Quantitative Biology 2022, 10(3): 287–298
https://doi.org/10.15302/J-QB-022-0303

RESEARCH ARTICLE

Identification of genomic regions distorting population structure inference in diverse continental groups

Qiuxuan Liu, Degang Wu, Chaolong Wang*

Department of Epidemiology and Biostatistics, Ministry of Education Key Laboratory of Environment and Health and State Key Laboratory of Environmental Health (Incubating), School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China
* Correspondence: chaolong@hust.edu.cn

Received January 28, 2022; Revised March 25, 2022; Accepted April 18, 2022

***Background*: Inference of population structure is crucial for studies of human evolutionary history and genome-wide association studies. While several genomic regions have been reported to distort population structure analysis of European populations, no systematic analysis has been performed on non-European continental groups and with the latest human genome assembly.**
***Methods*: Using the 1000 Genomes Project high coverage whole-genome sequencing data from four major continental groups (Europe, East Asia, South Asia, and Africa), we developed a statistical framework and systematically detected genomic regions with unusual contributions to the inference of population structure for each of the continental groups.**
***Results*: We identified and characterized 27 unusual genomic regions mapped to GRCh38, including 13 regions around centromeres, 2 with chromosomal inversions, 8 under natural selection, and 4 with unknown causes. Excluding these regions would result in a more interpretable population structure inferred by principal components analysis and ADMIXTURE analysis.**
***Conclusions*: Unusual genomic patterns in certain regions can distort the inference of population structure. Our compiled list of these unusual regions will be useful for many population-genetic studies, including those from non-European populations.**

Availability: The code to reproduce our results is available at the website of Github (/dwuab/UnRegFinder).

Keywords: population genetics; population structure; linkage disequilibrium; principal component analysis; natural selection

Author summary: We propose a systematical analysis framework based on principal component analysis (PCA) to identify such genomic regions. Based on whole-genome sequencing data from four major continental groups with no recent admixture from the 1000 Genomes Project, we compile a list of 27 unusual genomic regions and demonstrate that excluding these regions can lead to more interpretable population structure results. We recommend removing these regions as a routine in the analysis of population structure to avoid artifact results.

INTRODUCTION

Human population structure is shaped by evolutionary forces such as migration, admixture, and natural selection. Inference of population structure is the first step to understand the demographic history of different populations and depict their evolutionary relationship [1–4]. Moreover, in genome-wide association studies (GWAS) of complex traits and diseases, uncontrolled

population structure can lead to numerous spurious association signals [5–8]. This phenomenon, known as population stratification, has become more evident as genetic studies nowadays often include hundreds of thousands of samples from diverse populations [9,10]. Therefore, population structure of GWAS samples should be inferred accurately to avoid false positive signals, using multivariate dimensional reduction techniques, such as principal component analysis (PCA) [11–13]. Model-based clustering algorithms, such as STRUCTURE and ADMIXTURE [14–16], have been widely used to infer population structure from genome-wide SNP data [2–4,17].

One crucial assumption of population structure inference is that all SNPs in the genome have undergone the same demographic history, which drives the allele frequency drift. This assumption is likely to be violated if certain genomic regions are under strong selection. For this reason, several methods have been developed to detect selection signals by identifying regions with the unusual contribution to the inferred population structure [16,18–20]. Alternatively, genomic regions with unusual linkage disequilibrium (LD) might be overweighted in the inference of population structure, resulting in artifact patterns that are difficult to interpret. In an early study in 2008, Price *et al*. identified 24 autosomal regions spanning > 2 megabases (Mb) with outlier SNP loadings in PCA of European GWAS data, and argued that these regions were subjected to long-range LD rather than natural selection [21]. Nevertheless, several long-range LD regions reported by Price *et al.* were inferred as selection signals by a later study from the same research group [19]. SNP loadings from PCA alone cannot distinguish selection signals from long-range LD, but such distinction is irrelevant if the goal is to infer population structure. Thus, we refer to regions that distort population structure inference as "unusual regions", regardless of their origins.

It has become a common practice to remove the unusual regions reported by Price *et al.* [21] when analyzing population structure. Despite its high citation, we note several limitations of this early study. First, the analysis was restricted to European samples. It is unclear whether the proposed list of unusual regions is equally suitable for samples of non-European ancestry given the population difference in LD patterns [2,17]. Second, the study was based on old versions of genotyping microarrays, which suffered from ascertainment bias and insufficient genome coverage [22]. Third, the reported regions in Price *et al.* [21] were based on human genome assembly GRCh36, which has been replaced by GRCh38 in 2013. Last, the method to identify the unusual regions was not clearly described in the brief letter of Price *et al.* [21], hindering an updated analysis. Considering these limitations and the increasing volume of genetic data from non-European populations, we develop a formal analysis framework and apply it to high coverage whole-genome sequencing (WGS) data of diverse continental groups from 1000 Genomes Project [23].

RESULTS

Unusual regions in Europeans

After quality control (QC) procedures, 502 samples with 1,013,031 biallelic SNPs were included in subsequent analyses (materials and methods; Supplementary Fig. S1A). PCA yielded 11 significant PCs ($P < 0.001$, Tracy-Widom test; materials and methods), among which PC1 grouped the five populations into 3 clusters: northern Europeans (FIN), northern and western Europeans (CEU and GBR) and southern Europeans (IBS and TSI), while PC2 separated northern and western Europeans (CEU and GBR) from the rest (IBS, TSI and FIN) (Fig. 1A). No clear pattern associated with population structure could be identified in higher order PCs. After genomic control, we treated each significant PC as a quantitative trait and performed genome-wide association tests based on linear regression without covariates (materials and methods). We observed significant SNP associations in the *OCA2-HERC2* region with PC1 ($P = 7.17\times10^{-9}$ at rs12916300 on chromosome 15), *LCT* region with PC2 ($P = 1.49\times10^{-23}$ at rs182549 on chromosome 2), centromere regions of chromosome 5 ($P = 3.66\times10^{-8}$ at rs4865508) and chromosome 11 ($P = 1.97\times10^{-13}$ at rs61898393), the *KBTBD3-AASDHPPT* region ($P = 2.64\times10^{-8}$ at rs76871450 on chromosome 11) with PC3, centromere regions of chromosome 11 ($P = 3.06\times10^{-12}$ at rs7110003) and chromosome 12 ($P = 1.05\times10^{-252}$ at rs1400310636) with PC4 (Fig. 1B). Combining association signals across all 11 significant PCs (materials and methods), we obtained a well-controlled genomic inflation factor of $\lambda_{GC} = 1.002$ (Fig. 1C) and 15 regions showing excessively significant associations (Fig. 1D).

Next, we removed the unusual regions identified above and repeated PCA and PC association tests to detect additional unusual regions. We refer to the analysis with all biallelic SNPs as iteration 1. After removing the unusual regions, the number of significant PCs was reduced from 11 to 4 (iteration 2), and 3 significant regions were identified, including 1 new region (the centromere region labelled in Fig. 2A). We iterated this procedure again (iteration 3) and no new signals could be identified (Fig. 2B). Among signals identified across all iterations, the most significant ones

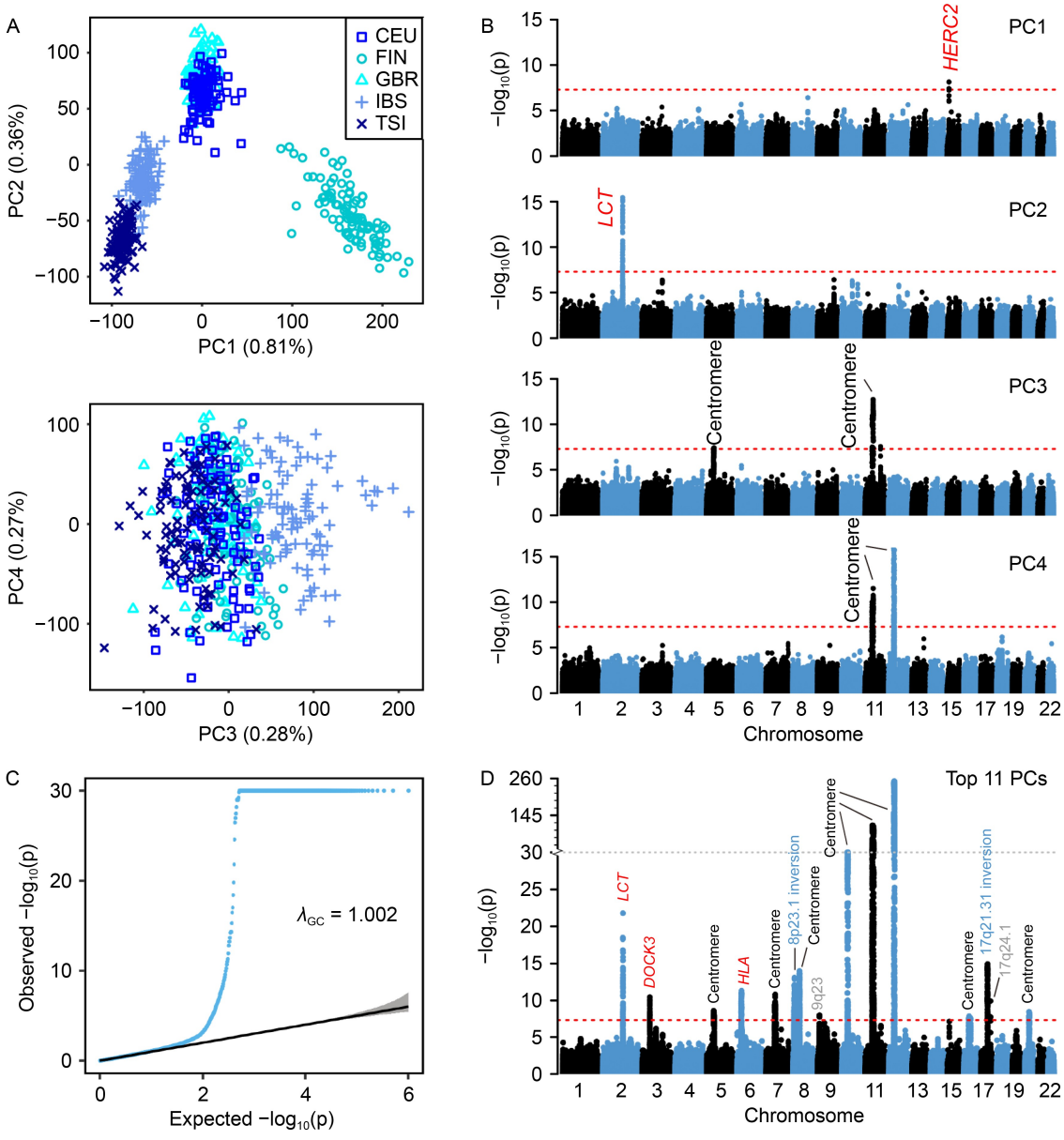

Figure 1. Unusual genomic regions in Europeans. (A) Top 4 PCs showing population structure of Europeans. Proportion of variance explained by each PC is indicated in the parentheses. (B) Manhattan plots of genetic associations with each PC after genomic control. The y-axes were truncated at $P = 10^{-15}$ for better visualization. (C−D) QQ plot and Manhattan plot of P_{Sidak}, combining signals across top 11 PCs. In (C), the black line indicates diagonal and the shaded area indicates 95% confidence interval. In (D), the red dash line indicates $P = 5\times10^{-8}$ and the gray dash line denotes break at $P = 10^{-30}$. Signals are labeled in colored texts based on potential causes: black, centromere; blue, inversion; red, selection; gray, unknown.

were three centromere regions on chromosomes 10, 11, and 12, all of which have signals with extremely small *P*-values, and the fourth signal was the *LCT* region on chromosome 2, a well-established positive selection signal in which the *LCT* gene encodes the lactase enzyme to digest lactose in milk [24]. In total, we identified 17 regions with unusual contribution to the inference of European population structure by PCA (Table 1), among which 4 were reported to harbor genes under selection (Supplementary Table S1), 9 were around the centromeres, 2 contained common inversion polymorphisms (8p23.1 and 17q21.31) [25–27], and 2 with unknown causes (Fig. 2C−D; chr9:11.35−11.88 Mb and chr17:64.89−64.93 Mb). After excluding all identified unusual regions, the number of significant PCs was reduced to 3. We compared the top 3 PCs before with those after excluding unusual regions (Supplementary Fig. S2). While the overall patterns

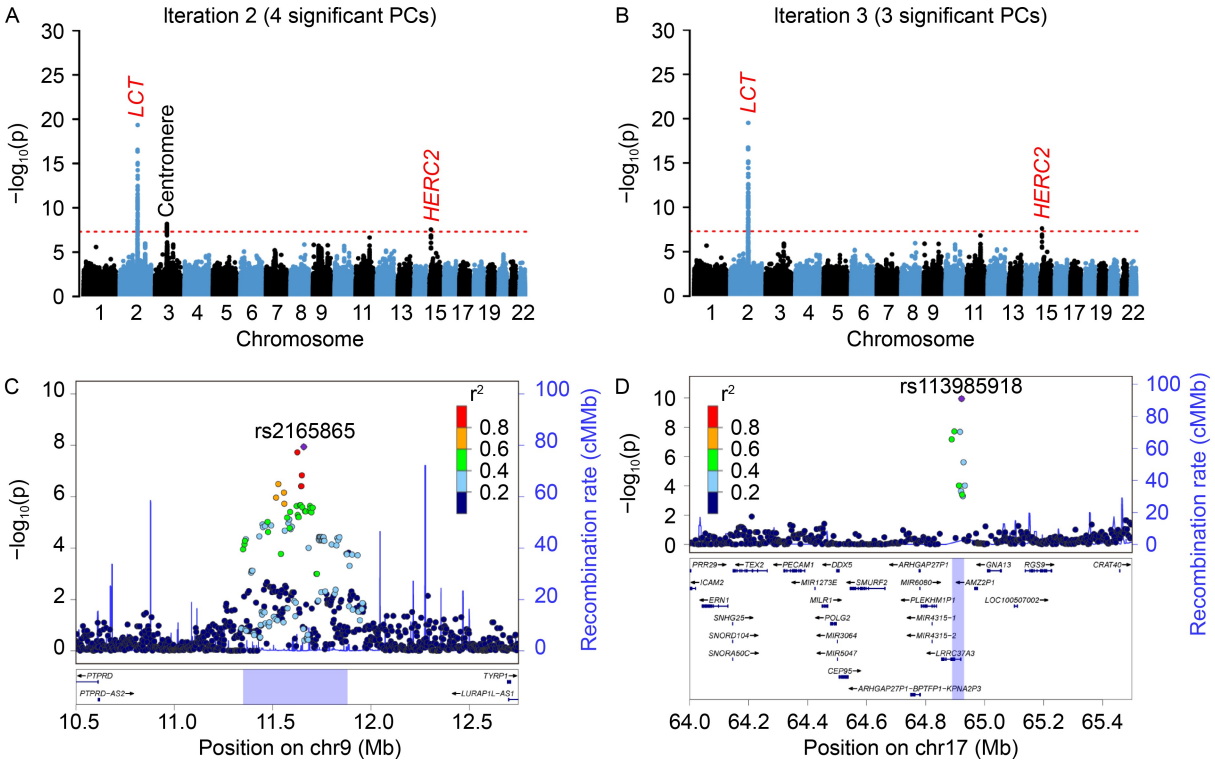

Figure 2. Manhattan and LocusZoom plots of unusual genomic regions identified in Europeans. We iteratively repeated analyses after excluding unusual genomic regions identified in previous iterations. (A−B) Manhattan plots showing P_{Sidak} in iteration 2−3. The red dash line indicates $P = 5\times10^{-8}$, and signals are labeled in colored texts based on potential causes: black, centromere; red, selection. (C−D) LocusZoom plots of two loci (9p23 and 17q24.1) with unknown causes identified in iteration 1 (Fig. 1D). The lead SNP in each locus is indicated by the purple diamond, with surrounding SNPs colored by LD to the lead SNP. The identified unusual region is shaded in purple at the bottom panel.

were similar, excluding unusual regions led to better separation of different populations, with the proportion of between-population variance Ψ (materials and methods) increased from 0.8854 to 0.9133 for the top 3 PCs.

The difference was more evident in the unsupervised ADMIXTURE analysis (Fig. 3). With the original data, two individuals with similar ancestral compositions when K = 4 could become drastically different in ancestral compositions when K = 5, as evident by the rugged ancestral compositions profiles in the lower panel of Fig. 3A, which is indicative of artifact components. After excluding the unusual regions, two individuals with similar ancestral compositions when K = 4 tend to have similar ancestral compositions when K = 5 (Fig. 3B). The proportion of the blue component, along with the green component, decreased from southern Europeans (TSI and IBS) to northern Europeans (FIN). The orange component prevalent in GBR and CEU reflected northern and western European ancestry. In summary, after excluding the unusual regions, the ADMIXTURE results reflect the geographic distribution of the various European populations better than those based on the original data.

Unusual regions in non-Europeans

We applied the same QC procedures to East Asians (Supplementary Fig. S1B), South Asians (Supplementary Fig. S1C) and Africans (Supplementary Fig. S1D). 501 (463, 481) samples with 983,988 (1,020,631, 1,108,267) biallelic SNPs remained for the three continental groups, respectively. Genomic inflation for the analyses in these continental groups remained well controlled, with λ_{GC} = 1.023, 1.017, and 1.023 for East Asians, South Asians, and Africans, respectively (Supplementary Fig. S3). For East Asians, we identified 13 unusual genomic regions in 3 iterations, including 8 regions around centromeres, 4 containing genes under selection, and 1 with unknown causes (Fig. 4). Three well-known selected loci, the *HLA* locus on chromosome 6, the *FADS* locus on chromosome 11, and the *IGH* locus on chromosome 14, remained significant in all iterations, indicating substantial allele frequency differences between East Asian subpopulations. After removing the unusual regions, the number of significant

Table 1 List of unusual genomic regions that may distort population structure inference

Index	Position (Mb)[a]	Size (Mb)[b]	Significant region in each continental group (Mb)[c]				Potential cause[d]
			Europe	East Asia	South Asia	Africa	
1	chr9: 40.61−67.74	27.13			40.61−67.74		Centromere: 42.2−45.5 Mb,constitutive heterochromatin
2	chr16: 31.86−47.04	15.18	32.18−46.94			31.86−47.04	Centromere: 35.3−38.4 Mb,constitutive heterochromatin
3	chr11: 46.46−57.20	10.74	46.46−57.20	47.98−56.89	46.57−57.14	47.53−56.86	Centromere: 51.0−55.8 Mb
4	chr2: 87.34−94.64	7.30			87.34−94.64		Centromere: 91.8−96.0 Mb
5	chr12: 33.14−39.77	6.63	33.23−39.24	34.46−38.28	33.38−39.14	33.14−39.77	Centromere: 33.2−37.8 Mb
6	chr10: 36.95−43.05	6.10	36.95−43.05	37.60−42.38			Centromere: 38.0−41.6 Mb
7	chr7: 57.37−63.46	6.09	57.37−63.46		57.41−63.10		Centromere: 58.1−62.1 Mb
8	chr5: 45.03−51.03	6.00	45.40−50.59	45.03−51.03	45.29−50.56		Centromere: 46.1−51.4 Mb
9	chr8: 42.46−48.29	5.83	43.87−47.70	42.67−46.94	43.90−47.39	42.46−48.29	Centromere: 43.2−47.2 Mb
10	chr3: 89.41−94.48	5.07	89.41−94.48				Centromere: 87.8−94.0 Mb
11	chr6: 58.41−63.17	4.76		58.41−61.98	58.41−63.17		Centromere: 58.5−62.6 Mb
12	chr20: 28.74−31.48	2.74	28.74−31.42	28.74−31.48			Centromere: 25.7−30.4 Mb
13	chr1: 122.96−124.64	1.68		122.96−124.64			Centromere: 121.7−125.1 Mb
14	chr8: 8.23−12.35	4.12	8.23−12.35				Chromosomal inversion: 8p23.1
15	chr17: 45.39−46.79	1.40	45.39−46.79				Chromosomal inversion: 17q21.31
16	chr6: 25.71−31.58	5.87	25.71−31.58	30.15−30.29			Selection: *HLA* group
17	chr2: 134.54−136.27	1.73	134.54−136.27				Selection: *LCT*
18	chr3: 50.27−51.72	1.45	50.39−51.72			50.27−51.71	Selection: *DOCK3/MAPKAPK3/CISH*
19	chr14: 66.17−67.42	1.25		66.17−67.42			Selection: *GPHN*
20	chr11: 61.62−61.90	0.28		61.62−61.90			Selection: *FADS1/FADS2*
21	chr14: 105.51−105.76	0.25		105.51−105.76			Selection: *IGH* group
22	chr15: 28.09−28.32	0.23	28.09−28.32				Selection: *OCA2/HERC2*
23	chr15: 48.10−48.22	0.12			48.10−48.22		Selection*: MYEF2/SLC24A5*
24	chr7: 64.85−66.80	1.95			64.85−66.80		Unknown
25	chr1: 173.49−175.06	1.57		173.49−175.06			Unknown
26	chr9: 11.35−11.88	0.53	11.35−11.88				Unknown
27	chr17: 64.89−64.93	0.04	64.89−64.93				Unknown

[a] Position, starting and ending positions after merging overlapping regions across continental groups. All coordinates are based on GRCh38.
[b] Size, length of the merged region.
[c] We listed the starting and ending position for significant regions in the continent-specific analysis.
[d] Centromere coordinates (labelled after colon) were obtained from the UCSC Genome Browser. There are constitutive heterochromatin regions near the centromeres of chromosomes 9 and 16, leading to large gaps with no SNPs. We merged two regions before and after these constitutive heterochromatin regions. For regions reported to be under selection, we listed the candidate selected genes reported in the literature (see Supplementary Table S1 for more information).

PCs decreased from 9 to 4, while the proportion of between-population variance Ψ increased from 0.8014 to 0.8305 for the top 4 PCs, suggesting better capture of population structure (Supplementary Fig. S4). In the unsupervised ADMIXTURE analysis of East Asian data, we observed a blue component prevalent in a substantial fraction of individuals in all populations when analyzing the original data (Fig. 5A). Such pattern disappeared in the analysis with unusual regions removed, in which the blue component was driven by 4 individuals from CHS, and the purple component is predominant in Japanese (Fig. 5B).

For South Asians, we identified 10 regions in 3 iterations (Fig. 6A−C), including 8 around centromeres, 1 under selection, and 1 with unknown cause. The region under selection contained *SLC24A5*, a well-known gene associated with skin pigmentation [28]. The region with unknown cause (chr7:64.85−66.80 Mb) was close to the centromere of chromosome 7 but showed independent association signals (Fig. 6D). Removal of the unusual regions resulted in a reduction in the number of significant PCs from 13 to 7 and a small increase of

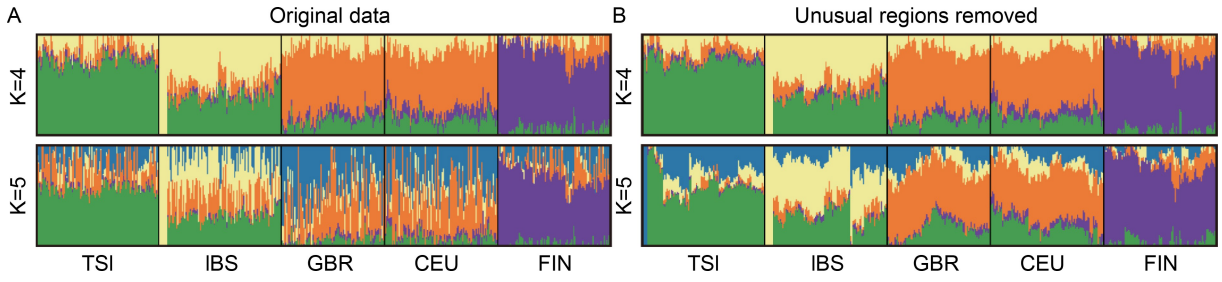

Figure 3. Impacts of unusual genomic regions on the ADMIXTURE analyses of Europeans. (A) Results based on original data. (B) Results based on data excluding unusual regions. We assumed K = 4 or 5 ancestral components, indicated by colors, in each analysis. Each vertical bar represents one individual, and the lengths of colored segments represent proportions of ancestral components. Individuals were plotted in the same order across panels, which was determined by hierarchical clustering on the ancestral proportions in panel B (K = 5)

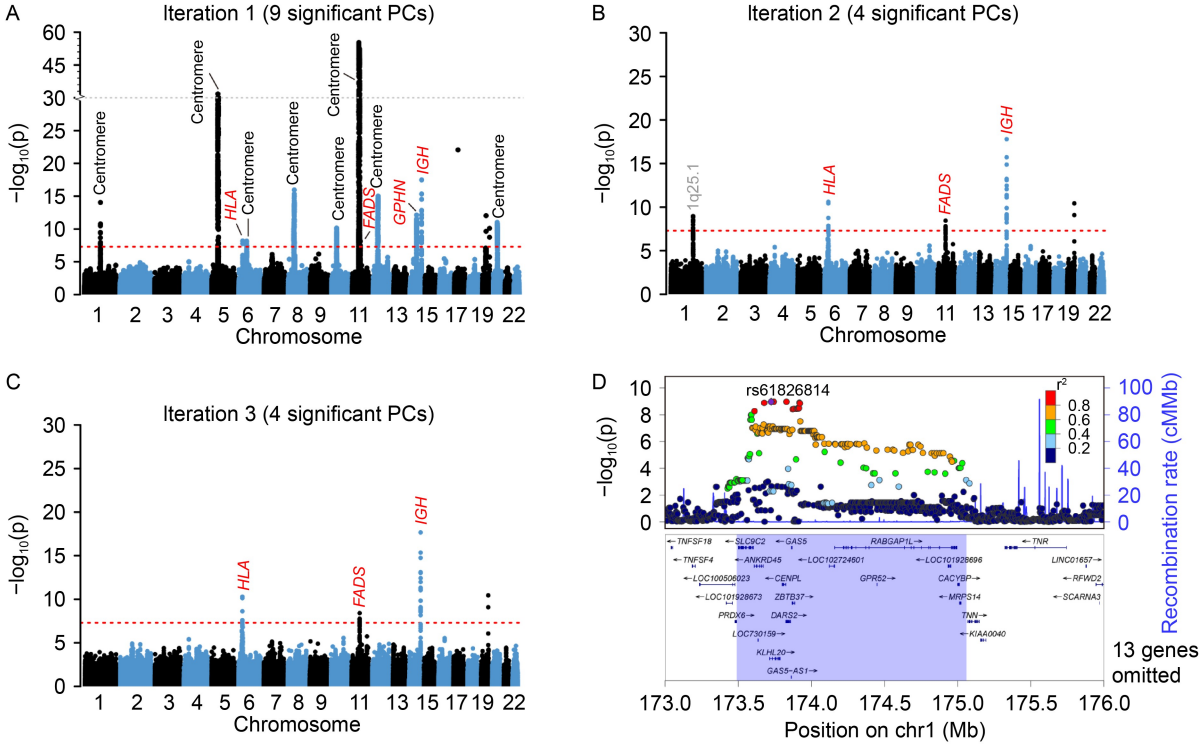

Figure 4. Manhattan and LocusZoom plots of unusual genomic regions identified in East Asians. We iteratively repeated analyses after excluding unusual genomic regions identified in previous iterations. (A−C) Manhattan plots showing P_{Sidak} in iterations 1−3. The red dash line indicates $P = 5\times10^{-8}$, and the gray dash line denotes break at $P = 10^{-30}$. Signals are labeled in colored texts based on potential causes: black, centromere; red, selection; gray, unknown. We labeled signals of unknown cause with the gene nearest to the leading SNP. (D) LocusZoom plot of the 1q25.1 locus with unknown causes identified in iteration 2. The lead SNP is indicated by the purple diamond, with surrounding SNPs colored by LD to the lead SNP. The identified unusual region is shaded in purple at the bottom panel.

Ψ for the top 7 PCs from 0.4057 to 0.4145. There was little change in the top PCs (Supplementary Fig. S5). Comparing ADMIXTURE analyses with and without removing unusual regions, no apparent difference could be observed until K = 7 (Supplementary Fig. S6).

For Africans, we identified 5 unusual regions in one iteration: 4 around centromeres and 1 under selection (Fig. 7). In addition, we observed 5 isolated significant SNPs in the first iteration, which were likely noise. After removing unusual regions, the number of significant PCs decreased from 9 to 4. The proportion of between-population variance Ψ increased substantially from 0.8623 to 0.9033 for the top 4 PCs. Consistently, we observed a clearer separation between different populations, especially between two West African populations, YRI and ESN (Supplementary Fig. S7).

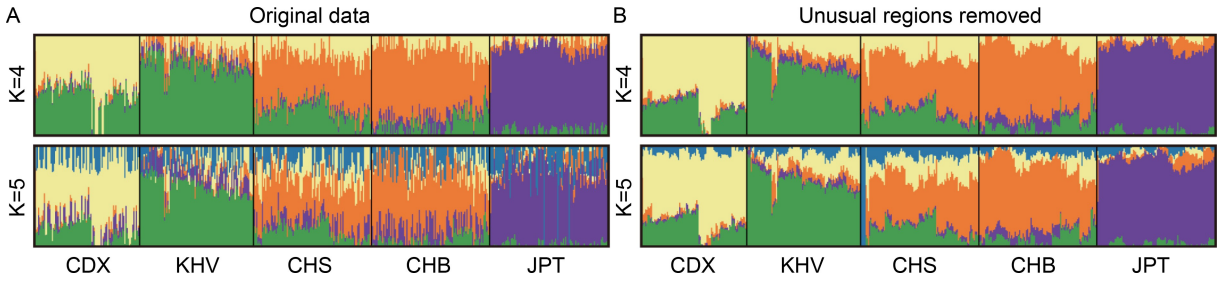

Figure 5. Impacts of unusual genomic regions on the ADMIXTURE analyses of East Asians. (A) Results based on original data. (B) Results based on data excluding unusual regions. We assumed K = 4 or 5 ancestral components, indicated by colors, in each analysis. Each vertical bar represents one individual, and the lengths of colored segments represent proportions of ancestral components. Individuals were plotted in the same order across panels, which was determined by hierarchical clustering on the ancestral proportions in panel (B) (K = 5)

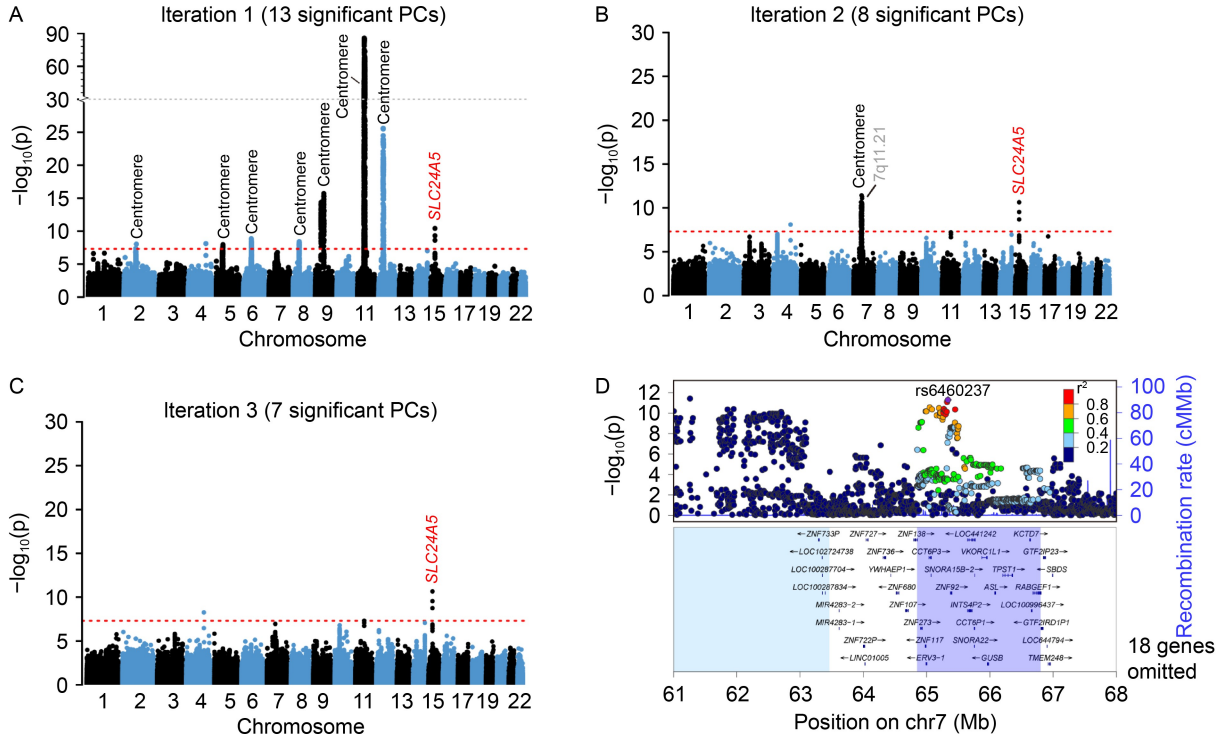

Figure 6. Manhattan and LocusZoom plots of unusual genomic regions identified in South Asians. We iteratively repeated analyses after excluding unusual genomic regions identified in previous iterations. (A−C) Manhattan plots showing P_{Sidak} in iterations 1−3. The red dash line indicates $P = 5\times10^{-8}$, and the gray dash line denotes break at $P = 10^{-30}$. Signals are labeled in colored texts based on potential causes: black, centromere; red, selection; gray, unknown. We labeled signals of unknown causes with the gene nearest to the leading SNP. (D) LocusZoom plot of the 7q11.21 locus with unknown cause identified in iteration 2. The lead SNP is indicated by the purple diamond, with surrounding SNPs colored by LD to the lead SNP. The identified unusual region is shaded in purple, while a neighboring region around the centromere is shaded in blue, at the bottom panel.

Similarly, in the ADMIXTURE analyses, we observed more distinct admixture fractions between YRI and ESN when K = 6 or 7 (Supplementary Fig. S8).

DISCUSSION

In this study, we developed a statistical framework based on PCA to identify genomic regions with unusual impact on the inference of population structure. We used this framework to systematically analyze high-coverage WGS data of 20 populations from four major continental groups. Taking together, we identified 27 unusual genomic regions (Table 1), among which, 17 were unique to one continental group and 10 were not

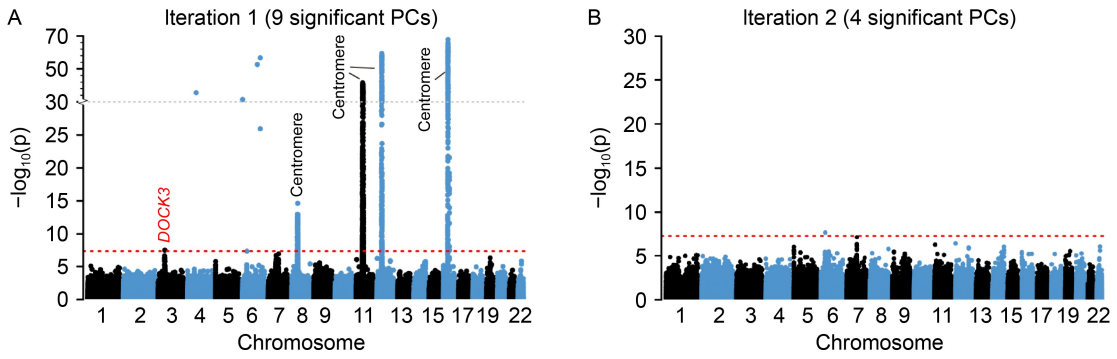

Figure 7. Manhattan plots of unusual genomic regions identified in Africans. We iteratively repeated analyses after excluding unusual genomic regions identified in previous iterations. (A−B) Manhattan plots showing P_{Sidak} in iterations 1−2. The red dash line indicates $P = 5\times10^{-8}$, and the gray dash line denotes break at $P = 10^{-30}$. Signals are labeled in colored texts based on potential causes: black, centromere; red, selection.

identified in Europeans (Supplementary Fig. S9), highlighting the strength of our study in analyzing diverse populations. Interestingly, the number of unusual regions detected in each continental group was 17 in Europeans, 13 in East Asians, 10 in South Asians, and 5 in Africans, showing a decreasing trend as the within-continent genetic diversity increases. This observation indicates that the impacts of these unusual regions are more evident when inferring population structure in samples with lower degrees of genetic diversity, as shown in our PCA and ADMIXTURE analyses with and without the unusual regions when compared across continental groups. Therefore, we hypothesize that using genotype data from relatively homogeneous subpopulations with increasing sample size will be helpful to detect more unusual regions [29]. Furthermore, a more dramatic example can be found in Supplementary Fig. S5 of Wang *et al*. [13], which showed that imputed SNPs around the centromere region of chromosome 11 could distort the PCA of European samples, resulting in two misleading clusters along PC2. To avoid the risk of misinterpretation, we recommend excluding all the unusual regions from standard analysis of population structure.

The 27 unusual regions included 13 around centromeres, 2 chromosomal inversions, 8 with genes reported to be under natural selection, and 4 with unknown causes. Unsurprisingly, the largest and most significant regions are around centromeres, where the LD was high due to the low recombination rate around centromeres [30]. In particular, centromere regions on chromosome 8, 11, and 12 were consistently identified in all four continental groups. The two chromosomal inversions, 8p23.1 and 17q21.31, are well-known large polymorphic inversions, which have been reported to associate with numerous diseases and have undergone natural selection in Europeans [27,31]. Regardless of the functional impacts, chromosomal inversions can disrupt recombination and induce extensive LD [32], and thus exert strong impacts on the inference of population structure. Similarly, strong selection sweeps can induce strong LD. We have detected 8 well-known selection loci (Supplementary Table S1), including *HLA* and *IGH* for immune response [33,34], *LCT* for lactase persistence [24], *OCA2* and *SLC24A5* for pigmentation [28,35], *DOCK3* for height [36], *GPHN* for neural development [37], and *FADS1/2* for the biosynthesis of long-chain polyunsaturated fatty acids (LC-PUFA) [38]. Notably, selection at *FADS1/2* was previously detected by comparing Japanese population with high LC-PUFA intake from their marine diet and other East Asian populations with agricultural diets [38,39], and we also detected this signal exclusively in the analysis of East Asians.

Finally, we identified 4 unusual regions with unclear causes, which might represent novel selection signals. The 9p23 and 17q24.1 were identified in Europeans (Fig. 2C−D). The 9p23 locus is intergenic with the nearest protein coding gene being *TYRP1*. While *TYRP1* is a pigmentation-associated gene under positive selection [40,41], our identified region is about 1 Mb upstream of *TYRP1* and thus their connection might be weak. The 17q24.1 locus spans only about 40 kilobases, overlapping with the *LRRC37A3* gene, which belongs to the *LRRC37* gene family, a core duplicon mapped to 18 distinct loci on the q arm of chromosome 17 [42]. The *LRRC37* family has expanded in the primate lineage since the divergence of the human lineage, and shows significant evidence of purifying selection, although the function of *LRRC37* genes remains unclear [42]. The 1q25.1 locus identified in East Asians spanned about 1.57 Mb (Fig. 4D), with the lead SNP rs61826814 residing in *KLHL20*, a key gene involved in the control of interferon responses [43]. Furthermore, the *CENPL* gene nearby is essential for proper kinetochore function and mitotic progression and has been implicated for

growth deficiency [44]. The 7q11.21 locus identified in South Asians spanned about 1.95 Mb (Fig. 6D). The lead SNP rs6460237 was close to *ERV3-1*, which might be involved in autoimmunity and cancer with an immunosuppressive role [45]. Another interesting gene within the 7q11.21 locus is *VKORC1L1*, which together with its paralog *VKORC1* encodes proteins with vitamin K oxidoreductase (VKOR) activity and plays important roles in hemostasis and coagulation [46]. While no selection signals have been reported for *VKORC1L1*, *VKORC1* at 16p11.2 have been reported to be under selection [4,47].

In conclusion, we have systematically identified and characterized 27 genomic regions with strong impact on population structure inference in diverse continental groups, providing a valuable resource for future population-genetic studies. While we do not focus on detecting selection signals, our method has the potential to help identify novel selection signals when applied to larger population-genetic datasets.

MATERIALS AND METHODS

Data and quality controls

We downloaded genotype data of the 1000 Genomes Project (1KGP), which were called from 30× high-coverage WGS data and aligned to the human genome assembly GRCh38 [23]. We extracted genotypes across 99,857,333 biallelic autosomal SNPs for 2,504 unrelated individuals from 26 populations worldwide [2]. We excluded populations with recent admixture history, including ACB (African Caribbeans in Barbados), ASW (African Americans in southwestern United States), and all four American populations. The remaining 20 populations were assigned to 4 continental groups (Europe, East Asia, South Asia, and Africa), each including 5 populations. As shown in Supplementary Fig. S1, we performed a series of QC procedures for each continental group using PLINK (v 2.0) [48]. We excluded SNPs with call rate < 0.95, minor allele frequency (MAF) < 0.05 or Hardy-Weinberg equilibrium (HWE) $P < 10^{-8}$, and then thinned the callset so that the closest pair of SNPs was at least 2 kb apart from each other. We performed PCA on each continental group and removed samples any of the top 10 PCs of which is more than 6 standard deviations (SD) away from the corresponding mean. Such outliers were excluded and this procedure was repeated until no more outliers could be identified. After QC, each continental group consisted of about one million biallelic autosomal SNPs among 463 to 502 unrelated individuals (Supplementary Fig. S1). The final sample sizes and excluded individuals were listed in Supplementary Table S2.

Identification of unusual regions

We considered "unusual regions" as genomic regions with outlier contribution to population structure inferred by PCA. First, we applied PCA on post-QC genotypes of each continental group separately and selected significant PCs ($P < 0.001$, Tracy-Widom test) [12]. We performed Tracy-Widom test using the ***t****racy.widom* function in the R package *LEA* [49]. Next, we treated each significant PC as a quantitative trait and performed genome-wide association tests based on linear regression without covariates to identify SNPs with unusual contribution to that PC. The background inflation due to population structure was adjusted by genomic control [50]. Because PCs are orthogonal to each other by construction, we combined association signals for each SNP across top PCs using the Sidak's method, where $P_{\text{Sidak}} = 1-(1-P_{\text{min}})^K$ and P_{min} is the minimum P value among top K significant PCs [51]. SNPs with $P_{\text{Sidak}} < 5\times10^{-8}$ were considered as genome-wide significant. Finally, we defined "unusual regions" by LD clumping based on P_{Sidak} (PLINK arguments:−clump −clump-p1 5e−8 −clump-p2 1e−3 −clump-kb 10000 −clump-r2 0.3 −clump-allow-overlap) [48]. Overlapping regions were merged as one. We excluded regions with fewer than 5 SNPs to reduce noise.

We then removed SNPs in the identified "unusual regions" from the genotype data and repeated the above procedure to identify additional unusual regions. Note that in subsequent iterations, PCs were computed from genotype data with unusual regions removed, but when performing GWAS of PCs, the original genotype data were used. We iterated the process until no new regions can be identified. We compiled the "unusual regions" identified in all iterations into a final list for each continental group.

Search for reported selection signals

For each gene located in the identified unusual regions, we searched in PubMed with the term "(natural selection [All Fields]) OR (positive selection [All Fields]) AND human [All Fields] AND *GENE-NAME* [All Fields]". We manually examined each publication to determine if an identified region harbors reported candidate selection signals.

Inference of population structure

We compared population structure inferences by PCA and ADMIXTURE [13,15] before and after removing the identified unusual regions. PCA was performed using the LASER software [13]. We quantified how well the top K PCs could separate populations by the proportion of between-population variance defined as:

$$\Psi = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\vec{x}_{ij}-\vec{\mu}_i)(\vec{x}_{ij}-\vec{\mu}_i)^T}{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\vec{x}_{ij}-\vec{v})(\vec{x}_{ij}-\vec{v})^T},$$

where the $\vec{x}_{ij}$ is a K-dimensional vector whose elements are top K PCs of individual j from population i, and $\vec{\mu}_i$ and $\vec{v}$ are the PC coordinate centroid of population i and the overall centroid, respectively [8]. m is the number of populations and n_i is the number of individuals in population i. Ψ ranges from 0 to 1, with larger values indicating better capture of population structure.

We performed unsupervised ADMIXTURE [15] analyses with the number of ancestral components K ranging from 4 to 7 for each continental group. We used CLUMPAK [52] to identify the optimal cluster alignment across different values of K.

SUPPLEMENTARY MATERIALS

The supplementary materials can be found online with this article at https://doi.org/10.15302/J-QB-022-0303.

ACKNOWLEDGEMENTS

This study was funded by the National Natural Science Foundation of China (No. 81973148).

COMPLIANCE WITH ETHICS GUIDELINES

The authors Qiuxuan Liu, Degang Wu, and Chaolong Wang declare that they have no conflict of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. and Feldman, M. W. (2002) Genetic structure of human populations. Science, 298, 2381−2385
2. The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. Nature, 526, 68−74
3. Wang, C., Zöllner, S. and Rosenberg, N. A. (2012) A quantitative comparison of the similarity between genes and geography in worldwide human populations. PLoS Genet., 8, e1002886
4. Wu, D., Dou, J., Chai, X., Bellis, C., Wilm, A., Shih, C. C., Soon, W. W. J., Bertin, N., Lin, C. B., Khor, C. C., *et al*. (2019) Large-scale whole-genome sequencing of three diverse Asian populations in Singapore. Cell, 179, 736−749.e15
5. Marchini, J., Cardon, L. R., Phillips, M. S. and Donnelly, P. (2004) The effects of human population structure on large genetic association studies. Nat. Genet., 36, 512−517
6. Price, A. L., Zaitlen, N. A., Reich, D. and Patterson, N. (2010) New approaches to population stratification in genome-wide association studies. Nat. Rev. Genet., 11, 459−463
7. Chen, H., Wang, C., Conomos, M. P., Stilp, A. M., Li, Z., Sofer, T., Szpiro, A. A., Chen, W., Brehm, J. M., Celedón, J. C., *et al*. (2016) Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet., 98, 653−666
8. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H. M., Stambolian, D., Chew, E. Y., Branham, K. E., Heckenlively, J., Fulton, R., Wilson, R. K., *et al*. (2014) Ancestry estimation and control of population stratification for sequence-based association studies. Nat. Genet., 46, 409−415
9. Wojcik, G. L., Graff, M., Nishimura, K. K., Tao, R., Haessler, J., Gignoux, C. R., Highland, H. M., Patel, Y. M., Sorokin, E. P., Avery, C. L., *et al*. (2019) Genetic analyses of diverse populations improves discovery for complex traits. Nature, 570, 514−518
10. Chen, J., Spracklen, C. N., Marenne, G., Varshney, A., Corbin, L. J., Luan, J., Willems, S. M., Wu, Y., Zhang, X., Horikoshi, M., *et al*. (2021) The trans-ancestral genomic architecture of glycemic traits. Nat. Genet., 53, 840−860
11. Zhu, C. and Yu, J. (2009) Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics, 182, 875−888
12. Patterson, N., Price, A. L. and Reich, D. (2006) Population structure and eigenanalysis. PLoS Genet., 2, e190
13. Wang, C., Zhan, X., Liang, L., Abecasis, G. R. and Lin, X. (2015) Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. Am. J. Hum. Genet., 96, 926−937
14. Falush, D., Stephens, M. and Pritchard, J. K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics, 164, 1567−1587
15. Alexander, D. H., Novembre, J. and Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res., 19, 1655−1664
16. Yang, W. Y., Novembre, J., Eskin, E. and Halperin, E. (2012) A model-based approach for analysis of spatial structure in genetic data. Nat. Genet., 44, 725−731
17. Jakobsson, M., Scholz, S. W., Scheet, P., Gibbs, J. R., VanLiere, J. M., Fung, H. C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., *et al*. (2008) Genotype, haplotype and copy-

number variation in worldwide human populations. Nature, 451, 998−1003
18. Tang, H., Choudhry, S., Mei, R., Morgan, M., Rodriguez-Cintron, W., Burchard, E. G. and Risch, N. J. (2007) Recent genetic selection in the ancestral admixture of Puerto Ricans. Am. J. Hum. Genet., 81, 626−633
19. Galinsky, K. J., Bhatia, G., Loh, P. R., Georgiev, S., Mukherjee, S., Patterson, N. J. and Price, A. L. (2016) Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. Am. J. Hum. Genet., 98, 456−472
20. Privé, F., Luu, K., Vilhjálmsson, B. J. and Blum, M. G. B. (2020) Performing highly efficient genome scans for local adaptation with R package pcadapt version 4. Mol. Biol. Evol., 37, 2153−2154
21. Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., *et al*. (2008) Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet., 83, 132–135., author reply 135–139
22. Lachance, J. and Tishkoff, S. A. (2013) SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. BioEssays, 35, 780−786
23. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., *et al*. (2021) High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv, 430068
24. Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S. F., Drake, J. A., Rhodes, M., Reich, D. E. and Hirschhorn, J. N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet., 74, 1111−1120
25. Broman, K. W., Matsumoto, N., Giglio, S., Martin, C. L., Roseberry, J. A., Zuffardi, O., Ledbetter, D. H. and Weber, J. L. (2003) Common long human inversion polymorphism on chromosome 8p. In: Statistics and Science: a Festschrift for Terry Speed. GOLDSTEIN, D. R., pp. 237–246. Beachwood, OH: Institute of Mathematical Statistics
26. Herva, R. and de la Chapelle, A. (1976) A large pericentric inversion of human chromosome 8. Am. J. Hum. Genet., 28, 208−212
27. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V. G., *et al*. (2005) A common inversion under selection in Europeans. Nat. Genet., 37, 129−137
28. Lamason, R. L., Mohideen, M. A., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., Humbert, J. E., *et al*. (2005) SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. Science, 310, 1782−1786
29. Abdellaoui, A., Hottenga, J.-J., de Knijff, P., Nivard, M. G., Xiao, X., Scheet, P., Brooks, A., Ehli, E. A., Hu, Y., Davies, G. E., *et al*. (2013) Population structure, migration, and diversifying selection in the Netherlands. Eur. J. Hum. Genet., 21, 1277−1285
30. Smith, A. V., Thomas, D. J., Munro, H. M. and Abecasis, G. R. (2005) Sequence features in regions of weak and strong linkage disequilibrium. Genome Res., 15, 1519−1534
31. Salm, M. P., Horswell, S. D., Hutchison, C. E., Speedy, H. E., Yang, X., Liang, L., Schadt, E. E., Cookson, W. O., Wierzbicki, A. S., Naoumova, R. P., *et al*. (2012) The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res., 22, 1144−1153
32. Stevison, L. S., Hoehn, K. B. and Noor, M. A. (2011) Effects of inversions on within- and between-species recombination and divergence. Genome Biol. Evol., 3, 830−841
33. Prugnolle, F., Manica, A., Charpentier, M., Guégan, J. F., Guernier, V. and Balloux, F. (2005) Pathogen-driven selection and worldwide HLA class I diversity. Curr. Biol., 15, 1022−1027
34. Watson, C. T. and Breden, F. (2012) The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease. Genes Immun., 13, 363−373
35. Yang, Z., Zhong, H., Chen, J., Zhang, X., Zhang, H., Luo, X., Xu, S., Chen, H., Lu, D., Han, Y., *et al*. (2016) A genetic mechanism for convergent skin lightening during recent human evolution. Mol. Biol. Evol., 33, 1177−1187
36. Jarvis, J. P., Scheinfeldt, L. B., Soi, S., Lambert, C., Omberg, L., Ferwerda, B., Froment, A., Bodo, J. M., Beggs, W., Hoffman, G., *et al*. (2012) Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. PLoS Genet., 8, e1002641
37. Climer, S., Templeton, A. R. and Zhang, W. (2015) Human gephyrin is encompassed within giant functional noncoding yin-yang sequences. Nat. Commun., 6, 6534
38. Ameur, A., Enroth, S., Johansson, A., Zaboli, G., Igl, W., Johansson, A. C. V., Rivas, M. A., Daly, M. J., Schmitz, G., Hicks, A. A., *et al*. (2012) Genetic adaptation of fatty-acid metabolism: a human-specific haplotype increasing the biosynthesis of long-chain omega-3 and omega-6 fatty acids. Am. J. Hum. Genet., 90, 809−820
39. Mathieson, S. and Mathieson, I. (2018) FADS1 and the timing of human adaptation to agriculture. Mol. Biol. Evol., 35, 2957−2970
40. Hudjashov, G., Villems, R. and Kivisild, T. (2013) Global patterns of diversity and selection in human tyrosinase gene. PLoS One, 8, e74307
41. Lao, O., de Gruijter, J. M., van Duijn, K., Navarro, A. and Kayser, M. (2007) Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann. Hum. Genet., 71, 354−369
42. Giannuzzi, G., Siswara, P., Malig, M., Marques-Bonet, T., Mullikin, J. C., Ventura, M. and Eichler, E. E., and the NISC Comparative Sequencing Program. (2013) Evolutionary dynamism of the primate LRRC37 gene family. Genome Res., 23, 46−59
43. Lee, Y. R., Yuan, W. C., Ho, H. C., Chen, C. H., Shih, H. M. and Chen, R. H. (2010) The Cullin 3 substrate adaptor KLHL20 mediates DAPK ubiquitination to control interferon responses. EMBO J., 29, 1748−1761

44. Burkardt, D. D., Rosenfeld, J. A., Helgeson, M. L., Angle, B., Banks, V., Smith, W. E., Gripp, K. W., Moline, J., Moran, R. T., Niyazov, D. M., *et al*. (2011) Distinctive phenotype in 9 patients with deletion of chromosome 1q24-q25. Am. J. Med. Genet. A., 155, 1336−1351
45. Bustamante Rivera, Y. Y., Brütting, C., Schmidt, C., Volkmer, I. and Staege, M. S. (2018) Endogenous retrovirus 3—history, physiology, and pathology. Front. Microbiol., 8, 2691
46. Lacombe, J., Rishavy, M. A., Berkner, K. L. and Ferron, M. (2018) VKOR paralog VKORC1L1 supports vitamin K-dependent protein carboxylation *in vivo*. JCI Insight, 3, e96501
47. Szpak, M., Mezzavilla, M., Ayub, Q., Chen, Y., Xue, Y. and Tyler-Smith, C. (2018) FineMAV: prioritizing candidate genetic variants driving local adaptations in human populations. Genome Biol., 19, 5
48. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M. and Lee, J. J. (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience, 4, 7
49. Frichot, E. and François, O. (2015) LEA: An R package for landscape and ecological association studies. Methods Ecol. Evol., 6, 925−929
50. Devlin, B. and Roeder, K. (1999) Genomic control for association studies. Biometrics, 55, 997−1004
51. Šidák, Z. (1967) Rectangular confidence regions for the means of multivariate normal distribution. J. Am. Stat. Assoc., 62, 626−633
52. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. and Mayrose, I. (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour., 15, 1179−1191