

RESEARCH

Open Access



Genetic ancestry, admixture and health determinants in Latin America

Emily T. Norris^{1,2,3}, Lu Wang¹, Andrew B. Conley³, Lavanya Rishishwar^{1,2,3}, Leonardo Mariño-Ramírez^{2,4}, Augusto Valderrama-Aguirre^{2,5} and I. King Jordan^{1,2,3*}

From Selected articles from the IV Colombian Congress on Bioinformatics and Computational Biology & VIII International Conference on Bioinformatics SolBio 2017
Santiago de Cali, Colombia. 13-15 September 2017

Abstract

Background: Modern Latin American populations were formed via genetic admixture among ancestral source populations from Africa, the Americas and Europe. We are interested in studying how combinations of genetic ancestry in admixed Latin American populations may impact genomic determinants of health and disease. For this study, we characterized the impact of ancestry and admixture on genetic variants that underlie health- and disease-related phenotypes in population genomic samples from Colombia, Mexico, Peru, and Puerto Rico.

Results: We analyzed a total of 347 admixed Latin American genomes along with 1102 putative ancestral source genomes from Africans, Europeans, and Native Americans. We characterized the genetic ancestry, relatedness, and admixture patterns for each of the admixed Latin American genomes, finding a spectrum of ancestry proportions within and between populations. We then identified single nucleotide polymorphisms (SNPs) with anomalous ancestry-enrichment patterns, i.e. SNPs that exist in any given Latin American population at a higher frequency than expected based on the population's genetic ancestry profile. For this set of ancestry-enriched SNPs, we inspected their phenotypic impact on disease, metabolism, and the immune system. All four of the Latin American populations show ancestry-enrichment for a number of shared pathways, yielding evidence of similar selection pressures on these populations during their evolution. For example, all four populations show ancestry-enriched SNPs in multiple genes from immune system pathways, such as the cytokine receptor interaction, T cell receptor signaling, and antigen presentation pathways. We also found SNPs with excess African or European ancestry that are associated with ancestry-specific gene expression patterns and play crucial roles in the immune system and infectious disease responses. Genes from both the innate and adaptive immune system were found to be regulated by ancestry-enriched SNPs with population-specific regulatory effects.

Conclusions: Ancestry-enriched SNPs in Latin American populations have a substantial effect on health- and disease-related phenotypes. The concordant impact observed for same phenotypes across populations points to a process of adaptive introgression, whereby ancestry-enriched SNPs with specific functional utility appear to have been retained in modern populations by virtue of their effects on health and fitness.

Keywords: Genetic ancestry, Admixture, Ancestry-enrichment, Adaptive introgression, Health, Disease, Population genetics, Immune system

* Correspondence: king.jordan@biology.gatech.edu

¹School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA

²PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia

Full list of author information is available at the end of the article



Background

The modern human species – *Homo sapiens sapiens* – originated in sub-Saharan Africa ~200,000 years ago and began to migrate out of Africa and spread throughout the world starting ~70,000 years ago [1, 2]. After heading north out of Africa, humans spread to the east and west, populating Melanesia, Asia, and Europe, and eventually made their way across the Bering Strait into the Americas ~20,000 years ago. As human populations occupied different parts of the globe, they often became geographically isolated in their new homelands. Thousands of years of geographic isolation were accompanied by population genetic diversification, giving rise to the diverse human population groups that can be seen around the world to this day [3, 4]. Distinct continental population groups – African, Asian, and European in particular – are the most obvious examples of this evolutionary process. There were, of course, a number of episodes of genetic admixture during this time [5], whereby previously isolated populations came into contact and began to mix, but for the most part, the dominant mode of human evolution since our emergence from Africa has been characterized by populations' physical isolation followed by genetic diversification.

This pattern of human evolution was turned upside down upon the arrival of Columbus in the New World a mere 500 years ago, which is less than 1% of the elapsed time since humans emerged from Africa. Columbus' voyages precipitated the so-called 'Columbian Exchange' – a massive transfer of life forms, which had evolved separately for millennia, between the Old and New World hemispheres [6, 7]. The human dimension of the Columbian Exchange entailed genetic admixture between previously isolated populations on an unprecedented scale, in terms of both scope and rapidity [8]. The conquest and colonization of the Americas, along with the trans-Atlantic slave trade, brought African, European, and Native American populations into close and sustained contact for the first time. As a consequence, these diverse population groups began to mix, giving rise to novel admixed American populations. This is particularly true for Latin America, where populations are characterized by high levels of genetic admixture among African, European, and Native American ancestral source populations [9–11].

Latin American genomes can thus be considered to represent a recent innovation in human evolution. Indeed, genomes from modern Latin American populations are evolutionarily novel in the sense that they contain combinations of genetic variants (haplotypes) that never previously existed together on the same genetic background. Our group is interested in trying to understand the implications of the recent advent of novel Latin American genomes, particularly as it relates

to the genetic determinants of health-related phenotypes. In other words, we are asking what it means when genomes that were separated for many thousands of years are suddenly brought back together and what the implications of this process are for human health and fitness.

Our group and others have employed an approach that we call ancestry-enrichment analysis to address these kinds of questions via population-level studies of admixed American genomes [8]. This approach relies on the characterization of local patterns of genetic ancestry for individual genomic loci. Local ancestry assignment, colloquially referred to as chromosome painting, entails the delineation of ancestral origins of specific haplotypes across the genome. The resulting chromosome paintings reveal the genomes of admixed individuals as mosaics of interspersed ancestry-specific haplotypes. When a population sample of admixed genomes is characterized in this way, the percent ancestry contributions from each ancestral source population can be computed for all haplotype loci genome-wide. Ancestry-enrichment analysis then entails the identification of specific haplotype loci that have anomalous patterns of local ancestry, i.e. levels of locus-specific ancestry that are significantly higher or lower than can be expected by chance given the overall ancestry profile of the population. Statistically significant signals of ancestry-enrichment are taken as evidence of adaptive introgression, whereby introgressed haplotypes increase in frequency by virtue of a selective advantage that they provide to individuals in an admixed population.

A number of recent studies have used ancestry-enrichment analysis to show evidence of adaptive introgression in admixed American genomes. The first study of this kind showed an excess of African ancestry at the major histocompatibility locus (MHC) in a sample of Puerto Rican genomes [12], and a follow up study several years later also found ancestry-enrichment at the same region in a Mexican population [13]. Since that time, several other studies have replicated the finding of ancestry-enrichment in admixed populations at this and other health related loci [14–18]. Our own more recent work on Colombian genome sequences revealed even more widespread ancestry-enrichment, which impacted various aspects of the immune system, including pathways involved in both innate and adaptive immunity [18].

All of the previous ancestry-enrichment studies were distinguished by their interrogation of the ancestral origins of individual haplotypes, i.e. physically linked sets of genetic variants that are inherited together. For this study, we developed and applied a novel method based on individual genetic variants – single nucleotide polymorphisms (SNPs) – in an effort to expand our view of the relationship between genetic ancestry, admixture and health in Latin American populations.

Our approach entails the detection of SNPs that are found at anomalously high frequencies in admixed populations compared to what is expected based on their frequencies in the ancestral source populations, i.e. ancestry-enriched SNPs. To find such ancestry-enriched SNPs, we consider the proportional contributions of ancestral source populations to admixed Latin American populations, together with SNP frequencies in the ancestral populations, to derive expected SNP frequencies for the Latin American populations. These expected frequencies are compared to observed frequencies in order to identify statistically significant ancestry-enriched SNPs; the connection between ancestry-enriched SNPs and health-related phenotypes is then explored via analysis of the functional annotations of the SNPs and their linked genes. In particular, we interrogated the impact of ancestry-enriched SNPs on disease, metabolism and immune system pathways. This approach uncovered signals of ancestry-enrichment in health- and disease-related traits shared

among all four of the Latin American populations that we analyzed, raising the possibility of shared selective pressures among them.

Materials and methods

Comparative genomic data sources

Whole genome sequences from four admixed Latin American populations – Colombia ($n = 94$), Mexico ($n = 64$), Peru ($n = 85$), and Puerto Rico ($n = 104$) – were taken from the 1000 Genomes Project (1KGP) phase 3 data release [19]: <http://www.internationalgenome.org/data/>. Genome sequences from the four Latin American populations were compared to whole genome sequences and whole genome genotypes of global reference populations from African, European, and Native American continental population groups to characterize their patterns of genetic ancestry and admixture (Fig. 1 and Table 1). The global reference whole genome sequences were also taken from the 1KGP, and reference whole genome genotype data was taken from the Human Genome Diversity

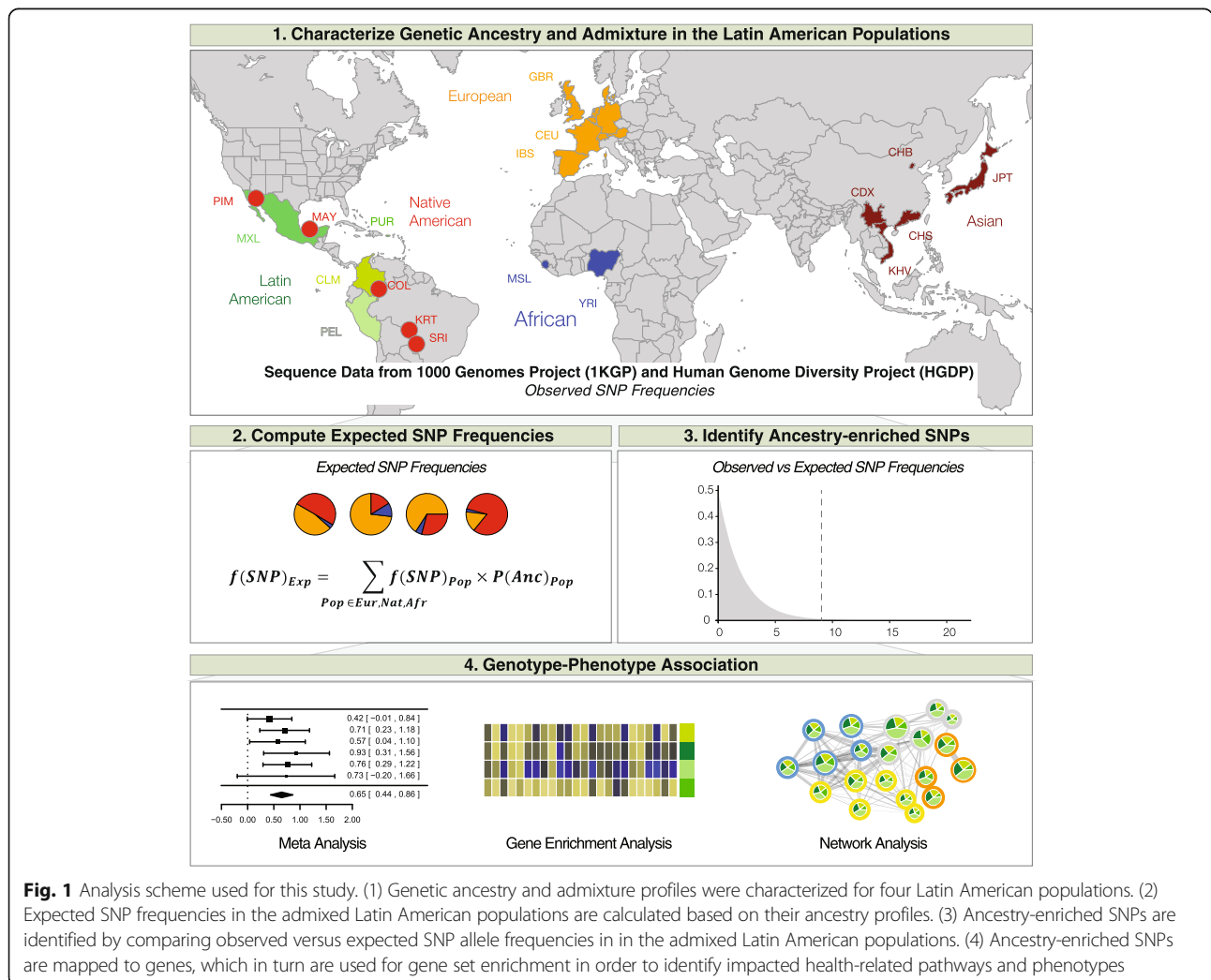


Table 1 Human populations analyzed in this study

| Dataset | Short | Full Description | n |
|---|-------|---|-----|
| 1KGP African (<i>n</i> = 193) | MSL | Mende in Sierra Leone | 85 |
| | YRI | Yoruba in Ibadan, Nigeria | 108 |
| 1KGP East Asian (<i>n</i> = 504) | CDX | Chinese Dai in Xishuangbanna, China | 93 |
| | CHB | Han Chinese in Beijing, China | 103 |
| | CHS | Southern Han Chinese, China | 105 |
| | JPT | Japanese in Tokyo, Japan | 104 |
| | KHV | Kinh in Ho Chi Minh City, Vietnam | 99 |
| 1KGP European (<i>n</i> = 297) | CEU | Utah residents with NW European ancestry | 99 |
| | IBS | Iberian populations in Spain | 107 |
| | GBR | British in England and Scotland | 91 |
| 1KGP Admixed American (<i>n</i> = 347) | CLM | Colombian in Medellin, Colombia | 94 |
| | MXL | Mexican Ancestry in Los Angeles, California | 64 |
| | PEL | Peruvian in Lima, Peru | 85 |
| | PUR | Puerto Rican in Puerto Rico | 104 |
| HGDP Native American (<i>n</i> = 108) | KRT | Karitiana in Brazil | 24 |
| | SRI | Surui in Brazil | 21 |
| | COL | Colombians in Colombia | 13 |
| | MAY | Maya in Mexico | 25 |
| | PIM | Pima in Mexico | 25 |

Populations are organized into continental groups: African, East Asian, European, and Admixed American from the 1000 Genomes Project (1KGP) and Native American from the Human Genome Diversity Project (HGDP). Short names, descriptions, and the numbers of genomes analyzed are provided for each individual population

Project (HGDP) [4]. The whole genome sequence and whole genome genotype data were merged, with sites that existed in both datasets retained for subsequent analysis, and PLINK v1.9 [20] was used to correct single nucleotide polymorphism (SNP) strand orientation as needed. This resulted in a dataset of 435,782 SNPs from 1449 individuals, across 19 populations. The final merged SNP dataset was phased with the 1KGP haplotype reference panel using the program SHAPEIT2 [21].

Genome ancestry assignment

ADMIXTURE [22] was run on individuals from both the global reference and admixed Latin American populations to infer their genome-wide ancestry profiles. ADMIXTURE was run with $K=3$ corresponding the three ancestral continental population groups: African, European, and Native American. The ADMIXTURE results for the admixed Latin American populations were used to infer individuals' percent ancestry contributions from each of these three continental ancestry groups. The program RFMix [23] was used to assign the continental ancestry origins of individual haplotypes across the genome, i.e. local ancestry. As with ADMIXTURE, African, European, and Native American populations were used as reference populations for RFMix. Ancestry-specific haplotypes were only called for regions where RFMix certainty was at least 99%.

Detection of ancestry-enriched SNPs

Ancestry-enriched SNPs were characterized as SNPs found in higher frequencies in admixed Latin American populations compared to what is expected based on (1) their frequencies in the ancestral source populations, and (2) the proportion of ancestry derived from each ancestral source population. For any given SNP, in any given Latin American population, the expected frequency of the SNP $f(SNP)_{Exp}$ can be calculated as:

$$f(SNP)_{Exp} = \sum_{Pop \in Eur, Afr, Nat} f(SNP)_{Pop} \times P(Anc)_{Pop} \quad (1)$$

where $f(SNP)_{Pop}$ is the frequency of the SNP in a specific ancestral source population and $P(Anc)_{Pop}$ is the proportion of ancestry in the modern Latin American population derived from that same ancestral population. Ancestry proportions were computed using the reduced, merged set of SNPs described above, with African and European reference populations from the 1KGP and Native American reference populations from HGDP. SNP frequencies were computed using whole genome sequences, in order to provide more complete coverage of variants genome-wide, using the YRI African and IBS European reference populations with the most closely

related East Asian population CHB taken as a surrogate for Native American ancestry.

The statistical significance of SNP ancestry-enrichment calculated this way was determined by comparing the observed (*Obs*) to expected (*Exp*) frequencies of the reference (*Ref*) and alternate (*Alt*) alleles for any given SNP as shown here:

$$\chi^2 = \frac{(Obs_{Ref} - Exp_{Ref})^2}{Exp_{Ref}} + \frac{(Obs_{Alt} - Exp_{Alt})^2}{Exp_{Alt}} \quad (2)$$

The χ^2 distribution was used to calculate *P*-values for each SNP, and false discovery rate (FDR) *q*-values were used to account for multiple statistical tests. The SNPs that had significant FDR values ($q < 0.05$) were considered to be ancestry-enriched in the Latin American population, i.e. present at a higher frequency than expected based on the population ancestry profile.

For ancestry-enriched SNPs, the individual ancestry components (*Anc*) that gave rise to the pattern of enrichment were also determined by jointly minimizing the frequency difference between the SNP in the Latin American population and a single ancestral source population while maximizing the distance between that single source population and the other two ancestral populations:

$$Anc_{Pop1} = \left[\frac{(f(SNP)_{Pop1} - f(SNP)_{Pop2}) + (f(SNP)_{Pop1} - f(SNP)_{Pop3})}{2 - |f(SNP)_{Obs} - f(SNP)_{Pop1}|} \right] \quad (3)$$

where $f(SNP)_{Popx}$ is the frequency of the SNP in each of the ancestral source populations and $f(SNP)_{Obs}$ is the frequency of the ancestry-enriched SNP in the Latin American population.

Gene set enrichment analysis

Ancestry-enriched SNPs were mapped to genes if they mapped within the NCBI RefSeq [24] gene models, i.e. between transcription start and stop sites, on the UCSC Genome Browser human genome reference sequence build GRCh37/hg19. Functionally coherent gene sets were curated from the Molecular Signatures Database (MSigDB) version 5.1 [25] for three broad functional categories: health- and disease-related phenotypes, metabolism, and immunity. Gene set enrichment analysis (GSEA) was performed for each Latin American population by adopting the MSigDB statistical framework to find functional gene sets that were enriched for genes with mapped ancestry-enriched SNPs. To do this, genes that harbor ancestry-enriched SNPs were overlapped

with genes from each functional gene set, and overlap enrichment was performed using the R limma package [26]. Overlap enrichment *P*-values were computed for each gene set using the Wilcoxon signed-rank test.

Expression quantitative trait loci (eQTL) analysis

RNA-seq data were taken from the GUEVADIS RNA sequencing (RNA-seq) project for 1KGP samples ftp://ftp.ebi.ac.uk/pub/databases/microarray/data/experiment/GEUV/E-GEUV-1/analysis_results/. A total of 445 RNA-seq samples were used in the analysis, including 87 African and 358 European individuals. The RNA-seq data correspond to gene expression levels for the same lymphoblastoid cell lines, i.e. Epstein–Barr virus (EBV) transformed B-lymphocytes, which were used for the 1KGP DNA-seq characterization. RNA-seq sample preparation, sequencing experiments and read-to-genome mapping were performed as previously described [27]. The read-to-genome mapping corresponds to human genome build GRCh37/hg19. Gene expression levels were quantified based on RNA-seq mapped reads and corresponded to ENSEMBL gene models [28]. Gene expression levels were quantified by using the reads per kilobase per million mapped reads (RPKM) approach in combination with the probabilistic estimation of expression residuals (PEER) method as previously described [29, 30].

Matched whole genome sequencing based SNP genotype calls were obtained for the same 445 individuals from the phase 3 release of the 1KGP [19]. Only SNPs with minor allele frequency (MAF) greater than 5% were used for the downstream analysis to ensure both the confidence of genotype calls and the reliability of the eQTL association analyses. Gene expression levels of 445 individuals were then regressed against their SNP genotypes to identify eQTLs using the program Matrix eQTL [31]. The Matrix eQTL program was run using the additive linear model option with gender and population labels included as covariates.

SNP pathway meta-analysis

Meta-analysis was used to evaluate the overall ancestry-enrichment for sets of SNPs that are implicated in specific health- or disease-related pathways. For any given ancestry-enriched SNP that mapped to a gene found in an overrepresented pathway, a log odds ratio was calculated as:

$$\text{Log Odds Ratio} = \ln \left[\frac{f(Obs)_{Ref} / f(Exp)_{Ref}}{f(Obs)_{Alt} / f(Exp)_{Alt}} \right] \quad (4)$$

where the observed (*Obs*) versus expected (*Exp*) frequencies (*f*) are compared for the SNP reference (*Ref*) versus alternate (*Alt*) alleles. Then for each pathway, the set of individual SNP log odds ratios was combined to

yield pathway-specific log odds ratio values along with 95% confidence intervals using the fixed-effect model with moderators via linear (mixed-effect) models implemented in the metafor package in R [32].

Results

Relating genome ancestry and health in Latin America

We developed and applied a single nucleotide polymorphism (SNP)-based approach to relate genome ancestry to genetic determinants of health in admixed Latin American populations (Fig. 1). First, patterns of genetic ancestry and admixture in Latin American populations were characterized via comparison with reference genome sequences of putative ancestral source populations from Africa, the Americas and Europe (Table 1). We then computed the expected SNP frequencies in Latin American populations by taking into consideration the SNP frequencies in the ancestral source populations along with the proportional contributions of each ancestral source population to the modern Latin American populations. Comparisons of observed versus expected SNP frequencies in admixed Latin American populations were used to identify what we refer to as 'ancestry-enriched' SNPs, which are SNPs found at anomalous frequencies in Latin American populations compared to what can be expected based on their ancestry profiles. Ancestry-enriched SNPs were mapped to genes, and then genes were used in gene set enrichment analysis to identify impacted health-related pathways and phenotypes.

Genetic ancestry and admixture in four Latin American populations

Genome sequences from four Latin American populations – Colombia, Mexico, Peru, and Puerto Rico – were compared to whole genome sequences and whole genome genotypes of global reference populations from African, European, and Native American continental population groups in order to characterize their patterns of genetic ancestry and admixture. Each Latin American population has a distinct pattern of three-way continental genetic admixture characterized by population-specific proportions of African, European and Native American ancestry (Fig. 2). Puerto Rico and Colombia are characterized by relatively high levels of three-way admixture, with substantial ancestry contributions from all three continental population groups, whereas Mexico and Peru have primarily Native American and European ancestry. Puerto Rico and Colombia also have the highest levels of European ancestry, while Peru and Mexico have majority Native American ancestry. The 80% Native American ancestry component for Peru is the single highest contribution of any ancestral population to an admixed Latin American population, and the 2% African ancestry fraction for this same population is the lowest. African source populations contribute the least amount

of ancestry to all four Latin American populations analyzed here. The continental ancestry proportions for each Latin American population were used as described in the following section to detect ancestry-enriched SNPs that exist in any given population at a higher frequency than expected based on its ancestry profile.

Ancestry-enriched SNPs in Latin American populations

Our approach to relating genetic ancestry to determinants of health and disease in modern Latin American populations relies on the detection of SNPs that are found at anomalously high frequencies in admixed populations compared to what is expected based on their frequencies in the ancestral source populations, i.e. ancestry-enriched SNPs. We reasoned that such ancestry-enriched SNPs are likely to have an outsized effect on health and disease in modern Latin American populations, perhaps related to an initial increase in population frequency via adaptive introgression.

We developed and applied a quantitative method to identify individual SNPs that are enriched in admixed Latin American populations with respect to ancestry from one of the three ancestral source populations: Africa, Europe, and the Americas. To do so, the expected frequencies for each SNP were calculated using the frequency of the given SNP in each of the three ancestral source populations conditioned upon the proportion of each ancestral source population in the Latin American population of interest. Observed SNP frequencies were compared to expected SNP frequencies to identify ancestry-enriched SNPs; the details of this approach are shown in the Materials and Methods section.

The distributions and median values of ancestry-specific SNP χ^2 values are shown in Fig. 3a. Peru shows the strongest overall signal of SNP ancestry-enrichment, followed by Mexico, Colombia, and Puerto Rico. Statistically significant ancestry-enriched SNPs for each population were mapped to genes for subsequent analysis of their impact on health- and disease-related phenotypes. There is a substantial amount of overlap of mapped genes among the four populations (Fig. 3b). Out of 156 total genes with mapped ancestry-enriched SNPs, 102 (65%) are shared among two or more populations compared to 54 (35%) that are population-specific. There are 40 genes that bear ancestry-specific SNPs in all four Latin American populations, which is by far the single largest component of shared versus unique genes. Lists of all SNPs that show significant ancestry-enrichment – along with details regarding their observed and expected allele frequencies, test-statistic values, and specific ancestry enrichment patterns – can be found in Additional file 1: Table S1.

Gene set enrichment analysis of overrepresented SNPs

The genes with mapped ancestry-specific SNPs were analyzed with gene set enrichment analysis (GSEA) to look

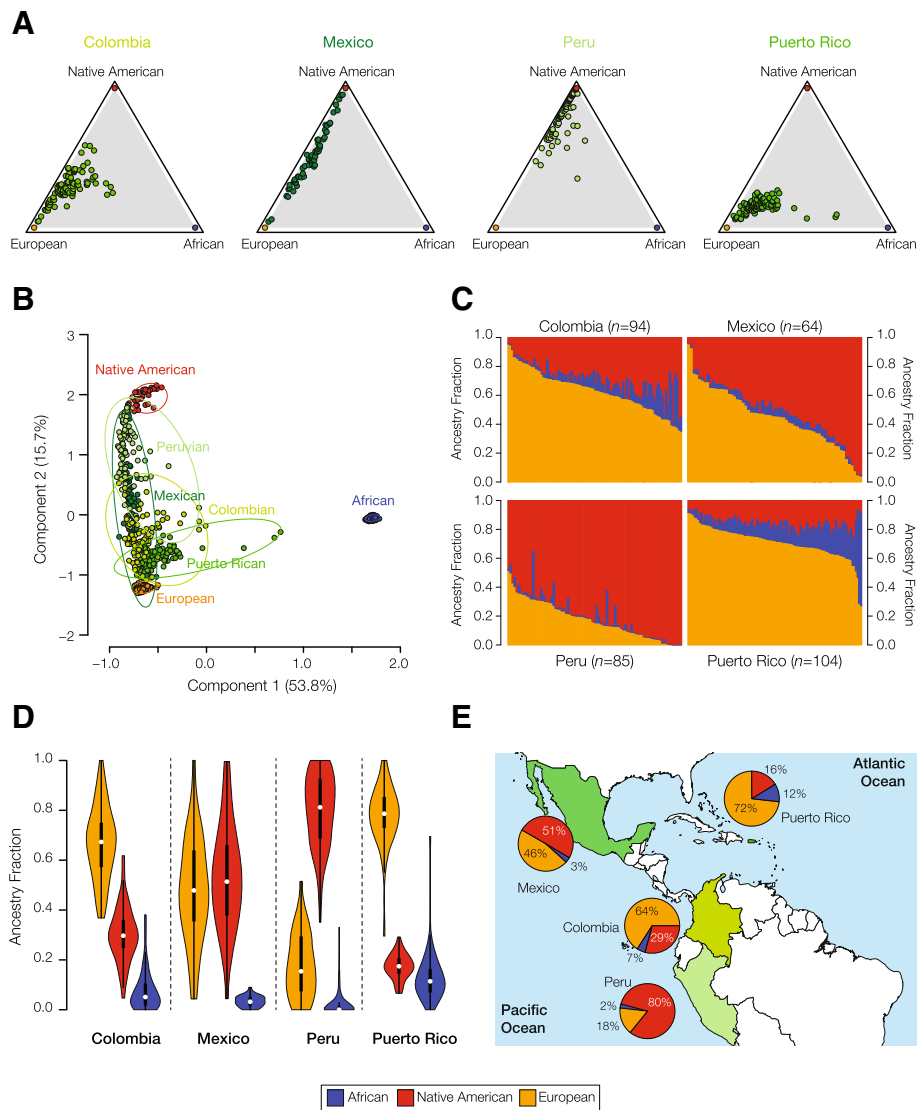
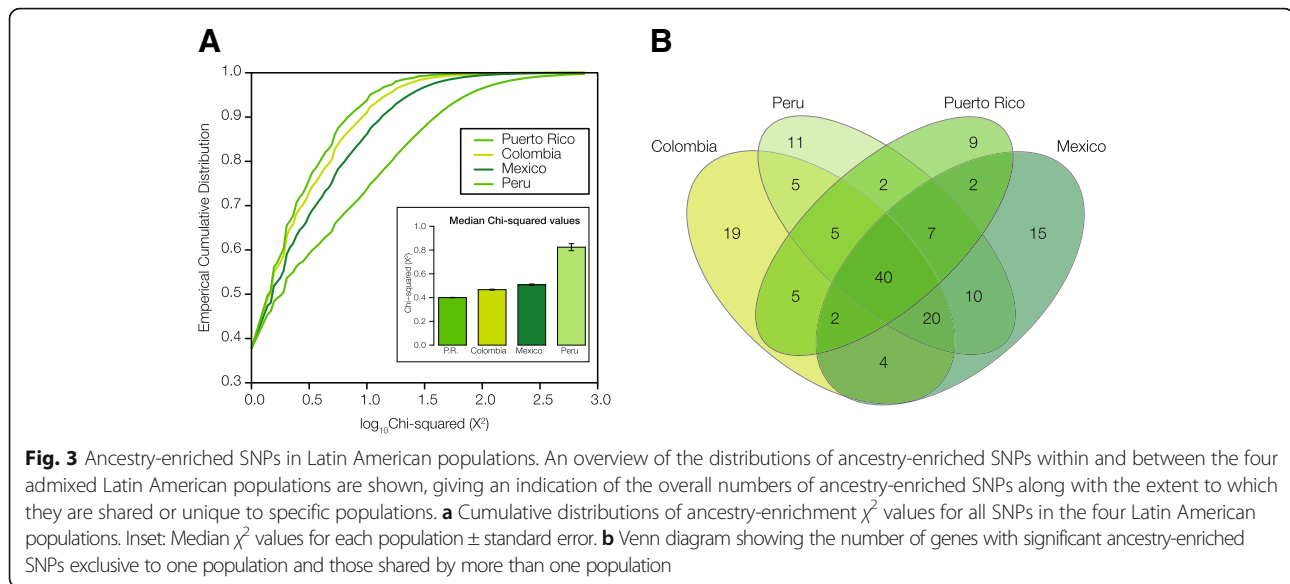


Fig. 2 Genetic ancestry and admixture in Latin American populations. The ancestry contributions of putative ancestral source populations to four modern, admixed Latin American populations are shown. **a** Triangle plots showing the relative ancestry contributions – African, European, Native American – to admixed individuals from four Latin American populations. **b** PCA plot showing the genetic relationships among individuals from admixed Latin American populations compared to putative ancestral source populations. Each population is bounded by a minimum spanning ellipse. **c** Admixture plots showing the fractions of African, Native American and European ancestry among admixed individuals from four Latin American populations. Each individual is represented as a column with the admixture fractions color coded as shown in the legend. **d** Violin plots showing distributions of ancestry fractions among individuals from four Latin American populations. **e** Pie charts showing the average ancestry values for each population next to its geographic location

for overrepresented health- or disease-related pathways and phenotypes (Fig. 4). This approach allowed us to identify the specific pathways and phenotypes that are most affected by ancestry-enriched SNPs. The presence of significantly overrepresented pathways and/or phenotypes in two or more populations was taken to indicate a higher likelihood of genetic ancestry effects on health and disease in modern Latin American populations.

A number of pathways and phenotypes have significantly overrepresented ancestry-enriched SNP genes in all four

populations (Fig. 4a). These include gene sets related to immunity (yellow) and metabolism (orange) as well as several disease-specific gene sets (blue) (Fig. 4b). Immune system pathways with ancestry-enriched SNPs include the cytokine receptor interaction, T cell receptor signaling, and antigen processing and presentation pathways. Implicated metabolic pathways include both drug and xenobiotic metabolism as well as steroid hormone biosynthesis. Diseases of note include several pathologies that are known to be found in high prevalence in Latin American populations: type I diabetes,



Alzheimer's disease and Leishmaniasis. A number of other signaling pathways were implicated by this analysis – calcium, MAPK, and GNRH signaling – pointing a role for ancestry-enriched SNPs in mediating human-environment interactions. Lists of all pathways that show significant enrichment of genes with mapped ancestry-enriched SNPs, for each admixed Latin American population – along with information regarding the overlapping genes and pathway enrichment statistical significance (FDR q -values) – are provided in Additional file 2: Table S2.

We focused on several notable examples of health- and disease-related pathways that were found to have significantly overrepresented ancestry-enriched SNP genes in all four Latin American populations (Fig. 5). For each of these pathways, and in each population studied, we performed additional meta-analyses of the sets of mapped ancestry-enriched SNPs in order to evaluate the pathway's overall ancestry enrichment. We also computed analogous overall observed versus expected allele frequencies for each pathway in all four populations. There are 15 genes from the Leishmaniasis immune response pathway with mapped ancestry-enriched SNPs, including a pair of Toll-like Receptor encoding genes as well as several interleukin genes (Fig. 5a). The meta-analysis for this pathway shows an overall ancestry-enrichment for all SNPs in each of the four populations analyzed here. Leishmaniasis is a parasitic disease with high prevalence in the tropics and subtropics including Latin America. Similar pathway-specific analysis revealed overall ancestry-enrichment for SNPs linked to drug metabolism (Fig. 5b), including multiple genes from the cytochrome P450 family, as well as the Jak-STAT signaling pathway, which is activated by cytokines as part of the innate immune response (Fig. 5c). The ancestry-enrichment observed for the drug metabolism pathway could represent an adaptation

based on detoxification linked to local dimensions of diet and environmental exposure in the New World.

Ancestry-specific expression quantitative trait loci (eQTL)

We explored the effects of ancestry-specific SNPs on gene regulation via expression quantitative trait loci (eQTL) analysis. eQTL are individual SNPs with genotype variants that are associated with gene expression levels; associations of this kind point to a role for SNP variants in gene regulation (e.g., via differential transcription factor binding affinities and/or allele specific expression levels) [33, 34]. To do this, we searched for ancestry-enriched SNPs that have ancestry-specific or shared genotype-expression associations. The first step of this analysis entailed the identification of the specific ancestry-components that predominantly contribute to the observed patterns of SNP ancestry-enrichment (see Materials and Methods). SNPs with highly asymmetric ancestry-enrichment patterns, i.e. predominant contributions from a single ancestral source population, were then chosen for eQTL analysis.

Using this approach, we found a number of cases of SNPs that show overrepresented African or European ancestry in modern Latin American populations and are also associated with ancestry-specific gene regulation (Fig. 6). A number of the genes regulated by ancestry-specific SNPs were found to play specific roles in the immune system and infectious disease responses. In particular, genes from both the innate and adaptive immune system were found to be regulated by ancestry-enriched SNPs that exert population-specific regulatory effects (Fig. 6a-d). For example, African ancestry-enriched SNPs were found to exert African-specific regulatory control over genes for both immunoglobulin receptors (PVR and TYROBP) and

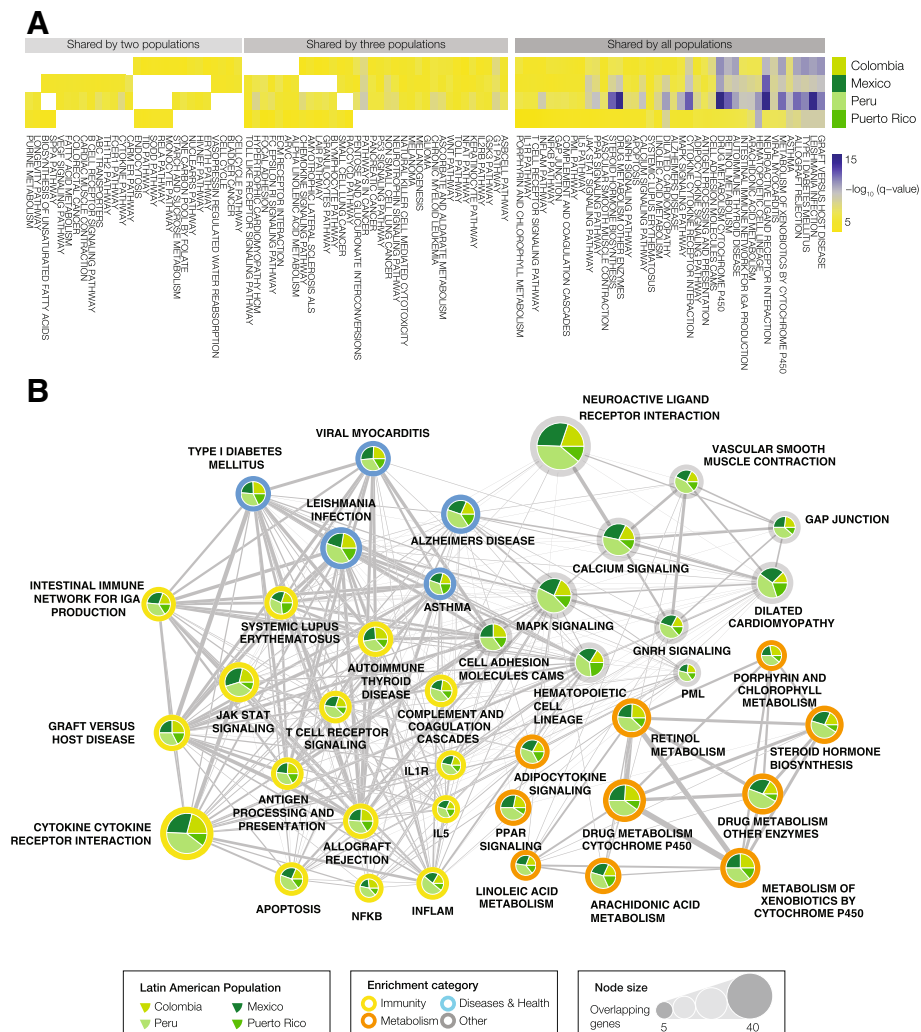


Fig. 4 Gene set enrichment analysis of ancestry-enriched SNP genes. Functionally coherent gene sets and pathways that are overrepresented with respect to ancestry-enriched SNPs are shown, giving an indication of the kinds of health-related phenotypes that have been shaped by genetic ancestry in the four admixed Latin American populations. **a** Heatmap showing significantly enriched functional gene sets (i.e., pathways and phenotypes) shared by two, three or all four Latin American populations. The cells are color coded by the log transformed statistical significance (FDR q -value) of gene set enrichment analysis results. **b** Network showing significantly enriched pathways and phenotypes shared by all four Latin American populations. Nodes sizes represent the number of ancestry-enriched genes in each set. Pie charts show how many genes in a given set are from each population. Color coding describes the functional enrichment category as shown in the legend

a downstream tyrosine kinase (ZAP70) involved in the adaptive immune response (Fig. 6e). Similarly, European ancestry-enriched SNPs were also found to act as population-specific eQTLs with regulatory effects on that were specific to the European populations. Analogous patterns of ancestry-specific SNP enrichment and gene regulatory control were found for genes involved in cytokine-receptor interactions, hematopoietic cell development, and cell-cell immunomodulatory interactions.

Discussion

Latin America has a unique genetic heritage with high levels of admixture from African, European, and Native

American ancestral source populations [9–11]. As such, the genome sequences of Latinos contain combinations of ancestry-specific genetic variants that never previously existed in the same genomic background. In other words, Latin American genomes represent a very recent evolutionary innovation in the long trajectory of human evolution and migration around the globe. Accordingly, the development and application of genomic approaches to healthcare in Latin America will require a deep understanding of the genetic ancestry and admixture profiles of Latin American populations. This issue is particularly pressing given the fact that the vast majority of studies aimed at uncovering genetic variants associated with health- and disease-related

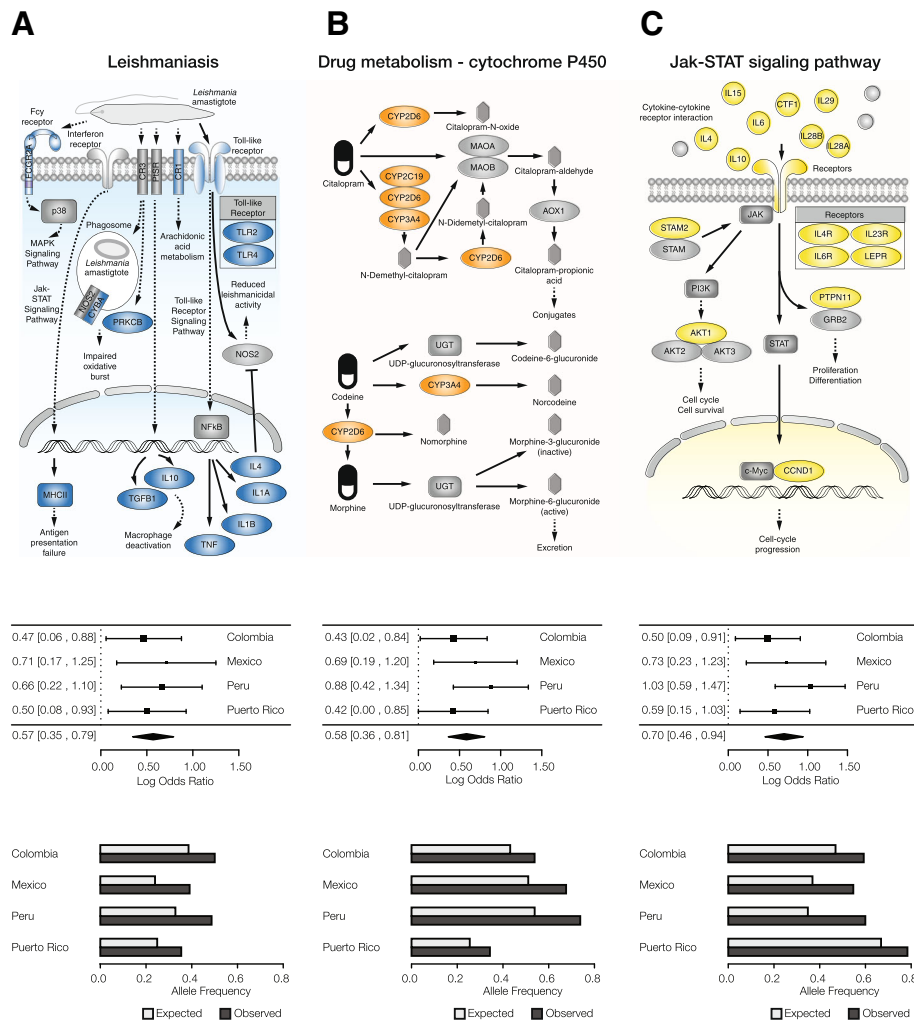


Fig. 5 Pathways with ancestry-enriched SNP genes in functional categories of interest. These results highlight examples of specific health-related functions and pathways that have been shaped by genetic ancestry in the four admixed Latin American populations. For each functional category, a pathway schematic is shown, indicating the pathway genes and their roles, along with meta-analysis results and observed versus expected SNP frequencies for each population. **a** Leishmaniasis, an example of a disease and health related pathway. **b** Cytochrome P450 drug metabolism, an example of a metabolism related pathway. **c** Jak-STAT signaling pathway, an example of an immune-related pathway

phenotypes have been conducted in populations with European ancestry [35, 36].

Here, we have tried to address this issue by relating patterns of genetic ancestry and admixture to health and disease determinants in Latin American genomes. To do so, we developed and applied a novel SNP-based approach to ancestry enrichment analysis. Our approach leverages information on the genetic ancestry of the modern Latin American populations to discover SNP variants that exist in a given population at higher frequencies than expected, i.e. ancestry-enriched SNPs. We found that specific sets of ancestry-enriched genetic variants, from each of the three ancestral source populations, have been preferentially retained in modern Latin American populations based on a variety of roles that they play in health and fitness. These

findings have relevance for the development of genomic approaches to healthcare, i.e. personalized or precision medicine, in Latin America.

Gene set enrichment analysis uncovered a number of immunity, metabolism, and disease-related pathways that are significantly overrepresented with respect to genes that contain ancestry-enriched SNPs (Figs. 4 and 5). These results suggest that these particular pathways, and their related phenotypes, could underlie population-specific health disparities in the four admixed Latin American populations studied here. They also give an indication that populations with particular ancestry profiles may be more or less disposed to some of these diseases and phenotypes; information of this kind could ultimately help to guide targeted health

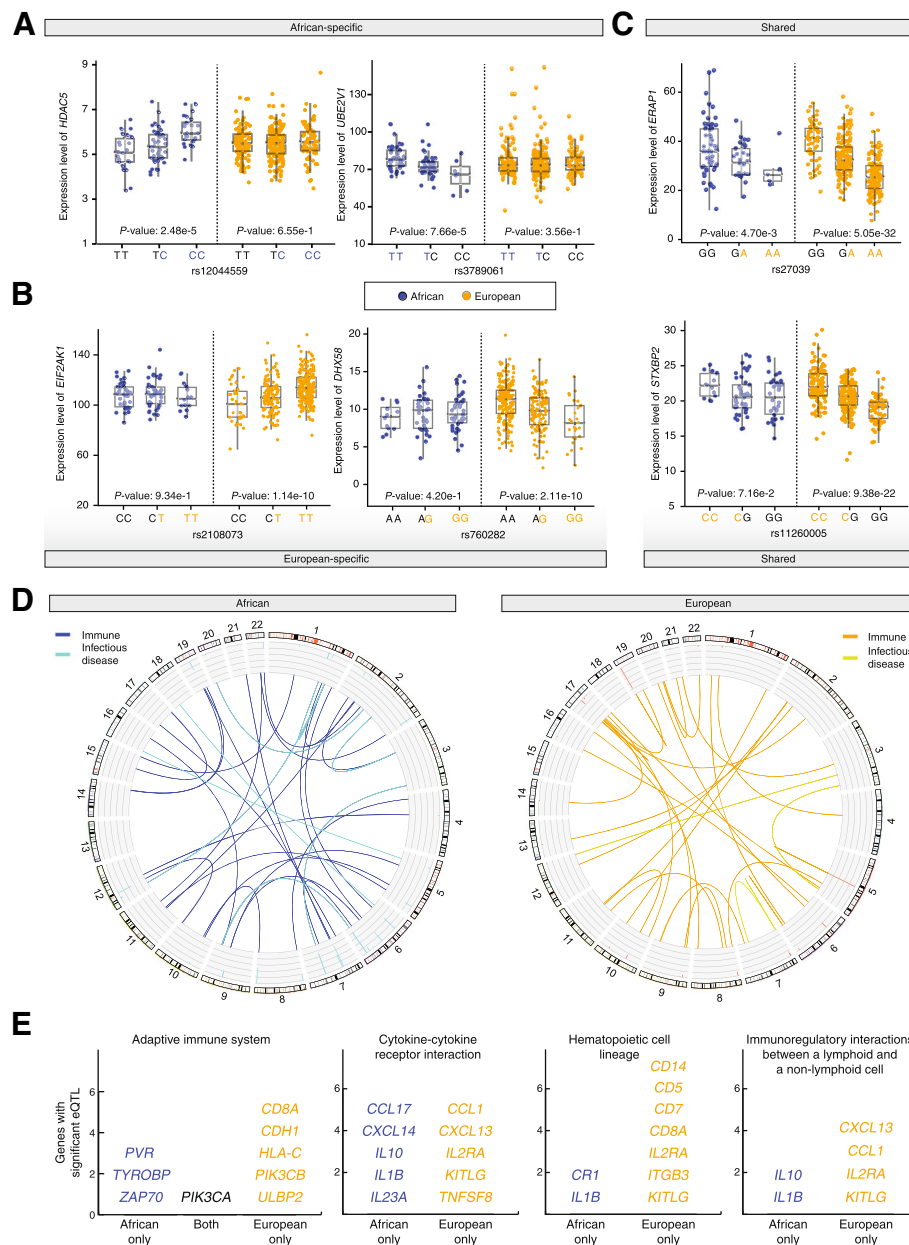


Fig. 6 Ancestry-specific effects on gene expression. These results give an indication of how ancestry-enriched SNPs can impact health-related phenotypes by virtue of their gene regulatory effects. SNP-by-ancestry interactions were characterized using an expression quantitative trait loci (eQTL) approach. Examples of **a** African-specific and **b** European-specific eQTL are shown along with **c** eQTL shared between populations. **d** eQTL related to immune system and infectious disease found in the African and European populations are shown in a CIRCOS plot with links indicated between eQTLs and their regulated genes. **e** Examples of immune-related pathways that include multiple eQTL-regulated genes for African and/or European populations

interventions. As these results represent basic research into the relationship between genetic ancestry and determinants of health, more clinically facing (translational) research will need to be done in order to precisely define the role of individual ancestry-enriched variants in disease etiology, prevention and treatment.

Expression quantitative trait loci (eQTL) analysis revealed ancestry-enriched SNPs in modern Latin American

populations that are associated with African- or European-specific patterns of gene regulation (Fig. 6). This includes SNPs that are associated with ancestry-specific regulation of genes involved in both the innate and adaptive immune systems as well as targeted infectious disease responses. These results underscore the relevance of gene regulatory control as an underlying driver of adaptive introgression in admixed populations.

One important caveat with respect to the interpretation of the results that we report is that they can only be taken to apply to the four specific populations analyzed here: Colombian in Medellin, Colombia (CLM), Mexican Ancestry in Los Angeles, California (MXL), Peruvian in Lima, Peru (PEL), and Puerto Rican in Puerto Rico (PUR). Given the diversity of Latin American populations, and in particular their distinct ancestry profiles, we should expect to see distinct ancestry enrichments for different countries in the region, such as Argentina, Chile, Brazil etc. This caveat not only applies to different countries but also applies to different populations within the same country. Colombia, for instance, is an extremely diverse country with populations from different regions that show very distinct ancestry profiles [9]. The population of Colombia analyzed here is from Medellín in the state of Antioquia, and this particular population shows averages of 64% European ancestry, 29% Native American ancestry, and 7% African ancestry. However, we have previously shown that the population from the neighboring state of Chocó has a totally distinct ancestry profile with 76% African ancestry, 13% European ancestry, and 11% Native American ancestry [37–39]. Accordingly, results from the analysis of the population from Medellín cannot be taken to represent the entire country of Colombia. Clearly, a deeper understanding of the relationship between genetic ancestry and health determinants in Latin America will require analysis of many more populations within and between the region's countries.

Additional files

Additional file 1: Table S1. Lists of all SNPs that show significant ancestry-enrichment. For each admixed Latin American population, all ancestry-enriched SNPs ($q < 0.05$) are given along with their meta-information, observed and expected reference and alternate allele frequencies, and significance values. The allele (reference or alternate) and enriched ancestry (African, European or Native American) are designated for each ancestry-enriched SNP. (XLSX 177 kb)

Additional file 2: Table S2. Lists of pathways that show significant enrichment of genes with mapped ancestry-enriched SNPs for each admixed Latin American population. For each KEGG pathway with significant enrichment of genes with mapped ancestry-enriched SNPs in at least one of the four populations, the overlapping genes and FDR q -values are given. (XLSX 29.6 kb)

Abbreviations

1KGP: 1000 genomes project; DNA-seq: DNA sequencing; eQTL: Expression quantitative trait locus; GEUVEDIS: Genetic European Variation in Health and Disease; HGDP: Human Genome Diversity Project; RNA-seq: RNA sequencing; SNP: Single nucleotide polymorphism

Funding

ETN, LR and IKJ were supported by Georgia Institute of Technology Bioinformatics Graduate Program, a Georgia Tech Global FIRE award, and the IHRC-Georgia Tech Applied Bioinformatics Laboratory (ABiL). LW was supported by Georgia Institute of Technology Bioinformatics Graduate Program. AVA was supported by Fulbright Colombia. This work was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine and National Center for Biotechnology

Information (NIH, NLM, NCBI) ZIA LM082713–05. The publication cost of this article was funded by a Georgia Tech Global FIRE award.

Availability of data and materials

All data supporting our findings can be accessed via the 1000 Genomes Project website <http://www.internationalgenome.org/> and the Human Genome Diversity Project website <http://www.hagsc.org/hgdp/>

About this supplement

This article has been published as part of *BMC Genomics Volume 19 Supplement 8, 2018: Selected articles from the IV Colombian Congress on Bioinformatics and Computational Biology & VIII International Conference on Bioinformatics SolBio 2017*. The full contents of the supplement are available online at <https://bmcbgenomics.biomedcentral.com/articles/supplements/volume-19-supplement-8>.

Authors' contributions

ETN and LR performed SNP ancestry-enrichment analysis and gene mapping. LW performed GSEA and eQTL analyses. ABC performed the genetic ancestry analyses. ETN, LW, LR and IKJ wrote the manuscript. All authors read and approved the final manuscript. IKJ, LMR and AVA conceived the study and topic. IKJ supervised all aspects of the project.

Ethics approval and consent to participate

This study uses publicly available, unrestricted, de-identified human genome sequence variant data from the 1000 Genomes Project and the Human Genome Diversity Project and therefore does not require ethics committee approval.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA 30332, USA. ²PanAmerican Bioinformatics Institute, Cali, Valle del Cauca, Colombia. ³IHRC-Georgia Tech Applied Bioinformatics Laboratory (ABiL), Atlanta, GA, USA. ⁴National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA. ⁵Biomedical Research Institute, Faculty of Health, Universidad Libre-Seccional Cali, Cali, Valle del Cauca, Colombia.

Published: 11 December 2018

References

- Schraiber JG, Akey JM. Methods and models for unravelling human evolutionary history. *Nat Rev Genet.* 2015;16(12):727–40.
- Veeramah KR, Hammer MF. The impact of whole-genome sequencing on the reconstruction of human population history. *Nat Rev Genet.* 2014;15(3):149–62.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovskiy LA, Feldman MW. Genetic structure of human populations. *Science.* 2002;298(5602):2381–5.
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, et al. Worldwide human relationships inferred from genome-wide patterns of variation. *Science.* 2008;319(5866):1100–4.
- Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. A genetic atlas of human admixture history. *Science.* 2014;343(6172):747–51.
- Mann CC. 1493: uncovering the new world Columbus created. New York: Alfred a Knopf Incorporated; 2011.
- Crosby AW. The Columbian exchange: biological and cultural consequences of 1492, vol. 2. Westport: Greenwood Publishing Group; 2003.
- Jordan IK. The Columbian exchange as a source of adaptive introgression in human populations. *Biol Direct.* 2016;11(1):17.

9. Ruiz-Linares A, Adhikari K, Acuna-Alonso V, Quinto-Sanchez M, Jaramillo C, Arias W, Fuentes M, Pizarro M, Everardo P, de Avila F, et al. Admixture in Latin America: geographic structure, phenotypic diversity and self-perception of ancestry based on 7,342 individuals. *PLoS Genet.* 2014;10(9):e1004572.
10. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, Poletti G, Mazzotti G, Hill K, Hurtado AM, et al. Geographic patterns of genome admixture in Latin American mestizos. *PLoS Genet.* 2008;4(3):e1000037.
11. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, Hammer M, Bustamante CD, Ostrer H. Colloquium paper: genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proc Natl Acad Sci U S A.* 2010;107(Suppl 2):8954–61.
12. Tang H, Choudhry S, Mei R, Morgan M, Rodriguez-Cintron W, Burchard EG, Risch NJ. Recent genetic selection in the ancestral admixture of Puerto Ricans. *Am J Hum Genet.* 2007;81(3):626–33.
13. Guan Y. Detecting structure of haplotypes and local ancestry. *Genetics.* 2014;196(3):625–42.
14. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo JM, Wambebe C, Tishkoff SA, et al. Genome-wide patterns of population structure and admixture in west Africans and African Americans. *Proc Natl Acad Sci U S A.* 2010;107(2):786–91.
15. Jin W, Xu S, Wang H, Yu Y, Shen Y, Wu B, Jin L. Genome-wide detection of natural selection in African Americans pre- and post-admixture. *Genome Res.* 2012;22(3):519–27.
16. Basu A, Tang H, Zhu X, Gu CC, Hanis C, Boerwinkle E, Risch N. Genome-wide distribution of ancestry in Mexican Americans. *Hum Genet.* 2008;124(3):207–14.
17. Zhou Q, Zhao L, Guan Y. Strong selection at MHC in Mexicans since admixture. *PLoS Genet.* 2016;12(2):e1005847.
18. Rishishwar L, Conley AB, Wigington CH, Wang L, Valderrama-Aguirre A, Jordan IK. Ancestry, admixture and fitness in Colombian genomes. *Sci Rep.* 2015;5:12376.
19. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. A global reference for human genetic variation. *Nature.* 2015;526(7571):68–74.
20. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK. Rising to the challenge of larger and richer datasets. *Gigascience.* 2015;4:7.
21. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods.* 2013;10(1):5–6.
22. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 2009;19(9):1655–64.
23. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am J Hum Genet.* 2013;93(2):278–88.
24. O'Leary NA, Wright MW, Brister JR, Ciuffo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016;44(D1):D733–45.
25. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A.* 2005;102(43):15545–50.
26. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43(7):e47.
27. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature.* 2013;501(7468):506–11.
28. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, et al. Ensembl 2013. *Nucleic Acids Res.* 2013; 41(Database issue):D48–55.
29. t Hoen PA, Friedlander MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, Laros JF, Buermans HP, Karlberg O, Brannvall M, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. *Nat Biotechnol.* 2013;31(11):1015–22.
30. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc.* 2012;7(3):500–7.
31. Shabalín AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* 2012;28(10):1353–8.
32. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw.* 2010;36(3):1–48.
33. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* 2008;24(8):408–15.
34. Gibson G, Powell JE, Marigorta UM. Expression quantitative trait locus analysis for translational medicine. *Genome Med.* 2015;7(1):60.
35. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature.* 2016; 538(7624):161–4.
36. Need AC, Goldstein DB. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* 2009;25(11):489–94.
37. Conley AB, Rishishwar L, Norris ET, Valderrama-Aguirre A, Marino-Ramirez L, Medina-Rivas MA, Jordan IK. A comparative analysis of genetic ancestry and admixture in the Colombian populations of Choco and Medellín. *G3 (Bethesda).* 2017;7(10):3435–47.
38. Chande AT, Rowell J, Rishishwar L, Conley AB, Norris ET, Valderrama-Aguirre A, Medina-Rivas MA, Jordan IK. Influence of genetic ancestry and socioeconomic status on type 2 diabetes in the diverse Colombian populations of Choco and Antioquia. *Sci Rep.* 2017;7(1):17127.
39. Medina-Rivas MA, Norris ET, Rishishwar L, Conley AB, Medrano-Trochez C, Valderrama-Aguirre A, Vannberg FO, Marino-Ramirez L, Jordan IK. Choco, Colombia: a hotspot of human biodiversity. *Rev Biodivers Neotrop.* 2016; 6(1):45–54.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

