Natl Acad Sci U S A 114:2195-2199.

Avila-Arcos MC, McManus KF, Sandoval K, Rodriguez-Rodriguez JE, Villa-Islas V, Martin AR, Luisi P, Penaloza-Espinosa RI, Eng C, Huntsman S, et al. 2020. Population History and Gene Divergence in Native Mexicans Inferred from 76 Human Exomes. Mol Biol Evol 37:994-1006.

Badillo Rivera KM, Nieves-Colón MA, Mendoza KS, Davalos VV, Lencinas LEE, Chen JW, Zhang ET, Sockell A, Tello PO, Hurtado GM. 2021. Clotting factor genes are associated with preeclampsia in high altitude pregnant women in the Peruvian Andes. medRxiv.

Balding DJ, Nichols RA. 1995. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96:3-12.

Basu A, Tang H, Zhu X, Gu CC, Hanis C, Boerwinkle E, Risch N. 2008. Genome-wide distribution of ancestry in Mexican Americans. Hum Genet 124:207-214.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN. 2004. Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111-1120.

Bhatia G, Patterson N, Sankararaman S, Price AL. 2013. Estimating and interpreting FST: the impact of rare variants. Genome research 23:1514-1521.

Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ. 2014. Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. The American Journal of Human Genetics 95:437-444.

Borrego F. 2013. The CD300 molecules: an emerging family of regulators of the immune system. Blood, The Journal of the American Society of Hematology 121:1951-1960.

Bouloc A, Bagot M, Delaire S, Bensussan A, Boumsell L. 2000. Triggering CD101 molecule on human cutaneous dendritic cells inhibits T cell proliferation via IL-10 production. European journal of immunology 30:3132-3139.

Calandra T, Roger T. 2003. Macrophage migration inhibitory factor: a regulator of innate immunity. Nature Reviews Immunology 3:791-800.

Cavalli-Sforza LL editor.; 1969.

Chacon-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuna-Alonzo V, Barquera R, Quinto-Sanchez M, Gomez-Valdes J, Everardo Martinez P, Villamil-Ramirez H, et al. 2018. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. Nat Commun 9:5388.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 4:s13742-13015-10047-13748.

Chen M-H, Raffield LM, Mousas A, Sakaue S, Huffman JE, Moscati A, Trivedi B, Jiang T, Akbari P, Vuckovic D. 2020. Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations. Cell 182:1198-1213. e1114.

Cheng JY, Stern AJ, Racimo F, Nielsen R. 2021. Detecting selection in multiple populations by modelling ancestral admixture components. Mol Biol Evol.

Consortium TGP. 2015. A global reference for human genetic variation. Nature 526:68.

Delaneau O, Marchini J, Zagury J-F. 2012. A linear complexity phasing method for thousands of genomes. Nature methods 9:179-181.

Deng L, Ruiz-Linares A, Xu S, Wang S. 2016. Ancestry variation and footprints of natural selection along the genome in Latin American populations. Sci Rep 6:21766.

Ettinger NA, Duggal P, Braz RF, Nascimento ET, Beaty TH, Jeronimo SM, Pearson RD, Blackwell JM, Moreno L, Wilson ME. 2009. Genetic admixture in Brazilians exposed to infection with Leishmania chagasi. Ann Hum Genet 73:304-313.

Fan S, Hansen ME, Lo Y, Tishkoff SA. 2016. Going global by adapting local: A review of recent human adaptation. Science 354:54-59.

Galinsky KJ, Bhatia G, Loh PR, Georgiev S, Mukherjee S, Patterson NJ, Price AL. 2016. Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. Am J Hum Genet 98:456-472.

Ghoussaini M, Mountjoy E, Carmona M, Peat G, Schmidt EM, Hercules A, Fumis L, Miranda A, Carvalho-Silva D, Buniello A. 2021. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. Nucleic acids research 49:D1311-D1320.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y. 2003. The international HapMap project.

Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, Torstenson ES, Kovesdy CP, Sun YV, Wilson OD. 2019. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. Nature genetics 51:51-62.

Gu S, Li H, Pakstis AJ, Speed WC, Gurwitz D, Kidd JR, Kidd KK. 2018. Recent Selection on a Class I ADH Locus Distinguishes Southwest Asian Populations Including Ashkenazi Jews. Genes (Basel) 9.

Guan Y. 2014. Detecting structure of haplotypes and local ancestry. Genetics 196:625-642.

Guo C, Ludvik AE, Arlotto ME, Hayes MG, Armstrong LL, Scholtens DM, Brown CD, Newgard CB, Becker TC, Layden BT. 2015. Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. Nature communications 6:1-8.

Hamid I, Korunes KL, Beleza S, Goldberg A. 2021. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. Elife 10.

Hancock AM, Witonsky DB, Ehler E, Alkorta-Aranburu G, Beall C, Gebremedhin A, Sukernik R, Utermann G, Pritchard J, Coop G. 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proceedings of the National Academy of Sciences 107:8924-8930.

Harris DN, Ruczinski I, Yanek LR, Becker LC, Becker DM, Guio H, Cui T, Chilton FH, Mathias RA, O'Connor TD. 2019. Evolution of hominin polyunsaturated fatty acid metabolism: from Africa to the New World. Genome biology and evolution 11:1417-1430.

Hayes MG, Urbanek M, Hivert M-F, Armstrong LL, Morrison J, Guo C, Lowe LP, Scheftner DA, Pluzhnikov A, Levine DM. 2013. Identification of HKDC1 and BACE2 as genes influencing glycemic traits during pregnancy through genome-wide association studies. Diabetes 62:3282-3291.

Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. Science 343:747-751.

Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, Soodyall H, Shriver MD, Perry GH. 2014. Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proc Biol Sci 281:20140930.

Hoffmann TJ, Ehret GB, Nandakumar P, Ranatunga D, Schaefer C, Kwok P-Y, Iribarren C, Chakravarti A, Risch N. 2017. Genome-wide association analyses using electronic health records identify new loci influencing blood pressure variation. Nature genetics 49:54-64.

Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America. PLoS Genet 11:e1005602.

Joffe GM, Esterlitz JR, Levine RJ, Clemens JD, Ewell MG, Sibai BM, Catalano PM. 1998. The relationship between abnormal glucose tolerance and hypertensive disorders of pregnancy in healthy nulliparous women. American journal of obstetrics and gynecology 179:1032-1037.

Kanthimathi S, Liju S, Laasya D, Anjana RM, Mohan V, Radha V. 2016. Hexokinase domain containing 1 (HKDC1) gene variants and their association with gestational diabetes mellitus in a south indian population. Annals of human genetics 80:241-245.

Karlsson EK, Kwiatkowski DP, Sabeti PC. 2014. Natural selection and infectious disease in human populations. Nature Reviews Genetics 15:379-393.

Kominsky DJ, Campbell EL, Colgan SP. 2010. Metabolic shifts in immunity and inflammation. The Journal of Immunology 184:4062-4068.

Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, Hasan S, Karamanis N, Maguire M, Papa E. 2017. Open Targets: a platform for therapeutic target identification and validation. Nucleic acids research 45:D985-D994.

Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, Sudmant PH, Schraiber JG, Castellano S, Lipson M, et al. 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature 513:409-413.

Lindo J, Haas R, Hofman C, Apata M, Moraga M, Verdugo RA, Watson JT, Llave CV, Witonsky D, Beall C. 2018. The genetic prehistory of the Andean highlands 7000 years BP though European contact. Science advances 4:eaau4921.

Loh PR, Kichaev G, Gazal S, Schoech AP, Price AL. 2018. Mixed-model association for biobank-scale datasets. Nat Genet 50:906-908.

Long JC. 1991. The genetic structure of admixed populations. Genetics 127:417-428.

Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N. 2013. The genotype-tissue expression (GTEx) project. Nature genetics 45:580-585.

Ludvik AE, Pusec CM, Priyadarshini M, Angueira AR, Guo C, Lo A, Hershenhouse KS, Yang G-Y, Ding X, Reddy TE. 2016. HKDC1 is a novel hexokinase involved in whole-body glucose use. Endocrinology 157:3452-3461.

Luisi P, García A, Berros JM, Motti JM, Demarchi DA, Alfaro E, Aquilano E, Argüelles C, Avena S, Bailliet G. 2020. Fine-scale genomic analyses of admixed individuals reveal unrecognized genetic ancestry components in Argentina. PloS one 15:e0233808.

Lumeng CN, Saltiel AR. 2011. Inflammatory links between obesity and metabolic disease. The Journal of clinical investigation 121:2111-2117.

Mathieson I. 2020. Limited evidence for selection at the FADS locus in Native American populations. Molecular biology and evolution 37:2029-2033.

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

Moreno-Estrada A, Gignoux CR, Fernandez-Lopez JC, Zakharia F, Sikora M, Contreras AV, Acuna-Alonzo V, Sandoval K, Eng C, Romero-Hidalgo S, et al. 2014. Human genetics. The genetics of Mexico recapitulates Native American substructure and affects biomedical traits. Science 344:1280-1285.

Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martinez RJ, Hedges DJ, Morris RW, et al. 2013. Reconstructing the population genetic history of the Caribbean. PLoS Genet 9:e1003925.

Norris ET, Rishishwar L, Chande AT, Conley AB, Ye K, Valderrama-Aguirre A, Jordan IK. 2020. Admixture-enabled selection for rapid adaptive evolution in the Americas. Genome Biol 21:29.

Osuna-Ramos JF, Reyes-Ruiz JM, Del Ángel RM. 2018. The role of host cholesterol during flavivirus infection. Frontiers in cellular and infection microbiology 8:388.

Pasaniuc B, Sankararaman S, Torgerson DG, Gignoux C, Zaitlen N, Eng C, Rodriguez-Cintron W, Chapela R, Ford JG, Avila PC. 2013. Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. Bioinformatics 29:1407-1415.

Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich D. 2012. Ancient admixture in human history. Genetics 192:1065-1093.

Pickup J, Crook M. 1998. Is type II diabetes mellitus a disease of the innate immune system? Diabetologia 41:1241-1248.

Pierron D, Heiske M, Razafindrazaka H, Pereda-Loth V, Sanchez J, Alva O, Arachiche A, Boland A, Olaso R, Deleuze JF, et al. 2018. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. Nat Commun 9:932.

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM. 2003. The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet 67:298-311.

Pulit SL, Stoneman C, Morris AP, Wood AR, Glastonbury CA, Tyrrell J, Yengo L, Ferreira T, Marouli E, Ji Y. 2019. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. Human molecular genetics 28:166-174.

Racimo F, Gokhman D, Fumagalli M, Ko A, Hansen T, Moltke I, Albrechtsen A, Carmel L, Huerta-Sánchez E, Nielsen R. 2017. Archaic adaptive introgression in TBX15/WARS2. Molecular biology and evolution 34:509-524.

Reynolds AW, Mata-Miguez J, Miro-Herrans A, Briggs-Cloud M, Sylestine A, Barajas-Olmos F, Garcia-Ortiz H, Rzhetskaya M, Orozco L, Raff JA, et al. 2019. Comparing signals of natural selection between three Indigenous North American populations. Proc Natl Acad Sci U S A 116:9312-9317.

Rishishwar L, Conley AB, Wigington CH, Wang L, Valderrama-Aguirre A, Jordan IK. 2015. Ancestry, admixture and fitness in Colombian genomes. Sci Rep 5:12376.

Robbins GR, Wen H, Ting JP-Y. 2014. Inflammasomes and metabolic disorders: old genes in modern diseases. Molecular cell 54:297-308.

Ruiz-Linares A, Adhikari K, Acuna-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, Fuentes M, Pizarro M, Everardo P, de Avila F, et al. 2014. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. PLoS Genet 10:e1004572.

Rumold CU, Aldenderfer MS. 2016. Late Archaic–Early Formative period microbotanical evidence for potato at Jiskairumoko in the Titicaca Basin of southern Peru. Proceedings of the National Academy of Sciences 113:13672-13677.

Santoscoy-Ascencio G, Baños-Hernández CJ, Navarro-Zarza JE, Hernández-Bello J, Bucala R, López-Quintero A, Valdés-Alvarado E, Parra-Rojas I, Illades-Aguiar B, Muñoz-Valle JF. 2020. Macrophage migration inhibitory factor promoter polymorphisms are associated with disease activity in rheumatoid arthritis patients from Southern Mexico. Molecular genetics & genomic medicine 8:e1037.

Schmiedel BJ, Singh D, Madrigal A, Valdovino-Gonzalez AG, White BM, Zapardiel-Gonzalo J, Ha B, Altay G, Greenbaum JA, McVicker G. 2018. Impact of genetic polymorphisms on human immune cell gene expression. Cell 175:1701-1715. e1716.

Shuai K, Liu B. 2003. Regulation of JAK–STAT signalling in the immune system. Nature Reviews Immunology 3:900-911.

Sibai BM. 2003. Diagnosis and management of gestational hypertension and preeclampsia. Obstetrics & Gynecology 102:181-192.

Sirugo G, Williams SM, Tishkoff SA. 2019. The Missing Diversity in Human Genetic Studies. Cell 177:1080.

Soares LR, Tsavaler L, Rivas A, Engleman EG. 1998. V7 (CD101) ligation inhibits TCR/CD3-induced IL-2 production by blocking Ca2+ flux and nuclear factor of activated T cell nuclear translocation. The Journal of Immunology 161:209-217.

Tan Y-X, Hu S-M, You Y-P, Yang G-L, Wang W. 2019. Replication of previous genome-wide association studies of HKDC1, BACE2, SLC16A11 and TMEM163 SNPs in a gestational diabetes mellitus case–control sample from Han Chinese population. Diabetes, metabolic syndrome and obesity: targets and therapy 12:983.

Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. 2007. Recent genetic selection in the ancestral admixture of Puerto Ricans. Am J Hum Genet 81:626-633.

Van Dijk M, Mulders J, Poutsma A, Könst AA, Lachmeijer AM, Dekker GA, Blankenstein MA, Oudejans CB. 2005. Maternal segregation of the Dutch preeclampsia locus at 10q22 with a new member of the winged helix gene family. Nature genetics 37:514-519.

van Dijk M, Oudejans C. 2011. STOX1: key player in trophoblast dysfunction underlying early onset preeclampsia with growth retardation. Journal of pregnancy 2011.

Vicente M, Priehodova E, Diallo I, Podgorna E, Poloni ES, Cerny V, Schlebusch CM. 2019. Population history and genetic adaptation of the Fulani nomads: inferences from genome-wide data and the lactase persistence trait. BMC Genomics 20:915.

Vicuna L, Klimenkova O, Norambuena T, Martinez FI, Fernandez MI, Shchur V, Eyheramendy S. 2020. Postadmixture Selection on Chileans Targets Haplotype Involved in Pigmentation, Thermogenesis and Immune Defense against Pathogens. Genome Biol Evol 12:1459-1470.

Villarreal-Molina MT, Flores-Dorantes MT, Arellano-Campos O, Villalobos-Comparan M, Rodríguez-Cruz M, Miliar-García A, Huertas-Vazquez A, Menjivar M, Romero-Hidalgo S, Wacher NH. 2008. Association of the ATP-binding cassette transporter A1 R230C variant with early-onset type 2 diabetes in a Mexican population. Diabetes 57:509-513.

Wang S, Lewis CM, Jakobsson M, Ramachandran S, Ray N, Bedoya G, Rojas W, Parra MV, Molina JA, Gallo C, et al. 2007. Genetic variation and population structure in native Americans. PLoS Genet 3:e185.

Warren HR, Evangelou E, Cabrera CP, Gao H, Ren M, Mifsud B, Ntalla I, Surendran P, Liu C, Cook JP. 2017. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. Nature genetics 49:403-415.

Warrington NM, Beaumont RN, Horikoshi M, Day FR, Helgeland Ø, Laurin C, Bacelis J, Peng S, Hao K, Feenstra B. 2019. Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. Nature genetics 51:804-814.

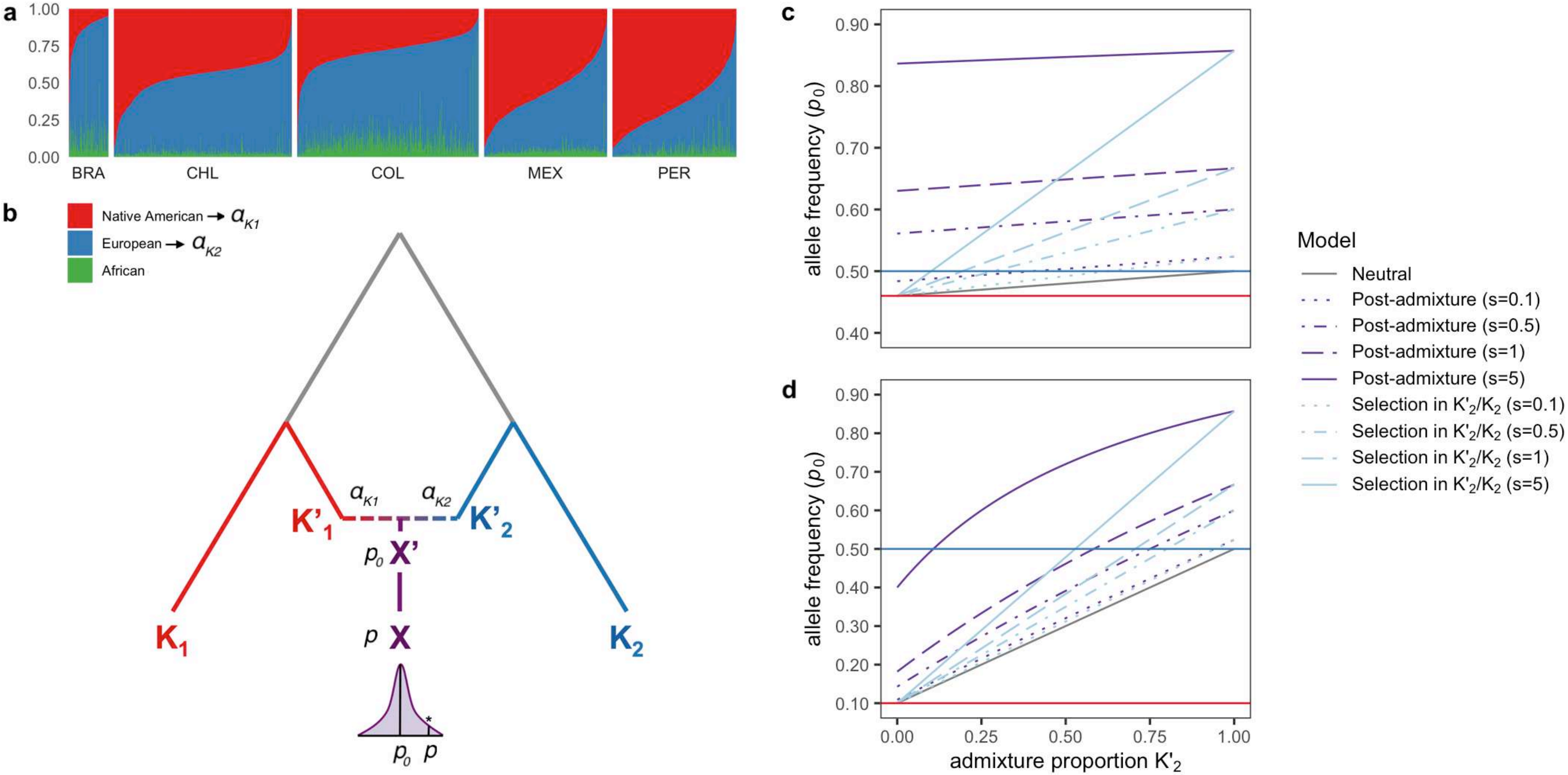
Weissgerber TL, Mudd LM. 2015. Preeclampsia and diabetes. Current diabetes reports 15:1-10.

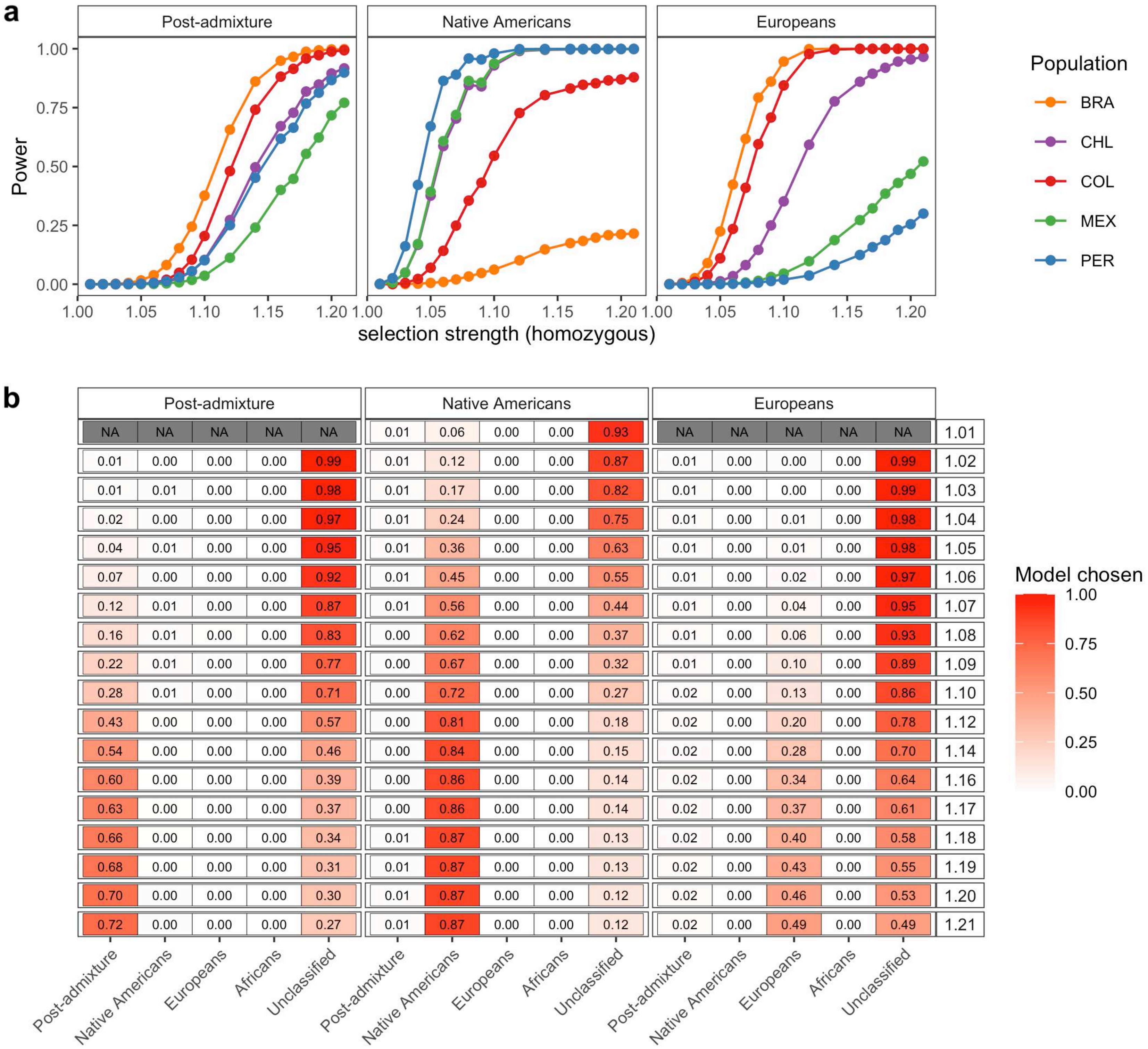
Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N, Korneliussen TS. 2010. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 329:75-78.

Zamudio S. 2007. High-altitude hypoxia and preeclampsia. Frontiers in bioscience: a journal and virtual library 12:2967.

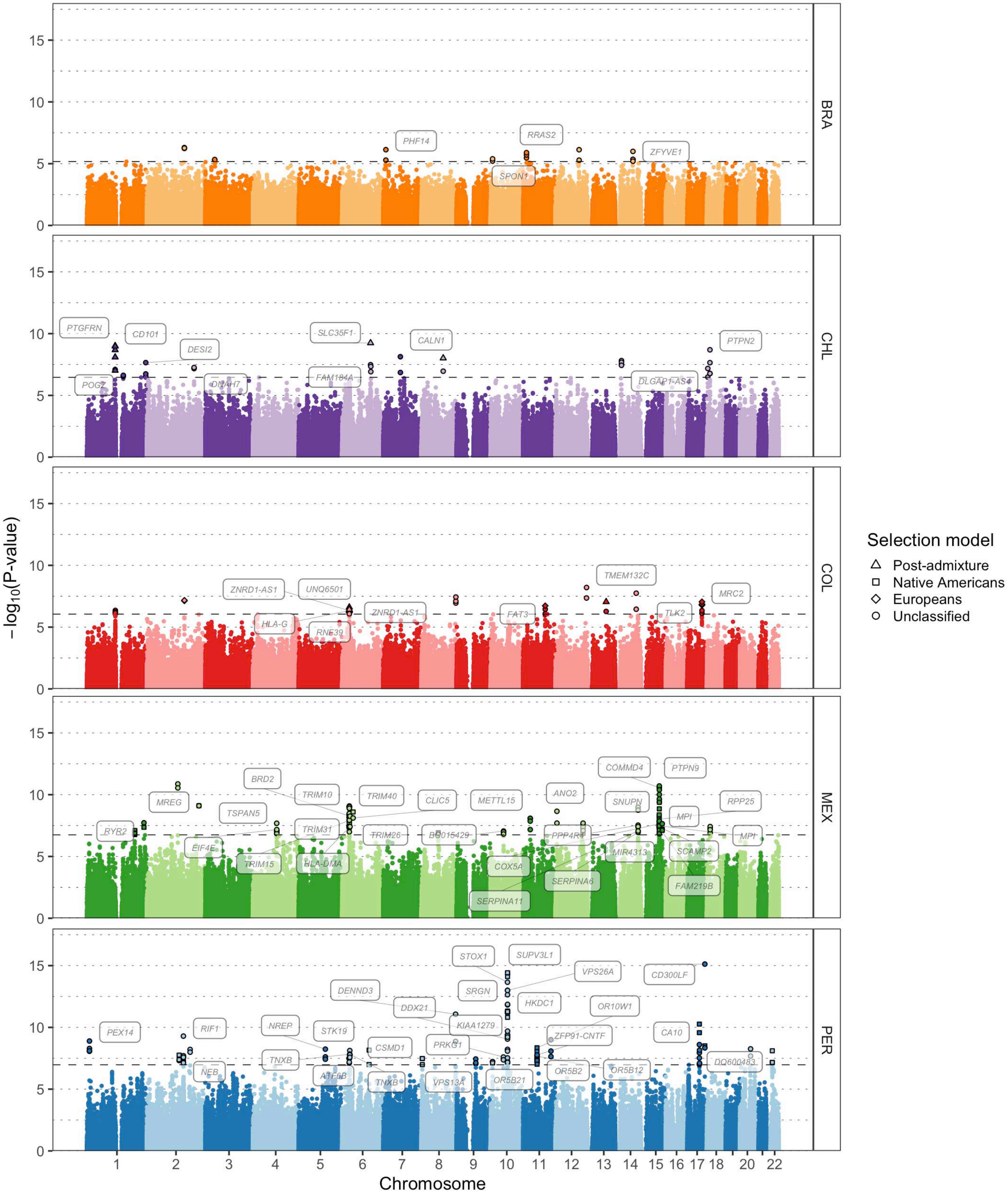
Zhou Q, Zhao L, Guan Y. 2016. Strong selection at MHC in Mexicans since admixture. PLoS genetics 12:e1005847.

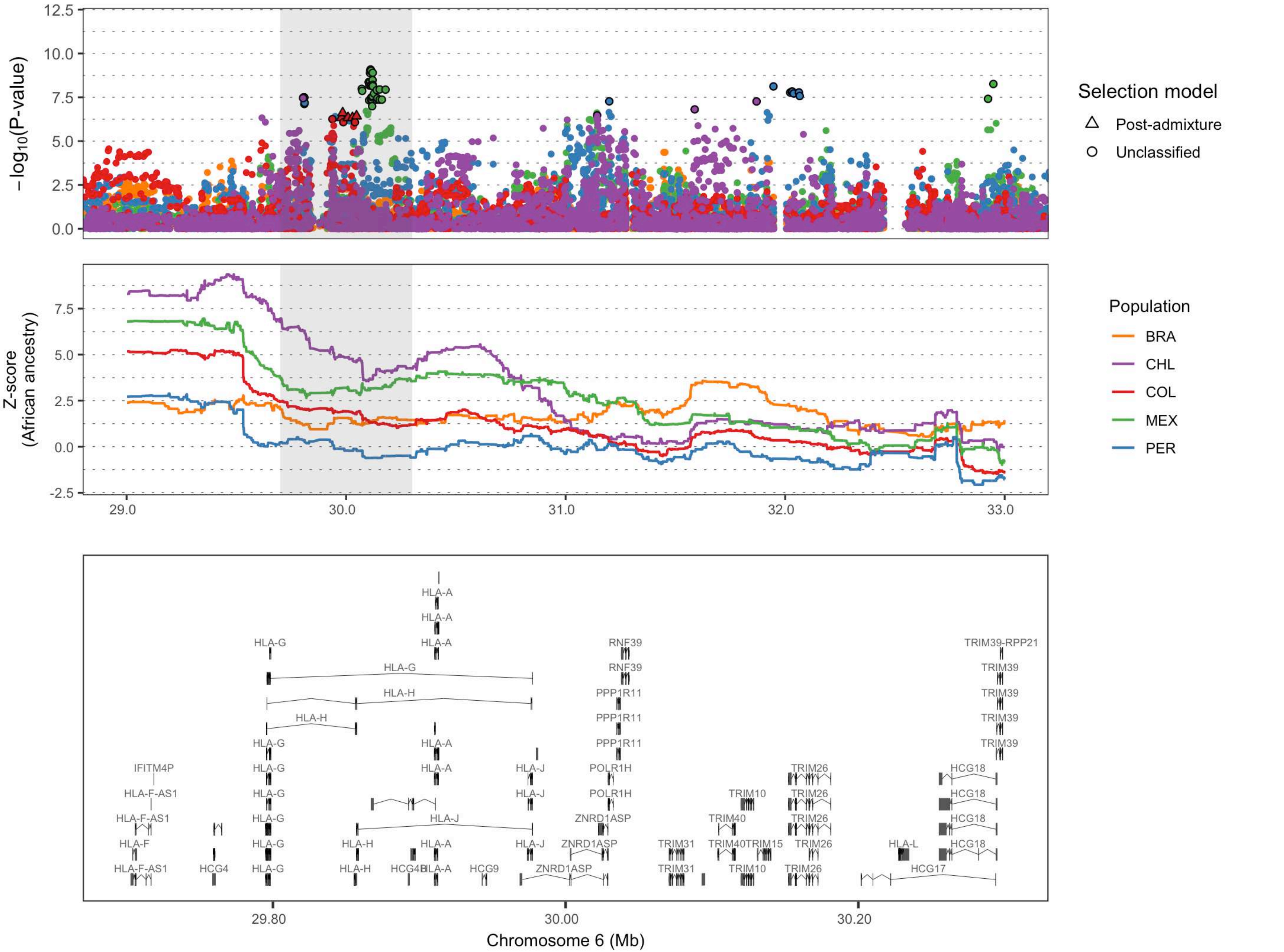
Zhu Z, Guo Y, Shi H, Liu C-L, Panganiban RA, Chung W, O'Connor LJ, Himes BE, Gazal S, Hasegawa K. 2020. Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank. Journal of Allergy and Clinical Immunology 145:537-549.

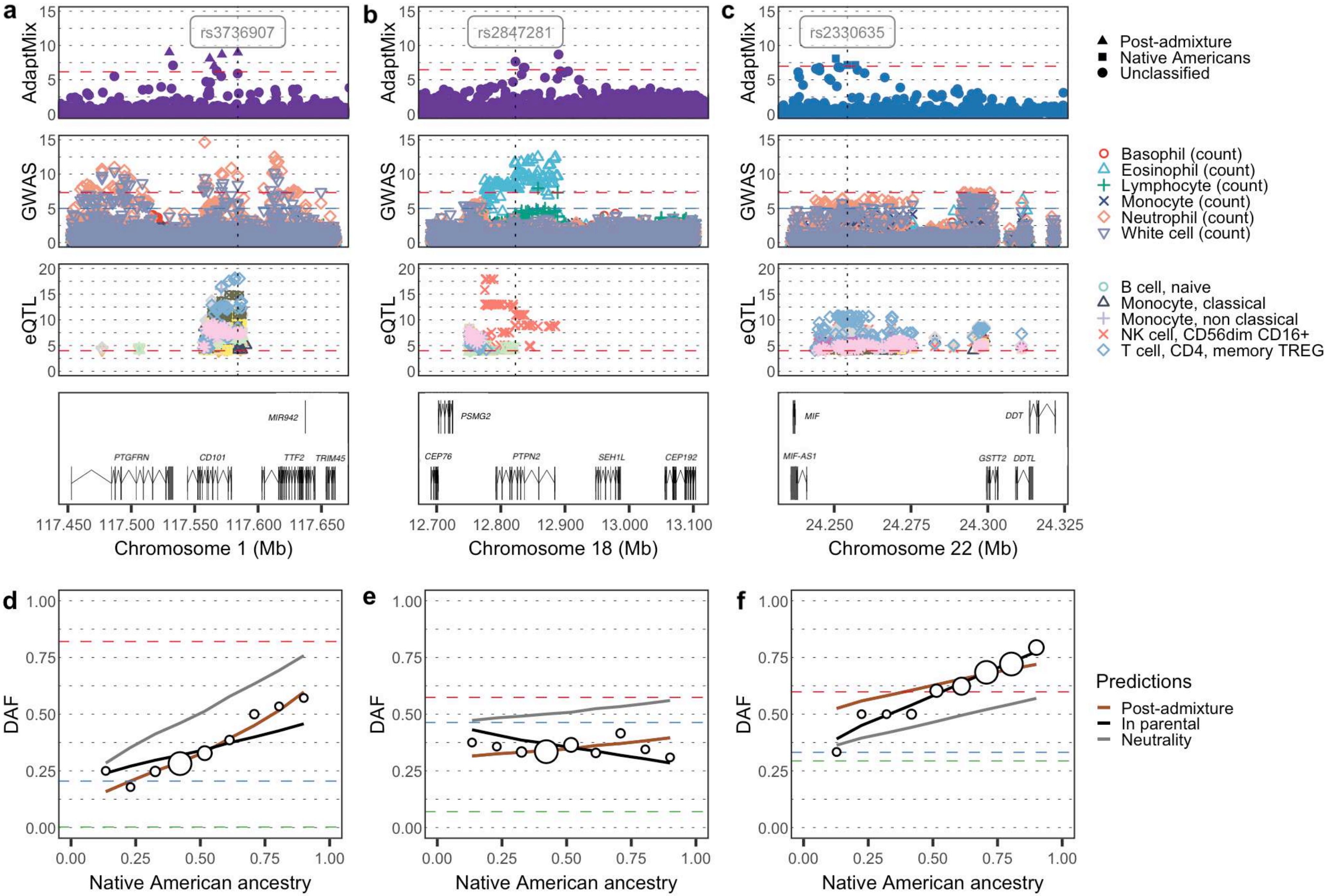

a
1.00
0.75
0.50
0.25
0.00
BRA
CHL
COL
MEX
PER
b
Native American → a_{K1}
European → a_{K2}
African
a_{K1}
a_{K2}
K'$_1$
K'$_2$
p_0 X'
p X
K$_1$
K$_2$
p_0
p
*
c
d
allele frequency (p_0)
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
0.25
0.50
0.75
1.00
admixture proportion K'$_2$
Model
Neutral
Post-admixture (s=0.1)
Post-admixture (s=0.5)
Post-admixture (s=1)
Post-admixture (s=5)
Selection in K'$_2$/K$_2$ (s=0.1)
Selection in K'$_2$/K$_2$ (s=0.5)
Selection in K'$_2$/K$_2$ (s=1)
Selection in K'$_2$/K$_2$ (s=5)

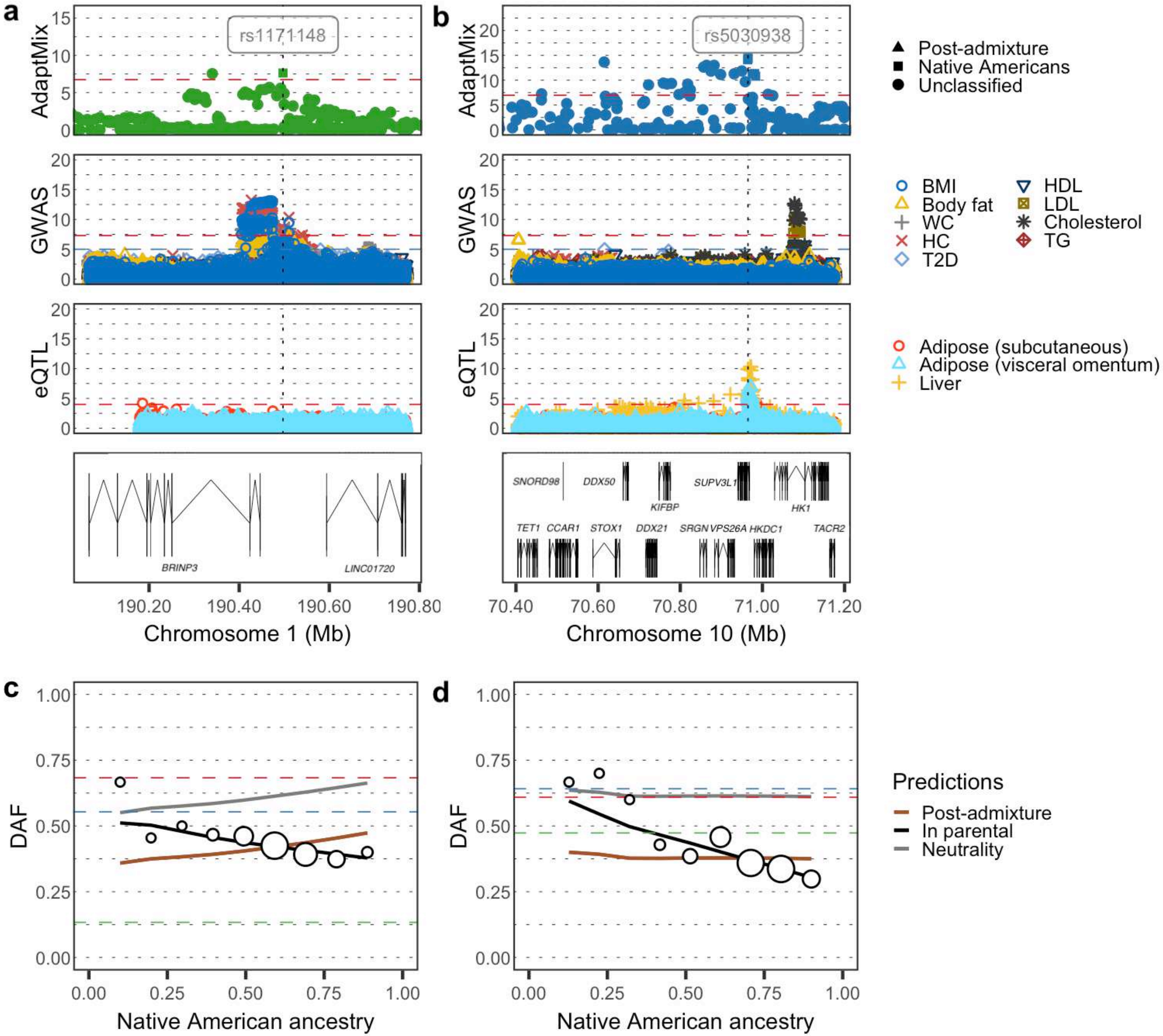


Post-admixture					Native Americans					Europeans					
Post-admixture	Native Americans	Europeans	Africans	Unclassified	Post-admixture	Native Americans	Europeans	Africans	Unclassified	Post-admixture	Native Americans	Europeans	Africans	Unclassified	
NA	NA	NA	NA	NA	0.01	0.06	0.00	0.00	0.93	NA	NA	NA	NA	NA	1.01
0.01	0.00	0.00	0.00	0.99	0.01	0.12	0.00	0.00	0.87	0.01	0.00	0.00	0.00	0.99	1.02
0.01	0.01	0.00	0.00	0.98	0.01	0.17	0.00	0.00	0.82	0.01	0.00	0.00	0.00	0.99	1.03
0.02	0.00	0.00	0.00	0.97	0.01	0.24	0.00	0.00	0.75	0.01	0.00	0.01	0.00	0.98	1.04
0.04	0.01	0.00	0.00	0.95	0.01	0.36	0.00	0.00	0.63	0.01	0.00	0.01	0.00	0.98	1.05
0.07	0.00	0.00	0.00	0.92	0.01	0.45	0.00	0.00	0.55	0.01	0.00	0.02	0.00	0.97	1.06
0.12	0.01	0.00	0.00	0.87	0.01	0.56	0.00	0.00	0.44	0.01	0.00	0.04	0.00	0.95	1.07
0.16	0.01	0.00	0.00	0.83	0.00	0.62	0.00	0.00	0.37	0.01	0.00	0.06	0.00	0.93	1.08
0.22	0.01	0.00	0.00	0.77	0.00	0.67	0.00	0.00	0.32	0.01	0.00	0.10	0.00	0.89	1.09
0.28	0.01	0.00	0.00	0.71	0.00	0.72	0.00	0.00	0.27	0.02	0.00	0.13	0.00	0.86	1.10
0.43	0.00	0.00	0.00	0.57	0.00	0.81	0.00	0.00	0.18	0.02	0.00	0.20	0.00	0.78	1.12
0.54	0.00	0.00	0.00	0.46	0.00	0.84	0.00	0.00	0.15	0.02	0.00	0.28	0.00	0.70	1.14
0.60	0.00	0.00	0.00	0.39	0.00	0.86	0.00	0.00	0.14	0.02	0.00	0.34	0.00	0.64	1.16
0.63	0.00	0.00	0.00	0.37	0.00	0.86	0.00	0.00	0.14	0.02	0.00	0.37	0.00	0.61	1.17
0.66	0.00	0.00	0.00	0.34	0.01	0.87	0.00	0.00	0.13	0.02	0.00	0.40	0.00	0.58	1.18
0.68	0.00	0.00	0.00	0.31	0.01	0.87	0.00	0.00	0.13	0.02	0.00	0.43	0.00	0.55	1.19
0.70	0.00	0.00	0.00	0.30	0.01	0.87	0.00	0.00	0.12	0.02	0.00	0.46	0.00	0.53	1.20
0.72	0.00	0.00	0.00	0.27	0.01	0.87	0.00	0.00	0.12	0.02	0.00	0.49	0.00	0.49	1.21


BRA
PHF14
RRAS2
ZFYVE1
SPON1
CHL
PTGFRN
CD101
DESI2
SLC35F1
CALN1
PTPN2
POGZ
DNAH7
FAM184A
DLGAP1-AS4
COL
TMEM132C
ZNRD1-AS1
UNQ6501
MRC2
ZNRD1-AS1
FAT3
TLK2
HLA-G
RNF39
MEX
COMMD4
PTPN9
BRD2
TRIM10
TRIM40
CLIC5
METTL15
ANO2
SNUPN
RPP25
MREG
TSPAN5
MPI
RYR2
TRIM31
TRIM26
BC015429
PPP4R4
MPI
EIF4E
MIR4313
SCAMP2
TRIM15
HLA-DMA
COX5A
SERPINA6
FAM219B
SERPINA11
PER
STOX1
SUPV3L1
VPS26A
CD300LF
DENND3
SRGN
HKDC1
OR10W1
DDX21
KIAA1279
PEX14
RIF1
NREP
STK19
CSMD1
PRKG1
ZFP91-CNTF
CA10
TNXB
DQ600483
NEB
ATF6B
TNXB
VPS13A
OR5B21
OR5B2
OR5B12
0
5
10
15
−log10(P-value)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
20
22
Chromosome
Selection model
△ Post-admixture
□ Native Americans
◇ Europeans
○ Unclassified


-log10(P-value)
12.5
10.0
7.5
5.0
2.5
0.0
Selection model
Post-admixture
Unclassified
Z-score (African ancestry)
7.5
5.0
2.5
0.0
-2.5
29.0
30.0
31.0
32.0
33.0
Population
BRA
CHL
COL
MEX
PER
HLA-A
HLA-A
HLA-G
HLA-A
RNF39
TRIM39-RPP21
HLA-G
RNF39
TRIM39
HLA-H
PPP1R11
TRIM39
HLA-H
PPP1R11
TRIM39
HLA-G
HLA-A
PPP1R11
TRIM39
IFITM4P
HLA-G
HLA-A
HLA-J
POLR1H
TRIM26
HCG18
HLA-F-AS1
HLA-G
HLA-J
POLR1H
TRIM10
TRIM26
HCG18
HLA-F-AS1
HLA-G
HLA-J
ZNRD1ASP
TRIM40
TRIM26
HCG18
HLA-F
HLA-G
HLA-H
HLA-A
HLA-J
ZNRD1ASP
TRIM31
TRIM40TRIM15
TRIM26
HLA-L
HCG18
HLA-F-AS1
HCG4
HLA-G
HLA-H
HCG4BHLA-A
HCG9
ZNRD1ASP
TRIM31
TRIM10
TRIM26
HCG17
29.80
30.00
30.20
Chromosome 6 (Mb)


a
b
c
AdaptMix
rs3736907
rs2847281
rs2330635
GWAS
eQTL
MIR942
PTGFRN
CD101
TTF2
TRIM45
PSMG2
CEP76
PTPN2
SEH1L
CEP192
MIF
DDT
MIF-AS1
GSTT2
DDTL
117.450
117.500
117.550
117.600
117.650
Chromosome 1 (Mb)
12.700
12.800
12.900
13.000
13.100
Chromosome 18 (Mb)
24.250
24.275
24.300
24.325
Chromosome 22 (Mb)
▲ Post-admixture
■ Native Americans
● Unclassified
○ Basophil (count)
△ Eosinophil (count)
+ Lymphocyte (count)
× Monocyte (count)
◇ Neutrophil (count)
▽ White cell (count)
○ B cell, naive
△ Monocyte, classical
+ Monocyte, non classical
× NK cell, CD56dim CD16+
◇ T cell, CD4, memory TREG
d
e
f
DAF
1.00
0.75
0.50
0.25
0.00
Native American ancestry
Predictions
Post-admixture
In parental
Neutrality


a
b
c
d
AdaptMix
GWAS
eQTL
DAF
rs1171148
rs5030938
▲ Post-admixture
■ Native Americans
● Unclassified
○ BMI
△ Body fat
+ WC
× HC
◇ T2D
▽ HDL
⊠ LDL
✱ Cholesterol
◈ TG
○ Adipose (subcutaneous)
△ Adipose (visceral omentum)
+ Liver
BRINP3
LINC01720
SNORD98
DDX50
KIFBP
SUPV3L1
HK1
TET1
CCAR1
STOX1
DDX21
SRGN
VPS26A
HKDC1
TACR2
190.20
190.40
190.60
190.80
Chromosome 1 (Mb)
70.40
70.60
70.80
71.00
71.20
Chromosome 10 (Mb)
0
5
10
15
20
0.00
0.25
0.50
0.75
1.00
Native American ancestry
Predictions
Post-admixture
In parental
Neutrality