Supplementary Information

Tracing Human Genetic Histories and Natural Selection with Precise Local Ancestry Inference

SIGuide .DOC

Supplementary Table 1

Supplementary Figures 1 - 28

Supplementary Methods

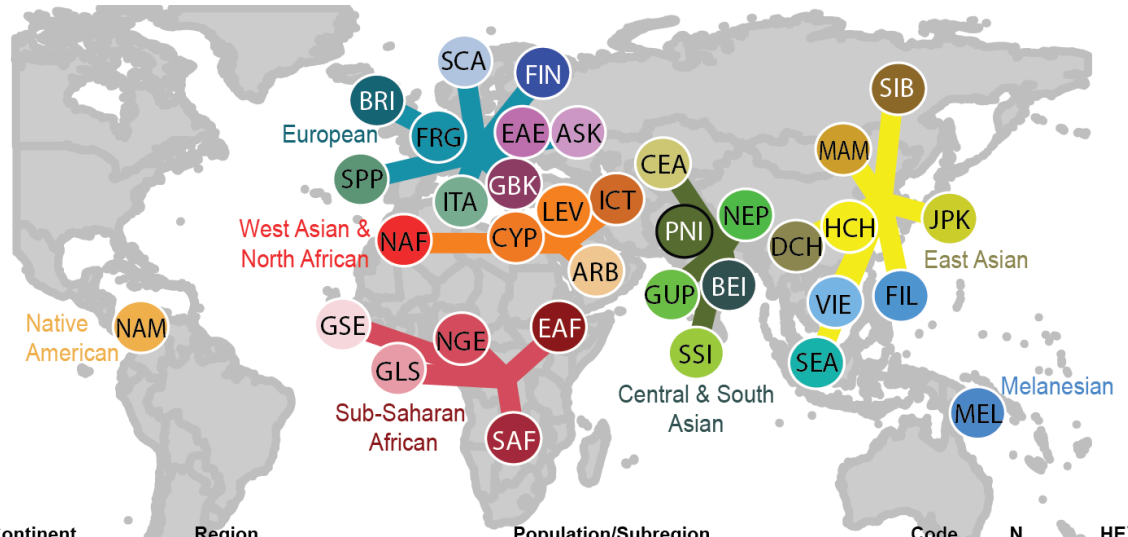
Supplementary Tables

Supplementary Table 1. Datasets included in the reference panel. Data source, genotyping technology and corresponding population in our study.

Data source	Description	Genotyping technology	Populations
Byrska-Bishop et al, 2022; Lowy-Gallego et al, 2019	1000 Genomes Project	WGS	SAF, GSE, GLS, NGE, DCH, HCH, VIE, JPK, FIN, ITA, SPP, BEI, GUP, PNI, SSI
Bergström et al, 2020	Human Genome Diversity Project	WGS	SAF, GSE, NGE, HCH, SEA, JPK, MAM, MEL, ITA, SPP, ARB, LEV, NAF, CEA, PNI
Mallick et al, 2016	Simons Genome Diversity Project	WGS	SAF, GSE, NGE, FIL, SIB, MEL, EAE, ITA, SPP, ARB, LEV, NAF, ICT, CEA
This study	Artificially constructed Native Americans	WGS (from 1KGP, HGDP and SGDP genomes)	NAM
Almarri et al, 2021	Middle Eastern populations	WGS	ARB
Malaria Genomic Epidemiology Network, 2019	Gambian Genome Variation Project (Fula, Jola, Mandinka and Wolof ethnic groups)	WGS	GSE
Zhang et al, 2014; Kim et al, 2018	Korean Personal Genome Project	WGS (Illumina Hiseq)	JPK
Carmi et al, 2014	The Ashkenazi Genome Consortium (healthy individuals of Ashkenazi Jewish descent)	WGS	ASK
Bycroft et al, 2018	UK Biobank project (participants from across the United Kingdom)	UK BiLEVE Axiom and UK Biobank Axiom arrays (imputed with the Haplotype Reference Consortium, UK10K and 1000 Genomes reference panels)	SAF, EAF, GSE, GLS, NGE, HCH, FIL, SEA, VIE, JPK, MAM, EAE, BRI, FIN, FRG, SCA, GBK, ITA, SPP, LEV, NAF, CYP, ICT, CEA, NEP, BEI, PNI, SSI
Wang et al, 2021; Jeong et al, 2019; Biagini et al, 2019; Vyas et al, 2017; Skoglund et al, 2017; Skoglund et al, 2016; Lazaridis et al, 2016; Lazaridis et al, 2014; Pickrell et al, 2012	Modern samples from the '1240K+HO' dataset provided by Dr David Reich laboratory	Affymetrix Human Origins array	SAF, EAF, HCH, FIL, MAM, SIB, ASK, EAE, GBK, SPP, ARB, LEV, NAF, ICT, CEA
Anagnostou et al, 2020	Berbers and Arabs from Southern Tunisia	Illumina Human OmniExpressExome v 8.1 array	NAF
Henn et al, 2012; Arauna et al, 2017	Berbers and Arabs from North Africa and Syria	Affymetrix 6.0 array	LEV, NAF
Hollfelder et al, 2017	Sudanese and South Sudanese populations	Illumina Human Omni5MExome array	EAF
Dobon et al, 2015	Populations (Arabs, Beja, Ethiopian and Nubian) from Sudanese region	Illumina Infinium Immunochip	EAF
Behar et al, 2013	Ashkenazi Jews and non-Ashkenazi populations from Eastern Europe, Northeast Africa and the Caucasus	Illumina Human610-Quad, Human660W-Quad, HumanOmniExpress-12v1 730K and HumanOmni1-Quad array	EAF, ASK, EAE, ICT
Yunusbayev et al, 2012	Caucasians and geographically nearby populations (Central Asia, Eastern Europe and Balkans)	Illumina 610K array	EAE, GBK, ICT, CEA
Yunusbayev et al, 2015	Turkic-speaking populations from regions across Eurasia and their geographic neighbors	Illumina 550k, 610k, 650k and Human1M-Duo BeadChips	MAM, EAE, ICT, CEA

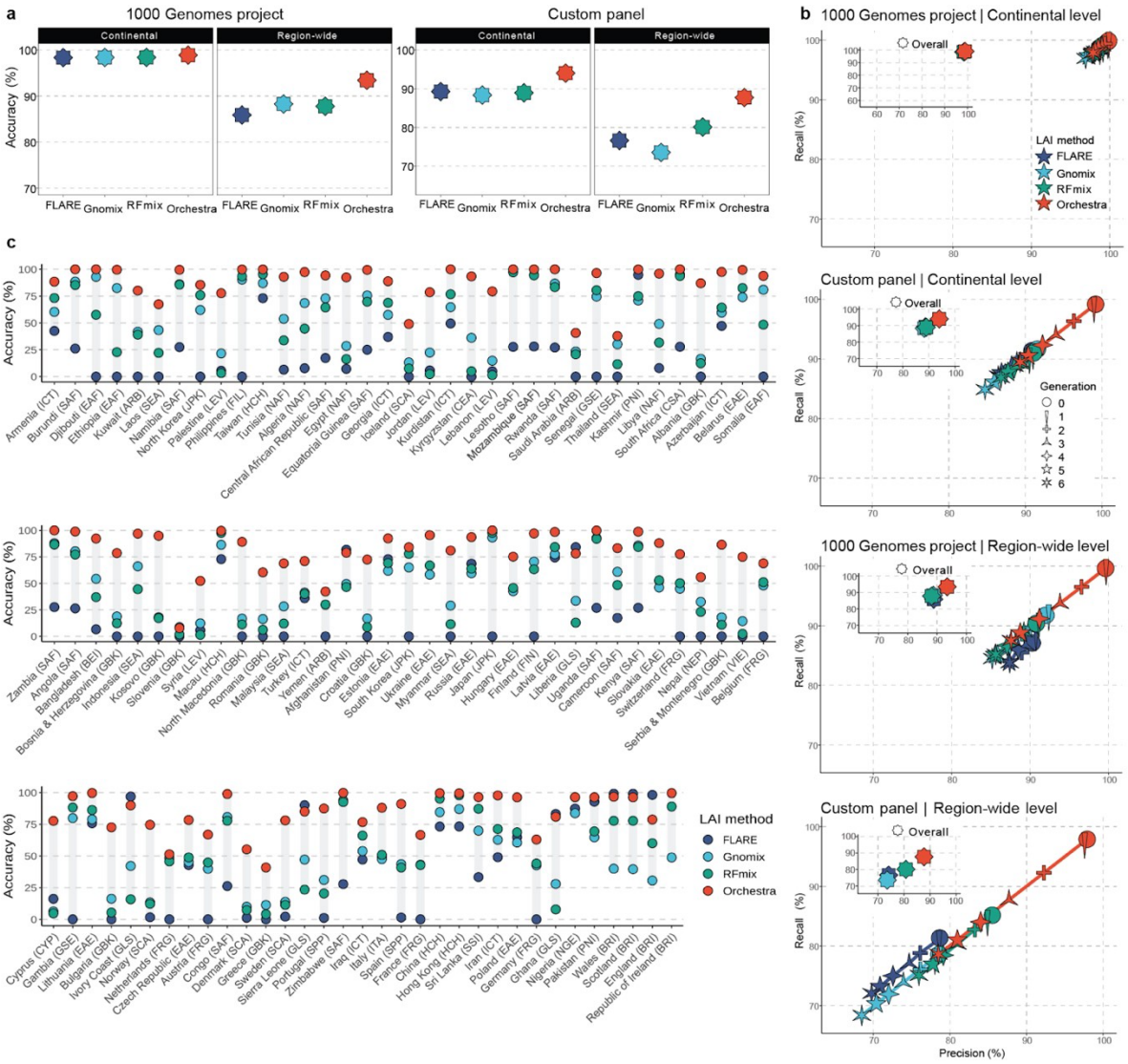
Behar et al, 2010	Jewish Diaspora communities and non-Jewish neighbor populations from Europe, Asia and Africa	Illumina Human610-Quad and Human660W-Quad bead arrays	EAF, ASK, EAE, SPP, ARB, LEV, NAF, ICT, CEA
Tambets et al, 2018	Uralic-speaking populations and local geographic neighbors	Illumina Human610-Quad, HumanHap650Y and Human660W-Quad BeadChip	EAE, FIN
Botigué et al, 2013	Spanish from South (Andalusian) and Northwest (Galician) Spain	Affymetrix 6.0 array	SPP
Flores-Bello et al, 2021	Basques (from France and Spain) and Spanish Peri-Basques	Axiom Genome-Wide Human Origins 1 Array	SPP
Henn et al, 2012	Basques from Spanish Basque country	Affymetrix 6.0 array	SPP
Pathak et al, 2018	Northwest Indian populations from Rajasthan and Haryana states (Gujjar, Kamboj, Ror)	Illumina HumanOmniExpress -24 BeadChip	PNI
Nelson et al, 2008	Indian Asians (Gujarati, Hindi, Punjabi, Pushto and Urdu) from the Population Reference Sample (POPRES) project	Affymetrix GeneChip 500K array	GUP, PNI
Changmai et al, 2022	Khmer and Kuy from mainland Southeast Asia (Thailand)	Affymetrix Human Origins SNP array	SEA
Tätte et al, 2019	Lao from Laos	Illumina OmniExpress BeadChips for 650k, 710k and 730k	SEA
Mörseburg et al, 2016	Island Southeast Asian populations (Malay, Igorot, Luz)	Illumina OmniExpress BeadChip for 730k	FIL, SEA

Supplementary Figures

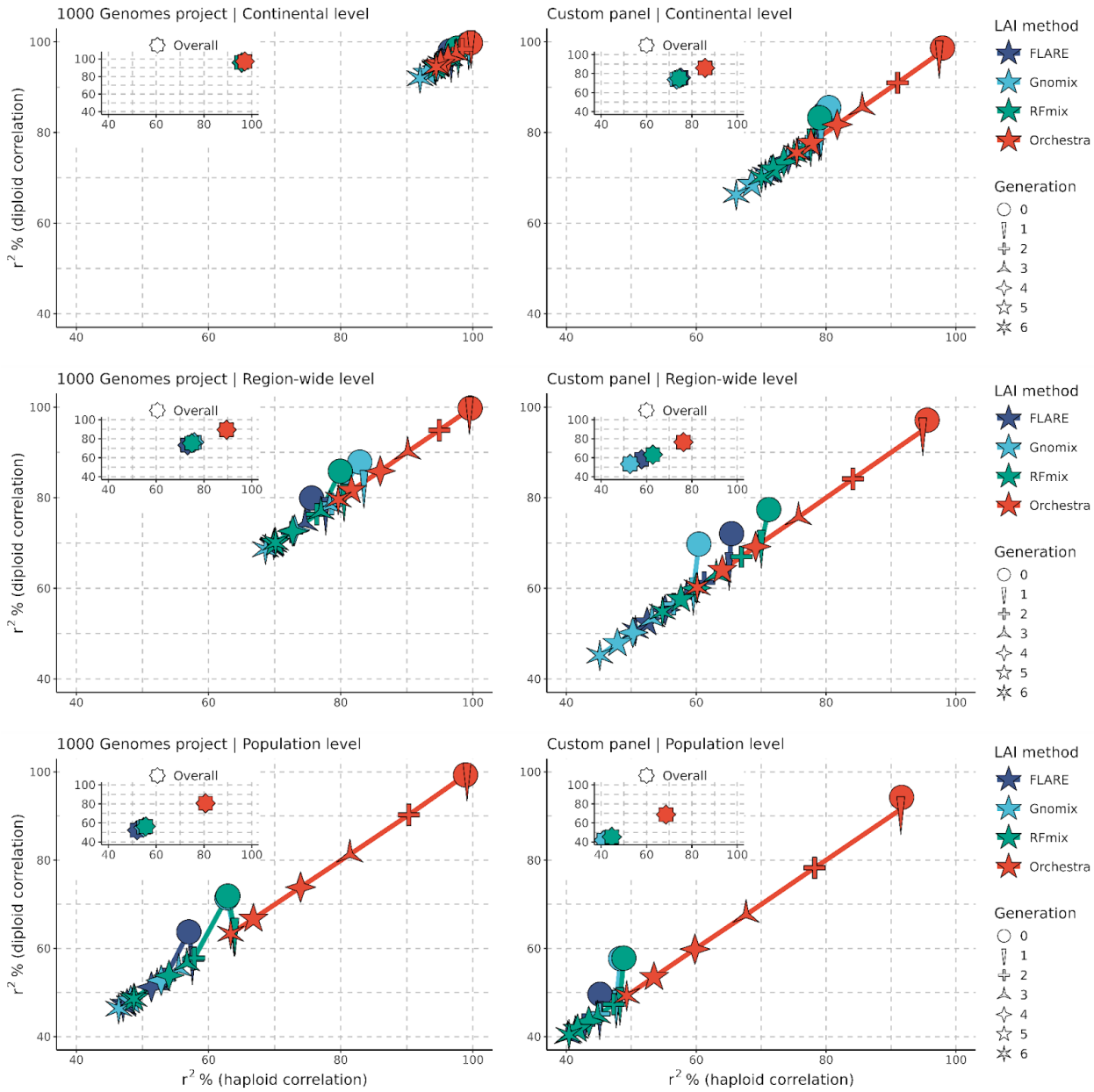


Continent	Region	Population/Subregion	Code	N	HEX
Sub-Saharan African	Central, South & Southeast African	Central, South & Southeast African	SAF	500	#a4293a
	West African	Ghanaian, Ivorian, Liberian & Sierra Leonean	GLS	157	#e59aa5
		Gambian & Senegalese	GSE	388	#f5d7db
		Nigerian	NGE	387	#d1495b
	Northern East African	Northern East African	EAF	500	#8B0000
Western Asian & North African	Arab & Levantine	Arab	ARB	206	#EEC591
		Levantine	LEV	196	#FF7F00
	Northern West Asian	Iraqi, Iranian, Caucasian & Turkish	ICT	451	#CD6600
		Cypriot	CYP	100	#EE7621
	North African	North African	NAF	500	#EE2C2C
Central & South Asian	Central Asian	Central Asian	CEA	182	#CDC673
	Northern South Asian	Pakistani & North Indian	PNI	500	#556B2F
		Gujarati Patel	GUP	87	#32CD32
		Bengali & East Indian	BEI	288	#2F4F4F
		Nepalese	NEP	111	#00FF00
	Southern South Asian	Sri Lankan & South Indian	SSI	412	#9ACD32
European	Northwest European	British & Irish	BRI	500	#006373
		French & German	FRG	500	#008fa6
		Scandinavian	SCA	167	#B0C4DE
		Finnish	FIN	220	#0000EE
	Eastern European	Eastern European	EAE	500	#CD69C9
	South European	Greek & Balkan	GBK	209	#8B3A62
		Italian	ITA	500	#76ab8f
		Spanish & Portuguese	SPP	500	#5a9375
	Ashkenazi Jewish	Ashkenazi Jewish	ASK	152	#CD96CD
East Asian	Chinese & Southeast Asian	Han Chinese	HCH	500	#FFFF00
		Dai Chinese	DCH	77	#8B864E
		Filipino	FIL	318	#4F94CD
		Southeast Asian	SEA	167	#20B2AA
		Vietnamese	VIE	94	#63B8FF
	Japanese & Korean	Japanese & Korean	JPK	462	#CDCD00
	Northern Asian	Siberian	SIB	62	#8B6508
		Manchurian & Mongolian	MAM	152	#CD9B1D
Native American	Native American	Native American	NAM	100	#edae49
Melanesian	Melanesian & Aboriginal Australian	Melanesian & Aboriginal Australian	MEL	24	#1E90FF

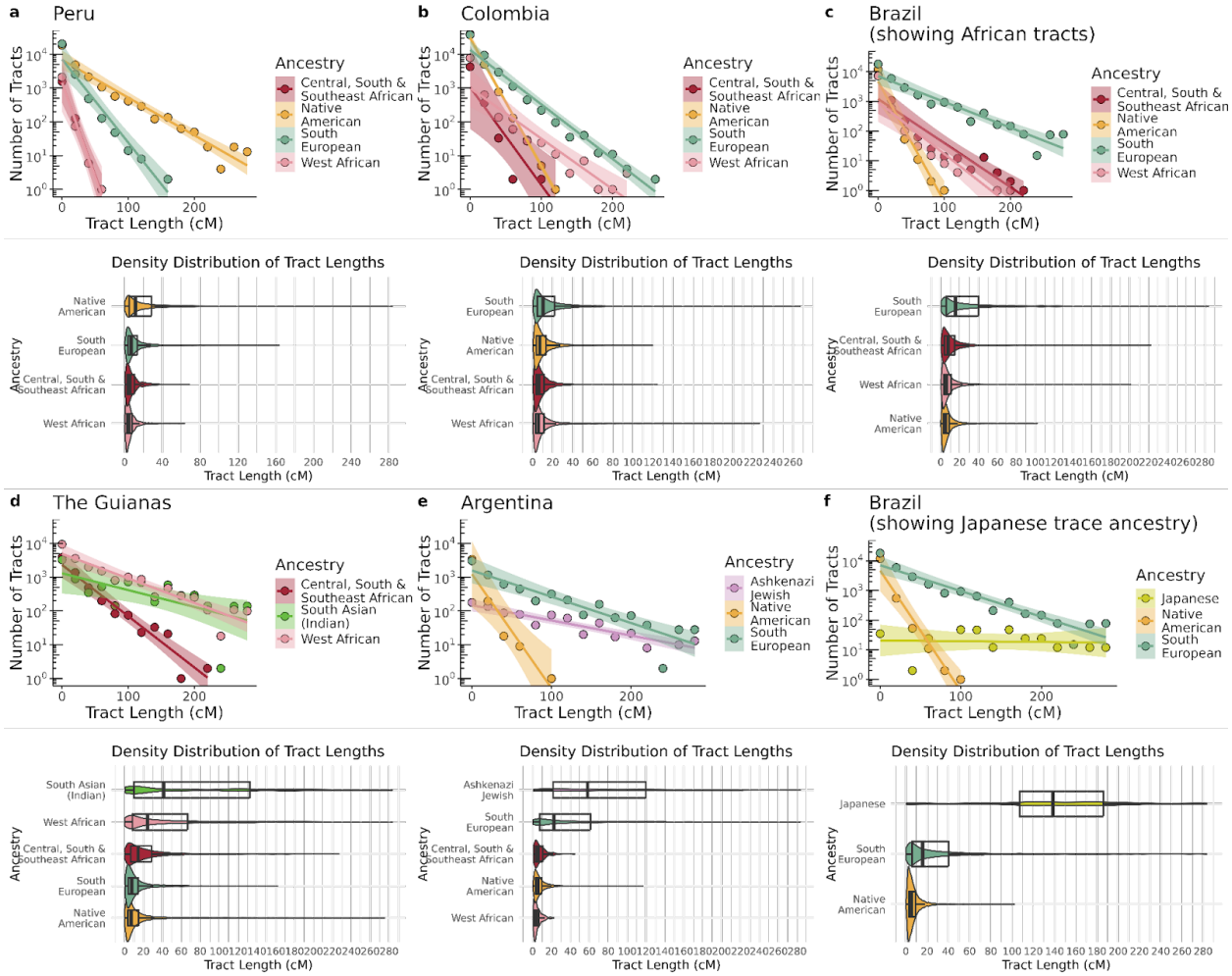
Supplementary Fig. 1 | 35 reference populations. Continent, region, code, sample number (N) and color codes (HEX) of the 35 reference populations used in the custom-35pops datasets.



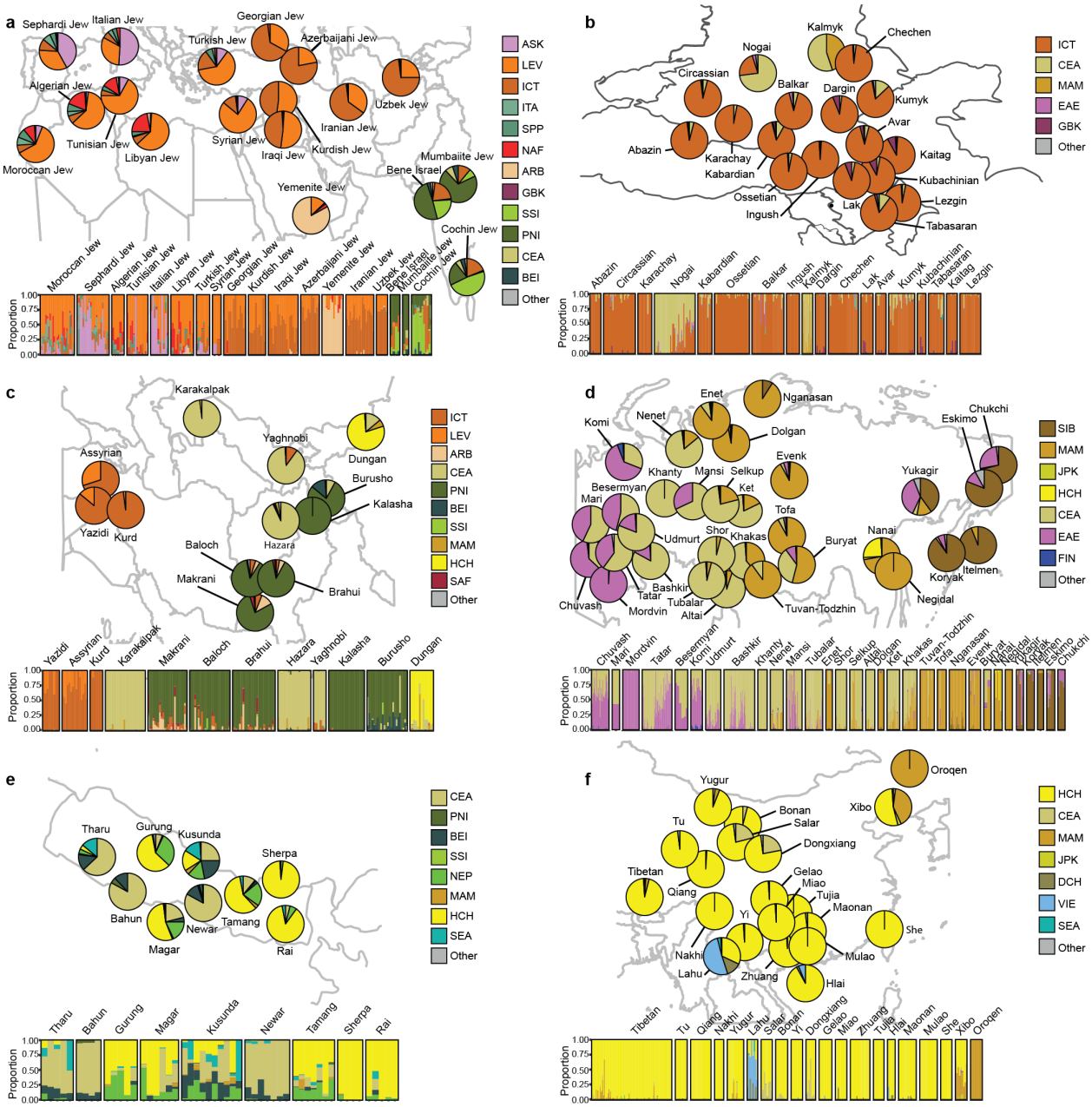
Supplementary Fig. 2 | Benchmarking results. (a-b) Performance of Orchestra and other LAI methods at continental and region-wide levels. (**a**) Overall accuracy (%) for 1KGP dataset (N=1,365) and the larger custom dataset with 35 populations (N=10,169), across continental and region-wide levels per method tested. (**b**) Recall and precision of Orchestra, RFmix, Gnomix and FLARE in ancestry deconvolution across 6 generations for 1KGP-16pops (N=3,276) and custom-35pops (N=24,408) reference panels at continental and region-wide levels. The generation is shown with star shapes and refers to the number of generations of simulated admixture (the more points the star has, the higher the generation). Results were calculated using chr17 to chr22. (**c**) Accuracy (%) per country in the UKBB external panel (N=10,241) with inferred single-origin individuals based on dimensionality reduction techniques. Countries are ordered by sample-size. The evaluated ancestry for each country is denoted in parentheses on the x-axis.



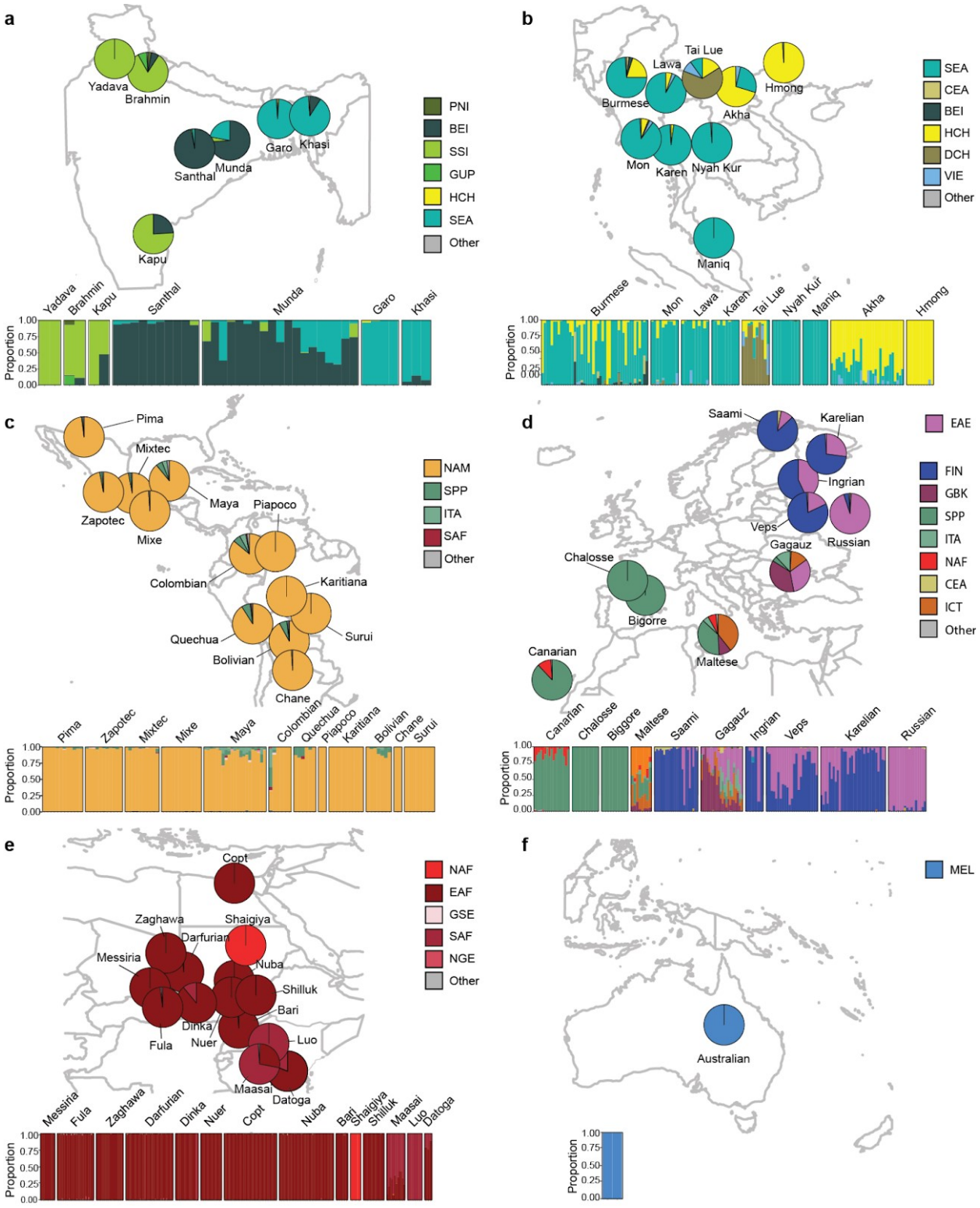
Supplementary Fig. 3 | Performance of Orchestra and other LAI methods using squared Pearson correlation as evaluation metric. Accuracy was measured for each method (Orchestra, RFmix, Gnomix and FLARE) with Pearson's r^2 for both 1KGP-16pops (N=3,276) (left column) and custom-35pops (N=24,408) (right column) reference panels across 6 generations at continental, regional (chromosomes 17-22) and population/subregional (chromosomes 15-22) levels. The difference between Orchestra and the second best model at the population/subregional level is approximately 24%, significantly higher than that estimated with recall and precision metrics; this is because we found that r^2 metric decays faster, so the differences for the first generations, where Orchestra is really accurate, are magnified.



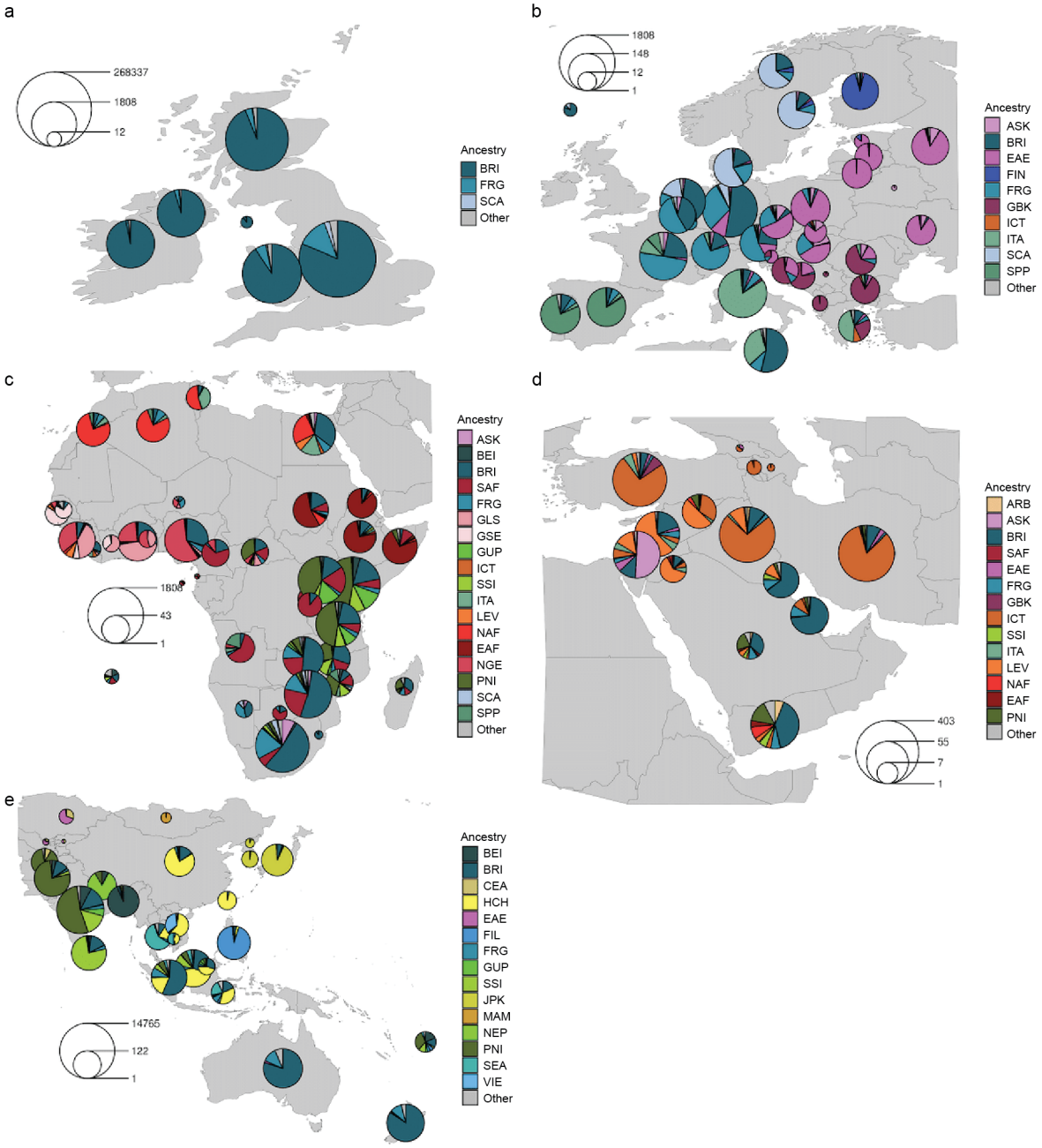
Supplementary Fig. 4 | Distribution of local ancestry tract lengths across South American populations. Each panel displays the number of tracts per ancestry as a function of tract length (in 20 cM bins) and their density distribution. Panels (**a**) (Peru (N=147)), (**b**) (Colombia (N=226)) and (**c**) (Brazil (N=234)) illustrate Native American, South European and African admixture, reflecting European colonization and the transatlantic slave trade. Panels (**d**) (The Guianas (N=311)), (**e**) (Argentina (N=69)) and (**f**) (Brazil (N=240)) depict trace ancestries from specific historical migrations: South Asian (Indian) tracts from 19th and early 20th-century indentured labor migration; Ashkenazi Jewish tracts from late 19th and early 20th-century Jewish immigration; and long Japanese tracts indicating immigration starting in 1908. The tract lengths correspond with the timing of these historical admixture events, with longer tracts indicating more recent admixture. Error bands show 95% CI. Boxplots over violin plots show tract density distribution. The central line of the boxplot denotes the median, box boundaries represent the first and third quartiles, the whiskers range from minimum to maximum values.



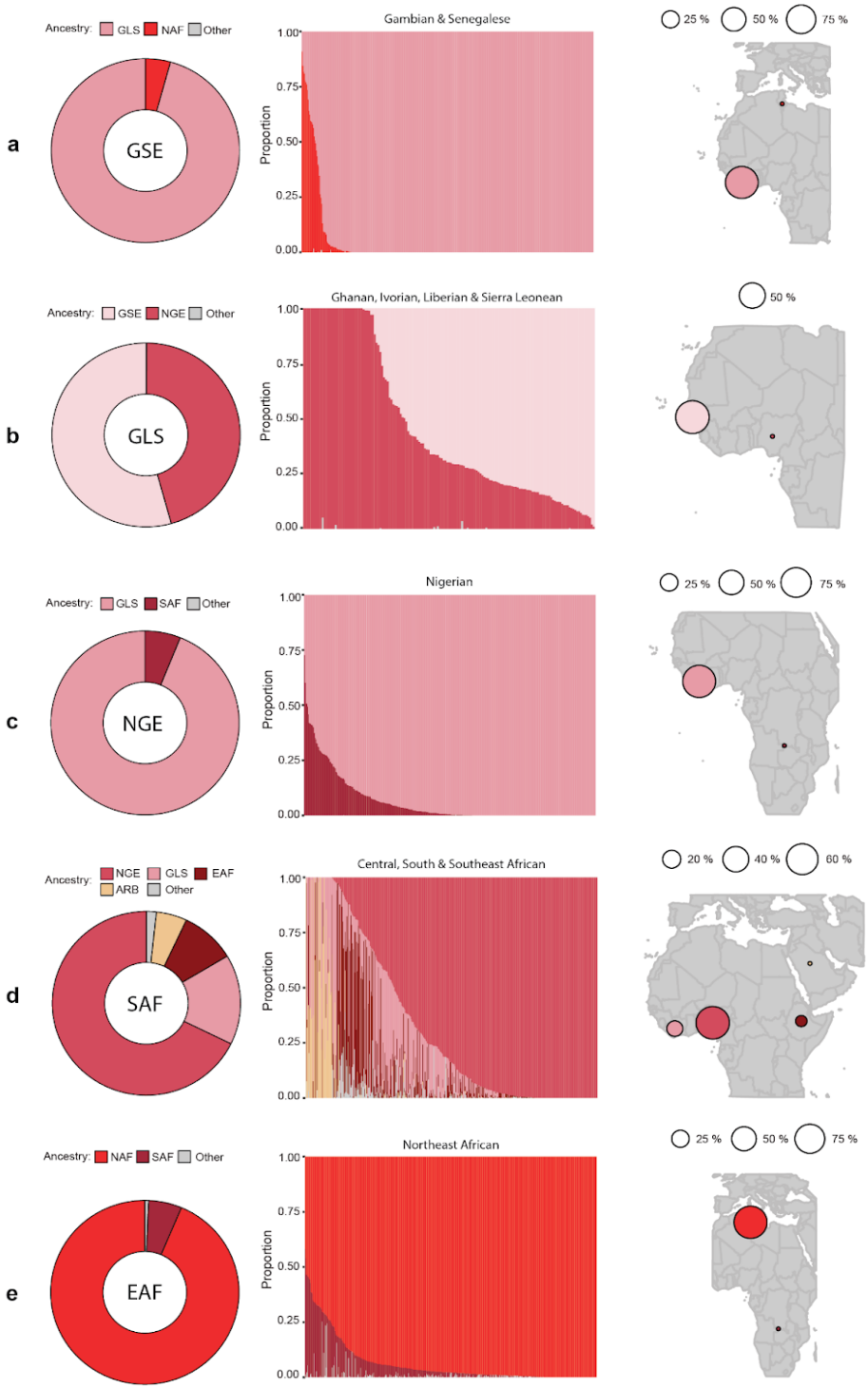
Supplementary Fig. 5 | Ancestry inference for ethnic groups and minorities. Orchestra's ancestry inference for ethnic groups not included in our 35 reference populations. Each column in the bar plot shows ancestral decompositon for a single individual. (**a**) Jewish groups (N=214), and ethnic groups in: (**b**) the Caucasus (N = 333), (**c**) Western and South Asia (N=209), (**d**) Siberia (N=589), (**e**) Nepal (N=53) and (**f**) China (N=304). Ancestry proportions were infered for chromosomes 17-22.



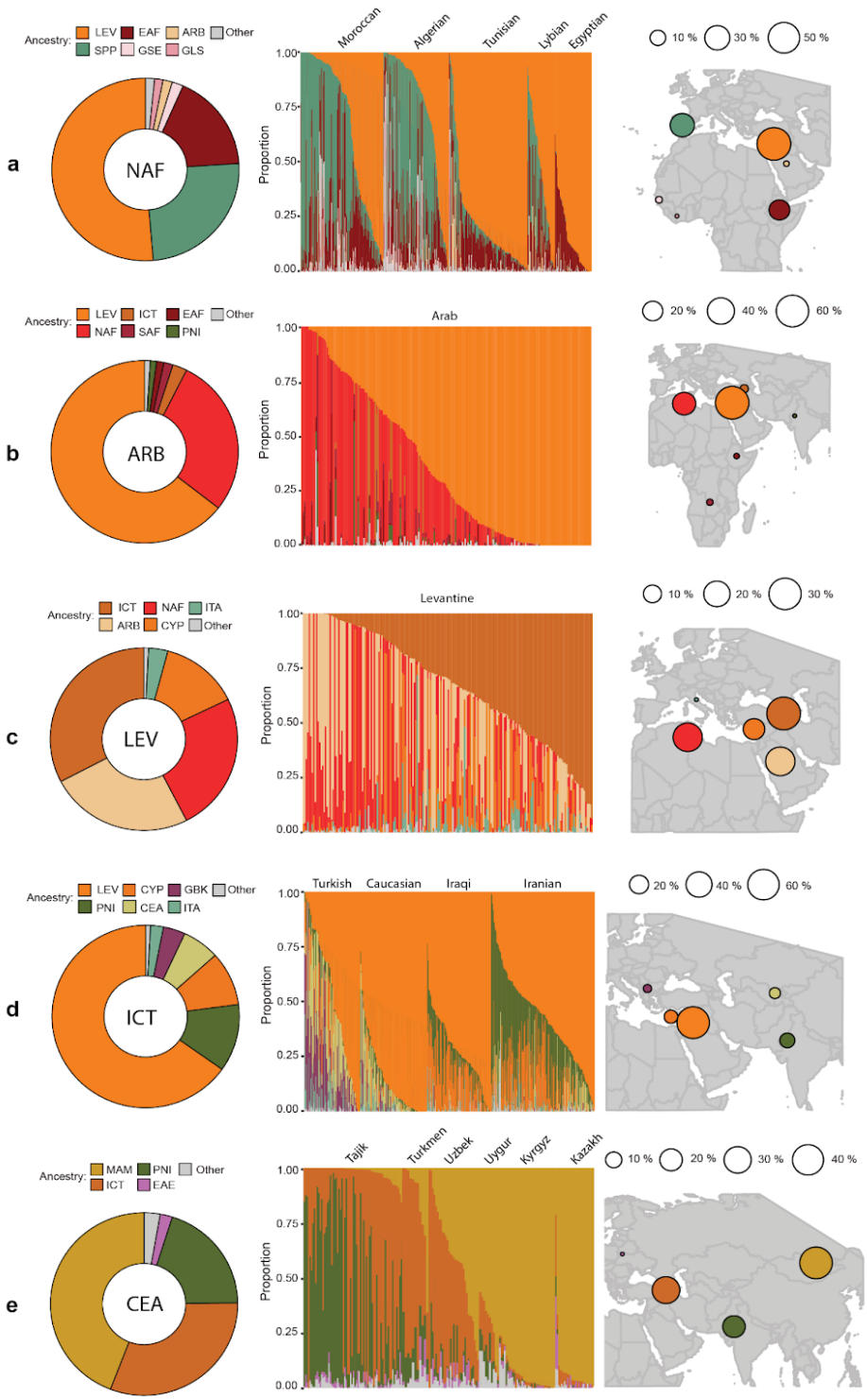
Supplementary Fig. 6 | Ancestry inference for ethnic groups and minorities. Orchestra's ancestry inference for ethnic groups not included in our 35 reference populations: in (**a**) South Asia (N=42), (**b**) Southeast Asia (N=147), (**c**) Americas (N=107), (**d**) Europe (N=169), (**e**) Africa (N=308) and (**f**) Australia (N=2). Each column in the bar plot shows ancestral decompositon for a single individual. Ancestry proportions were infered for chromosomes 17-22.



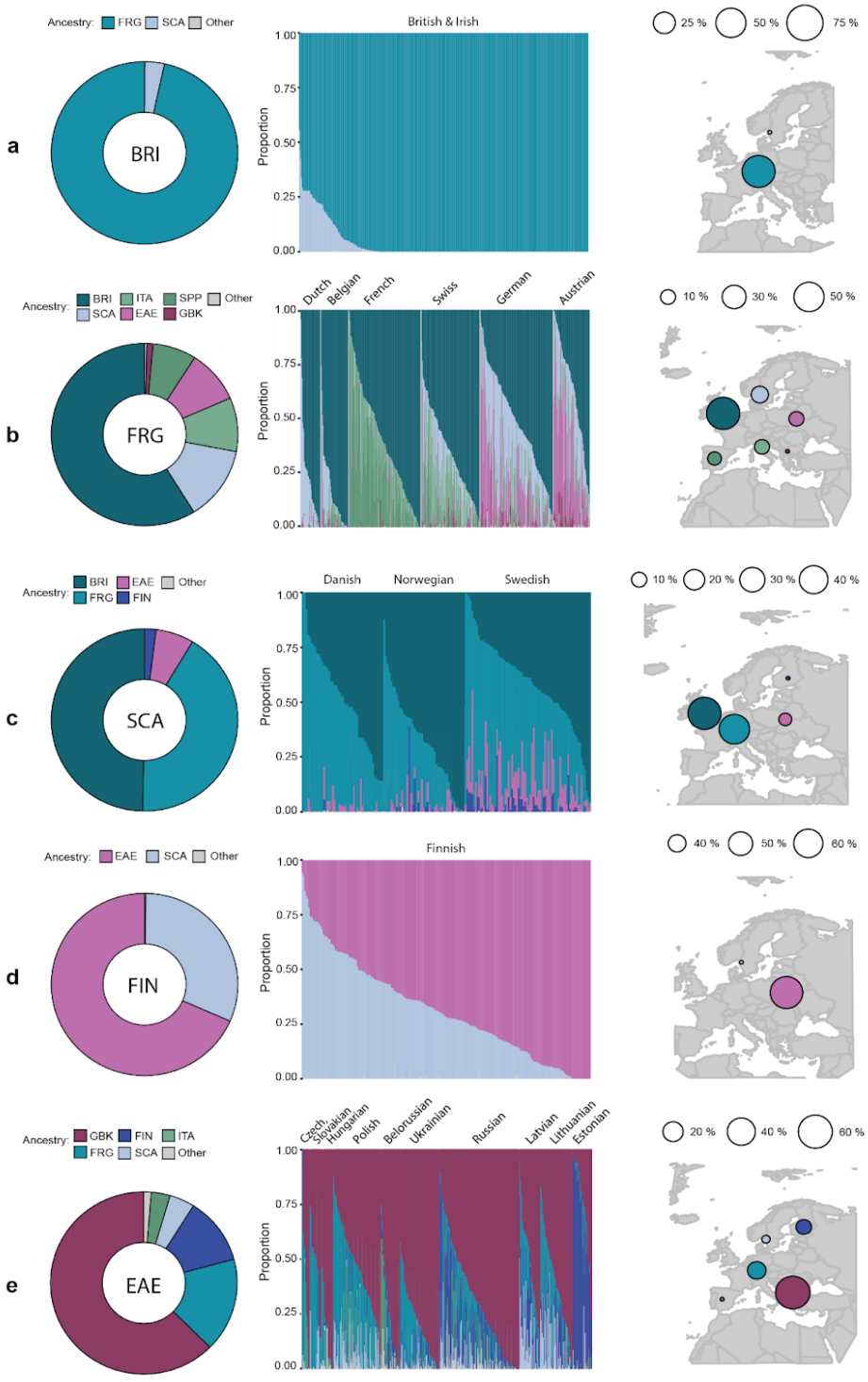
Supplementary Fig. 7 | Inferred ancestry for UKBB samples mapped by country of birth. Samples that were added to our reference panel were excluded. A bias towards excess British & Irish (BRI) ancestry is noted across all regions. ARB: Arab, ASK: Ashkenazi Jewish, BEI: Bengali & East Indian, BRI: British & Irish, CEA: Central Asian, EAE: Eastern European, EAF: Northeast African, FIL: Filipino, FIN: Finish, FRG: French & German, GBK: Greek & Balkan, GLS: Ghanaian, Ivorian, Liberian & Sierra Leonean, GSE: Gambian & Senegalese, GUP: Gujarati Patels, HCH: Han Chinese, ICT: Iraqi, Iranian, Caucasian & Turkish, ITA: Italian, JPK: Japanese & Korean, LEV: Levantine, MAM: Manchurian & Mongolian, NAF: North African, NEP: Nepalese, NGE: Nigerian, PNI: Pakistani & North Indian, SAF: Central, South & Southeast African, SCA: Scandinavian, SEA: Southeast Asian, SPP: Spanish & Portuguese, SSI: Sri Lankan & South Indian, VIE: Vietnamese.



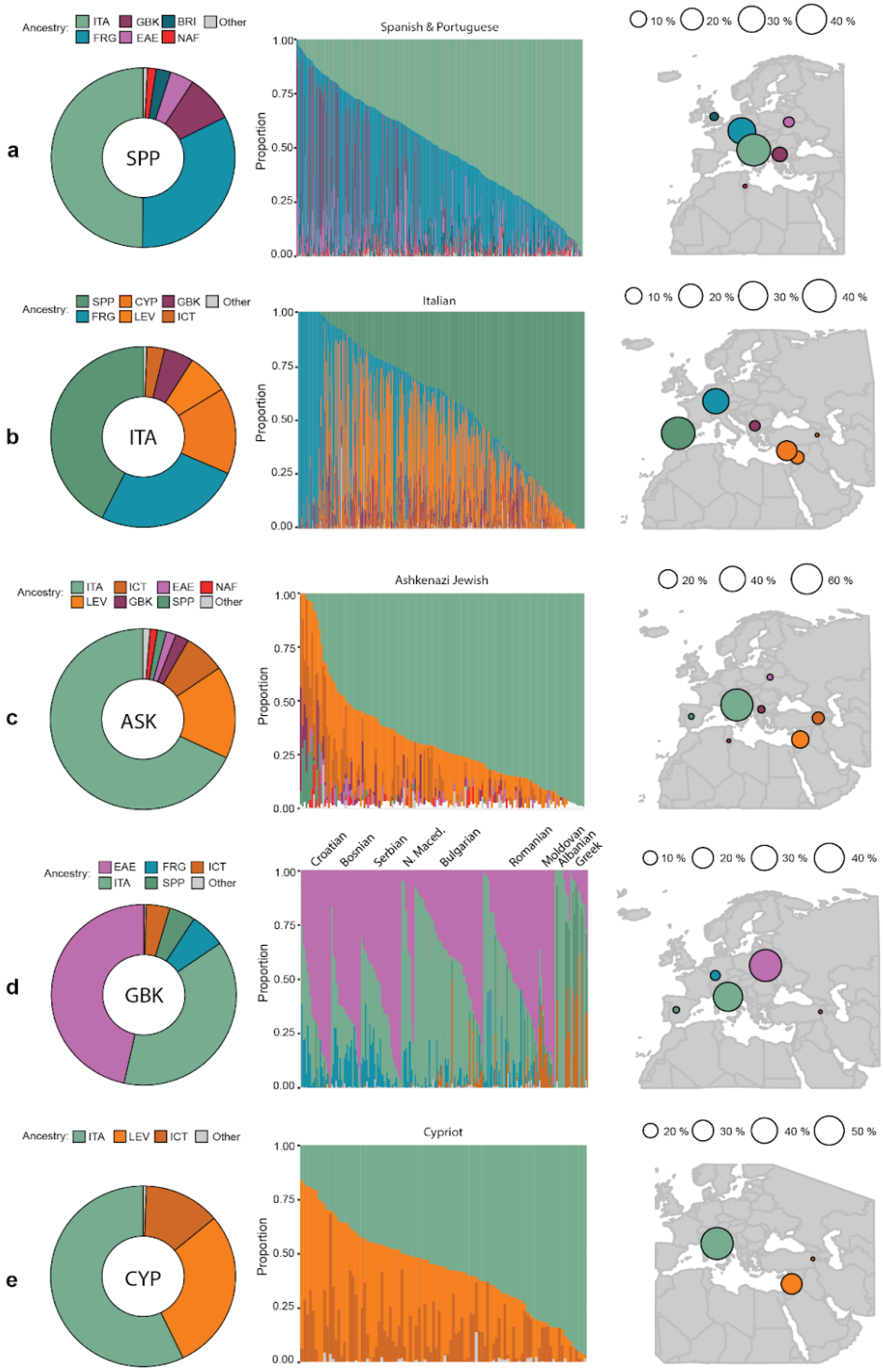
Supplementary Fig. 8 | Ancestral mapping for Sub-Saharan African populations: (a) Gambian & Senegalese (GSE, N=388), **(b)** Ghanan, Ivorian, Liberian & Sierra Leonean (GLS, N=157), **(c)** Nigerian (NGE, N=387)**, (d)** Central, South & Southeast African (SAF, N=500) and **(e)** Northeast African (EAF, N=500). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.



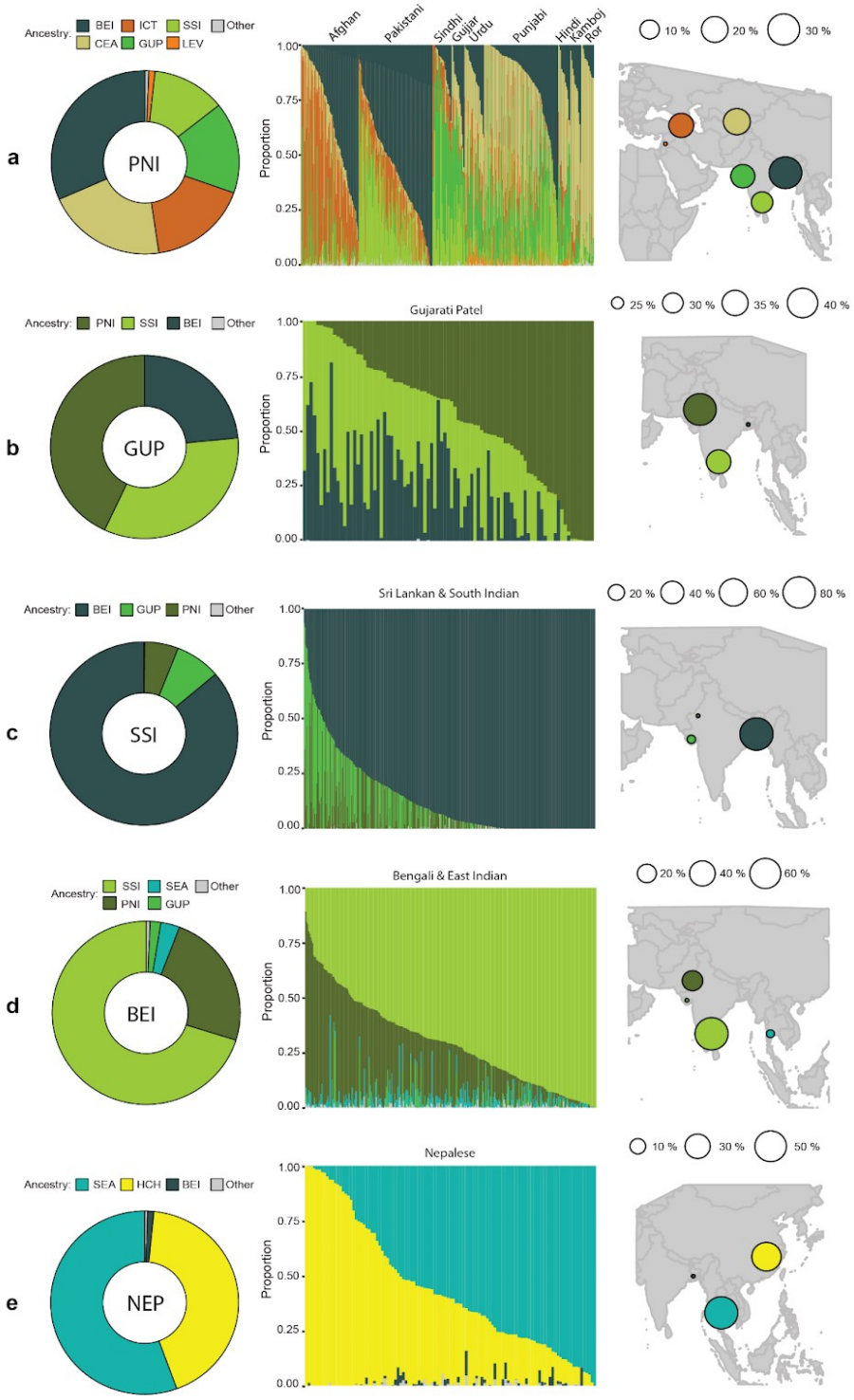
Supplementary Fig. 9 | Ancestral mapping for North African and West and Central Asian populations: (a) North African (NAF, N=500), **(b)** Arab (ARB, N=206), **(c)** Levantine (LEV, N=196), **(d)** Iraqi, Iranian, Caucasian & Turkish (ICT, N=451) and **(e)** Central Asian (CEA, N=182). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.



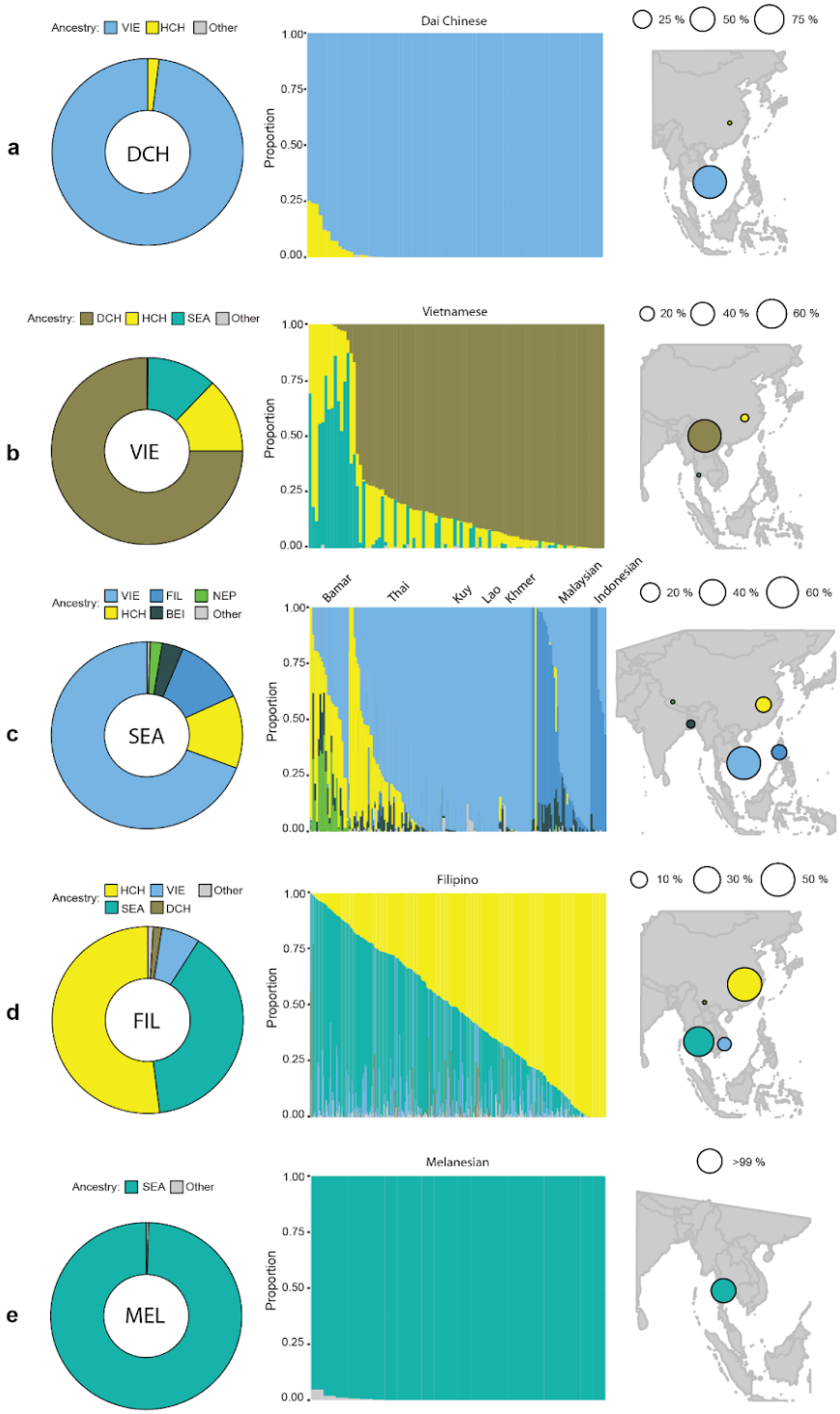
Supplementary Fig. 10 | Ancestral mapping for North, West and Eastern European populations: (a) British & Irish (BRI, N=500), **(b)** French & German (FRG, N=500), **(c)** Scandinavian (SCA, N=167), **(d)** Finish (FIN, N=220) and **(e)** Eastern European (EAE, N=500). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.



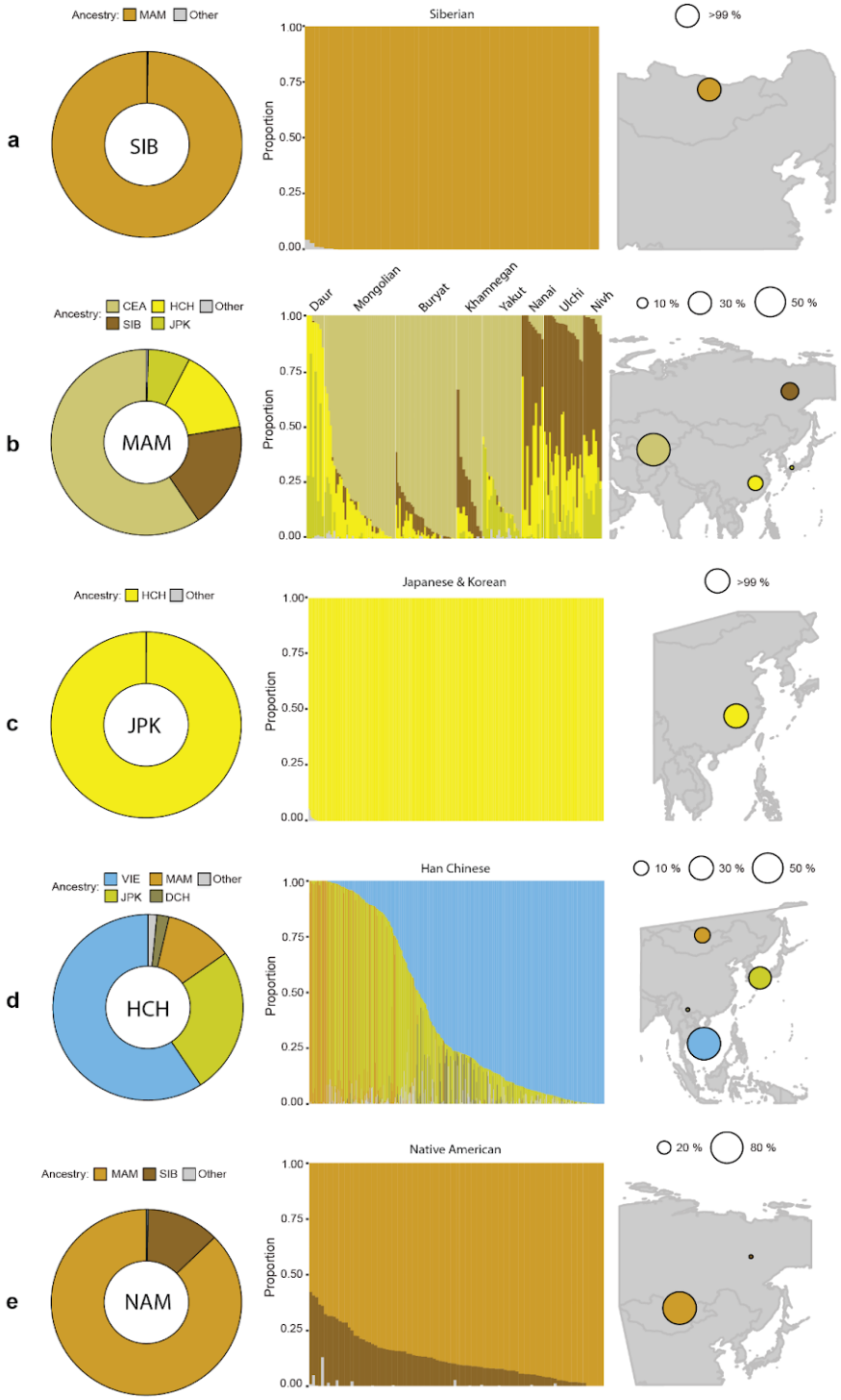
Supplementary Fig. 11 | Ancestral mapping for South European and Ashkenazi Jewish populations: **(a)** Spanish & Portuguese (SPP, N=500), **(b)** Italian (ITA, N=500), (c) Ashkenazi Jewish (ASK, N=152), (d) Greek & Balkan (GBK, N=209) and (e) Cypriot (CYP, N=100). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.



Supplementary Fig. 12 | Ancestral mapping for South Asian populations: (a) Pakistani & Northern Indian (PNI, N=500), **(b)** Gujarati Patel (GUP, N=87), **(c)** Sri Lankan & South Indian (SSI, N=412), **(d)** Bengali & East Indian (BEI, N=288) and **(e)** Nepalese (NEP, N=111). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.

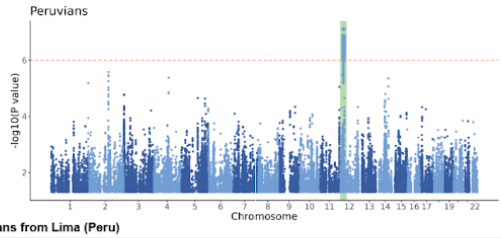


Supplementary Fig. 13 | Ancestral mapping for Southeast Asian populations: (a) Dai Chinese (DCH, N=77), **(b)** Vietnamese (VIE, N=94), **(c)** Southeast Asian (SEA, N=167), **(d)** Filipino (FIL, N=318) and **(e)** Melanesian (MEL, N=24). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.



Supplementary Fig. 14 | Ancestral mapping for North and East Asian and Native American populations: (a) Siberian (SIB, N=62), **(b)** Manchurian & Mongolian (MAM, N=152), **(c)** Japanese & Korean (JPK, N=462), **(d)** Han Chinese (HCH, N=500) and **(e)** Native American (NAM, N=100). Overall ancestry proportions (left); individual ancestry proportions shown grouped by ethnicity/subregion, with each bar representing a single individual (middle); ancestry proportions shown on a map (right). Ancestry proportions were infered for chromosomes 17-22.

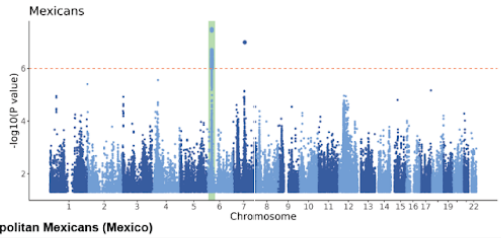
a



Peruvians from Lima (Peru)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	1KGP (The 1000 Genomes Project Consortium, Nature 2015)	85	East Peruvian Native Americans (Quechua, Aymara) [0.77] / Iberian populations from Spain (IBS) [0.20] / Sub-Saharan Africans (GWD, LWK, YRI) [0.03]south Bantu speakers (Luhya, Sotho, Zulu) [0.60] / Mandarese [0.40]
This study	Peruvians from 1KGP	87	South Native Americans [0.60] / Iberians [0.38] / Sub-Saharan Africans [0.2]

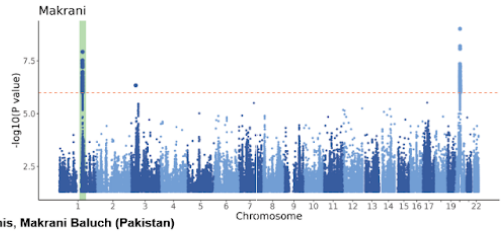
b



Cosmopolitan Mexicans (Mexico)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	Moreno-Estrada et al., 2014	418	Mexican Native Americans (Maya Campeche, Tepehuano, north Zapotec) [0.54] / Iberian populations from Spain (IBS) [0.42] / Sub-Saharan Africans (YRI) [0.04]
This study	Mexicans from 1KGP	65	North Native Americans [0.54] / Iberians [0.44] / SubSaharanAfrican [0.02]

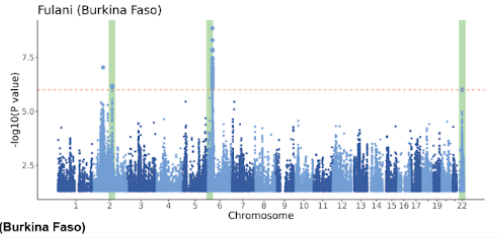
c



Makranis, Makrani Baluch (Pakistan)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	EGAS00001002558 (Laso-Jadart et al. 2017)	46	Baluch [0.85] /east and south Bantu speakers (Luhya, Sotho) [0.15]
This study	Makrani samples from Reich lab	24	Proxy for Baluch (individuals from Pakistan) [0.98] / Bantu and Luhya people, individuals from Lesotho [0.02]

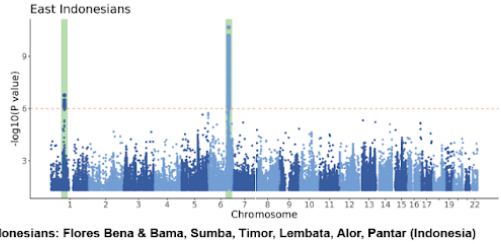
d



Fulani (Burkina Faso)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	E-MTAB-8434 (Vicente et al., 2019)	53	West Africans (Jola, Igbo) [0.77] / north Africans (Mozabite) + Europeans (CEU, Czech) [0.23]
This study	E-MTAB-8434 (Vicente et al., 2019)	53	Jola people [0.87] / North Africans (Mozabite) + Europeans (Czech, French and British individuals) [0.13]

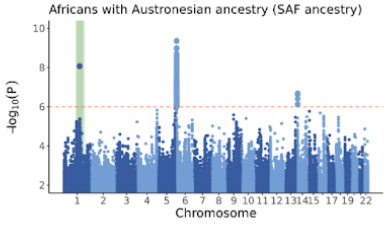
e



East Indonesians: Flores Bena & Bama, Sumba, Timor, Lembata, Alor, Pantar (Indonesia)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	GSE80534 (Hudjashov et al., 2017)	201	Filipinos [0.58] / Papuans [0.42]
This study	GSE80534 (Hudjashov et al., 2017)	201	Custom-35pops panel: SEA [0.92] / FIL [0.04] / MEL [0.03]

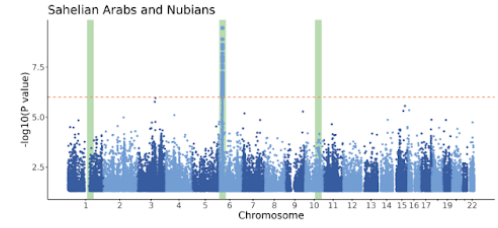
f



Malagasy (Madagascar)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	EGAS00001002549 (Pierron et al., 2017)	700	East and south Bantu speakers (Luhya, Sotho, Zulu) [0.60] / Mandarese [0.40]
This study	Southeast African UKBB participants with >1% Austronesian heritage	221	BEI [0.31] / SSI [0.25] / SAF [0.17] / PNI [0.08] /FRG [0.03] / BRI [0.03] / SEA [0.03]

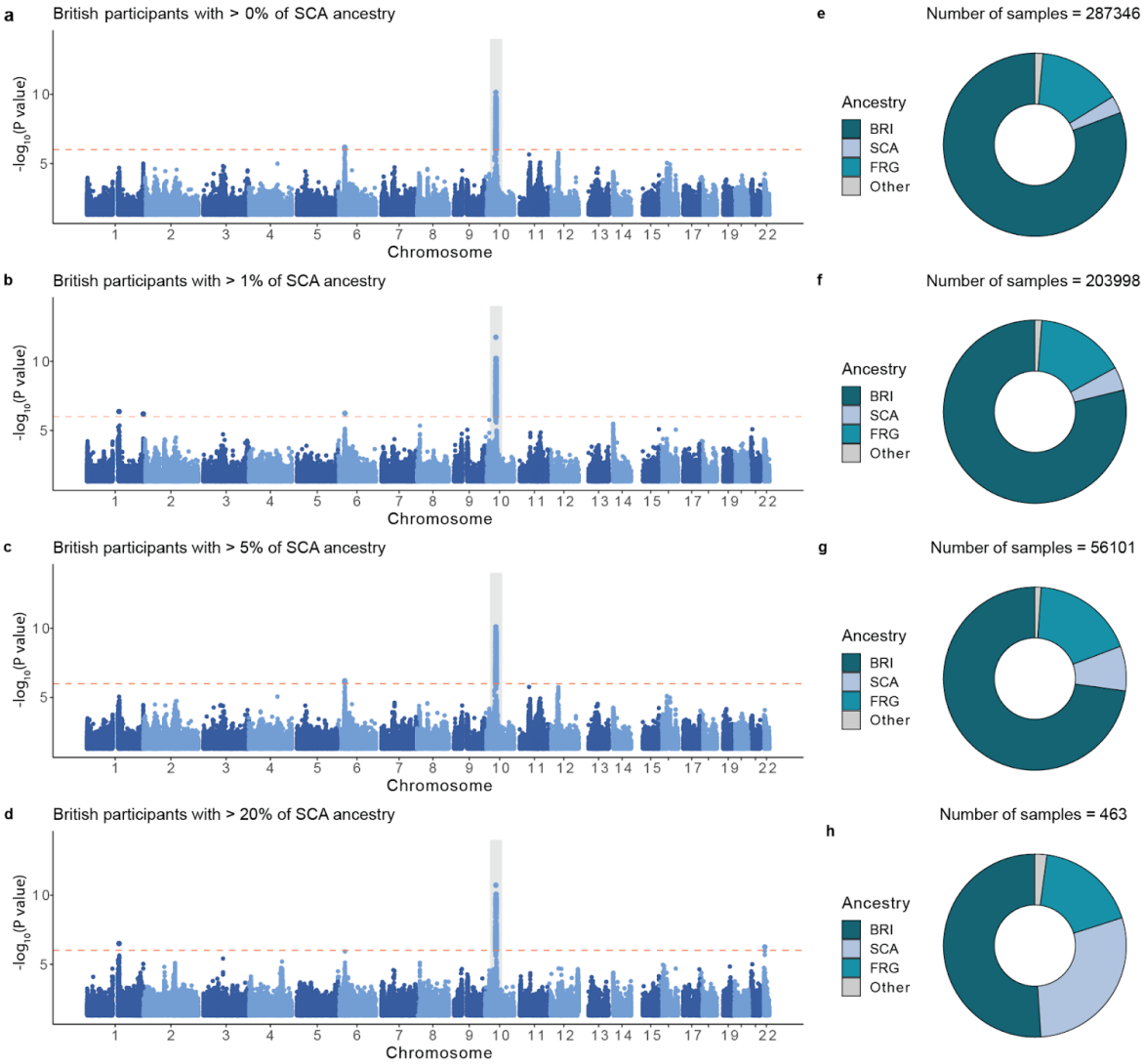
g



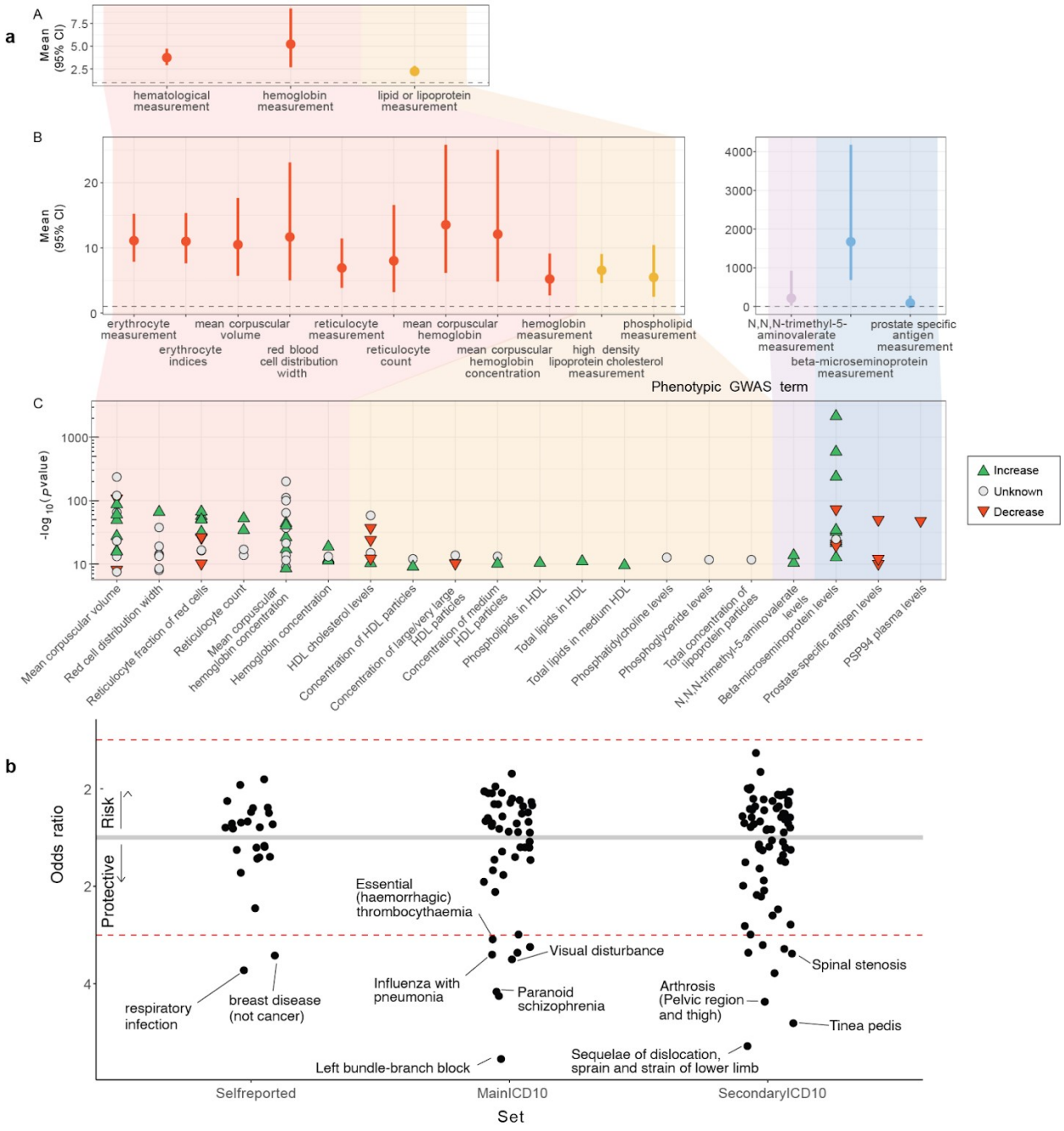
Sahelian Arabs: Bataheen, Gaalien, Shaigia, Messiria & Nubians: Danagla, Mahas, Halfawieen (Sudan)

Study	Dataset	N	Source populations vs. main inferred ancestries [admixture proportions]
Cuadros-Espinosa et al. 2022	http://jakobsson-lab.iob.uu.se/data/ (Hollfelder et al., 2017)	97	East Africans (Gumuz, Somali) and Nile peoples (Baria, Dinka, Nuer, Shilluk) [0.50] / South Europeans (Tuscans, Sardinians) and Druze and Bedouins [0.50
This study	http://jakobsson-lab.iob.uu.se/data/ (Hollfelder et al., 2017)	97	Nile peoples [0.56] / South Europeans (Tuscans, Sardinians) and Druze and Bedouins [0.44]

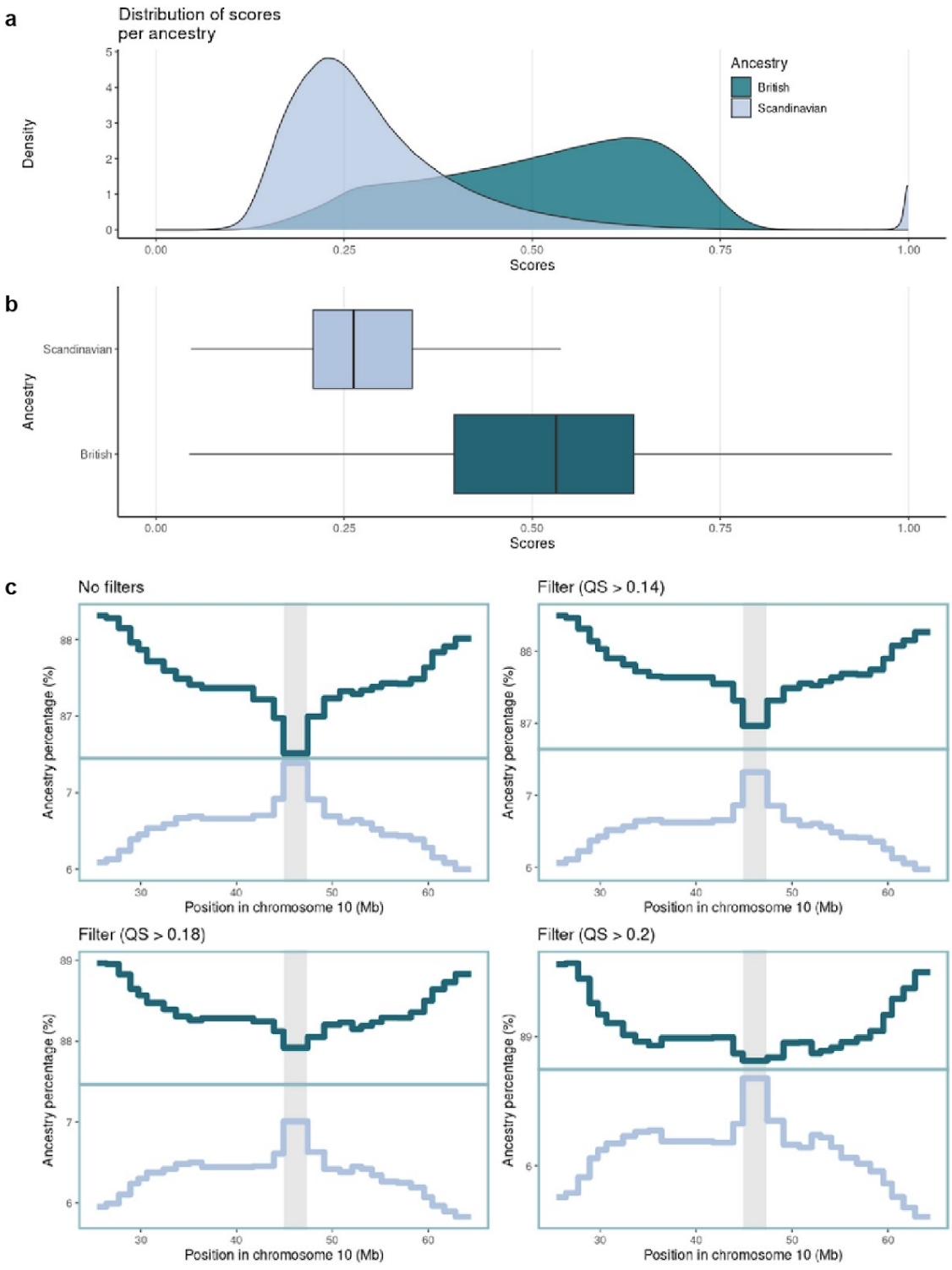
Supplementary Fig. 15 | Genomic signals of adaptive admixture form Cuadros-Espinoza et al (2022) replicated in this study. Regions shaded in green indicate local signatures of adaptation described by Cuadros-Espinoza et al (2022). Admixed populations tested include (**a**) Peruvians, (**b**) Mexicans, (**c**) Makranis, (**d**) Fulanis, (**e**) East Indonesians (**f**) Africans with Austronesian ancestry as a proxy of Malagasy and (**g**) Sahelian Arabs and Nubians. The assessed ancestry for each population is indicated in parentheses within the title of each panel. *P* values were derived from F_{adm} and LAD scores based on SNP rank and integrated using Fisher's method and evaluated with a two-sided chi-squared test. Larger points indicate variants that pass the established significance threshold, $P<10^{-6}$, shown by a horizontal dotted line. Study, dataset, sample size and reference ancestries are given in the table under each image. Reference panels used as proxies or to attempt replication of Cuadros-Espinoza et al. (2022) include South Native Americans (Maya, Zapotec, Mixe, Pima) and North Native Americans (Quechua, Karitiana, Surui, Chane) from the HGDP and SGDP; Sub-Saharan Africans for Mexicans (Yoruba from 1KGP) and for Peruvians (Yoruba, Luhya, Gambian from 1KGP); Iberians from 1KGP; Proxy Baluch individuals from UK Biobank samples born in Pakistan; Bantu speakers from South Africa, Tswana, and Kenya (HGDP, SGDP), Luhya from 1KGP and UK Biobank individuals born in Lesotho; Jola from GGVP; Mozabite from HGDP; Czech, French, and British from UK Biobank and Reich lab; South Europeans (Tuscan, Sardinians) from 1KGP; Druze and Bedouins from HGDP and SGDP; Nile peoples referenced from Hollfelder et al. (2017).



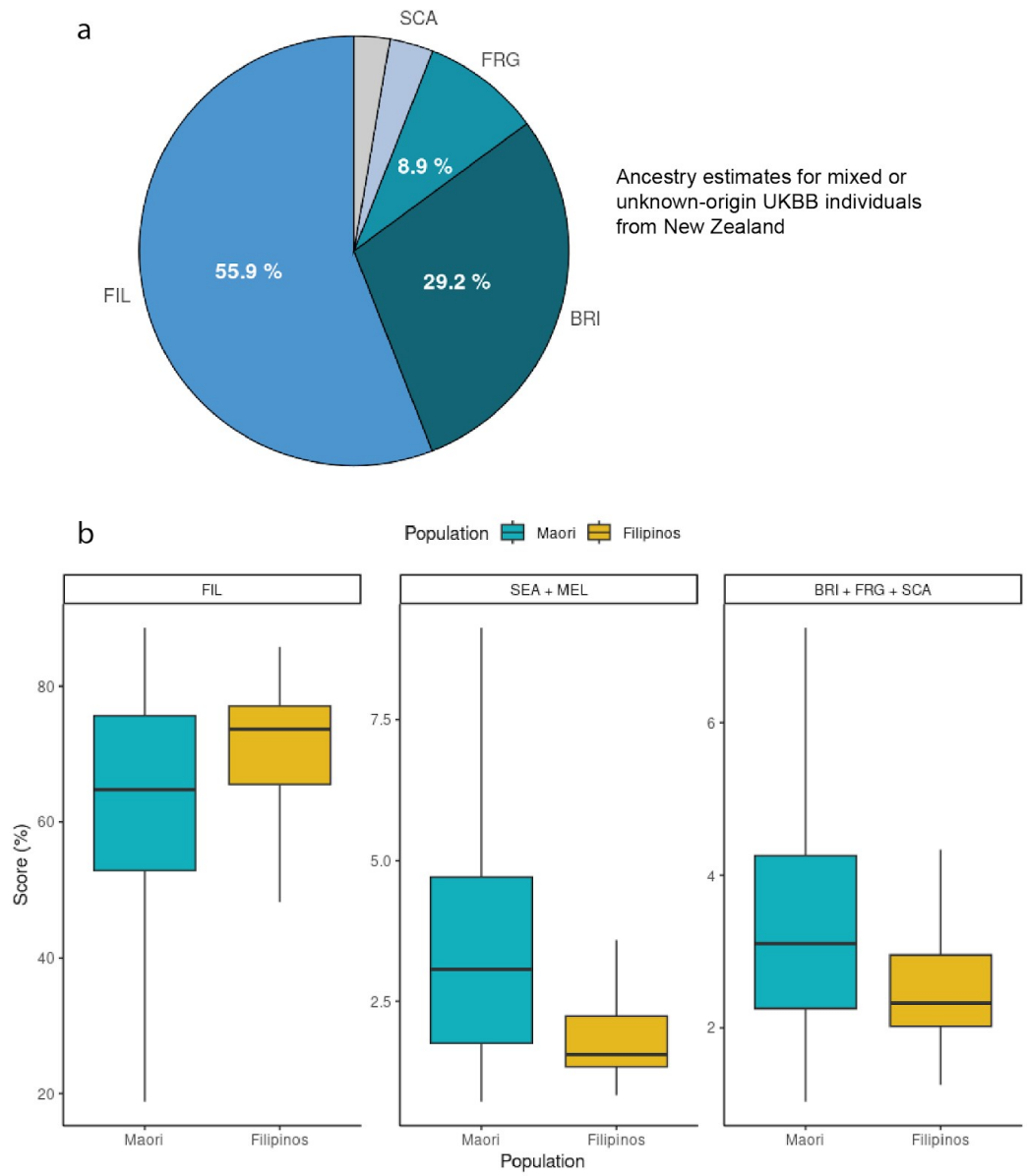
Supplementary Fig. 16 | Genome-wide adaptive admixture signal in British UKBB participants with varying levels of Scandinavian ancestry. *P* values, derived from F_{adm} and LAD scores (based on SNP rank), were integrated using Fisher's method and evaluated with a two-sided chi-squared test. Larger points indicate variants that pass the established significance threshold, shown by a horizontal dotted line.The highlighted area in gray pinpoints the chr10 signal discovered. Genome-wide signals are presented for British sample sets with increased Scandinavian heritage (**a,b,c,d**), alongside their corresponding ancestry proportions and the total number of individuals selected (**e,f,g,h**). BRI: British & Irish; SCA: Scandinavian; FRG: French & German.



Supplementary Fig. 17 | (**a**) Excess of GWAS hits located in the chromosome 10 region corresponding to the identified adaptive signal. Enrichment of parent categories (A) and GWAS phenotypes (B). The direction of the GWAS effects resulting from the presence of SCA instead of BRI ancestry was inferred from SNP frequencies (C). Discrepancies may be related to SNP frequency inaccuracies. (**b**) Nominally significant UKBB phenotypes (encompassing self-reported illness codes and primary/secondary ICD10 codes) are ranked by odds ratio. 'Protective' indicates an excess of BRI ancestry among cases or SCA among controls, while 'risk' points to a higher proportion of SCA ancestry in cases.

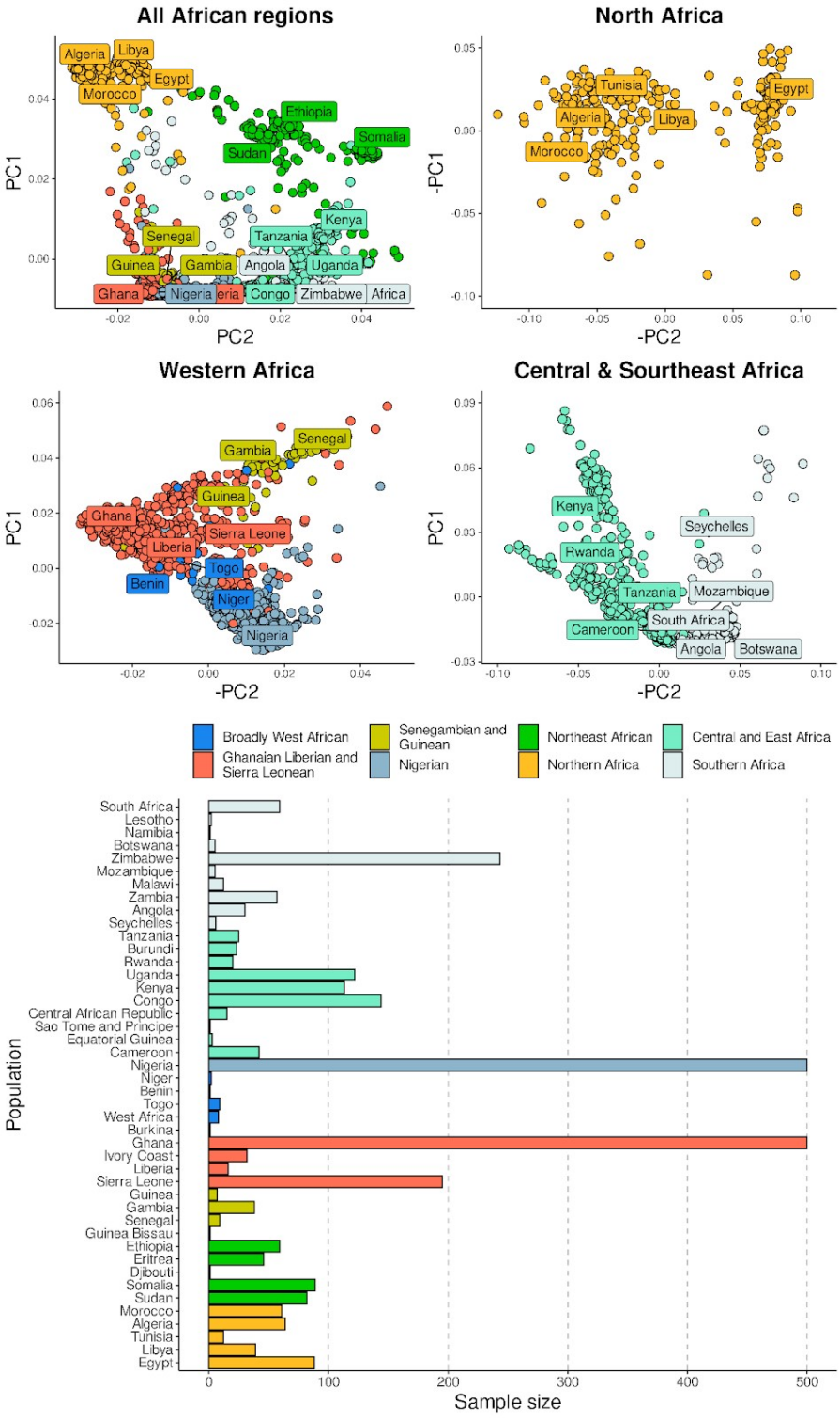


Supplementary Fig. 18 | (a,b) Distribution of Orchestra's quality scores for British and Scandinavian ancestry calls within the UKBB British samples. The central line of the boxplot denotes the median, box boundaries represent the first and third quartiles, the whiskers range from minimum to maximum values. **(c)** Local percentages of British and Scandinavian ancestries within the chromosome 10 window, using different filtering thresholds to exclude low-quality haplotypes. We applied filters to exclude haplotypes with quality scores below 14%, 18% and 20%, corresponding to the exclusion of 1%, 5% and 10% of Scandinavian haplotypes, respectively.

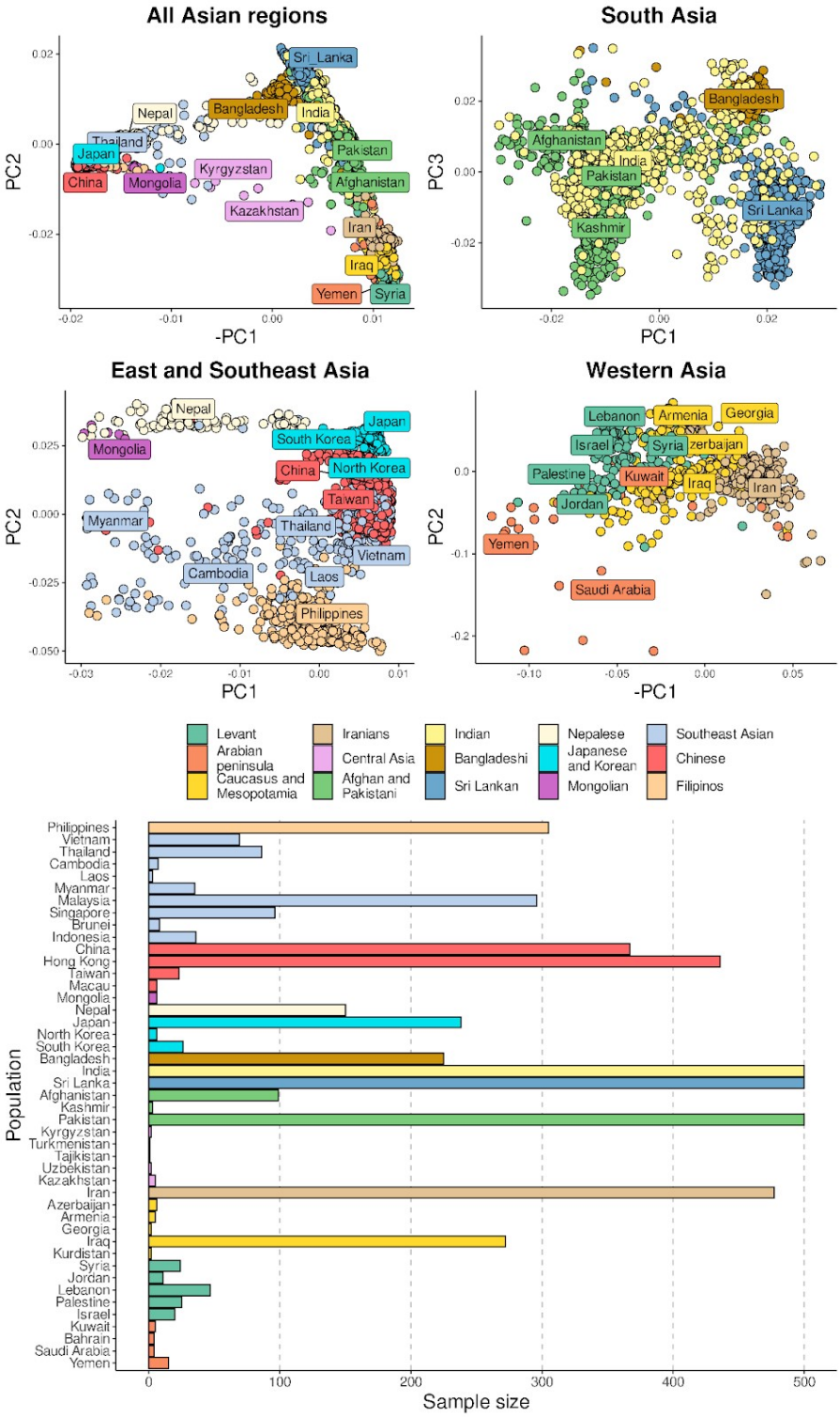


Supplementary Fig. 19 | (**a**) Distribution of ancestry estimated for mixed and unknown origin UKBB individuals from New Zealand (N=10), after removing individuals of predominantly European ancestry. FIL: Filipino; BRI: British & Irish; FRG: French & German; SCA: Scandinavian. (**b**) Distribution of Orchestra's probability scores for windows that classify as FIL in individuals of presumed Māori (N=10) and Filipino (N=10) descent in UKBB. The central line of the boxplot denotes the median, box boundaries represent the first and third quartiles, the whiskers range from minimum to maximum values. The data indicates that probability scores for FIL in the presumed Māori are slightly lower compared to those in Filipinos.

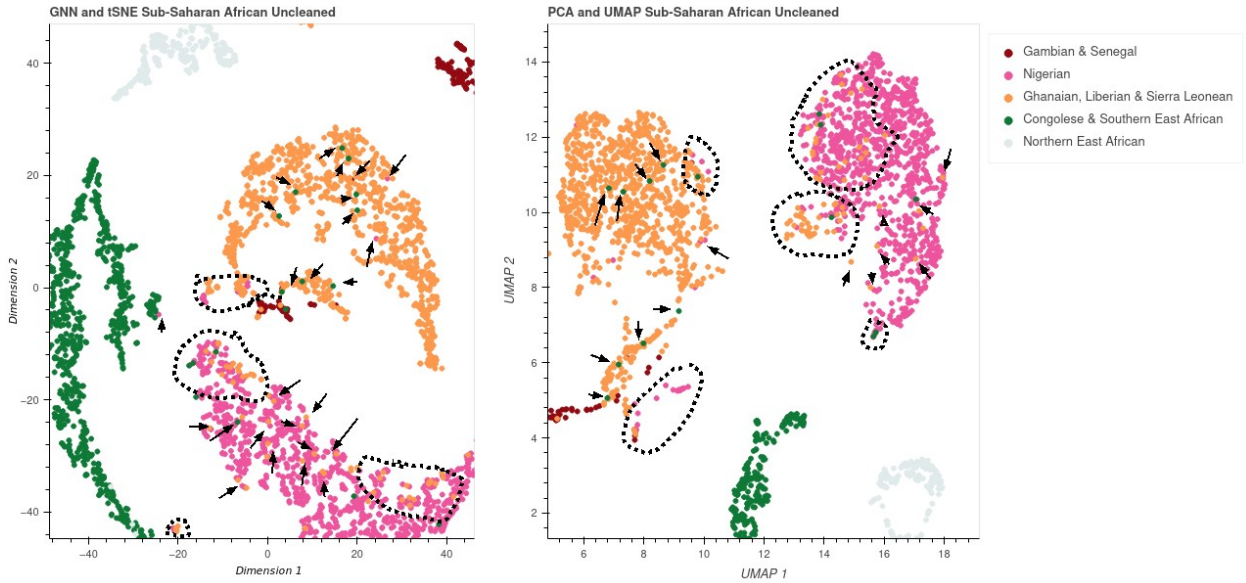
Supplementary Fig. 20 | PCA of European UKBB participants used in the reference panel and the number of samples per population. Countries and regions were randomly down-sampled to 500 individuals if their sample size was larger, in order to retain comparable numbers between populations for illustration purposes. Some country labels were not shown and rearranged around the sample centroids so they do not overlap. Note that this means that those populations capped at 500 have more participants from these regions.



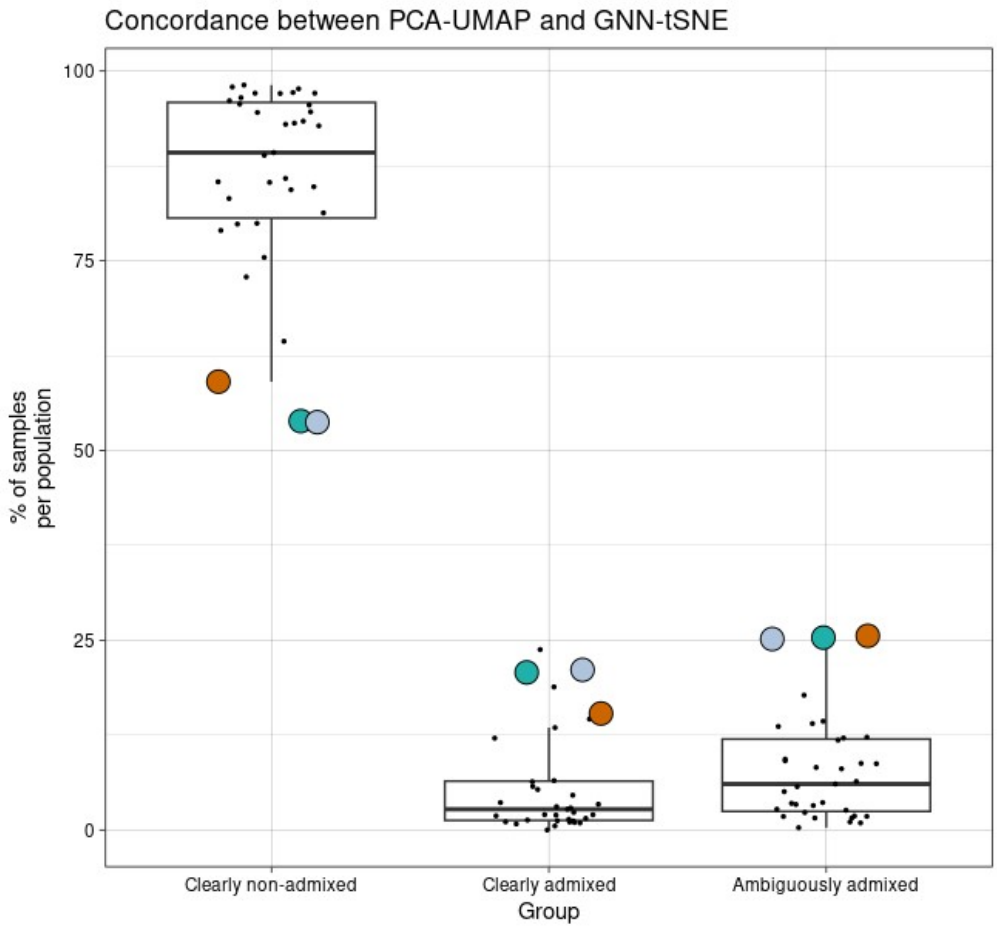
Supplementary Fig. 21 | PCA of African UKBB participants used in the reference panel and the number of samples per population. Countries and regions were randomly down-sampled to 500 individuals if their sample size was larger, in order to retain comparable numbers between populations for illustration purposes. Some country labels were not shown and rearranged around the sample centroids so they do not overlap. Note that this means that those populations capped at 500 have more participants from these regions.



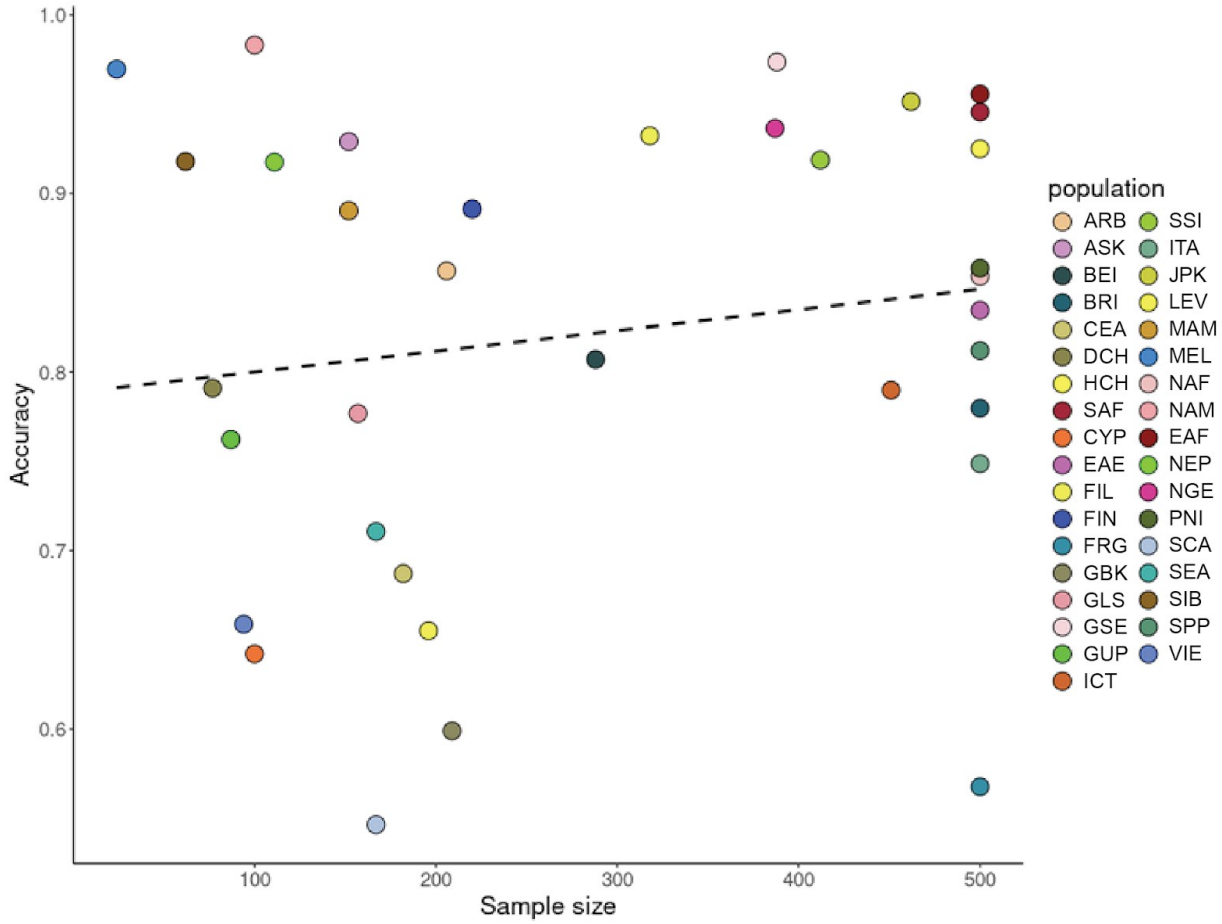
Supplementary Fig. 22 | PCA of Asian UKBB participants used in the reference panel and the number of samples per population. Countries and regions were randomly down-sampled to 500 individuals if their sample size was larger, in order to retain comparable numbers between populations for illustration purposes. Some country labels were not shown and rearranged around the sample centroids so they do not overlap. Note that this means that those populations capped at 500 have more participants from these regions.



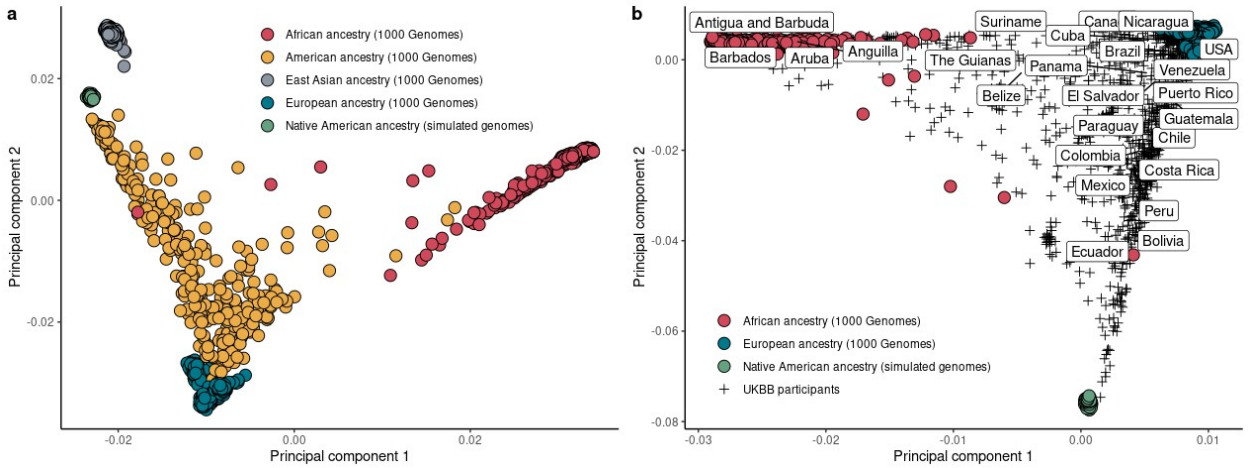
Supplementary Fig. 23 | 2D visualization of Sub-Saharan UKBB samples using two different dimensionality reduction techniques: GNN-tSNE and PCA-UMAP. Arrows and circles highlight samples that are located in different population clusters than expected, indicating potential admixture or incorrectly assigned/reported ethnicity.



Supplementary Fig. 24 | Concordance between PCA-UMAP and GNN-tSNE techniques. The "Clearly nonadmixed" group encompasses samples from the reported population correctly clustered according to expectations. The "Clearly admixed" group contains individuals identified as admixed by both techniques, showing clustering in unexpected populations. The "Ambiguously admixed" group includes samples flagged as admixed by one technique but not the other. "Ambiguously admixed" samples were included in the reference panel if fewer than 500 samples were available for a population; otherwise they were excluded. Highlighted populations are represented by distinct colors: light blue for Scandinavian, light green for Southeast Asian, and brown for Iraqi, Iranian, Caucasian & Turkish populations. These highlighted populations exhibit significant discrepancies in admixture classification with over 25% of their samples categorized as "Ambiguously admixed". The central line of the boxplot denotes the median, box boundaries represent the first and third quartiles, whiskers denote the range without outliers, and individual points denote outliers.

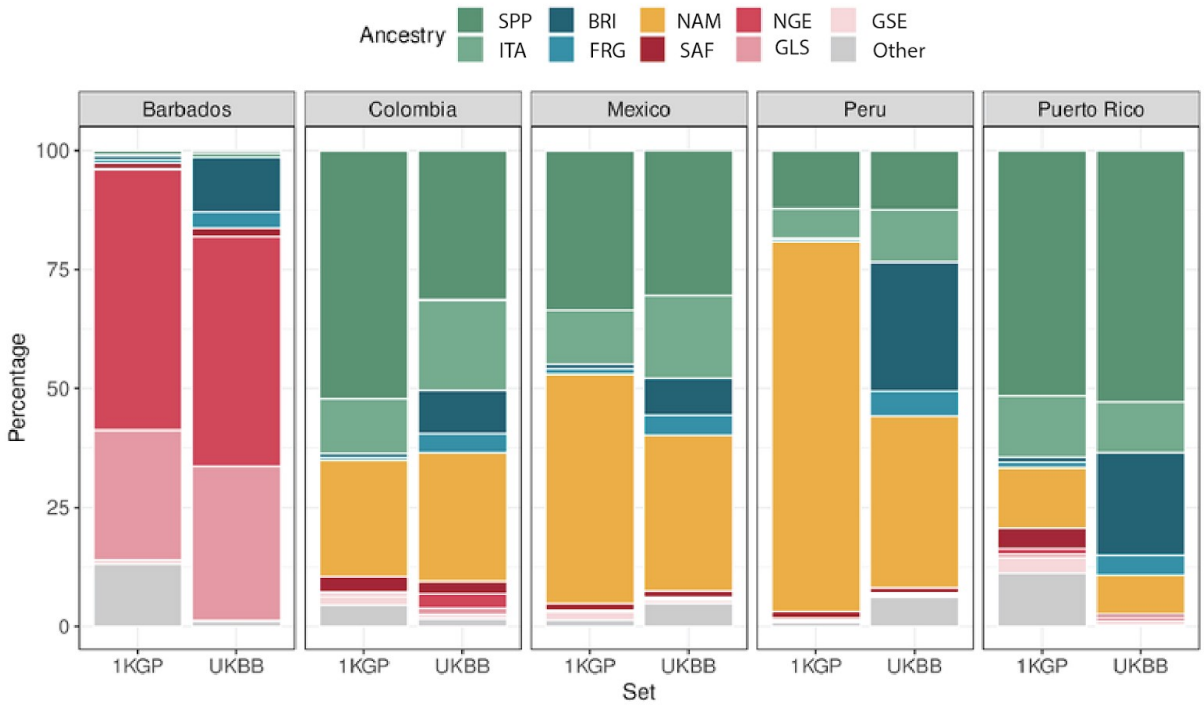


Supplementary Fig. 25 | Correlation between reference panel sample size and accuracy. Accuracy was evaluated across generations of admixture. ARB: Arab, ASK: Ashkenazi Jewish, BEI: Bengali & East Indian, BRI: British & Irish, CEA: Central Asian, CYP: Cypriot, DCH: Dai Chinese, EAE: Eastern European, EAF: Northeast African, FIL: Filipino, FIN: Finish, FRG: French & German, GBK: Greek & Balkan, GLS: Ghanaian, Ivorian, Liberian & Sierra Leonean, GSE: Gambian & Senegalese, GUP: Gujarati Patels, HCH: Han Chinese, ICT: Iraqi, Iranian, Caucasian & Turkish, ITA: Italian, JPK: Japanese & Korean, LEV: Levantine, MAM: Manchurian & Mongolian, MEL: Melanesian & Aboriginal Australian, NAF: North African, NAM: Native American, NEP: Nepalese, NGE: Nigerian, PNI: Pakistani & North Indian, SAF: Central, South & Southeast African, SCA: Scandinavian, SEA: Southeast Asian, SIB: Siberian, SPP: Spanish & Portuguese, SSI: Sri Lankan & South Indian, VIE: Vietnamese.



Supplementary Fig. 26 | Population genetic structure of artificially-constructed Native American genomes. (**a**) Principal component analysis (PCA) of generated pure Native American genomes compared to 1KGP samples as a reference. A gradient from European to American poles can be seen in those 1KGP admixed genomes that were sampled in the Americas. (**b**) UKBB participants that were born in American countries were included together with our set of pure Native American genomes and African and European individuals from 1KGP. Countries were indicated in the samples' centroid PC locations according to the UKBB participant labels. The same Europe-America gradient can be seen for admixed American UKBB participants.

Supplementary Fig. 27 | Simulated Latin American genomes for training and benchmarking purposes. The simulated Latin American genomes were designed to reflect the genetic patterns observed in contemporary Latin Americans from three major American regions: Caribbean (upper row) (N=1,356), Central American (middle) (N=1,356) and South America (lower row) (N=1,356). We conducted intermixing for 12 generations using SLiM (**a**). Percentage of each ancestry per simulated sample in a subset of the simulations (**b**). Illustration of a simulated chromosome (15) showing the lengths of each segment derived from each ancestral origin (**c**).



Supplementary Fig. 28 | Comparison of ancestral composition between Latin American samples from 1KGP and UKBB. UKBB participants that were born in the Americas show a higher proportion of British ancestry compared to 1KGP. BRI: British & Irish; SAF: Central, South & Southeast African; FRG: French & German; GLS: Ghanaian, Ivorian, Liberian & Sierra Leonean; GSE: Gambian & Senegalese; ITA: Italian; NAM: Native American; NGE: Nigerian; SPP: Spanish & Portuguese.

Supplementary Methods

Dataset Preparation

For non-UKBB samples: Autosomal genotypes from array-based studies were retrieved, converted to Variant Call Format (VCF) using PLINK v2.0[1], processed to remove variants with missing genotypes and no allele information, and then lifted-over to GRCh Build 38 with Picard toolkit[2]. All whole genome and genotyping datasets were phased with Beagle v5.4[3]. All genotyping datasets were imputed with Beagle v5.4[3] using the 1000 Genomes 30x (1KG)[4] as reference panel. We made that choice based on previous analyses, see De Marino et al. 2022[9], to maximize the comprehensive diversity and balance across different populations in our imputation reference panel. At the time we curated our dataset, this was the optimal option, as more advanced reference panels were not yet available or accessible to our team. In the case of Behar et al (2010)[5] and Tambets et al (2018)[6] datasets, variation profiling data needed to be preprocessed according to Illumina's TOP/BOT nomenclature for allele designation (illumina.com/documents/products/technotes/technote_topbot.pdf). To exclude possible relative pairs (up to second degree relatedness), we applied the software KING v2.3[7] to the entire dataset. In case of duplicated individuals, we selected the samples with the highest number of genotyped SNPs. All datasets were merged together and missing genotypes were filled with homozygous reference for sequenced genomes, since no variants (alternative alleles) were called at those locations. Lastly, we filtered variants with a minor allele frequency (MAF) below 5%, since these are likely to be SNPs with poor imputation quality[8,9]. For that, we used allele frequency computed in unrelated individuals from the 1KGP due to the high sequencing depth and reliable genotype calling. Multiallelic loci and indels were also removed.

For UKBB samples: The majority of UKBB participants have British and Irish ancestors, but this dataset is also a source of many diverse ancestries beyond the UK. We thereby focused on identifying individuals with other European birthplaces following the filtering steps detailed in Gilbert et al. (2022)[10] with some modifications, and also applied the same strategy to extract Africans and Asians (Supplementary Figs. 20-22). UKBB data was imputed using the Haplotype Reference Consortium[11], the UK10K Cohorts[12] and the 1KGP as reference panels, which together form a large and reliable haplotype resource that ensures an optimal performance[13]. For Europeans, we selected individuals with an ethnicity (f.21000 UKBB code) that was either ‘White’, ‘White British’, ‘White Irish’, ‘any other white background’ or ‘other ethnic group’. These individuals were further filtered based on birthplace (f.20115). For those participants who were born in the UK, we separately identified individuals with British or Irish birthplace (f.1647), selecting samples from England, Wales, Scotland, Northern Ireland and the Republic of Ireland. We performed an initial PCA using PLINK to identify ancestry outliers. For that, we employed the raw genotypes from UKBB array and applied a series of steps: (1) filter for autosomal genotypes; (2) removed strand ambiguous SNPs; (3) removed SNPs with missingness of > 5%, minor allele frequency of < 2% or Hardy-Weinberg equilibrium of $P < 10^{-6}$; and (4)

removed individuals with missingness of > 5%. Finally, instead of using estimated kinship with KING as in Gilbert et al. (2022)[10], to remove related individuals, we restricted the analysis to individuals used for computing the principal components in the UKBB analyses (f.22020), since these individuals are already unrelated and have passed several quality controls, including removing samples having a mismatch between inferred sex and self-reported sex and outliers based on heterozygosity[13]. PCA was calculated on pruned SNPs with PLINK. We manually identified and removed overall outlier samples indicative of non-European ancestry. Similarly, we identified participants with Asian and African ancestry. In this regard, we extracted those that were born in African and Asian countries and that passed the same quality control checks. We also removed those samples that had signs of being admixed; in particular, we removed (1) an isolated group of participants that were born in African countries but had South Asian ancestry (possibly due to Indian diaspora); (2) Caribbeans, since they fell along a genetic continuum from Sub-Saharan Africa to North Africa and Europe with no clear grouping (potential admixture with Europeans); (3) self-reported Indians that were born in Western Asia or East and Southeast Asia; and (4) all Oceanians were removed since these participants clustered with Europeans. The UKBB set was then merged with the rest.

To minimize the impact of potential biases due to imputation, we maximized the presence of directly genotyped SNPs in our final panel by filtering by the union of all genotyped SNP sets contained in the array-based studies employed. In total, this composite set consisted of 35 ancestral populations and contained 5,770,170 unique loci before batch effect and outlier removal.

Batch Effect Removal

We mitigated unwanted technical variation by filtering SNPs associated with differences among study groups, while retaining biological information. This association analysis was conducted between genotypes in each locus and study labels with an analysis of variance (ANOVA) test. The resulting F-values represent the ability of each variant to segregate samples from distinct studies and were used to rank variants from highest to lowest impact on batch effect. Those SNPs in the top 70% were marked to be discarded. Since a fraction of these SNPs may be showing a significant association due to diverse ancestry compositions within each dataset, we recovered those variants that were also highly associated with population labels as biologically important (those within the top 30 percentile of associations). In total, we retrieved 1,202,443 variants.

Quality Control

PCA-UMAP: Genotype data was pre-processed using LD pruning with PLINK and a principal component analysis (PCA) was performed on this set to extract meaningful population structure and filter stochastic noise. Next, uniform manifold approximation and projection (UMAP)[14], a non-linear dimension reduction technique designed to emphasize local features, was applied on

top of the genetic principal components (PCs). PCA-UMAP places the samples in a twodimensional space that preserves finer-scale distances, while displaying clusters that correspond with ancestral populations. PCA and UMAP were performed using run_pca and run_umap from the pca python package, respectively.

GNN-tSNE: A specific sample's GNN[15] is a vector of 35 values, each one describing the proportion of its immediate neighbors within the genealogical tree from each of the given 35 reference populations in our panel. We used a soft LD threshold ($r^2 > 0.8$) on our dataset to fit all chromosomes as input for tsinfer, maintaining a compromise between computational burden and accuracy in the reconstruction of ancestral haplotypes by the software. We next applied tdistributed stochastic neighbor embedding (t-SNE)[16], another non-linear dimensionality reduction technique, on these GNN vectors ('GNN-tSNE').

Outliers (i.e. admixed genomes) were removed automatically using a K-Nearest Neighbor (KNN) algorithm. Specifically, those samples with fewer than 'K' neighbors coming from the sample's population were excluded. To find the optimal K for KNN for population clustering (search range: from 8 to 24), along with the hyper-parameters for both PCA-UMAP and GNNtSNE that need to be tuned - including the number of neighbors (search range: from 8 to 32), minimum distance (0.0 - 1.0, step size 0.1) and the number of PCs (2 - 25) used for UMAP, and perplexity (5 - 50) and exaggeration level for t-SNE (6 - 18)-, we used Optuna[17]. Data was split into training and testing sets (80 and 20% of the samples, respectively) and the indicated ranges of values for the hyper-parameters were used to run both PCA-UMAP and GNN-tSNE in the training set. Resulting models were applied to the testing set and evaluated using balance accuracy (i.e. average of recall obtained on each reference population) as the objective function. Those values achieving the highest accuracy were selected for automatic data cleaning with KNN. This process was repeated for population labels after continental/subcontinental level cleaning.

We are aware that UMAP can over-discretize the continuous nature of ancestry into discrete clusters, creating artificial boundaries. However, this characteristic was strategically employed to enhance the curation of our reference panel. Specifically, we applied UMAP to the first principal components of PCA to accentuate sample relationships that might be obscured by PCA alone, as UMAP tends to amplify denser data regions. This combined approach leveraged PCA's ability to capture large-scale genetic structures and UMAP's capacity to elucidate finer-scale interactions, aiding in the identification and exclusion of admixed samples. We did not use this technique for sample selection but rather to remove potentially admixed individuals. Moreover, to enhance the robustness of our selection process, we complemented the PCA-UMAP analysis with t-SNE applied to genealogical nearest neighbor (GNN) statistics. This 'dual-filtering' strategy enabled us to cross-validate the ancestry consistency of each sample and reduce the biases inherent in each dimensionality reduction method (Supplementary Figs. 23-24).

Due to unbalanced sample sizes across our reference populations, we were more stringent in populations that had more samples. Thus, if the sample size was larger than 300 individuals, we

only took those samples that passed both PCA-UMAP and GNN-tSNE filters; if not, we only removed individuals that were excluded by both approaches. Other suspicious outliers that appeared after changing selected hyper-parameters in the automatic pipeline were also excluded by manual inspection with Orange3[18]. Lastly, we down-sampled populations to 500 individuals if their sample size was larger to retain comparable numbers, prioritizing individuals from studies with higher coverage of genotyped variants. We then employed random selection to ensure a broad representation of the reference population, choosing samples from various positions within the cluster rather than focusing solely on those at the center.

Analytical Framework Details and Orchestra Implementation Guide

To train Orchestra, we require a reference panel consisting of genotype data paired with a sample mapping file that links each sample ID to its population membership. This panel should consist exclusively of non-admixed samples. For benchmarking in the present study, we utilized two reference panels: (i) 1KGP-16pops, a WGS dataset from the 1000 Genomes Project; and (ii) custom-35pops, a broader and more diverse panel containing 35 populations (see ‘Reference Panel’ in the main Materials and Methods). Orchestra’s model, however, is designed to be compatible with any reference panel, with no restrictions on granularity – allowing flexibility to accommodate continental, sub-continental, population, or sub-population level distinctions – and with any number of populations. Additionally, Orchestra can be trained with adjustable window sizes and a wide range of SNP inputs, including our ancestry-specific SNP set (as used with custom-35pops), array-based genotyping data or WGS, providing adaptability for a range of research applications.

Given the challenge of obtaining datasets with perfectly annotated ground truth, we split the reference panel into two independent subsets for benchmarking: 80% for training (serving as the reference panel for model development) and 20% for testing (serving as the target for accuracy benchmarking). To create synthetic admixed individuals for training, we simulated six generations of random admixture within the training subset using SLiM[19] (see 'Simulated Admixed Individuals' in the main Materials and Methods). During Orchestra’s base layer phase, initial non-admixed samples from the training set were used as reference to reconstruct the synthetic admixed haplotypes generated with SLiM, generating recombination distance vectors (see Fig. 1a). To prevent overfitting, given that reference and target haplotypes overlapped within the training set, we excluded the best-matching haplotypes, assuming these represented the original reference genomes from which synthetic individuals were generated. These vectors served in turn to train the parameters of Orchestra’s smoothing layer (weights and bias terms), with the simulated admixed genomes as target haplotypes with known local ancestry information. Once trained, the model could then be applied to any target cohort.

For accuracy benchmarking, the resulting trained model was applied to the defined testing cohort, where the original non-admixed samples were considered as generation 0. Additionally, synthetic admixed samples were generated across six generations using SLiM to test varying

levels of admixture. This setup maintained a clear separation of haplotypes by using independent training and testing cohorts, thus avoiding any potential inflation in accuracy metrics.

Following benchmarking, Orchestra was trained on the full reference panel (without splitting) to maximize accuracy in downstream analyses. The model leverages both non-admixed individuals from the reference panel and simulated admixed individuals, which are necessary as input to train the smoothing layer parameters; thus, generating synthetic admixed samples is a required preliminary step before model training. Once trained, these models are fully portable and can be applied to any independent target cohort using Orchestra's inference module. Detailed usage instructions are available in our GitHub repository at *https://github.com/omicsedge/orchestrapaper*.

Pseudo-probability vector

Orchestra currently provides a quality metric derived from the deep learning model. For each genomic window, Orchestra outputs a pseudo-probability vector indicating its similarity to each reference population, biased toward the most likely ancestry. This metric was trained to identify the most likely ancestry per genomic region rather than to accurately quantify the probability of each potential ancestry.

We show examples of how this vector can be used in Supplementary Figures 18-19. Supplementary Figure 18 shows this metric employed in the analyses of chromosome 10 in British samples, where we looked for an enrichment of Scandinavian ancestry. By applying different thresholds to exclude low-quality calls, we confirmed that our findings remained robust, even when we filtered out less reliable data according to the metric.

Supplementary Figure 19 shows the metric applied to windows classified as Filipino (FIL) in 10 presumed samples of Māori descent from New Zealand and 10 samples from the Philippines found in the UK Biobank dataset. This was done in order to confirm that when an ancestry is not available in the reference population (Polynesian), windows of that ancestry classify as the phylogenetically closest available population – in this case Filipino (FIL).

It is important to note that this metric is affected by a number of factors, including the size and quality of a reference population, and how genetically close that population is to others in the reference panel. We recognize the need to further develop and calibrate the confidence metrics, and we plan to address this in the future.

In Silico Native American Reference Panel

We used 195 whole-genome sequenced Latin American admixed participants from 1KGP, SGDP and HGDP datasets. Children were excluded from trios. We then created *in silico* genomes that minimized the presence of other DNA segments inherited from non-Native ancestors (Supplementary Fig. 26). As current Latin American populations can be considered a mix of

African, European and Native American ancestry, we first ran Orchestra at a continental level to obtain local ancestry estimates using Africans, Europeans and East Asians found in 1KGP as a reference panel. East Asians served as a proxy for inferring Indigenous American ancestry due to their close genetic relatedness. Once the origin of haplotype blocks were called, we combined them to generate complete genomes using only those DNA sequences inferred as East Asian, i.e. indigenous. Since some regions did not have enough unique Native American segments, we reused haplotypes until only 5% of all available segments were left unused. Each chromosome was reconstructed independently and the smallest number of samples obtained for all chromosomes was selected. In total, we obtained 100 artificially-constructed Native American genomes.

Simulation of Latin American Genomes

We tailored simulations to represent the genetic makeup found today in three broad regions within Latin America: the Antilles (referred to here as Caribbean), comprised of 55% European (40% South and 15% North), 40% African (specifically NGE) and 5% Native American ancestry; Mexico and Central America (here, Central America), made up of 50% European (40% South and 10% North), 10% African (GLS and GSE) and 40% Native American ancestry; and South America, reflecting 65% European (50% South and 15% North), 15% of African (GLS, GSE and SAF) and 20% of Native American heritage (Supplementary Fig. 27)[20]. For that, we generated 2172/280/1084 Native Americans, 2172/2172/2172 South Europeans, 540/804/812 North Europeans and 540/2168/816 Africans for the Central American/Caribbean/South American simulations in the first generation with SLiM to obtain the specific admixture proportions detailed above. We took single-origin samples from the custom panel randomly and, if there were not enough to cover the specified number (for instance, artificial Native Americans), those genomes were repeated until that number was achieved. This resulted in 4,068 simulated genomes (1,356 per region).

Detecting Natural Selection Signatures

To detect natural selection in admixed populations we focused on the F_{adm} and LAD statistics described in Cuadros-Espinoza et al. (2022)[21]. We used the imputed dataset, filtering lowfrequency variants (MAF < 1%). We derived the allele frequency-based statistic, F_{adm}, for each SNP by calculating the difference between the observed allele frequency in the admixed population and the expected allele frequency. The expected frequency was estimated using allele frequencies observed in modern populations and inferred from our custom-made ancestry reference panel. To minimize imputation inaccuracies or frequency biases arising from using contemporary populations as a surrogate for the ancestral populations that led to the formation of admixed groups, variants were excluded if their observed allele frequency in the admixed population fell outside the range observed in the source populations. Moreover, variants were excluded if the disparity between their observed and expected frequencies exceeded 40%, due to

potential allele annotation artifacts or allele switches. The LAD statistic was quantified by the deviation of local ancestry percentage in each genomic window from the average found in its chromosome pack. In cases where high deviations from mean LAD scores were observed at extreme chromosome ends, we adjusted the 10 windows at the tips of chromosome peaks by the ratio between these peaks and areas outside, to address potential biases in these regions. We employed Orchestra to estimate both local ancestry for LAD scores and global admixture proportions for F_{adm} scores. To enhance detection accuracy, empirical *P* values derived from F_{adm} and LAD scores — calculated as the SNP rank for a given statistic divided by the total number of analyzed SNPs (for LAD, all SNPs within a window were assigned with the lowest corresponding *P* value) — were integrated using Fisher's method and subsequently evaluated with a chi-squared test. To ascertain statistical significance, we established a stringent *P* value threshold at 10^{-6}, exceeding the rigor of a Bonferroni correction, which was based on the number of Orchestra-inferred haplotypes ($0.05 / 1992 / 2 = 1.255 \times 10^{-5}$). This helped mitigate potential false positives arising from the extensive number of variants present in the imputed dataset.

Supplementary References

1. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet*. **81**, 559–575 (2007). doi: 10.1086/519795; pmid: 17701901

2. Picard v.2.26.7 v.3.0.0 (The Broad Institute Picard, 2020 & 2021). http://broadinstitute.github.io/picard/

3. Browning, B. L., Zhou, Y. & Browning, S. R. A One-Penny Imputed Genome from NextGeneration Reference Panels. *Am. J. Hum. Genet*. **103**, 338-348 (2018). doi: 10.1016/j.ajhg.2018.07.015; pmid: 30100085

4. Byrska-Bishop, M. et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell* **185**, 3426-3440.e19 (2022). doi: 10.1016/j.cell.2022.08.004; pmid: 36055201

5. Behar, D. M. et al. The genome-wide structure of the Jewish people. *Nature* **466**, 238-242 (2010). doi: 10.1038/nature09103; pmid: 20531471

6. Tambets, K. et al. Genes reveal traces of common recent demographic history for most of the Uralic-speaking populations. *Genome Biol*. **19**, 139 (2018). doi: 10.1186/s13059-018-1522-1; pmid: 30241495

7. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867-2873 (2010). doi: 10.1093/bioinformatics/btq559; pmid: 20926424

8. 1000 Genomes Project Consortium, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015). doi: 10.1038/nature15393; pmid: 26432245

9. De Marino, A. et al. A comparative analysis of current phasing and imputation software. *PLoS One*. **17**, e0260177 (2022). doi: 10.1371/journal.pone.0260177; pmid: 36260643

10. Gilbert, E., Shanmugam, A. & Cavalleri, G. L. Revealing the recent demographic history of

Europe via haplotype sharing in the UK Biobank. *Proc. Natl. Acad. Sci. U. S. A*. **119**, e2119281119 (2022). doi: 10.1073/pnas.2119281119; pmid: 35696575

11. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016). doi: 10.1038/ng.3643; pmid: 27548312

12. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun*. **6**, 8111 (2015). doi: 10.1038/ncomms9111; pmid: 26368830

13. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018). doi: 10.1038/s41586-018-0579-z; pmid: 30305743

14. McInnes, L. et al. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, **3**, 861 (2018). doi: 10.21105/joss.00861

15. Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330-1338 (2019). doi: 10.1038/s41588-019-0483-y; pmid: 31477934

16. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008). https://www.jmlr.org/papers/v9/vandermaaten08a.html

17. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M.. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; 2623-2631 (2019).

18. Demšar, J. et al. Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* **14**, 2349−2353 (2013). https://jmlr.org/papers/v14/demsar13a.html

19. Haller, B. C. & Messer, P. W. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model. *Mol. Biol. Evol*. **36**, 632-637 (2019). doi: 10.1093/molbev/msy228; pmid: 30517680

20. Ongaro, L. et al. The Genomic Impact of European Colonization of the Americas. *Curr. Biol*. **29**, 3974-3986.e4 (2019). doi: 10.1016/j.cub.2019.09.076; pmid: 31735679

21. Cuadros-Espinoza, S., Laval, G., Quintana-Murci, L. & Patin, E. The genomic signatures of natural selection in admixed human populations. *Am. J. Hum. Genet*. **109**, 710-726 (2022). doi: 10.1016/j.ajhg.2022.02.011; pmid: 35259336