Continental-scale genomic analysis suggests shared post-admixture adaptation in the Americas

Linda Ongaro[1,*], Mayukh Mondal[1], Rodrigo Flores[1], Davide Marnetto[1], Ludovica Molinaro[1], Marta E. Alarcón-Riquelme[2], Andrés Moreno-Estrada[3], Nedio Mabunda[4], Mario Ventura[5], Kristiina Tambets[1], Garrett Hellenthal[6], Cristian Capelli[7,8], Toomas Kivisild[9], Mait Metspalu[1], Luca Pagani[1,10,#], Francesco Montinaro[1,5,#]

[1]*Estonian Biocentre, Institute of Genomics, Tartu, Riia 23b, 51010, Estonia.*

[2]*Department of Medical Genomics, GENYO. Centro Pfizer - Universidad de Granada - Junta de Andalucía de Genómica e Investigación Oncológica, Av de la Ilustración 114, Parque Tecnológico de la Salud (PTS), 18016, Granada, Spain.*

[3]*National Laboratory of Genomics for biodiversity (LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico.*

[4]*Instituto Nacional de Saúde, Distrito de Marracuene, Estrada Nacional N°1, Província de Maputo, Maputo, 1120, Mozambique.*

[5]*Department of Biology-Genetics, University of Bari, Bari, 70126, Italy.*

[6]*Department of Genetics, Evolution and Environment and UCL Genetics Institute, University College London, London WC1E 6BT, UK.*

[7]*Department of Zoology, University of Oxford, Oxford, UK.*

[8]*Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy.*

[9]*Department of Human Genetics, KU Leuven, Herestraat 49 - box 602, B-3000, Leuven, Belgium.*

[10]*Department of Biology, University of Padua, Padua, Italy.*

[#]*These authors contributed equally as Senior authors.*

[*]*To whom correspondence may be addressed. Linda Ongaro, University of Tartu. Email: linda.ongaro@ut.ee.*

Abstract

American populations are one of the most interesting examples of recently admixed groups, where ancestral components from three major continental human groups (Africans, Eurasians and Native Americans) have admixed within the last 15 generations. Recently, several genetic surveys focusing on thousands of individuals shed light on the geography, chronology and relevance of these events. However, even though gene flow could drive adaptive evolution, it is unclear whether and how natural selection acted on the resulting genetic variation in the Americas.

In this study, we analyzed the patterns of local ancestry of genomic fragments in genome-wide data for ~6,000 admixed individuals from ten American countries. In doing so, we identified regions

characterized by a Divergent Ancestry Profile (DAP), in which a significant over or under ancestral representation is evident.

Our results highlighted a series of genomic regions with Divergent Ancestry Profiles (DAP) associated with immune system response and relevant medical traits, with the longest DAP region encompassing the Human Leukocyte Antigen locus. Furthermore, we found that DAP regions are enriched in genes linked to cancer-related traits and autoimmune diseases. Then, analyzing the biological impact of these regions, we showed that natural selection could have acted preferentially towards variants located in coding and non-coding transcripts and characterized by a high deleteriousness score.

Taken together, our analyses suggest that shared patterns of post admixture adaptation occurred at a continental scale in the Americas, affecting more often functional and impactful genomic variants.

Introduction

The genomic variation of a substantial proportion of the individuals living in the Americas is the result of admixture involving Native American, European and African populations, together with minor recent contributions from Asia, as the results of deportation and mass migrations followed by admixture episodes (1–4).

Although many studies uncovered the complexity of the admixture dynamics in the continents (1,3–5), addressing the role of adaptive introgression in shaping the modern-day variation of American populations has been particularly challenging. In fact, the high variance in individual continental ancestries, due to very recent and still ongoing admixture, makes it hard to apply classical natural selection tools based on the distribution of genetic variation along the genome (6–9). Following the advent of high-throughput sequencing and the development of statistical tools aimed at inferring the ancestry of specific genomic regions, a commonly used approach to tackle this question focuses on loci showing ancestral proportions across the entire population that significantly diverge from whole-genome estimates (10). Regions enriched or depleted for a given ancestry are usually interpreted as targets of natural selection, driven by post admixture adaptive pressures. Despite the large number of studies harnessing these methods, the results have not been consistent and were often not replicated.

For example, the analysis of ~2000 African Americans found putatively enriched loci (deviating more than three standard deviations) for European and African ancestries (11). However, a subsequent study analyzing ~29,000 individuals and applying a genome-wide significance threshold did not find any statistically significant diverging loci, highlighting the possibility of false-positive detection when less conservative thresholds are applied (12).

Furthermore, most of the selection scans performed so far focused either on single populations or multiple groups of small sample sizes, increasing the chance of collecting false positives.

Nevertheless, some investigations of local ancestry tracts distributions in individuals from Peru, Puerto Rico, Mexico, and Colombia are concordant in suggesting rapid natural selection in genomic loci associated with immune response, such as the Major Histocompatibility Complex (MHC) (13–18).

In this study, we analyze local ancestry tracts distribution for 5,828 American individuals from 19 populations to elucidate the role of post-admixture selection in shaping the genetic variation of the Americas.

In doing so, we focused only on signals shared across multiple populations, reducing the inference of false positives and at the same time highlighting shared or convergent episodes of selection as proposed in Yelmen et al. 2019 (19). Moreover, given the Columbian Exchange phenomenon, the asymmetric resistance to pathogens such as Measles and Smallpox in Europeans and not in Americans, and Syphilis in Americans but not in Europeans, we specifically looked for signals in genes associated with the biology of these diseases (20–22).

Our results highlighted a series of genomic regions with Divergent Ancestry Profiles (DAP) associated with relevant medical traits such as the MHC, and others with recurring association to cholesterol and triglyceride levels, systemic lupus erythematosus and blood protein levels. Furthermore, we found that the SNPs belonging to DAP windows are enriched in genes linked to cancer-related traits and autoimmune diseases. Lastly, the analysis of the functional impact and annotation of the DAP windows revealed a more prominent role of natural selection on variants located in coding and non-coding transcripts and characterized by a stronger annotated impact.

Results

Evaluation of signals identified in multiple populations

We aimed to identify genomic regions showing significant deviations in local ancestry assignment from the average ancestry proportion in a given population. We applied the Local Ancestry RFMix software (23) on data from the Americas (Table 1) as presented in Ongaro et al. (2019) (1), in which the ancestry composition of the admixed individuals was deconvoluted using four putative source groups (Africa: 190 individuals; Europe: 289 individuals; Americas: 67 individuals; Asia: 213 individuals; Supplementary Table 1A). We have here excluded the latter, given the low proportion of Asian ancestry detected in our previous analysis (1).

We evaluated the local ancestry output using two differently assembled datasets, as described in the Methods section. From this point on, we performed all the analyses for each ancestry separately. Briefly, to take into account possible errors when populations or individuals have extreme ancestry proportions, for the first dataset (20Pop dataset), we focused only on populations with more than 20% of a given ancestry; while for the second one (1090Ind dataset), we excluded individuals characterized by extreme proportions of any given ancestry (< 10% and > 90%).

We estimated the ancestry specific Z-score for each SNP (data available at https://doi.org/10.5281/zenodo.4446628), as explained in the Methods section.

We aggregated the SNP-based output in 100kb non-overlapping windows, retaining only windows containing more than 5 SNPs, and annotated as DAP (Divergent Ancestry Profile) those with at least one SNP with a significant Z score (>|3|). We selected only DAP signals replicated in at least two populations to reduce the number of false positives and focus on shared or convergent signatures. Consecutive 100kb DAP windows were grouped and are hereafter referred to as *DAP regions*.

We further distinguished between two types of DAP: Inter-Samples DAP (IS-DAP), when they are shared between populations in our dataset that contains individuals from the same countries (for example, PEL and Peru, both Peruvian populations); and Inter-Populations DAP (IP-DAP) when they are shared across populations more distantly related, for instance, European Americans and Caribbean. The populations and the number of haplotypes retained in the two datasets are reported in Supplementary Table 1B.

African DAP

We identified only a single genomic window with an African divergent ancestry profile present in more than one population (Figure 1 and Table 2). This IP-DAP window (chr15: 65,613,654-65,695,283) contains 10 SNPs within two immunoglobulin genes: *IGDCC3* and *IGDCC4*. These SNPs show significant underrepresentation of African ancestry in ACB (Barbados Island) and Dominicans when analyzed in the 1090Ind dataset.

European DAPs

We identified 7 DAP genomic regions with underrepresented European ancestry in the 20Pop dataset, of which 2 are replicated in the 1090Ind dataset; in addition, we found one overrepresented region in the 20Pop dataset. More details about the content of the DAP genomic regions are reported in Table 2. Strikingly, the major signal was found for IP-DAP region in chromosome 6 extending for a total of 7.8 Mb (chr6:25,312,755-33,098,966) with underrepresented European ancestry in two populations, Caribbean and European Americans (EuroAme). This region contains 373 genes (Supplementary Table 2) including the large locus of the Human Leukocyte Antigen (HLA). Nineteen of the 1079 SNPs typed in our dataset are predicted to have a high functional impact, according to the Combined Annotation Dependent Depletion (CADD) annotation (PHRED-scaled C-score > 20, Supplementary Table 3, PHRED hereafter). Interestingly, six of the 100kb windows included in this region carry high proportions of SNPs with large PHRED values (>= 20; 26% of all SNPs in chr6:31,602,967-31,688,217, 10% in both chr6:26,413,088-26,484,376 and chr6:29,407,970-29,483,911, 8.3% in both chr6:28,300,336-28,391,465 and chr6:32,019,769-32,083,175 and 7.7% in chr6:30,823,630-30,897,022), corresponding to the top 5% of the distribution estimated on all the genomic windows (26% corresponds to 0.0006, 10% to 0.025, 8.3% to 0.039 and 7.7% to 0.039). Overall, the variants in this DAP region belong to 35 different genes, of which 32 are related to HLA.

In this context, a variant (*rs3130618, PHRED=32,* Supplementary Table 3, Caribbean Z=-5.5, EuroAme Z=-4.8) belonging to *GPANK1* gene is associated with MMR (measles, mumps and rubella) vaccination-related febrile seizure (24) and Membranous Glomerulonephritis (25). Inside this region,

two SNPs, *rs140973961* and *rs78331658*, not present in our genome-wide dataset, have been associated with measles and immune response to the measles vaccine, respectively, in the GWAS catalog (26,27). To the best of our knowledge, these results might be related to the phenomenon of the "Columbian Exchange".

Notably, we observed a neighbouring IP-DAP region (chr6:34,102,061-35,495,811), distant ~1Mb from the one reported above, with underrepresented American ancestry for 20Pop and 1090Ind in PEL and Mexican, as shown in Table 2. We registered a similar region, 1.1 Mb (chr6:34,332,179-35,495,811), with an overrepresentation of African ancestry (20Pop dataset) in Caribbean (Min Z= 3.2, Max Z= 3.4) and with values close to the significance threshold in the African Americans (AfroAme, Min Z=2.4, Max Z=2.6).

The important role of HLA in shaping the immunological response in humans may raise the possibility that the new environmental dynamics have triggered adaptive forces that ultimately led to the selection of some African allelic variants.

We identified an additional IP-DAP region with underrepresented European ancestry in chromosome 9. It is identical in both datasets, extending for 157kb and containing 44 SNPs (chr9:38,615,175-38,771,831) located in five genes, shared among African Americans, (AfroAme 20 Pop dataset: Min Z=-3, Max Z=-3.6; 1090Ind: Min Z=-3.2, Max Z=-3.4), Mexicans (20Pop: Min Z=-3.2, Max Z=-3.3; 1090Ind: Min Z=-3, Max Z=-3.3) and European Americans (20Pop: Min Z=-5.2, Max Z=-6.5). Notably, *rs7039377* (chr9:38,675,465; AfroAme: 20Pop Z=-3, 1090ind Z=-2.8; Mexican: 20Pop Z=-2.99, 1090Ind Z=-3; EuroAme: 20Pop Z=-5.5) has been associated with obesity-related traits in Hispanic children (28).

Moreover, we observed, in both the analyzed datasets (although characterized by different size), a consistent signal of underrepresented European ancestry in chromosome 15 (Figure 1 and Table 2), shared across seven populations: African Americans, Argentina, Caribbean, European Americans, Mexican, MXL and PUR. Among the others, a SNP (rs2278458, chr15:22,999,857; AfroAme Z=-8, Argentina Z=-5, Caribbean Z=-7.4, EuroAme Z=-31, Mexican Z=-4.8, PUR Z=-4.6) in *CYFIP1* gene

is associated with Triglyceride levels and the response to diuretics in individuals of European and African ancestry (29).

The remaining four IP-DAP regions with significantly lower European ancestry in Caribbean and European American individuals were observed only in the 20Pop dataset and are located in chromosomes: 8, 13 (underrepresented also in African Americans), 14 and 21 (Supplementary Text).

Additionally, we identified an IP-DAP region in European ancestry of the 20Pop dataset overrepresented in Dominican and Puerto Ricans (Puerto, Table 2). This region is in chromosome 8 and contains the *rs4871180* (chr8:122,259,074; Dominican Z=3.1 Puerto Z=3) that could be associated with Diverticular disease in British individuals (30). The frequency of the risk allele (T) of this variant is very high in the analyzed Native American source populations, ranging from 0.87 to 0.9, while it has frequencies of 0.33 and 0.23 in Dominican and Puerto, respectively**.**

American DAPs

When the American ancestry profiles were evaluated besides the IP-DAP region of chromosome 6 discussed above, we identified one depleted and three enriched genomic regions in the 1090Ind dataset. Three different IS-DAP regions with observed overrepresentation in Caribbean and Puerto Rico (PUR) were found on chromosome 11 (Figure 1, Supplementary Text and Table 2). They are defined as IS-DAPs because some of the samples included in the Caribbean population are from Puerto Rico (31). The distance between the first two regions is approximately 2Mb, and could in fact be part of the same selection event. One of these two regions (chr11:46,210,259-46,297,631; 87kb) host a *SNP* (the *rs10437653;* Caribbean Z=2.97, PUR Z=2.88) that is associated with birth weight in Europeans (32).

The last region maps on chromosome 4 (chr4:24,512,590-25,679,209, Table 2) and is underrepresented for American ancestry in two Peruvian populations, PEL and Peru. This is also an IS-DAP region. Interestingly, we identified two SNPs within this region that are associated with diseases/traits: *rs12500612* (chr4:24,740,958; PEL Z=-3.1, Peru Z=-3.1) with Major depressive disorder in Europeans (33); and *rs1395221* (chr4:24,626,903; PEL Z=-3.1 Peru Z=-3.1), with Apolipoprotein A1 and HDL cholesterol levels in European individuals (34). The mean frequency of the risk allele (G) of *rs1395221*

in the analyzed Native American source populations is 0.90 and 0.60 in the European ones, whereas PEL and Peru have a frequency of 0.77 and 0.78, respectively.

Detecting divergent ancestry profile SNPs in simulated data

We simulated two datasets with stdpopsim (35) to assess the robustness of our results, as described in the Methods section. For the first set, *10SIM*, we performed 10 independent runs that simulated two admixed American populations (ADM1 and ADM2) to assess how many SNPs with Z>|3| are replicated in the two simulated populations. For the second set, *100SIM*, we conducted 100 independent runs to simulate one admixed population to calculate the corresponding false-positive rates (Supplementary Figure 1). For the *10SIM* simulated dataset, we did not find any SNP with Z>|3| in more than one population for any of the 10 independent runs (simulation results are reported in Supplementary Table 4A-B). We obtained single populations signals in some of the admixed simulated populations; however, the amount of significant SNPs in the simulated data is lower than in the real populations, as shown in Supplementary Table 4C. These results suggest that the DAP regions that we detected could not be the result of chance. However, they could be the result of natural selection.

The functional impact of DAP regions in the Americas

We explored the functional impact measured as SNP deleteriousness (PHRED-scaled C-score) and functional annotation (Annotype), both derived from the CADD annotation, focusing on the multiple population signals described above.

When we compared the distribution of all the PHRED values belonging to windows showing divergent ancestry profiles (DAP) with the non-divergent ones (non-DAP), European ancestry values were significantly higher in SNPs from DAP windows than those from non-DAP (Wilcoxon test, p-value=2.8E-29). The same result was observed for the American ancestry (Wilcoxon test, p-value=0.00014) of the 20Pop dataset (Figure 2 and Supplementary Table 5A). No comparison within the 1090Ind dataset showed any significance (Supplementary Figure 2).

Then, in DAP and non-DAP windows, we analyzed the distribution of the proportion of the five types of functional annotation: Coding Transcript (variants in protein-coding exons), Untranslated Transcript

(variants in UTRs and introns), Non-Coding Transcript (mature miRNA and non-coding transcript exon variants), Regulatory Feature and Intergenic. In detail, for each 100kb window, we annotated the relative functional composition of all the SNPs and evaluated the proportion of each category (Figure 3, Supplementary Figure 3 and Supplementary Table 5B). Every SNP was assigned to at least one Annotype.

We observed that the proportion of "Coding transcript" (p-value=8.3E-11) and "Non-coding Transcript" (p-value=4.12E-16) is higher in DAP windows than non-DAP in the European ancestry (20Pop Dataset, Wilcoxon test, p-values reported in Supplementary Table 5B, Figure 3). In contrast, the proportion of "Untranslated Transcript" is higher in non-DAP SNPs than DAP (European ancestry 20Pop Dataset, Wilcoxon test, p-value=0.00014, Figure 3).

Moreover, when we looked at the Native American ancestry, the proportion of "Coding Transcript" was higher in DAP SNPs than in non-DAP in the 1090Ind dataset (Wilcoxon test, p-value=0.0025, Supplementary Figure 3).

Gene Set Enrichment Analysis

We performed gene set enrichment analysis (GSEA) for the genes encompassing the DAP SNPs to understand if specific health/disease pathways or phenotypes are significantly enriched or depleted in the continental ancestries of multiple populations. We explored five different libraries: Human 2019 Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Ontology (GO) 2018, GTEx Tissue Sample Gene Expression Profiles down and up and Genome-wide Association Studies (GWAS) Catalog 2019. Interestingly, we found signals exclusively related to underrepresented variants in the European ancestry in the 20Pop dataset from four libraries (GWAS, KEGG, GO, and GTEx up) and related to overrepresented variants in the Native American ancestry in the 1090Ind dataset from only one library (KEGG).

In detail, for the European ancestry results, we found 46 significantly overrepresented traits in GWAS Catalog, where the most significant one is related to the *Autism spectrum disorder or schizophrenia* (p-value=3.9E-217) term followed by *Blood protein levels* (p-value=4E-84) and *Ulcerative colitis* (p-value=7.7E-49). Six significant terms are associated with types of cancer: *Lung cancer* (p-value=1.9E-

33), *Squamous cell lung carcinoma* (p-value=2.7E-31), *Lung cancer in ever smokers* (p-value=1.3E-29), *Small cell lung carcinoma* (p-value=4.1E-08), *Prostate cancer* (p-value=0.00011) and *Cervical cancer* (p-value=0.00016). Four are connected with Hepatitis B, both related to the chronic infection and the response to the vaccine. Then, a group of significant traits is linked to autoimmune diseases such as *Sarcoidosis*, *Psoriasis*, *Lupus*, *Behcet's disease, Type I diabetes and Inflammatory Bowel disease*.

Eighteen traits are overrepresented when the KEGG library is explored, where the *Systemic lupus erythematosus* (p-value=2.6E-46), which is in common with the GWAS library list of terms, has the lowest adjusted p-value followed by *Alcoholism* (p-value=7.8E-27) and *Viral carcinogenesis* (p-value=6E-14). Other terms in common with the GWAS catalogue are *Inflammatory Bowel disease* (KEGG: p-value=0.00022, GWAS: p-value=2.7E-29) *and Type I diabetes* (KEGG: p-value=1.4E-11, GWAS: p-value=5.9E-19). Finally, significant GO enrichment was found for *MHC class II receptor activity (*p-value=2.29E-05) and GTex *lung expression in females of 60-69 years (*p-value=0.00036). The results are presented in Figure 4, where only the first ten most significant terms for each annotation library are shown, while in Supplementary Table 6A-D all the results are reported.

When we excluded the gene list related to the chromosome 6 DAP region with underrepresented European ancestry, the GSEA gave no significant results.

For the Native American ancestry, we identified one significant enriched term related to *Olfactory Transduction* (p-value=0.00047) according to the KEGG library (Supplementary Table 6E).

Discussion

In this study, we evaluated the presence of post-admixture regions with a divergent ancestry proportion on a broad dataset composed of 5,828 individuals from 19 recently admixed populations. As previous studies have focused on small sample sizes and a single population, our approach aimed to identify repeated signals at a Pan-American scale, allowing at the same time, the reduction of false-positive rate, and the discovery of selective forces acting at the scale of the entire continent. Furthermore, we

considered possible errors by the Local Ancestry algorithm harnessed here, performing the same analysis discarding individuals with extreme values of a given ancestry.

In doing so, despite the potential for African admixtures to enrich the genetic diversity of any recipient human group, among the five populations in the 20Pop dataset and the seven populations in the 1090Ind dataset having a relatively high contribution from Africa, we identified only a single underrepresented and no overrepresented region. Although this scenario is in contrast with surveys identifying DAP regions of African inheritance (17,18), our results might be explained by the high degree of the conservativeness of our approach. Moreover, the recently documented high genetic variation of the African populations contributing to the modern-day American populations, coupled with the relatively low African ancestry proportion in the analyzed samples, may have provided differential bases for the subsequent selective forces (1,36–38).

On the other hand, we identified 7 underrepresented regions which consistently showed deviation from the ancestral proportions of the European ancestry, including a ~8Mb region in chromosome 6 encompassing the Major Histocompatibility Complex II. A neighbouring region was identified as underrepresented also for the American ancestry. Although we failed to identify a replicated signal for African ancestry, it is worth noting that the same HLA has been found overrepresented for the African ancestry in Caribbean, and close to the significant threshold in African American, confirming previous results (13–18).

In this scenario, it might be possible that African variants of the HLA locus might have been the target of selective pressure, also given the new diseases brought by the European populations colonizing the area. Nevertheless, besides this important observation, we did not find consistent and replicated signals for alleles conferring protection for newly introduced pathogens, with the only possible exception of measles. It may be possible that the selection pressure acted independently among different populations, or alternatively, the American populations might already harbour in their genome adaptive variants.

Although a similar signal has been widely replicated, both using genotype arrays and whole-genome sequencing technologies, the high diversity of the region would require an additional confirmation through the analysis of sequence data that uses novel approaches, such as long reads sequencing.

Replicated signals for overrepresented regions that were found in the Americas are consistent with a scenario in which the 15,000 years of the peopling of the continents resulted in convergent adaptation. Our evaluation of the biological impact of post-admixture adaptation in the Americas revealed that irrespective of their direction, SNPs having high functional or phenotypic consequences tend to be "selected" more often than those with a mild effect. This is also confirmed by the fact that SNPs in UTRs and introns are not a preferential target of natural selection, in contrast to what has been observed for SNPs in coding regions (Figure 3 and Supplementary Figure 3).

Lastly, the gene set enrichment analysis performed here revealed that selection acted predominantly on regions associated, among the others, to the onset of autoimmune disease, various protein levels in blood and several different kinds of cancer. The fact that we did not obtain significant results when we removed from the GSEA the underrepresented IP-DAP region in chromosome 6 of European ancestry could be a sign of reduced polygenic enrichment in our data.

Overall, our research suggests that common selective pressure in the Americas had a non-negligible impact on shaping the genetic variation of the two continents, while the Columbian Exchange phenomenon seems to have played just a minor role. Our results also indicate that given the limitations of genetic scan for natural selection algorithms implemented so far, the analysis of multiple population datasets characterized by high sample size will be essential for both the identification and characterization of post admixture adaptation at a more local scale.

Material and Methods

Genome-wide data

The analyzed genome-wide dataset was recovered from Ongaro et al. 2019 (1). This dataset was filtered using PLINK ver. 1.9 (39) to include only SNPs and individuals with genotyping success rate > 97%, retaining a total of 251,548 autosomal markers. For this study, we used the reference genome version b37. We removed 22,295 SNPs belonging to centromeric regions or the first or last 5 Mb of the chromosomes based on the information retrieved from the UCSC browser (https://genome.ucsc.edu/cgi-bin/hgTables); after this step, we kept a total of 229,253 variants. In the current study, we analyzed

6,587 individuals, of which 5,828 belong to 19 admixed American populations, while the remaining 759 samples are from populations from Africa, Asia, Europe and America. The first set of individuals represents the so-called "targets", while the second set represents the "sources". The details are reported in Supplementary Table 1A-C.

Phasing

Germline phase was inferred using the Segmented Haplotype Estimation and Imputation tool (ShapeIT2) software (40), using the HapMap37 human genome build 37 recombination map. We used the options --thread 16 and the default value of --effective-size 15,000 as suggested in the manual.

Local ancestry

We estimated local ancestry assignation for genomic fragments of the target American individuals with RFMIX software(23), using the following reference source populations: Yoruba (YRI), Gambia (GWDwg) and Mozambique for Africa, Chinese Han (CHB) and Japanese (JPT) for Asia, Spanish (IBS), British (GBR) and Tuscany (TSI) for Europe and Tepehuano, Wichi and Karitiana for Native American ancestry (Supplementary Table 1A). We used "PopPhased", "-n 5", and "--forward-backward" options as recommended in the RFMix manual.

Starting from the RFMix output files, we built four PLINK file sets, one for each of the four source ancestries, masking in each one the SNPs assigned to any of the other three ancestries. In details, an allele was assigned as missing in the PLINK file of ancestry A when that allele was not assigned to ancestry A in the "Viterbi" output file or the probability of belonging to ancestry A (as reported in the "forward-backward" output) was less than a defined threshold (< 0.9). In this and the following analyses, we considered the samples as separated into the two phased haplotypes.

Given that Asian ancestry was consistently found at low frequency (range:0.2-9.4%), we decided not to include this ancestry in our search for DAP windows.

At this point, we adopted two different analysis approaches and assembled two datasets (Supplementary Table 1B):

1. 1090Ind dataset: this dataset was assembled considering the ancestral proportions of each individual. In detail, we used the PLINK command --missing to obtain a missing report for each individual; then, using the information from ".imiss" files, we kept those individuals with more than 0.1 (10%) and less than 0.9 (90%) of F_MISS (representing the missingness of the individual) for each ancestry.
2. 20Pop dataset: this was assembled, taking into consideration the ancestral proportions averaged by populations. In this case, we kept for the analyses only those populations with at least 20% of a specific ancestry. These proportions were recovered from the global ancestry analysis (SOURCEFIND (4)) reported in Ongaro et al. 2019 (1).

SNPs annotation

We annotated all our SNPs with CADD (Combined Annotation Dependent Depletion, (41)), a tool for scoring the deleteriousness of single-nucleotide variants in the human genome.

Detecting deviation from the expected ancestral proportions

We estimated the per-SNP average population assignation to detect deviation from the expected ancestral proportions in Local Ancestry Inference.

In doing so, starting from the ancestry-specific PLINK files described above, we initially calculated the population-specific missingness using --missing --family in PLINK 1.9 (39). We then estimated the proportion of Local ancestry assignment as 1-missing. The obtained proportions were finally standardized, calculating the Z-score for each SNP. Finally, we partitioned each chromosome into windows of 100,000 bp, obtaining 16,857 windows, and defined as Divergent Ancestry Profile (DAP) windows the ones that contained more than five SNPs of which at least one presented a significant Z-score. We decided to set as statistically significant values of $|Z|>3$ (data available at https://doi.org/10.5281/zenodo.4446628). To further reduce the identification of false positives, we focused exclusively on signals replicated in at least two populations. We distinguished between two types of DAP: Inter-Samples DAP (IS-DAP), when they are shared between populations in our dataset that contains individuals from the same countries (for example, PEL and Peru, both Peruvian

populations); and Inter-Populations DAP (IP-DAP) when they are shared across populations more distantly related, for instance, European Americans and Caribbean.

To visualize the results, we used the UpsetR package in R (Conway et al. 2017) for Figure 1A; whereas, the chromosome map (Figure 1B) was obtained using the chromoMap package in R.

Validation of our method through simulations

We performed two separate simulations with stdpopsim (35). Each simulation analysis starts with simulating three source populations (a proxy for African, European and Asian populations), each comprising 50 individuals, then proceed using these populations to create a three-way admixed population. We ran the analyses to simulate the chromosome 1 for 100 individuals per admixed population. The first simulation, named *10SIM*, used a slightly modified version of the American population admixture model (AmericanAdmixture_4B11), already implemented in stdpopsim (35). We modified the code to get two independent admixed populations as outputs, ADM1 and ADM2, each comprising 100 individuals. We performed 10 independent runs. The second set, named *100SIM*, used the American population admixture model (AmericanAdmixture_4B11) available in stdpopsim (35). We simulated one admixed population with 100 individuals, and we performed 100 independent runs. The subsequent steps were the same for both simulated datasets. The admixed population had an initial size of 30,000 and grew at a rate of 5% per generation, with 1/6 of the population of African ancestry, 1/3 European, and 1/2 Asian. We used the HapMapII_GRCh37 genetic map, and we set the parameter ploidy=2. To compare the simulation results with the real data, we filtered the simulated SNPs based on the site frequency spectrum of the Caribbean population from our dataset because of the high number of individuals (1,112) available for this population, obtaining 19,732 SNPs. Then, we applied the same analyses ran on the real data, starting from the Local Ancestry inference (LAI) using RFMix. From the LAI results, we created three masked datasets by removing the SNPs assigned to the other two ancestries. After running the --missing command with PLINK1.9, we detected those SNPs that diverged from the expected ancestral proportions in the Local Ancestry assignment. We used Z>|3| as the threshold for significance. From the results of *10SIM*, we looked at the signals shared between the two simulated populations ADM1 and ADM2. Whereas from the results of *100SIM*, we estimated the false-

positive rate by dividing the number of significant SNPs by the total number of SNPs examined (19,732) as reported in Supplementary Table 4B. Supplementary Figure 1 was obtained using the geom_jitter option of the ggplot2 package in R.

PHRED-scaled C-score analysis

We explored the differences in values of PHRED in the so-called DAP and non-DAP windows. We first extracted from the 100kb windows all the SNPs with a corresponding PHRED value retrieved from the CADD annotation. Then, we compared the distribution of all the PHRED values belonging to the DAP windows with the ones from the non-DAP using a paired Wilcoxon test with R. We used a corrected cutoff threshold of 0.01.

Annotype analysis

We investigated the differences in the distribution between DAP and non-DAP windows of the five types of functional annotations (Annotype; Coding Transcript, Intergenic, Non-Coding Transcript, Regulatory Feature and Transcript). In CADD annotation, the broad category of transcripts is divided into Coding Transcript, Non-Coding Transcript and Transcript. In details, "Coding Transcript" refers to several types of variants like missense, synonymous, stop-gained, stop-lost, initiator codon, stop-retained, frameshift, inframe insertion, inframe deletion, incomplete terminal codon and protein-altering. The "Non-Coding Transcript" category contains mature miRNA and non-coding transcript exon variants, while "Unspecific Transcript" refers to UTRs and introns.

In detail, we extracted all the SNPs from the windows of interest, recovering the corresponding Annotype, then we estimated the Annotype proportions for each window (DAP and non-DAP), and we compared the proportion distributions using a paired Wilcoxon test with R. We used a Bonferroni corrected cutoff threshold of 0.002.

Gene set enrichment analysis (GSEA)

GSEA was performed using the GSEAPY enrichr python module on gene lists obtained from the results of the detection of the DAP windows. We focused on the over- and under-represented SNPs from the

replicated DAP windows related to the three continental ancestries; then, we extracted the genes assigned to those SNPs by the CADD annotation. In detail, we compiled 7 different gene lists to use as inputs for GSEA. These lists were extracted from multiple population signal results and assembled as follows: a) 20Pop and 1090Ind European ancestry negative (Z < -3) DAP SNPs, b) 20Pop European ancestry positive (Z>3) DAP SNPs, c) 20Pop and 1090Ind Native American ancestry negative DAP SNPs, d) 1090Ind Native American ancestry positive DAP SNPs and e) 1090Ind African ancestry negative DAP SNPs.

Only the combinations of datasets and ancestries with at least one DAP window are present.

To identify different aspects of the biological impact of selected DAP windows, we used five diverse libraries:

- Human 2019 Kyoto Encyclopedia of Genes and Genomes (KEGG) that is a database resource for understanding high-level functions and utilities of the biological system.
- Gene Ontology (GO) Molecular Function 2018.
- Genotype-Tissue Expression (GTEx Tissue Sample Gene Expression Profiles down and up) both for down and up regulated genes.
- Genome-wide Association Studies (GWAS) Catalog 2019.

We applied a significance cutoff threshold of 0.001, and any adjusted p-value below this cutoff was considered significant.

Acknowledgements

We thank the people working at the High Performance Computing Center of the University of Tartu for the help and support provided. We thank Marco Rosario Capodiferro for useful discussions. This work was supported by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.16-0030 to LO, MMe, FM; Project No. 2014-2020.4.01.16-0271 to RF; Project No. 2014-2020.4.01.16-0125 to RF; Project No. 2014-2020.4.01.16-0024 to DM, LP). This work was supported by the Estonian Research Council grant PUT (PRG243) (to RF, MMe, LP). This work was supported by institutional research funding IUT (IUT24-1) of the Estonian Ministry of Education and Research (to TK). This research was supported by the European Union through Horizon 2020 grant no. 810645 (to MMe). This research was supported by the European Union through the Horizon 2020 research and innovation programme under grant no 810645 and through the European Regional Development Fund project no. MOBEC008 to MMo.

Conflict of Interest Statement

The authors declare no conflict of interests.

References

1. Ongaro, L., Scliar, M.O., Flores, R., Raveane, A., Marnetto, D., Sarno, S., Gnecchi-Ruscone, G.A., Alarcón-Riquelme, M.E., Patin, E., Wangkumhang, P., et al. (2019) The Genomic Impact of European Colonization of the Americas. *Curr. Biol.*, **29**, 3974–3986.e4.
2. Bryc, K., Durand, E.Y., Macpherson, J.M., Reich, D. and Mountain, J.L. (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. *Am. J. Hum. Genet.*, **96**, 37–53.
3. Montinaro, F., Busby, G.B.J., Pascali, V.L., Myers, S., Hellenthal, G. and Capelli, C. (2015) Unravelling the hidden ancestry of American admixed populations. *Nat. Commun.*, **6**, 6596.
4. Chacón-Duque, J.-C., Adhikari, K., Fuentes-Guajardo, M., Mendoza-Revilla, J., Acuña-Alonzo, V., Barquera, R., Quinto-Sánchez, M., Gómez-Valdés, J., Everardo Martínez, P., Villamil-Ramírez, H., et al. (2018) Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nat. Commun.*, **9**, 5388.
5. Homburger, J.R., Moreno-Estrada, A., Gignoux, C.R., Nelson, D., Sanchez, E., Ortiz-Tello, P., Pons-Estel, B.A., Acevedo-Vasquez, E., Miranda, P., Langefeld, C.D., et al. (2015) Genomic Insights into the Ancestry and Demographic History of South America. *PLoS Genet.*, **11**, e1005602.
6. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature*, **449**, 913–918.
7. Prohaska, A., Racimo, F., Schork, A.J., Sikora, M., Stern, A.J., Ilardo, M., Allentoft, M.E., Folkersen, L., Buil, A., Moreno-Mayar, J.V. (2019) Human Disease Variation in the Light of Population Genomics. *Cell*, **177**, 115–131.
8. Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature*, **419**, 832–837.
9. Hejase, H. A., Dukler, N. and Siepel, A. (2020) From Summary Statistics to Gene Trees: Methods for Inferring Positive Selection. *Trends Genet.*, **36**, 243–258.
10. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., et al. (2010) Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 786–791.
11. Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B. and Jin, L. (2012) Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.*, **22**, 519–527.
12. Bhatia, G., Tandon, A., Patterson, N., Aldrich, M.C., Ambrosone, C.B., Amos, C., Bandera, E.V., Berndt, S.I., Bernstein, L., Blot, W.J., et al. (2014) Genome-wide scan of 29,141 African Americans finds no evidence of directional selection since admixture. *Am. J. Hum. Genet.*, **95**, 437–444.
13. 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.
14. Rishishwar, L., Conley, A.B., Wigington, C.H., Wang, L., Valderrama-Aguirre, A. and Jordan, I.K. (2015) Ancestry, admixture and fitness in Colombian genomes. *Sci. Rep.*, **5**, 12376.
15. Deng, L., Ruiz-Linares, A., Xu, S. and Wang, S. (2016) Ancestry variation and footprints of natural selection along the genome in Latin American populations. *Sci. Rep.*, **6**, 21766.
16. Zhou, Q., Zhao, L. and Guan, Y. (2016) Strong Selection at MHC in Mexicans since Admixture. *PLoS Genet.*, **12**, e1005847.
17. Norris, E.T., Wang, L., Conley, A.B., Rishishwar, L., Mariño-Ramírez, L., Valderrama-Aguirre, A. and Jordan, I.K. (2018) Genetic ancestry, admixture and health determinants in Latin America. *BMC Genomics*, **19**, 861.
18. Norris, E.T., Rishishwar, L., Chande, A.T., Conley, A.B. and Ye, K. (2020) Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol.*, **21,** 29**.**
19. Yelmen, B., Mondal, M., Marnetto, D., Pathak, A.K., Montinaro, F., Gallego Romero, I.,

Kivisild, T., Metspalu, M. and Pagani, L. (2019) Ancestry-Specific Analyses Reveal Differential Demographic Histories and Opposite Selective Pressures in Modern South Asian Populations. *Mol. Biol. Evol.*, **36**, 1628–1642.
20. Nunn, N. and Qian, N. (2010) The Columbian Exchange: A History of Disease, Food, and Ideas. *J. Econ. Perspect.*, **24**, 163–188.
21. Crosby, A. W. (1987) The Columbian Voyages, the Columbian Exchange, and Their Historians. Essays on Global and Comparative History. *The Columbian Voyages, the Columbian Exchange, and Their Historians. Essays on Global and Comparative History*; American Historical Association, 400 A Street SE, Washington, DC 20003 ($3.50, plus $1.00 shipping)., (1987) .
22. Harper, K.N., Ocampo, P.S., Steiner, B.M., George, R.W., Silverman, M.S., Bolotin, S., Pillay, A., Saunders, N.J. and Armelagos, G.J. (2008) On the origin of the treponematoses: a phylogenetic approach. *PLoS Negl. Trop. Dis.*, **2**, e148.
23. Maples, B.K., Gravel, S., Kenny, E.E. and Bustamante, C.D. (2013) RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.*, **93**, 278–288.
24. Feenstra, B., Pasternak, B., Geller, F., Carstensen, L., Wang, T., Huang, F., Eitson, J.L., Hollegaard, M.V., Svanström, H., Vestergaard, M., et al. (2014) Common variants associated with general and MMR vaccine–related febrile seizures. Common variants associated with general and MMR vaccine–related febrile seizures. *Nature Genetics* **(2014)** , *46*, 1274–1282.
25. Stanescu, H.C., Arcos-Burgos, M., Medlar, A., Bockenhauer, D., Kottgen, A., Dragomirescu, L., Voinescu, C., Patel, N., Pearce, K., Hubank, M., et al. (2011) Risk HLA-DQA1 and PLA(2)R1 alleles in idiopathic membranous nephropathy. *N. Engl. J. Med.*, **364**, 616–626.
26. Haralambieva, I.H., Ovsyannikova, I.G., Kennedy, R.B., Larrabee, B.R., Zimmermann, M.T., Grill, D.E., Schaid, D.J. and Poland, G.A. (2017) Genome-wide associations of CD46 and IFI44L genetic variants with neutralizing antibody response to measles vaccine. *Hum. Genet.*, **136**, 421–435.
27. Tian, C., Hromatka, B.S., Kiefer, A.K., Eriksson, N., Noble, S.M., Tung, J.Y. and Hinds, D.A. (2017) Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun.*, **8**, 599.
28. Comuzzie, A.G., Cole, S.A., Laston, S.L., Voruganti, V.S., Haack, K., Gibbs, R.A. and Butte, N.F. (2012) Novel genetic loci identified for the pathophysiology of childhood obesity in the Hispanic population. *PLoS One*, **7**, e51954.
29. de Las Fuentes, L., Sung, Y.J., Sitlani, C.M., Avery, C.L., Bartz, T.M., Keyser, C. de, Evans, D.S., Li, X., Musani, S.K., Ruiter, R., Smith, A.V., et al. (2020) Genome-wide meta-analysis of variant-by-diuretic interactions as modulators of lipid traits in persons of European and African ancestry. *Pharmacogenomics J.*, **20**, 482–493.
30. Maguire, L.H., Handelman, S.K., Du, X., Chen, Y., Pers, T.H. and Speliotes, E.K. (2018) Genome-wide association analyses identify 39 new susceptibility loci for diverticular disease. *Nat. Genet.*, **50**, 1359–1365.
31. Ghani, M., Sato, C., Lee, J.H., Reitz, C., Moreno, D., Mayeux, R., St George-Hyslop, P. and Rogaeva, E. (2013) Evidence of recessive Alzheimer disease loci in a Caribbean Hispanic data set: genome-wide survey of runs of homozygosity. *JAMA Neurol.*, **70**, 1261–1267.
32. Warrington, N.M., Beaumont, R.N., Horikoshi, M., Day, F.R., Helgeland, Ø., Laurin, C., Bacelis, J., Peng, S., Hao, K., Feenstra, B., et al. (2019) Maternal and fetal genetic effects on birth weight and their relevance to cardio-metabolic risk factors. *Nat. Genet.*, **51**, 804–814.
33. Investigators, G., Investigators, M. and Investigators, S. D. (2013) Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. *Am. J. Psychiatry*, **170**, 207–217.
34. Richardson, T.G., Sanderson, E., Palmer, T.M., Ala-Korpela, M., Ference, B.A., Davey Smith, G. and Holmes, M.V. (2020) Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomization analysis. *PLoS Med.*, **17**, e1003062.
35. Adrion, J. R., Cole, C. B., Dukler, N., Galloway, J.G., Gladstein, A.L., Gower, G., Kyriazis, C.C., Ragsdale, A.P., Tsambos, G., Baumdicker, F., et al. (2020) A community-maintained standard library of population genetic models. *Elife*, **9**.

36. Gouveia, M.H., Borda, V., Leal, T.P., Moreira, R.G., Bergen, A.W., Aquino, M.M., Araujo, G.S., Araujo, N.M., Kehdy, F.S.G., Liboredo, R., et al. (2020) Origins, admixture dynamics and homogenization of the African gene pool in the Americas. *Molecular Biology and Evolution*.
37. Micheletti, S.J., Bryc, K., Ancona Esselmann, S.G., Freyman, W.A., Moreno, M.E., Poznik, G.D., Shastri, A.J., 23andMe Research Team, Beleza, S. and Mountain, J.L. (2020) Genetic Consequences of the Transatlantic Slave Trade in the Americas. *Am. J. Hum. Genet.*, **107**, 265–277.
38. D'Atanasio, E., Trionfetti, F., Bonito, M., Sellitto, D., Coppa, A., Berti, A., Trombetta, B. and Cruciani, F. (2020) Y haplogroup diversity of the Dominican Republic: reconstructing the effect of the European colonization and the trans-Atlantic slave trades. *Genome Biol. Evol.*
39. 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.
40. Delaneau, O., Marchini, J. and Zagury, J.-F. (2011) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.
41. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.

Legends to Figures

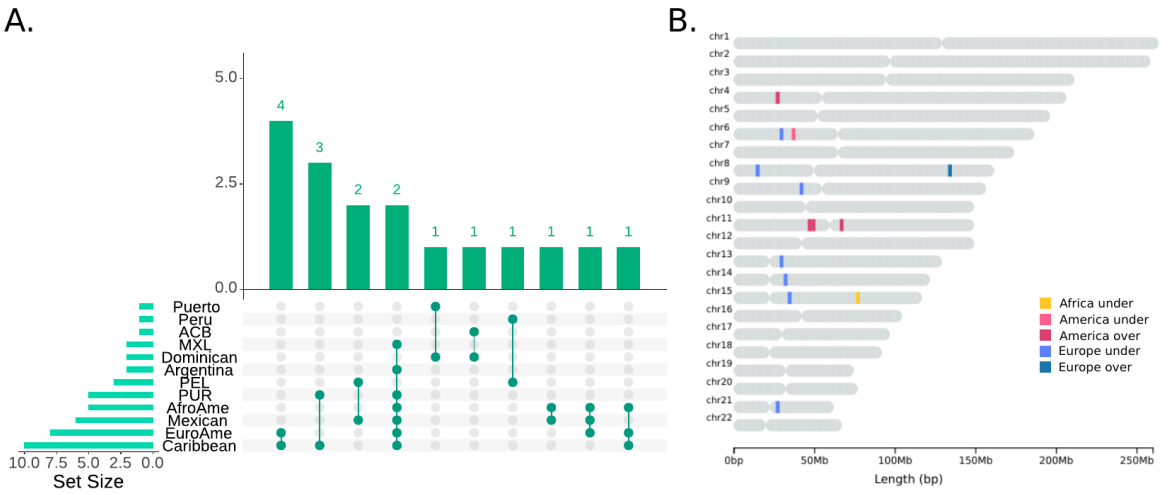

Figure 1: An overview of Divergent Ancestry Profile regions inferred by local ancestry profiles for all ancestries. A) Population distribution of DAPs. The x axis shows DAPS in single populations while y axis shows DAP sharing among groups. B) Genomic location of shared DAPs. Different colors refer to the ancestry and direction of the divergence, as indicated in the legend.

Figure 2: Comparison of the distribution of all the PHRED-scaled C-score values belonging to windows with divergent ancestry profiles (DAP) with the ones from the non-divergent for European and American ancestries in the 20Pop dataset. The asterisk refers to a statistically significant p-value (Wilcoxon test, Bonferroni corrected alpha=0.01). The number of analyzed windows is reported in Supplementary Table 5A.

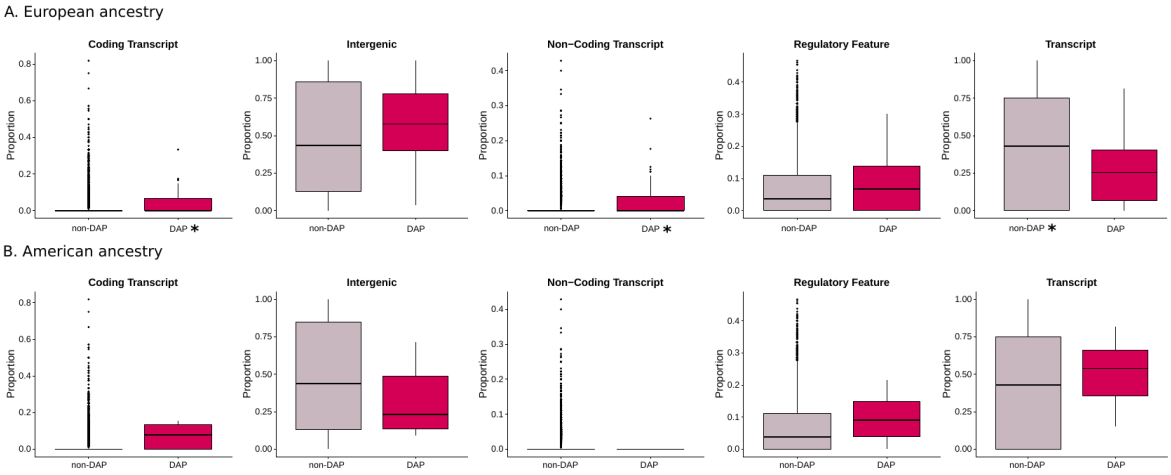

Figure 3: Comparison of the distribution of Annotypes (Coding Transcript, Intergenic, Non-coding Transcript, Regulatory Feature, Transcript) belonging to divergent ancestry profiles (DAP) windows with the ones from the non-divergent for European (A) and American (B) ancestries in the 20Pop dataset. The asterisk refers to a statistically significant p-value (Wilcoxon test, Bonferroni corrected alpha=0.002). The number of analyzed windows is reported in Supplementary Table 5B.

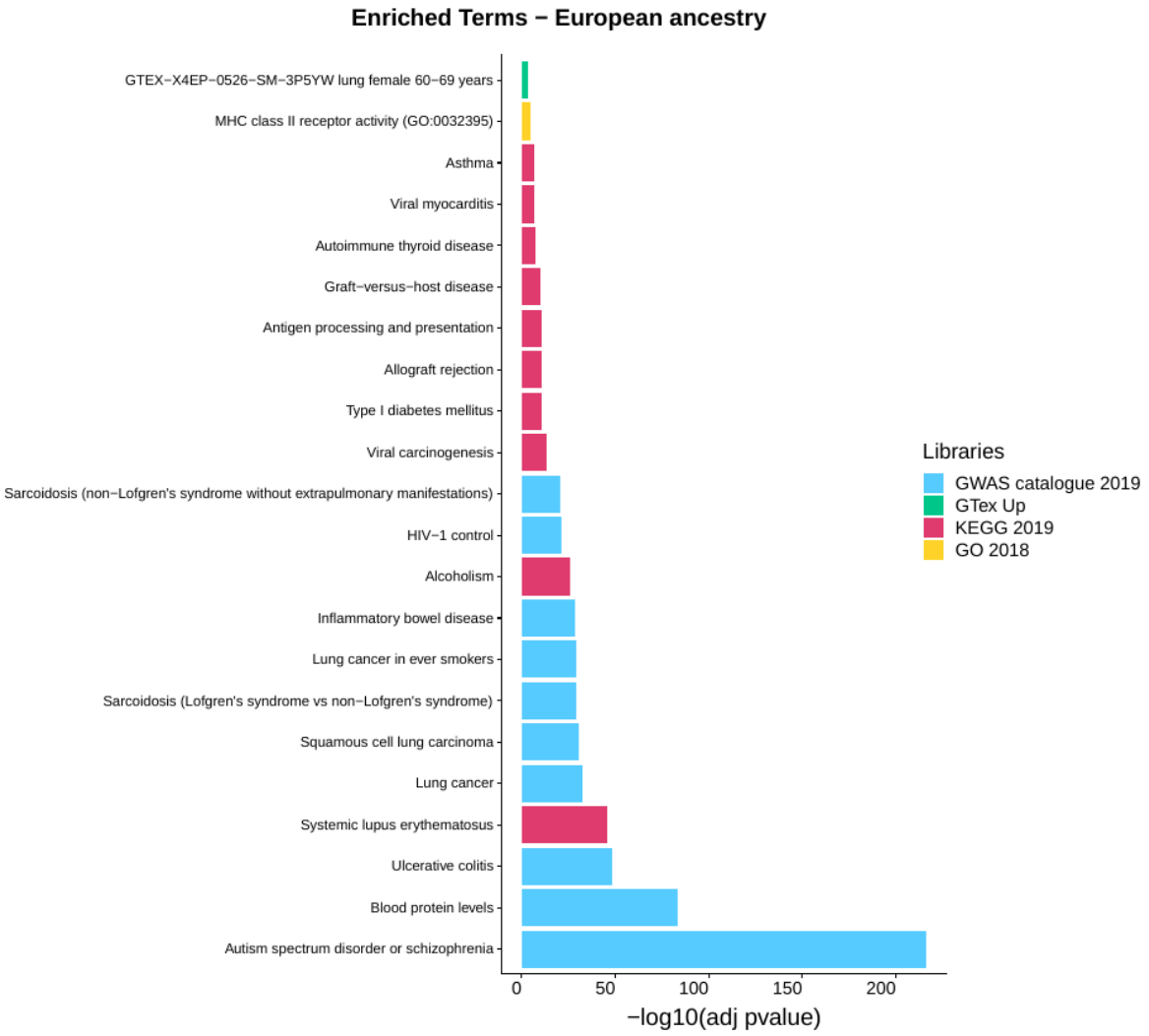

Figure 4: Gene-set enrichment analysis results for the European ancestry of the 20Pop dataset (Supplementary Table 6A-D). Only the first ten enriched terms for library are shown. Libraries: Genome-wide Association Studies (GWAS) Catalog 2019, GTEx Tissue Sample Gene Expression Profiles up, Human 2019 Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) 2018.

Tables

Population	N	Country	References
ACB	68	Barbados	1000 Genomes Project
AfroAme	2004	USA	IlluminaI Control Database
Argentina	133	Argentina	Lopez Herraez et al 2013
ASW	55	SW_USA	1000 Genomes Project
Caribbean	1112	Puerto Rico & Dominican Republic	Ghani et al 2013
Chile	25	Chile	Lopez Herraez et al 2013
CLM	72	Colombia	1000 Genomes Project
Colombian	26	Colombia	Bryc et al 2015
Dominican	27	Dominican Republic	Bryc et al 2015
Ecuadorian	19	Ecuador	Bryc et al 2015
EuroAme	1562	USA	IlluminaI Control Database
Maya	25	Mexico	Moreno-Estrada et al 2014
Mayas	21	Mexico	Li et al 2008
Mexican	364	Mexico	Moreno-Estrada et al 2014
MXL	63	Mexico	1000 Genomes Project
PEL	68	Peru	1000 Genomes Project
Peru	85	Peru	Lopez Herraez et al 2013
Puerto	26	Puerto Rico	Bryc et al 2015
PUR	73	Puerto Rico	1000 Genomes Project

Table 1. Admixed American populations under study. N refers to the number of individuals included in the dataset.

Chr	Start	End	Wind	SNPs	Length (kb)	Data set	Ancestry	Type	Populations (Z)	Nge nes	Genes
4	24512590	25679209	9	91	1167	1090 Ind	-America	IS-DAP	PEL (Z=-3.1) Peru (ZMin=-3.1, Max=-3.5)	15	MIR573;DHX15;RP11-496D24.2;AC006390.4;SOD3;CCDC149;SEPSECS;PI4K2B;ZCCHC4;AC108218.1;RP11-206P5.1;RP11-302F12.3;RP11-302F12.2;SLC34A2;RP11-302F12.1
6	25312755	33098966	68	1079	7786	20Pop	-Europe	IP-DAP	Caribbean (ZMin=-3.1, Max=-6.5) EuroAme (ZMin=-3, Max=-7)	373	View Supplementary Table 2
6	34102061	35495811	8	64	1394	20Pop, 1090 Ind	-America	IP-DAP	Mexican (ZMin=-3.1, Max=-3.4) PEL (ZMin=-4.3, Max=-4.8)	20	GRM4;PACSIN1;SPDEF;RP3-391O22.3;C6orf106;RPL7P25;UHRF1BP1;TAF11;ANKS1A;HSPE1P11;TCP11;SCUBE3;RP3-329A5.1;ZNF76;DEF6;MKRNP2;FANCE;RPL10A;TEAD3;TULP1
8	121904711	122498253	6	66	594	20Pop	+Europe	IP-DAP	Dominican (Z=3.1) Puerto (Z=3)	5	RP11-369K17.1;AC011626.1;RP11-369K17.2;AC027238.1;RPL35AP19
8	10706801	11499967	8	108	793	20Pop	-Europe	IP-DAP	Caribbean (ZMin=-	23	RP11-177H2.2;XKR6;MIR598;AF131215.6;AF131215.9;AF131215.2;AF131215.4;AF131215.8;LINC00529;RPL19P13;MTMR

									3.1, Max=-3.4) EuroAme (Z Min=-9, Max=-9.5)		9;AF131216.6;SLC35G5;TDH;AF131216.5;C8orf12;RN7SL293P;RNU6-1084P;FAM167A;BLK;RP11-148O21.3;RP11-148O21.6;LINC00208
9	38615175	38771831	2	44	157	20Pop	-Europe	IP-DAP	AfroAme (Z Min=-3, Max=-3.6) EuroAme (Z Min=-5.2, Max=-6.5) Mexican (Z Min=-3.2, Max=-3.3)	5	GLIS3;ANKRD18A;FAM201A;RP13-198D9.3;RNU6-765P
9	38615175	38771831	2	44	157	1090Ind	-Europe	IP-DAP	AfroAme (Z Min=-3.2, Max=-3.4) Mexican (Z Min=-3.2, Max=-3.3)	5	GLIS3;ANKRD18A;FAM201A;RP13-198D9.3;RNU6-765P
11	43908230	44296591	4	68	388	1090Ind	+America	IS-DAP	Caribbean (Z=3) PUR (Z Min=3.	12	OR52B3P;TRIM21;ALKBH3;C11orf96;RP11-613D13.4;ALKBH3-AS1;RP11-613D13.8;ACCSL;ACCS;CTD-2609K8.3;EXT2;ALX4

									1, Max=3.3)		
11	46210259	46297631	1	10	87	1090Ind	+America	IS-DAP	Caribbean (Z=3.1) PUR (Z=3.1)	5	TRIM68;RP11-702F3.4;CTD-2589M5.5;CTD-2589M5.4;CREB3L1
11	56000288	56184888	2	21	185	1090Ind	+America	IS-DAP	Caribbean (Z=3.1) PUR (ZMin=3.1, Max=3.3)	12	OR5T2;OR8K3;OR8K2P;FAM8A2P;OR8K1;OR8J1;RPL5P29;OR8U1;OR8L1P;OR5AL2P;OR5AL1;OR5R1
13	19612262	19690836	1	9	79	20Pop	-Europe	IP-DAP	AfroAme (Z=-3.3) Caribbean (Z=-3.3) EuroAme (ZMin=-14.4, Max=-16)	2	PHF2P2;RNA5SP24
14	20445618	20697600	3	46	252	20Pop	-Europe	IP-DAP	Caribbean (ZMin=-3, Max=-3.1) EuroAme (ZMin=-6.1, Max=-8)	17	OR4K15;OR4Q2;OR4K14;OR4K13;AL359218.1;OR4U1P;OR4L1;RNA5SP380;OR4T1P;OR4K17;OR11G1P;RP11-98N22.6;OR11G2;OR11H5P;OR11H6;AL356019.1;OR11H7
15	22837	23975	6	69	113	20Po	-	IP-	AfroA	7	TUBGCP5;CYFIP1;NIPA2;NIPA1;MIR4508;MKRN3;RP11-73C9.1

	143	482			8	p	Europe	DAP	me (ZMin=-3.4, Max=-9.2) Argentina (ZMin=-4.9, Max=-5.2) Caribbean (ZMin=-7.1, Max=-8.2) EuroAme (ZMin=-4.2, Max=-35.1) Mexican (ZMin=-4.8, Max=-5.4) MXL (ZMin=-3.2, Max=-3.6) PUR (ZMin=-4.3, Max=-4.8)		
15	22837143	23053839	3	43	217	1090 Ind	-Europe	IP-DAP	AfroAme (ZMin=-7.2, Max=-8.4)	4	TUBGCP5;CYFIP1;NIPA2;NIPA1

									Argentina (ZMin=-4.8, Max=-5) Caribbean (ZMin=-6.1, Max=-7.1) EuroAme (ZMin=-3.4, Max=-3.9) Mexican (ZMin=-4.8, Max=-5.3) MXL (ZMin=-3.2, Max=-3.6) PUR (ZMin=-4.3, Max=-4.8)		
15	65613654	65695283	1	10	82	1090Ind	-Africa	IP-DAP	ACB (Z=-3.1) Dominican (Z=-3)	2	IGDCC3;IGDCC4
21	15412399	15599963	2	17	188	20Pop	-Europe	IP-DAP	Caribbean (Z=-3.2) EuroA	6	AP001347.6;RNA5SP488;ANKRD20A18P;LIPI;ERLEC1P1;RBM11

									me (Z Min=-8.2, Max=-9.5)		

Table 2. Details about inferred genomic regions with Divergent Ancestry Profiles. “Wind” refers to the number of DAP 100kb windows inside the genomic region; “SNPs” shows the number of SNPs contained in the genomic region; “Ancestry” refers to the specific ancestry for which we found an underrepresented (-) or overrepresented (+) DAP window or region; “Type” describes the DAP regions as Inter-Samples (IS-DAP) or Inter-Populations (IP-DAP); “Populations (Z)” reports Z scores for the significant populations; “N genes” shows the number of genes related to SNPs part of the genomic region; “Genes” refers to the Gene/clone identifiers provided in ENSEMBL annotation.

Abbreviations

CADD - Combined Annotation Dependent Depletion

DAP - Divergent Ancestry Profile

GTEx - GTEx Tissue Sample Gene Expression Profiles down and up

GWAS - Genome-wide Association Studies (GWAS)

HLA – Human Leukocyte Antigen

IS-DAP – Inter-Samples DAP

IP-DAP – Inter-Populations DAP

KEGG - Human 2019 Kyoto Encyclopedia of Genes and Genomes

MHC - Major Histocompatibility Complex

SNP – Single Nucleotide Polymorphism