FAPESP / Thematic Project
Human Population Genomics: a view from admixed populations

Coordinator
Prof. Dr. Diogo Meyer - University of São Paulo (USP)

Principal Investigators
Prof. Dr. Diogo Meyer - University of São Paulo (USP)
Prof. Dr. Erick C. Castelli - São Paulo State University (Unesp)
Profa. Dra. Carla Luana Dinardo - Fundação Pró-Sangue Hemocentro de São Paulo

Associate Investigators
Profa. Dra. Regina Célia Mingroni Netto - University of São Paulo (USP)
Prof. Dr. Michel S. Naslavsky - University of São Paulo (USP)
Prof. Dr. Celso T Mendes-Junior - University of São Paulo (USP)
Dra. Cibelle Masotti - Hospital Sírio-Libanês, São Paulo

International collaborators
Dr. Bruce Spencer Weir - University of Washington, USA.
Dr. Nicolas Vince - Université de Nantes, France.

Admixture among populations produces genomes that are a mosaic of different ancestries. This mixing can originate novel genetic combinations, which are the potential substrate for adaptive evolution and can also modulate phenotypes, including diseases. As a consequence, understanding the genetic consequences of admixture is of both evolutionary and biomedical interest. Here, we investigate the genetics of admixed populations with an emphasis on two aspects: the role of admixture on Brazilian populations, and its impact in two genomic regions involved in immunity and human health: the Major Histocompatibility Complex (MHC) and the Leukocyte Receptor Complex (LRC). Our research is structured around three main themes. First, we propose to develop bioinformatic tools to solve the problems of alignment bias and genotype calling which are common in the MHC and LRC regions. We will then apply these tools to large samples of sequenced genomes, including admixed Brazilians, documenting the diversity of these genomic regions and discussing the effects of variation to gene function and transplantation. Second, we will use the newly generated data to investigate how natural selection has shaped differences among populations in the MHC and LRC, and will test whether recent admixture increases diversity in these immunity-related genes. Third, we will investigate how admixture contributes to the prevalence of deleterious and disease-causing alleles throughout the genome. We will specifically investigate the severity of sickle-cell disease, the most common Mendelian disease in Brazil, which was originally associated with African ancestry genomes but has become increasingly present in genomes with extensive non-African ancestry. Our research will contribute to an understanding of how admixture is shaping genomic diversity in Brazilians and will highlight the consequences of this process to genes involved in immunity. Our findings will be relevant to the understanding of evolutionary processes and biomedical applications, and will broaden our understanding of genetic diversity in understudied populations.

FAPESP / Projeto temático

Genômica populacional humana: uma perspectiva de populações miscigenadas

Coordenador
Prof. Dr. Diogo Meyer - University of São Paulo (USP)

Pesquisadores principais
Prof. Dr. Diogo Meyer - University of São Paulo (USP)
Prof. Dr. Erick C. Castelli - São Paulo State University (Unesp)
Profa. Dra. Carla Luana Dinardo - Fundação Pró-Sangue Hemocentro de São Paulo

Pesquisadores associados
Profa. Dra. Regina Célia Mingroni Netto - University of São Paulo (USP)
Prof. Dr. Michel S. Naslavsky - University of São Paulo (USP)
Prof. Dr. Celso T Mendes-Junior - University of São Paulo (USP)
Dra. Cibelle Masotti - Hospital Sírio-Libanês, São Paulo

Colaboradores internacionais
Dr. Bruce Spencer Weir - University of Washington, USA.
Dr. Nicolas Vince - Université de Nantes, France.

A miscigenação entre populações produz genomas que são um mosaico de diferentes ancestralidades, podendo originar novas combinações genotípicas e afetar a diversidade genética. Essa diversidade é um substrato para a evolução adaptativa e pode modular fenótipos, inclusive de doenças. Dessa forma, compreender o impacto genético da miscigenação em populações humanas é de interesse evolutivo e biomédico. Neste projeto, investigaremos as consequências da miscigenação enfatizando dois aspectos: o papel da miscigenação em populações brasileiras e seu impacto em duas regiões genômicas envolvidas na imunidade: o Complexo Principal de Histocompatibilidade (MHC) e o Complexo de Receptores de Leucócitos (LRC). Nossa pesquisa se estrutura em três temas principais. Primeiro, desenvolveremos ferramentas bioinformáticas capazes de resolver problemas que tipicamente afetam a análise de genes do MHC e LRC, que são consequência da extensa paralogia e alto polimorfismo desses genes. Aplicando nossas ferramentas, faremos um levantamento preciso da variação genética nesses genes em brasileiros e outras populações miscigenadas, discutindo os efeitos da variação para a função dos genes e para transplantes. Segundo, usaremos os dados recém-gerados para investigar como a seleção natural moldou as diferenças entre as populações no MHC e LRC e como a mistura recente contribui para a diversidade desses genes. Terceiro, iremos estender nossa pesquisa das consequências evolutivas da miscigenação além dos genes do MHC e LRC, investigando como a mistura contribui para a prevalência de alelos deletérios e causadores de doenças. Investigaremos especificamente como a miscigenação afeta a severidade da anemia falciforme, a doença mendeliana mais comum no Brasil, que tem se tornado cada vez mais presente em genomas com extensa ancestralidade não africana. Nossa pesquisa contribuirá para o entendimento de como a miscigenação está moldando a diversidade genômica dos brasileiros, um grupo ainda sub-representado em estudos genômicos, e destacará as consequências da miscigenação para os genes envolvidos na imunidade.

A. Statement of the problem

Human genetic variation

Two human haploid genomes differ at approximately 0.1% of all nucleotide positions [1]. These differences correspond to up to 3 million genetic differences that can contribute to the phenotypic variation among individuals. Understanding how these genetic differences are distributed across the genome, how they vary among populations and geographic regions, and which evolutionary processes shape these observations constitute the central questions of population genetics [2]. Patterns of variation are informative about the historical movements of populations [3], the intensity and timing of population size changes [4], and the occupation of new territories [5]. Genomic data also allows us to understand how natural selection drives adaptive change [6], how deleterious mutations can build up in the genome [7], and how demographic and selective processes interact [8].

Admixed populations carry alleles originally present in distinct geographic regions, but combined into novel genotypic combinations, bringing additional challenges and opportunities to the study of human variation. The dynamics of microevolutionary processes such as positive selection, the purging of deleterious variants, and the extent of linkage disequilibrium need to be modeled to account for the impact of the admixture process. This requires reliable inferences about the contribution of parental populations, the timing of admixture events, and the interactions between admixture and natural selection [9].

Admixed populations can provide insight into disease associations, identifying variants that contribute to disease, and can be used to tease apart the contribution of local ancestry (i.e., the geographic origin) of putative causal variants from that of the genome-wide ancestry [10]. A powerful approach to identify genes associated with complex phenotypes is admixture mapping, which searches for genomic regions associated with a phenotype and which present an excess of a geographic ancestry [11]. Admixture mapping can also be used to identify recent natural selection, which increases the proportion of an ancestry that carries advantageous variants [12].

Here, we propose to investigate genetic variability at a genomic scale in samples of admixed individuals, with an emphasis on Brazil. Analysis of genomic data from admixed Brazilians will provide insights into how the combination of African, European, and Native American ancestry have collectively shaped existing patterns of genetic diversity [13]; how this diversity influences the magnitude of genetic load; how natural selection has influenced the genome of Brazilians; how admixture has shaped genetic diversity at specific loci of key importance in the immune response; and how knowledge about Brazil's history of admixture can contribute to biomedical research. Brazil is now on the cusp of an expansion in the number of population-level surveys of genetic diversity [14, 15, 16] making these investigations timely. We will study samples from various regions of Brazil (with different admixtures histories), and compare our findings to those for other publicly available admixed populations, which likewise have distinct demographic histories.

We will use two complementary strategies in our study of admixed Brazilians. First, we will investigate genome-wide patterns of variation, which are informative about genetic ancestry, demographic dynamics, and the accumulation of deleterious variants. Second, we will carry out an in-depth investigation of two specific genomic regions, the Leukocyte Receptor Complex (LRC) and the Major Histocompatibility Complex (MHC), which harbor genes that play a critical role in the adaptive immune response.

The MHC region and HLA genes

The MHC is a 5Mb region on chromosome 6 that harbors over 200 loci, most of which are involved in immunity. Among these loci, the Human Leukocyte Antigen (HLA) genes code for proteins that participate in the antigen processing and presentation pathway. Classical HLA molecules bind peptides of intra and extracellular origin and present these on the cell surface to T-cell receptors, triggering an immune response if the peptide-HLA complex is recognized as being "non-self" [17]. The association between HLA molecules and peptides requires interactions at specific amino-acids, and the individual's genotype at HLA genes will therefore determine the set of peptides to which will trigger a response. This results in a coevolutionary dynamic, where host-pathogen interactions drive recurrent changes in both HLA genes and pathogen proteins [18,19]. The history of such coevolutionary dynamics is visible in the extreme polymorphism of HLA genes, the high density of GWAS hits [20] for autoimmune, psychiatric, and infectious diseases [21], and the strongest genome-wide evidence of balancing selection [22].

Important questions remain to be addressed regarding the evolution of genes in the MHC region. These include an understanding of the timescale and mode of selection, the influence of admixture and selection on polymorphism and population structure, and an understanding of the genetic control of HLA expression. Population genetic analyses of HLA genes also have implications for medicine. For instance, population structure may influence the chances of finding compatible donors for hematopoietic cell transplantation (see section 2.3), and HLA data on admixed samples may improve the quality of imputation (see section 1.3), and aid the interpretation of disease association studies. More generally, quantifying HLA variation in admixed populations and understudied populations can uncover as yet undescribed variants (see section 1.1), which are medically important. We will also explore mechanisms underlying HLA expression levels and alternative splicing, processes that are also relevant to understanding disease (section 1.2).

However, studying variation in the MHC, and of HLA genes in particular, presents methodological challenges. Standard bioinformatic pipelines fail to provide accurate genotypes and haplotypes in this region because HLA genes are extremely polymorphic, are part of a multigene family whose genes have high sequence similarity, and are in high linkage disequilibrium. As a consequence, analysis of the MHC and of HLA genes requires specialized bioinformatic pipelines, which we propose to develop.

The LRC and KIR genes

The Leukocyte Receptor Complex (LRC), located at 19q13.4 [23], is a gene cluster encoding receptors of Natural Killer (NK) cells and molecules related to the immunoglobulin superfamily, including killer-cell Ig-like-receptors (KIR). When these receptors bind their ligands on the cell surface, they can send inhibitory or activating signals, depending on the combination ligand/receptor. The molecules encoded in the LRC are associated with important tissue homeostatic function, with the innate immune response during disturbing situations [24], and with the activation and modulation of the adaptive immune responses [24,25]. The KIR receptors are expressed in NK cells and some cytotoxic T cells [26,27], and they all bind HLA molecules. KIR genes are also highly polymorphic at the sequence and structural (copy number) levels. While some KIR genes present inhibitory properties (e.g., KIR3DL2), others may activate NK cell response (e.g., KIR2DS4).

The polymorphic nature of HLA and KIR systems, associated with the fact that different KIRs recognize different HLA molecules, results in substantial heterogeneity in response among individuals with different HLA/KIR genotypes, with implications in viral infections [28,29], autoimmune diseases [30,31], and cancer [31–33]. Moreover, HLA and KIR compatibility may influence the outcome of transplants [34]. The genetic diversity of both systems and their interaction indicates an ongoing co-evolution between them [35]. For instance, in sub-Saharan Africans, each individual has a unique compound genotype of HLA and KIR, probably driven by balancing selection [36]. Globally, there is a negative correlation between the presence of activating KIR genes and their corresponding HLA ligand groups across populations [37].

Motivation

There is an increasing awareness that human genetics has not adequately studied human diversity, in particular that of admixed populations of South America [38,39]. The fact that Brazil has initiated, and is likely to further develop, large scale genome sequencing projects, adds urgency to the need to develop analytical tools and to address theoretical questions regarding evolutionary and biomedical consequences of admixture.

We are motivated by challenges on two scales. First, in-depth studies of genes involved in immunity, of medical and evolutionary interest, provide crucial insight into how humans have adapted to their environments, and aid in the development of biomedical resources. Accordingly, we will investigate the MHC and LRC regions, developing bioinformatic methods to address the shortcomings of existing resources.

At a second level, we are motivated by the need to understand how admixture is shaping genetic diversity of Brazilians. Investigation about the buildup of deleterious mutations in admixed populations, the role of admixture in decoupling African alleles from a genomic background of predominantly African ancestry, will provide answers about how admixture is shaping the genetic identity of admixed populations in general, and Brazilians in particular.

B. Expected Results for three projects

Our study of the population genomics of admixed populations is structured in three main projects and an outreach initiative.

The first project provides a state-of-the-art survey of genetic variation in the MHC and LRC regions at the genomic level. This requires developing bioinformatic resources to process NGS data to make calls for the MHC and LRC regions. We will generate a map of MHC and LRC variation at a resolution and scale that is currently not available for admixed Brazilians. Variability at HLA loci will also be interpreted in light of the challenges of identifying donors for hematopoietic stem cell transplantation in an admixed population. In addition, we will develop new panels for HLA imputation, a valuable resource for the community.

In the second project, we use the MHC and LCR data from the first project to address evolutionary and population genetic questions. We first address a simple question: what is the impact of admixture on the genetic diversity of MHC and LRC loci in admixed Brazilians? We will also explore the patterns of kinship and population structure at MHC and LRC loc, and use these to make inferences about how selection has shaped similarities and differences among populations in these genomic regions. Finally, we will investigate how admixture has contributed to the adaptive evolution of these genomic regions, with particular emphasis on understanding why alleles of African ancestry are overrepresented among admixed Brazilians in the MHC region when compared to the rest of the genome.

In the third project, we explore the consequences of admixture to genetic variation and evolutionary processes in regions beyond the MHC and LRC. First, we will investigate how admixture influences the magnitude of genetic load (i.e., the burden of deleterious alleles present per genome in a population). Next, we will tackle how admixture and recombination creating genomes which are a mosaic of ancestries, in such a way that phenotypes originally associated with African ancestry are increasingly found in non-African backgrounds. This is directly relevant to understanding how Sickle Cell Disease (SCD) has become a condition no longer exclusively associated with African individuals. Finally, we will investigate how admixture is informative about disease phenotypes, using admixture mapping approaches to understand the genetic basis of severity of sickle-cell disease.

Project 1. The genomics and transcriptomics of the MHC and LRC regions: structure, variation, and function

Understanding the immune phenotypes associated with infectious disease susceptibility and autoimmunity requires a detailed understanding of genetic variation at the MHC and LRC loci. For admixed populations in general and understudied ones in particular, a survey of genetic variation can uncover unique patterns of linkage disequilibrium, new variants and haplotypes.

Existing surveys of HLA variation in Brazil [reviewed in [40,41]] provide a broad picture of variation but are limited by the level of genotyping resolution used, which often lacks intronic and regulatory sequences and the absence of genome-wide data for the same individuals. Genome-wide data is critical to place HLA and KIR variation in the context of other potentially epistatically interacting variants, and polymorphisms that are informative about ancestry and demography. We have started to survey genomic datasets in Brazilians for HLA class I polymorphism [14], but there is still a lack of extensive surveys of MHC variation built on NGS data of large population samples. In addition, few studies address KIR diversity in Brazil, the vast majority surveying only the presence or absence of specific KIR genes.

Population genomic studies based on NGS use well established computational methods to align short reads to reference genomes, evaluate data quality, infer the reliability of genotype calls [42], and use population genetic methods (such as Hardy-Weinberg testing) as sanity checks [43]. However, the unusually high polymorphism of genes from the MHC and LRC, and their extensive paralogy, leads to inaccuracies in the alignment of reads to a reference genome [44], causing extensive genotyping errors and misestimation of expression levels. These issues call for methods that are specifically designed to deal with the features of MHC and LRC polymorphism.

With the advent of the NGS technology, raw DNA and RNA sequence data are available for thousands of samples in public databases. However, in most cases, the analysis of HLA and KIR genes is unreliable [44] or absent because suitable bioinformatic tools were not used. Even newly released datasets such as the high-coverage sequencing of the 1000 Genomes samples eliminated large regions encompassing HLA and KIR genes, because of technical difficulties in making calls [45]. Given the importance of (a) well-curated HLA and KIR data for large population samples, (b) the need to extend investigations to admixed populations, and (c) the importance of placing HLA and KIR variation in the context of genome-wide variation, we will develop and apply computational pipelines to reconstruct sequences at the MHC and LRC loci from WGS data. Our protocols will be tailored to obtain reliable and unbiased genotypes and haplotypes and discover novel variants.

1.1 Generation of molecular-level data on the MHC and LCR loci using novel bioinformatic approaches: genomics and transcriptomics

Standard pipelines for the alignment of short reads rely on a single reference genome, which does not cover the extensive polymorphism of MHC and LRC genes. As a consequence, short reads with multiple nucleotide differences compared to the reference genome fail to align (being classified as unmapped reads) leading to genotyping errors. The extensive paralogy of MHC and LRC genes is another challenge since reads from one gene align to more than one locus or only to the wrong gene. These problems affect genotyping and haplotyping in WGS and exome studies, as well as the estimation of expression levels in RNAseq studies. Thus, it is essential to develop a bioinformatic pipeline that addresses these challenges.

Research in Diogo Meyer's and Erick Castelli's groups has addressed these difficulties and provided solutions. DM's group developed a strategy to map RNAseq reads to personalized genomes, minimizing mapping bias and allowing accurate expression level estimates [46,47]. Over the past eight years, ECC has worked on computational strategies to obtain reliable HLA SNP and haplotype calls from NGS data. In 2018, ECC developed *hla-mapper* [48], a program which uses a multi-referenced genome to correct alignments for HLA class I genes, together with pipelines for genotype calling, variant refinement, and haplotyping of HLA class I genes and some KIR genes. With this strategy, the ECC team surveyed HLA class I genetic diversity in samples from Brazil and Benin [49–51,14].

We will extend the hla-mapper strategy to genotype other genes within the MHC and LRC, and to support large datasets and high-capacity computers. The LRC loci are particularly challenging to genotype due to copy number variation and the presence of large introns, often containing large indels. The strategies to achieve this goal will be discussed later (section C). We will publicly share programs and pipelines to estimate HLA expression, and assemble sequences from MHC and LRC genes from WGS, exome, and amplicon sequencing, as we have done before (see https://github.com/erickcastelli/, http://www.castelli-lab.net/Downloads.html and https://github.com/genevol-usp/HLApers).

We will generate a state-of-the-art survey of genetic variation at MHC and LRC genes in Brazilians at various levels of resolution, ranging from genotypes of individual SNPs and indels, to coding region haplotypes (which define "HLA alleles"), to extended haplotypes, defined by introns, promoters and UTRs, and intergenic sequences. The WGS-based approaches will allow novel HLA and KIR variants to be placed in the context of genomic coordinates.

Finally, the same issues that jeopardize HLA genotyping using standard workflows also hinder HLA expression estimation in RNA-seq studies. We will update the hla-mapper strategy to evaluate RNA-seq data. This approach will allow SNP and haplotype-level genotyping in RNA-seq data and more accurate expression estimates (including isoform diversity). Together with the estimates provided by HLApers developed by DM's group [46], this data will allow the mapping of eQTLs and estimation of heritability of HLA expression.

1.2 The variability of MHC and LRC genes in Brazilians and other admixed populations: insights regarding diversity and function

We will investigate genetic variation and genomic organization of MHC and LRC genes in Brazilians and other admixed populations at the molecular level, providing a detailed assessment of nucleotide level variation in all genes and intergenic sequences. We will provide an in-depth map of regulatory diversity and putative effect of polymorphism on TFBS, the relationship between neighboring genes and haplotype diversity, and the genetic variability of all genes across populations, including less studied genes such as *TAP1*, *TAP2*, *MICA*, *MICB*. We will also quantify the degree to which current accounts of HLA and KIR variation (based on traditional non-NGS typing of a subset of exons and public WGS datasets not applying methods to correct HLA and KIR genotyping) underestimate the true genetic diversity of these loci.

Our study will provide a resource for the scientific community (complete sequences, genotype calls, allele calls, encoded proteins, functional sites), with a thorough survey of genomic variation in the MHC and LRC. Our survey of polymorphism in MHC and LRC loci will provide information relevant to understanding the function of these genes. This includes uncovering protein-level variation that can influence peptide binding and receptor-ligand interactions; predicting the effect of variation on binding of transcription factors and microRNAs; compiling a curated resource with the frequency of clinically-relevant alleles in different geographic regions of Brazil.

An additional layer of information concerns the measure of expression levels for these genes. Our previous work has shown that appropriate processing of RNA-seq assays allows reliable estimates of the expression levels of HLA class I genes, with identification of expression of individual alleles [46,47]. Here, we will use the hla-mapper and HLApers approaches, extended as described above, to obtain new estimates of HLA expression controlling for isoform diversity. We will address open questions concerning HLA expression: To what degree does HLA expression vary among alleles? Does the expression level of a specific HLA allele depend on the surrounding sequences? Do structural variants influence HLA expression levels? What are the causal polymorphisms driving variation in HLA expression? Do intronic variants influence splicing patterns and isoform expression?

1.3 Applications of data on HLA and KIR diversity: imputation and disease-association

Population-level surveys of HLA and KIR variation in datasets with WGS are a valuable resource for biomedical research. The joint HLA and dense SNP data in flanking regions surrounding HLA loci provide a resource to improve the quality of HLA imputation. Imputation explores the existence of linkage disequilibrium between HLA alleles and flanking SNPs outside the loci to make calls for datasets based on genotyping arrays [52]. Arrays are still widely used in association studies, and imputation allows reliable HLA calls to be made without the costs of direct HLA typing. However, the Brazilian population (as is the case for other admixed groups) presents specific challenges for imputation: a large number of poorly documented African and Native American alleles, the presence of unusual genotypic combinations due to admixture, and the presence of previously described alleles at frequencies that differ substantially from other world regions. The development of reference panels that are specifically designed for admixed populations can therefore improve the accuracy of imputation [52,53].

As part of our efforts to document HLA variation for Brazilians, we will develop new reference panels, enriching the representation of admixed Brazilians and thus improving accuracy. This will be carried out as part of an ongoing international initiative, the SNP-HLA Reference Consortium - SHLARC [54], which we are members of. Our analysis will include developing the HLA and SNP repositories and testing imputation accuracy in different populations, particularly Brazilians.

The in-depth characterization of the MHC and LRC genes among Brazilians will also allow us to use this data as a resource for disease-association studies. Since we will characterize every individual's genomic and local ancestries, SNPs and haplotypes across the MHC and LRC, we can use this dataset as population-based controls and compare allele frequencies with groups of individuals presenting specific phenotypes, controlling for population stratification. For example, we have compared allele frequencies for genes in the MHC and LRC in groups of individuals presenting symptomatic COVID-19 and their household exposed but uninfected partners, detecting important associations in genes from the MHC (*MICA* and *MICB*) and LRC (*LILRB1* and *LILRB2*) with infection resistance. We used the HLA data from a population-based Brazilian sample to validate the findings and to quantify how much the frequencies observed in both groups deviate from that expected in the general population [55].

Project 2. Population Genetics of the MHC and LRC in admixed populations

Using NGS data for large samples, we will investigate how admixture has shaped genetic diversity of MHC and LRC loci. Our research will focus on three questions: (1) how has admixture contributed to adaptation of genes in the MHC and LRC? (2) How has admixture influenced the degree of relatedness of individuals within and between populations at loci in the MHC and LRC; (3) What is the effect of admixture on the chances of finding a compatible donor for hematopoietic stem cell transplantation (HSCT)?

2.1 Natural selection in the MHC and LRC regions

Admixture can contribute to adaptation by introducing novel advantageous variants into a population [56]. We and others have shown that the MHC region in many South American populations has an excess of African ancestry with respect to the rest of the genome [57–59], consistent with the hypothesis that African alleles are advantageous in the admixed population . This is an instance of *adaptive admixture*, the process where advantageous alleles are introduced into a population through admixture and contribute to rapid adaptation [14,60].

Here, we will ask if populations with distinct sources of African ancestry (i.e., with migrants from different regions from Africa) share the signal of recent admixture-mediated natural selection. We will attempt to identify the specific African HLA alleles that have increased in frequency in the admixed population. To do so, we will use tests based on the extent of linkage disequilibrium around HLA loci (e.g., integrated haplotype homozygosity, iHS) [61], which will also provide an estimate of the timescale of selection on the HLA loci, allowing us to distinguish between the role of recent selection from that which took place in Africa.

Our study of adaptive admixture in the MHC will rely on the detailed map of diversity. We will also use our well curated data for the LRC to test for ancestry deviations in this region, an effort that has yet not been undertaken in large admixed population samples. Together, these investigations will provide a detailed survey of how selection on very recent timescales (i.e.,after the onset of admixture, fewer than 20 generations) has shaped the diversity of key genes of the immune system.

2.2 Effects of admixture on diversity, relatedness, and differentiation

How does admixture influence the genetic diversity of a population? Although the notion that "admixture can increase genetic diversity" is intuitive [62], it is essential to place this expectation in a quantitative framework. For a population originating from two parental sources in a single pulse of admixture, the conditions necessary for an increase in heterozygosity were expressed by Boca and coworkers [63] as a function of the heterozygosities of source populations, the F_{ST} between them, and the contributions of sources to the admixed populations. These findings imply that for SNPs segregating at low frequencies and with a low F_{ST} between sources, the heterozygosity of the admixed population will not be higher than that of the most polymorphic parental. However, at extremely polymorphic loci higher heterozygosity can readily be attained as an outcome of admixture [63].

We propose to use this framework to analyze how admixture has shaped heterozygosity in different genomic regions. We will contrast the impact of admixture on both genome-wide and at HLA heterozygosity, confront the findings with model predictions, and make predictions about the relative contribution of admixture as a microevolutionary process shaping genetic diversity of humans. This will allow us to revisit Templeton's (2016) argument that admixture in humans will result in individuals carrying a greater number of nucleotide positions in a heterozygous state.

In addition to variability within populations, we are also interested in population structure. We previously showed that SNPs within HLA genes have significantly lower population pairwise F_{ST} than those from the rest of the genome [64]. However, for recently diverged populations (those from the same continental regions) F_{ST} of SNPs in HLA genes are higher than genome-wide, suggesting that the mode and effect of selection vary depending on the timescale of population divergence [65]. Here, we will draw on the relatedness and structure framework of Weir and Goudet [66] to systematically explore the effects of selection at different geographic scales (by using different reference populations) and to address the relationship between population structure and relatedness. We are motivated by the fact that kinship at HLA loci is biologically relevant to both evolutionary dynamics (e.g., the ability of pathogens to spread throughout a population) and medical applications (e.g., the ability to find matching HLA individuals for transplantation). Our previous analyses addressed SNP-level polymorphism, but it is the combination of variants (which define "HLA alleles") that defines the repertoire of peptides that are bound and presented to the immune system. Accordingly, we will use the F_{ST} framework to compare relatedness patterns at these different biological levels.

2.3 Admixture and Hematopoietic Stem-Cell transplantation (HSCT)

The ideal setting for HSCT is a match between donor and patient in 10 out of 10 alleles (over 5 HLA loci: *HLA-A*, *-B*, *-C*, *-DRB1*, *-DQB1*). Because of the high HLA polymorphism, the first attempt to find matching donors is to investigate family members. When these fail to provide matches, unrelated donors are investigated by querying bone marrow registries. We previously showed that when querying REDOME, Brazil's marrow donor registry with close to 5 million enrolled, individuals with greater African ancestry, on average, have fewer potential donors than those with higher European ancestry [67]. This has been attributed to the fact that two random African individuals are on average less genealogically (and therefore genetically) related to one another than two European, Asian, or Native American individuals. In addition, REDOME is underrepresented for individuals with African ancestry, exacerbating the difficulty in finding donors for individuals with high African ancestry [67].

Here, we will quantify the relative contribution of two factors to the difficulty in finding donors for individuals with African ancestry: the underrepresentation of Africans in REDOME and the higher diversity of African HLA loci. We will model these factors to estimate how much the African ancestry component in REDOME would have to be increased to allow individuals with either African or European ancestry to have similar chances of finding a compatible donor. These research questions have the potential to guide donor recruitment policies in REDOME, and will be carried out in close collaboration with the REDOME scientific advisory committee.

Project 3. Evolutionary genetics of admixed populations: insights into disease

We will investigate how admixture influences the accumulation of deleterious mutations in the Brazilian population and how a disease-predisposing mutation of African origin has spread onto non-African backgrounds. We will also show how admixture can leverage the mapping of disease predisposing mutations.

3.1 The history of deleterious and disease mutations in an admixed population

Each human genome can have up to hundreds of mutations that are deleterious, reducing the fitness of its carrier and contributing to disease phenotypes. This burden of deleterious mutations is referred to as the genetic load, and is formally defined as the fitness decrease in a population when compared to the value of the theoretically "fittest" genotype [68]. Load arises as a consequence of the influx of deleterious mutations, and its magnitude within a population is shaped by the patterns of inbreeding, the efficacy of selection in removing deleterious mutations, and demographic history. Populations with a history of intense genetic drift accumulate proportionally more deleterious mutations as a consequence of the reduced efficacy of selection in removing these low fitness variants [69,70]. However, there is relatively little theory developed for load in admixed populations [71].

We will investigate how admixture determines the amount of genetic load in humans. We will quantify the relative contribution of deleterious variants originating in different ancestries. In the case of Brazilian populations, this will allow us to quantify the relative contribution of African and European populations to the current load, as well as the possible effect of admixture reducing the intensity of load, by masking recessive deleterious alleles.

In addition to the intensity of genetic drift, mating patterns can also influence the degree of genetic load [72]. Using genome-wide data we will quantify the degree of inbreeding and deviation from random mating in Brazilian admixed populations [73]. This will allow us to quantify how non-random mating contributes to the accumulation of deleterious alleles in admixed Brazilians. Previous results showed that inbreeding affects individuals with African ancestry disproportionately, increasing the burden of homozygous deleterious genotypes (a consequence of the higher number of deleterious variants segregating in Africans) [74].

We will also carry out an in-depth study of the impact of admixture on a specific condition, Sickle Cell Disease (SCD), which is the most common Mendelian disease in Brazil, with 3000 annual births, and arises in individuals who are homozygous for the Hemoglobin S mutation (occurring in the beta-hemoglobin locus, *HBB*). The S mutation of *HBB* originated in Africa, and was brought to Brazil by Africans forcibly transported as slaves. Originally, individuals in Brazil carrying the HBB-S mutation were of African ancestry genome-wide. However, over time, admixture has spread the mutation to different genomic contexts. Despite this history of admixture, SCD remains considered as an "African condition" [75]. We will survey the rate at which HBB-S has, through admixture, spread onto non-African backgrounds.

Our research on the dynamics of HBB-S will be useful from an epidemiological and medical perspective, increasing awareness of how genetic and socially constructed identities interact and are perceived (by patients and medical practitioners, see [75]). Our findings will also provide an estimate of the rate at which SCD, due to ongoing admixture, will become a disease that cannot be considered an "African condition", from the genetic point of view.

3.2 How admixture influences severity of genetic disease: a case study with SCD

Admixture mapping can be used when genetic variants associated with a phenotype differ in frequency between the populations that contribute to the admixed sample [11]. We propose to analyze a large sample of admixed Brazilians with sickle-cell disease (SCD, see section B.3.1) to identify genetic variants associated with disease severity. We will use our clinical knowledge of SCD, along with our understanding of the evolutionary history of the HBS mutation, to argue that admixture mapping is particularly powerful to understand this disease.

Sickle-cell disease has a very broad spectrum of clinical manifestations, and various methods have been proposed to classify disease severity on a quantitative scale [76]. Previous studies have identified the expression of fetal hemoglobin (HbF) as modifier of the disease phenotype, with continued expression of HbF into adulthood contributing to long-term survival of SCD patients. Genetic variants within the *BCL11A*, *HBS1L-MYB* and *β-globin* loci have been shown to be associated with HbF levels, and thus the clinical phenotype of SCD [77].

The strong advantageous effect of high HbF expression levels on survival has led us to hypothesize that this trait has been under positive selection in African populations in which SCD is prevalent. This would occur because in sub-Saharan Africa SCD is prevalent, and modifiers that result in a less severe SCD phenotype would be strongly favored. These same modifiers of HbF expression outside Africa are unlikely to be under positive selection, since there is no comparable selective advantage to continued HbF expression in the absence of SCD. This provides an ideal scenario for admixture mapping, since genetic modifiers of SCD, leading to the less severe phenotype, are expected to be of predominantly African origin within admixed populations. While our discussion above emphasizes HbF, several other modifiers of SCD severity have been described, and can be addressed using the admixture mapping strategy.

Our approach will consist in searching for genomic regions with an enrichment of African ancestry among individuals with a phenotype of reduced severity. Because of our prior hypothesis involving HbF expression levels with the non-severe phenotype, we will specifically query whether sets of genes which have been shown to modulate HbF expression, or regions harboring eQTLs for HbF, show an excess of African ancestry when analyzed jointly (providing a polygenic test for enrichment of African ancestry in regulators of HbF expression).

Through our research, we expect to find answers to an evolutionary and a medical question. From a gene mapping perspective, we hope to identify genes that contribute to the reduced severity of the SCD disease phenotype. We expect that this finding will, in turn, be informative about past selection favoring modifiers for reduced severity of SCD in regions of Africa where it is a common condition.

Project 4. Outreach: discussing genetics, diversity, and race

Throughout this project we will be dealing with topics that are socially sensitive: the effect of admixture on human health, the history of a disease of African origin, the effect of African ancestry on the chances of finding a donor, and the genetic similarities and differences between humans, including those living in different geographic regions. All these themes are particularly sensitive in a country with a history of structural racism, as is the case of Brazil.

Our genetic research questions are not placed in the context of "racial concepts", which have been largely abandoned in the scientific community after Lewontin's landmark work [78]. Instead, we will be addressing human genetic diversity from a geographic perspective. While the biological basis of racial categories has been firmly rejected, socially constructed views of race are still very much present in our society [79,80]. As such, it is timely to consider how views on genetic diversity interact with discussions of race, which take place outside the field of genetics. These include addressing questions such as the relationship, in Brazil, between self identified race and ethnicity and genetics [81], and the relevance of self-assigned categories from a medical perspective. In addition, we will be in a position to present to the general audience, in an accessible form, information about the genetic consequences of admixture in Brazil. To deliver this information it is critical to be tuned to what the general audience seeks to understand, and how groups actively involved in questions regarding racial equity can benefit from this empirical knowledge. We are assuming that our scientific efforts, which are grounded in observation, data collection, and hypothesis testing, are not produced or delivered in a vacuum, but within a social context that calls for critical examination.

To this end, we have planned an additional dimension to our research project, which includes establishing interactions with researchers working on human diversity from different perspectives. We will work to establish a multidisciplinary research group, based within the University of São Paulo, involving medical practitioners, anthropologists, social scientists, as well as geneticists, with the goal of discussing connections among fields. In addition, we plan to present our research results in a clear and accessible way, which is sensitive to the implications these themes may have. To this end, we are proposing within the grant a "Jornalismo Cientifico" scholarship, to contribute to the development of media and written material that can help summarize in an accessible manner the technical results of our research.

C. Methods and samples

Project 1 - The genomics and transcriptomics of the MHC and LRC regions: structure, variation, and function

Samples included in our study

We will analyze admixed and parental populations, with approximately 9,000 samples (Figure 1). The TOPMed cohorts, because of the multicentric recruitment for SCD patients, will allow an enrichment of African ancestry and will complement the census-based SABE dataset, with proportionally more Europeans (Figure 2). As parental populations, we will use samples from the 1000 Genomes project and HGDP-CEPH, with African, European, East Asian, and Native American (listed under the admixed samples from Figure 2), with more than 95% of ancestry inferred for that region. Ancestry inferences will be performed with supervised analysis (K = 4) after applying linkage disequilibrium filters (r^2 = 0.1) within a sliding window of 50Kb and a shift step of 10Kb.

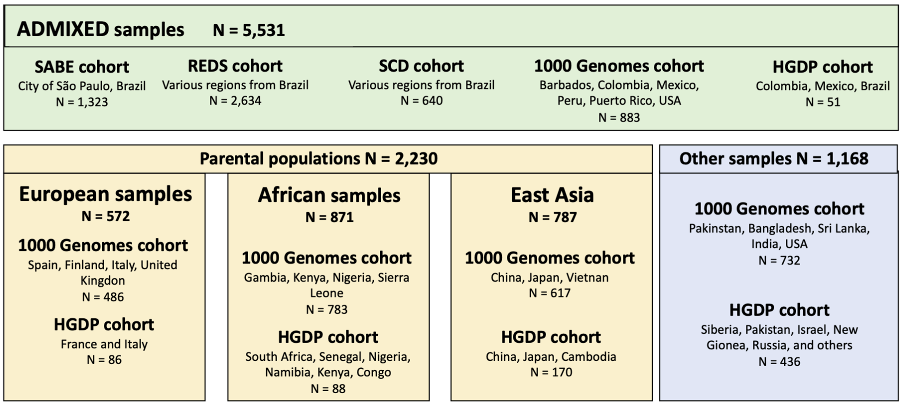

Figure 1. Cohorts and population samples. All present WGS, depth of 30x, 150bp read, paired-end sequencing. The SABE cohort [14,82] data made available through direct transfer and collaboration with Dr. Michel S. Naslavsky. The REDS cohort [REDS-III Sickle Cell Disease cohort] [82]ata is available through the TOPMed Trans-Omics for Precision Medicine (TOPMed), sponsored by the National Heart, Lung and Blood Institute (NHLBI).

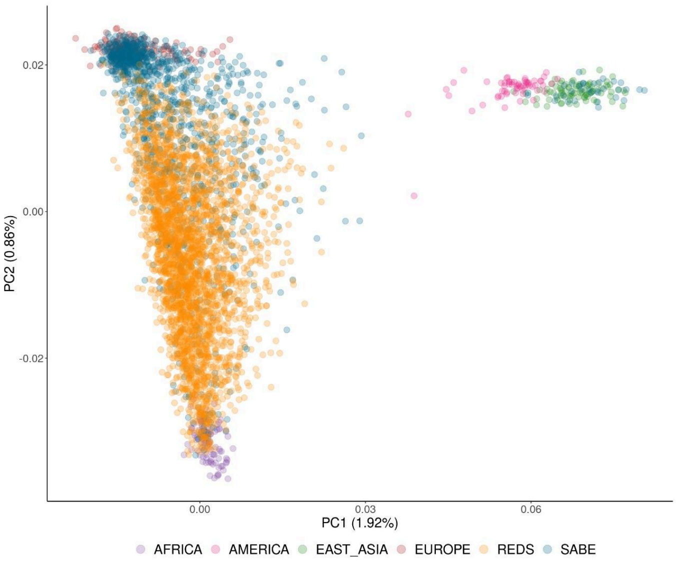

Figure 2: PCA illustrating the relationship among samples included in our study. The broad area covered by the REDS cohort (orange) and SABE (blue) highlights the presence of admixed individuals, critical to our study.

MHC and LRC genotyping and haplotyping

Many bioinformatic tools have been proposed to genotype alleles and estimate expression for HLA loci [83–89]. Their primary focus is HLA typing (Optitype, HLA-VBSeq, HLA-LA, and others), focusing on identifying previously described alleles for a subset of HLA loci. Except for MHC-PRG, which uses a Population Reference Graph of human MHC to improve genome inference, existing methods perform read alignment considering the reference genome. However, MHC-PRG is not suitable for discovering new variants. Regarding polymorphism in the LRC, there are two main methods that perform KIR typing: KPI and PING [87,90]. None of these tools align reads to the reference genome, and they are compatible with exon sequences only, not inferring promoter and intronic variants. Thus, none of these methods are suitable to obtain reliable genotypes across the MHC and LRC regions, which is necessary for our goals.

Therefore, it is mandatory to develop applications that align sequences to the MHC and LRC reliably. In 2008, Castelli developed the software hla-mapper [48], which minimizes alignment errors in these two gene regions, allowing more accurate downstream analysis. This tool, differently from others, optimizes the alignments over the hg38 reference, allowing the inference of SNPs and indels across these regions. The development of hla-mapper will provide a single framework for de novo characterization of HLA and KIR variation and expression, with a strong focus on the description of novel variants. Given our proposal to develop a new computational resource, we will also dedicate extensive attention to evaluating the reliability of calls made with hla-mapper.

The current version of hla-mapper allows an accurate alignment of reads from some HLA genes from the MHC class I region, and has been used to study the genetic diversity of these loci among Brazilians and other populations [49–51,91–93]. As part of this project we will extend hla-mapper to analyze HLA class II genes, other MHC and LRC genes, and to use RNA-seq data.

To extend and update the hla-mapper database and software, we need to update its code (written in C++) and its internal database. This database is composed of known sequences of each gene, used by the software to guide alignment. These sequences were obtained from the IMGT/HLA Database and GENBANK, and from predicted sequences generated in-house by Sanger and NGS sequencing. This process needs serial evaluation of thousands of samples, in which we search for regions with low-depth of coverage or unbalanced heterozygous genotypes (indicating alignment problems), allowing a manual inspection and inference of the sequences observed in each gene. This process was done for HLA class I genes and must be performed for other MHC and LRC genes.

ECC group has developed many pipelines to call genotypes and haplotypes for HLA and KIR genes (please refer to https://github.com/erickcastelli/ and www.castelli-lab.net) after hla-mapper. In brief, we call SNP and indel genotypes by using GATK HaplotypeCaller [94], with a further refinement step by using vcfx (www.castelli-lab.net/apps/vcfx) and VQSR from GATK. The refined unphased VCF file is the input for the inference of the physical phase between heterozygous genotypes in trios or unrelated samples using WhatsHap [95]. To call haplotypes, we combine the read-aware phasing (with WhatsHap) with probabilistic models using Shapeit4. The phased VCF file is the input for downstream analysis, including the call of HLA and KIR alleles, the inference of complete sequences, and the prediction of the protein sequences encoded by each chromosome. We will apply this method to call genotypes and haplotypes for the entire MHC region, and for all genes from the LCR, considering the samples from Figure 2. We will also call HLA and KIR alleles for all samples.

Detection of structural variants

Methods for detecting structural variants are constantly developing, particularly for regions with repetitive sequences such as the MHC. The alignment optimization applied by hla-mapper allows comparing read depth within genes and reference regions to detect the number of gene copies. We will use this method to characterize the number of copies of specific loci such as *HLA-DRB5* and *KIR2DS4*. However, this method is less efficient in intergenic regions. We will therefore apply several methods to characterize structural variants in intergenic regions from the MHC. One of the strategies is the simultaneous use of the Breakdancer, BreakSeq2, CNVnator, Delly, Lumpy, and Manta algorithms, implemented in Partiament2 (https://github.com/dnanexus/parliament2), and comparing the results to avoid false-positive variants. We may also develop pipelines as necessary, always integrating the results of many different algorithms. We will integrate these structural variants with the VCF file, correcting genotypes when necessary.

Allele frequencies, functional annotations, discovery of new variants

We will investigate genetic variation and genomic organization of MHC and LRC genes in Brazilians and other admixed populations at the molecular level, which includes detailed lists of SNPs, indels, MNPs, structural variants, and haplotypes, and their frequencies. Allele frequencies will be estimated by direct count. We will use this data to provide an in-depth map of regulatory diversity, focused on the SNPs and haplotypes observed in promoters, enhancers, and 3'UTRs, and the putative effect of polymorphism on TFBS and miRNA binding. Many computational tools can be used to assess the influence of polymorphisms on TFBS and miRNA binding, including HaploReg[96], mrSNP[97], and miRWalk[98]. The phasing method described earlier also allows us to characterize haplotypes between neighboring genes, such as *HLA-B* and *HLA-C* haplotypes. This data will be used to characterize the haplotype diversity within and among populations, especially for less-characterized genes such as *TAP1*, *TAP2*, *MICA*, *MICB*. Linkage disequilibrium will be assessed using classical D, D', and correlation-based methods, implemented in Haploview[99] and Tomahawk (https://mklarqvist.github.io/tomahawk/).

We will also annotate all variants regarding their possible effects (missense variants, stop-gain, etc.) using SNPeff and Annovar. Our approach allows the detection of new SNPs and HLA/LRC alleles. For new HLA/LRC alleles in Brazil, when possible, we will clone and Sanger sequence the locus, submitting their sequences to the IPD-IMGT database for validation. This will allow commercial typing methods for transplantation include these new alleles.

Imputation

We will use the SNP data and allele calls to create new reference panels for HLA imputation, using HIBAG-HLA[100]. This is part of an international effort to improve HLA imputation, the SNP-HLA Reference Consortium, in which ECC and DM are members[54]. Today, HLA imputation accuracy is very low for admixed populations and demands reference panels with more non-European samples. To evaluate the accuracy of the new reference panels, we will impute HLA alleles from Brazilian samples that have both SNP-Array data and HLA typing. We will address HLA imputation accuracy using different metrics, including the F1 score.

Measuring expression levels of HLA genes

We will extend the hla-mapper approach to RNA-seq data. The beta version of the hla-mapper for RNA sequences uses the STAR aligner[101] to detect possible splicing sites and applies a scoring system to define where each read should be aligned. This version is under development but has achieved excellent results for HLA class I genes (Figure 4).

With 50 simulated transcriptomes presenting different HLA alleles and transcripts for all HLA genes, cross-alignments between *HLA-B* and *HLA-C* and unmapped reads are common (Figure 4, panel A). Because of these alignment errors, *HLA-C* expression levels would be highly underestimated, while *HLA-B* levels might be lower or higher depending on the sample's genotype. After hla-mapper RNA optimization, cross-alignments were removed and unmapped reads recovered, allowing accurate genotyping and expression estimation (Figure 4, panel B).

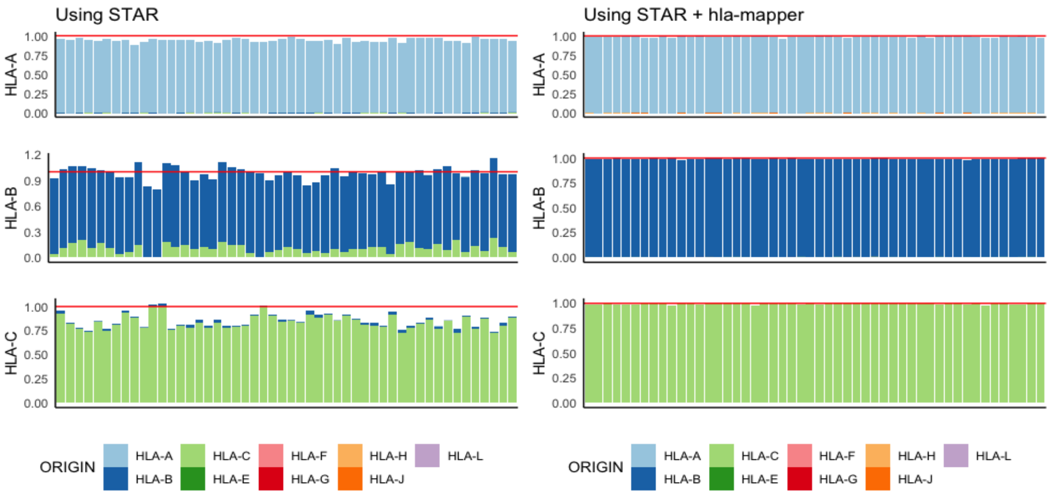

Figure 4: Alignment accuracy for classical HLA class I genes when using STAR 2.7.10a with default parameters (panel A) and the beta version of hla-mapper RNA (panel B) in 50 simulated transcriptomes (2x100bp) with transcripts from all HLA genes. The red line indicates the expected proportion of reads to each locus. ORIGIN defines the HLA gene that gave origin to the read. We excluded reads aligned as secondary.

We will explore different datasets to measure HLA expression and correlate the expression levels with genotypes. The first consists of 445 individuals from the Geuvadis project, sequenced by the 1000Genomes Project in high coverage. The second, is a cohort of 96 individuals recruited in a project led by D.M, in collaboration with Mary Carrington, from the NCI. After hla-mapper optimization, we will search for eQTLs using QTLtools [102] , which implements PCA-based strategies to remove confounding variables (e.g. batch effects) and uses a simple linear model with permutations to more efficiently define significance thresholds (see also our previous work on eQTL mapping, [46]). We will also perform read counts using FeatureCounts, normalizing the data accordingly. The alignment optimization also allows the detection of allele-specific expression levels, comparing the expression levels of each chromosome within and between samples.

Project 2 - Population Genetics of the MHC and LRC in admixed populations

Local and Global ancestry estimation

Assessing the genetic ancestry of individuals will be key to our studies of recent selection in HLA and LRC genes, to carrying out admixture mapping, and to study the history of *HBB-S* in post-admixture Brazil. Global ancestry (i.e., the proportion of an individual's genome that traces to a particular ancestral group) will be estimated using ADMIXTURE [103], based on likelihood models and the information about allele frequencies of the parental populations. We will infer local ancestry using RFMix, which requires prior phasing, to be carried out using SHAPEIT [104]. Tests for recent selection will be based on extended haplotype homozygosity [61], an approach that identifies regions of high linkage disequilibrium surrounding an allele, using the LD of other alleles at that locus as demographic controls. Evolutionary models that incorporate selection and demography will be tested using various simulators, including msms [105] and msprime [106] .

The performance of local ancestry methods such as RFMIX, in the context of highly polymorphic regions which are challenging to phase (such that containing HLA anr LRC genes), poses important challenges [107]. We will therefore study the performance of LAI using simulated admixture scenarios, where haplotypes of known ancestry are allowed to admix and then resulting mosaic chromosomes are reconstructed.

Population structure, relatedness and inbreeding

Our survey of genetic variation will follow the work of Weir and Goudet [66] and use the degree of sharing of alleles among individuals within and between populations to estimate F_{ST} for the *i*-th population, as follows:

$$F_{st,\,i} = \frac{S_{w} - S_{B}}{1 - S_{B}}$$

where S_w refers to the average degree of sharing of alleles between alleles randomly sampled from two distinct individuals *within* a population, and S_B refers to the average degree of sharing of alleles between individuals *between* populations. This provides the population-specific F_{ST}, which is a measure of how much a population has progressed in the fixation process (i.e., attaining genetic homogeneity due to identity-by descent) with respect to the larger group to which that population belongs. An overall F_{ST} can be obtained by simply averaging over population-specific values:

$$F_{ST} = \sum_{i=1}^{n} = \frac{F_{st}, i}{n}$$

where n refers to the number of populations studied. The value of F_{ST} provides an estimate of how the average kinship among individuals within populations differs from that of individuals between populations. This approach can be extended to the contrast of sharing within and among individuals, thus providing a measure of inbreeding [66]. Because our datasets are for WGS, we can use the distribution of F_{ST} over the non-MHC regions to create a null distribution, with which to compare that found within the MHC.

To estimate inbreeding we will use estimators based on the number and size distributions of Runs of Homozygosity (ROHs), stretches of the genome which are identical by state in the two homologous copies of the chromosomes within individuals. Theory shows that distributions of ROH provide robust estimators of inbreeding, as well as of remote population history (associated to bottlenecks in the ancestry of the samples) [108,109]. Estimates of ROHs can be obtained from dense array or WGS data, and estimators are implemented in PLINK [110] and GARLIC [111].

We have previously shown that although F_{ST} provides useful quantifications of coancestry at the SNP level for HLA genes, the unusually high polymorphism results in cases where populations may have sets of alleles with a little or now overlap, but with relatively low F_{ST} estimates [112]. In these situations, alternative approaches to quantify the degree of ancestry (or similarity) among populations may be necessary. We therefore propose to apply and critically assess the utility of various standardized population differentiation measures [113], which correct for constraints imposed by degrees of polymorphism.

Our analysis of population structure at HLA loci will compare results across four levels of biological organization, all of which comprise possible targets of natural selection: (a) individual nucleotide positions, (b) HLA alleles (defined by combinations of nucleotide variants), (c) multi-locus haplotypes, comprising syntenic combinations of alleles spanning different HLA loci (e.g., HLA-A, HLA-B, DRB1).

HLA diversity and effect on finding compatible donors

We are interested in understanding how an individual's ancestry affects their chances of finding a compatible donor in REDOME. To do so, we will first generate data for approximately 5531 admixed individuals (Figure 1) containing information about their HLA genotype at HLA-A, B, C, DRB1, DQB1, and DPB1 (using the approaches outlined in methods for Project 1). We will then estimate their global genetic ancestry, and their local ancestry within the MHC region (as described above). We will then query whether each individual finds compatible donors in REDOME using in-house scripts and files containing the genotypes of approximately 5 million REDOME volunteers. We will use parametric statistics to quantify the degree to which ancestry estimates differ among groups where matches are found, with respect to those where they are not found. Analysis of REDOME data will be carried in collaboration with Luis Cristóvão Porto (UERJ) who has authorized access to REDOME HLA and demographic information.

Our previous studies, using a smaller dataset and imputed HLA alleles, indicated that individuals with African ancestry on average are less likely to find compatible donors in REDOME [67]. Two factors contribute to this result: the smaller size of the African component in REDOME, and the higher diversity of african HLA genes. To tease apart the relative contribution of these two factors, we will subsample the Brazilian registry to produce replicates with varying ancestry compositions, thus allowing us to isolate the contribution of registry composition to the differences in matching rates.

Project 3 - Evolutionary genetics of admixed populations: insights into disease

Admixture mapping

Admixture mapping is an approach to identify genetic associations, and is particularly powerful if the disease risk allele frequency differs across groups [11]. We are interested in using the admixed nature of Brazilian populations to identify genetic variants that contribute to complex phenotypes . We hypothesize that genetic variants that decrease the severity of sickle cell disease (SCD) will have been favored in populations living in regions of Africa where malaria is endemic, driving the HBS mutation to high frequencies. We therefore predict that alleles which reduce the severity of the disease phenotype will be of predominantly African ancestry. Our analysis will therefore search for regions enriched in African ancestry among SCD patients with less severe phenotypes.

We will initially estimate local ancestry (see "Local and global ancestry estimation"), a genetic relatedness matrix (GRM) among individuals, and classify the individuals according to the severity of their disease phenotype [76]. We will then use a mixed modelling approach, treating severity as a fixed effect and the genomewide relatedness (measured by the GRM) as a random effect, using a score test to test the association between linkage disequilibrium blocks and the null model of no association, with critical values defined as in [114]. This admixture analysis will be implemented using the GENESIS R package [115], and the GRM will be estimated by controlling for population structure within the sample [116]. If an association is found, a secondary admixture analysis treats each ancestry separately so as to identify which one accounts for the significant result.

Genetic load measurement and analysis

In order to understand how admixture shapes the burden of deleterious mutations a population carries, we will initially quantify the number of deleterious alleles present in various populations, including admixed Brazilians. To do so, we will use the CADD approach [117], which has the advantage of being a composite score and of providing a deleteriousness score, rather than an assignment to a fixed number of categories. We will additionally use our information regarding the distribution of local ancestry (see above) and the distribution of runs of homozygosity (see above).

Previous studies have shown that inbreeding has the potential to create runs of homozygosity that harbor deleterious variants in homozygous state, a condition which would be unlikely without inbreeding given the low frequency of deleterious alleles [74]. Specifically, because African populations have a larger absolute number of segregating variants (including deleterious alleles), they are expected to be disproportionately affected by inbreeding (i.e., show a higher gain of deleterious alleles in homozygous state).

Here, we will investigate if recent inbreeding has contributed to the homozygosis of deleterious variants (i.e., if runs of homozygosity are enriched for deleterious alleles in a homozygous state) among admixed Brazilians, and if this effect varies according to the ancestry of individuals. This effect will be quantified using a regression approach, which quantifies how increase in total homozygosity affects deleterious homozygosity [74]. By examining the impact of inbreeding across ancestries, we will test how inbreeding can contribute to disease phenotypes. We will also investigate if recent admixture, by mixing haplotypes of different ancestries, can reduce the number of runs of homozygosity, and thus alleviate the average burden of deleterious homozygous genotypes.

The history of the hemoglobin-S mutation in admixed Brazilians

To investigate the history of the hemoglobin S (HBS) mutation in Brazil, we will use the whole genome sequencing data from the REDS cohort, comprised of 2700 individuals clinically defined as having sickle-cell disease, most of whom are homozygous for the HBS mutation.

Using the local ancestry information (see section above), we will investigate how the degree of African ancestry on chromosome 11 declines as a function of the distance with respect to the HBS mutation. Our working hypothesis is that, because our cohort is of SCD patients, African ancestry will be enriched at the HBS locus, and will decline with distance. Although this result is expected, the rate of decay of African ancestry is of interest, since it can provide insights into the demographic context in which populations received the HBS mutation: time, number of inputs, proportion of admixture. Here, we will use a simulation-based approach (inspired in that of Kehdy et al. [9], to identify the population parameters that best the HBS data. The approach consists in exploring the combination of parameters which provides the best fit to the observed distribution of ancestry in the true data [9].

D. Ethical aspects

The raw data for the samples from the 1000Genomes project and HGDP are publicly available in online repositories.

All the Brazilian participants from the SABE cohort, evaluated here in collaboration with Dr. Michel S. Naslavsky, were asked for specific consent on taking part in genomic studies from the year 2010 and beyond, which was approved by CEP/CONEP (Brazilian local and national ethical committee boards) under the following protocols: COEP FSP USP OF.COEP/23/10, CONEP 2044/2014, CEP HIAE 1263-10.

E. Project timeline

	Semesters									
Milestone	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
whole-genome sequencing of Brazilian samples from trios	🟩	🟩								
Development/update of the hla-mapper software and database	🟩	🟩								
MHC genotyping and haplotyping of ~9000 samples			🟩	🟩						
LRC genotyping and haplotyping, copy number detection of ~9000 samples			🟩	🟩						
Confirmation of new alleles					🟩					
Development of imputation panel and accuracy test					🟩					
Estimation of the expression levels of each HLA gene for RNAseq data						🟩	🟩			
In-depth description of the genetic diversity of each HLA and LRC gene						🟩	🟩	🟩	🟩	🟩
Global and local ancestry analyses				🟧	🟧					
Survey of Population structure in MHC and LRC					🟧	🟧	🟧			
Quantification of how ancestry affects transplantation matching					🟧	🟧	🟧			
Prediction of genetic load /diversity in admixed populations: theory							🟦	🟦		
Prediction of genetic load in admixed populations: analyses								🟦	🟦	
Description of admixture effects on disease allele context	🟦	🟦					🟦	🟦	🟦	🟦

F. Project management and data dissemination

We will be attending bi-monthly meetings with the presentation (primarily online) of partial results from researchers and students involved in the project and monthly management meetings with all the researchers from the project to plan activities and use of resources.

The data produced by this project will be disseminated in different ways and formats. First, we will present preliminary/partial results in international scientific conferences, including conferences from the American Societies of Human Genetics and Histocompatibility, the European Federation for Immunogenetics, and the European Society of Human Genetics. Second, we will publish our results in high-impact scientific periodicals, with social and traditional media announcements. Third, we will provide a database of variants, sequences, tools, and pipelines for MHC and LRC genes hosted on the project website.

G. Other resources and support for the project

Besides the research groups cited in section D (principal investigators), which include different groups from USP and UNESP, this project is also supported by other research facilities and parallel grants. ECC laboratory is located in a multi-user research facility from the UNESP medical school, with all the structure and staff for molecular analysis. Dr. Nicolas Vince, from the University of Nantes, is a collaborator of this project regarding the development of reference panels for imputation in the context of the SNP-HLA Reference Consortium (which DM and ECC are members). Dr. Vince will provide a Ph.D. scholarship for a Brazilian student to work in the development and application of these reference panels for imputation in cotutel with ECC. FAPESP has supported the development of the hla-mapper strategy in two previous grants (2013/17084-2 for HLA class I genes and 2017/19223-0 for some KIR genes). The DM lab is located within the Institute of Biosciences within the University of São Paulo, and has the appropriate infrastructure and environment for the proposal. DM is currently funded by an NIH grant (RO1 GM 075091) in a collaborative project with Bruce Weir (University of Washington).

H. References

1. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
2. Gillespie, J. H. *Population Genetics: A Concise Guide*. (Johns Hopkins University Press, 2004).
3. Nielsen, R. *et al.* Tracing the peopling of the world through genomics. *Nature* **541**, 302–310 (2017).
4. Gravel, S. *et al.* Demographic history and rare allele sharing among human populations. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11983–11988 (2011).
5. Hey, J. On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLoS Biol.* **3**, e193 (2005).
6. Harris, E. E. & Meyer, D. The molecular signature of selection underlying human adaptations. *Am. J. Phys. Anthropol.* **Suppl 43**, 89–130 (2006).
7. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220 (2012).
8. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. *Nat. Genet.* **46**, 220–224 (2014).
9. Kehdy, F. S. G. *et al.* Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8696–8701 (2015).
10. Chi, C. *et al.* Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. *PLoS Genet.* **15**, e1007808 (2019).
11. Winkler, C. A., Nelson, G. W. & Smith, M. W. Admixture mapping comes of age. *Annu. Rev. Genomics Hum. Genet.* **11**, 65–89 (2010).
12. Secolin, R. *et al.* Distribution of local ancestry and evidence of adaptation in admixed populations. *Sci. Rep.* **9**, 13900 (2019).
13. Pena, S. D. J., Santos, F. R. & Tarazona-Santos, E. Genetic admixture in Brazil. in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* vol. 184 928–938 (Wiley Online Library, 2020).
14. Naslavsky, M. S. *et al.* Whole-genome sequencing of 1,171 elderly admixed individuals from the largest Latin American metropolis (São Paulo, Brazil). *Cold Spring Harbor Laboratory*

2020.09.15.298026 (2020) doi:10.1101/2020.09.15.298026.

15. Patrinos, G. P. *et al.* Roadmap for Establishing Large-Scale Genomic Medicine Initiatives in Low- and Middle-Income Countries. *Am. J. Hum. Genet.* **107**, 589–595 (2020).
16. Rocha, C. S., Secolin, R., Rodrigues, M. R., Carvalho, B. S. & Lopes-Cendes, I. The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. *NPJ Genom Med* **5**, 42 (2020).
17. Klein, J. & Sato, A. The HLA system. First of two parts. *N. Engl. J. Med.* **343**, (2000).
18. Hedrick, P. W. What is the evidence for heterozygote advantage selection? *Trends Ecol. Evol.* **27**, 698–704 (2012).
19. Borghans, J. A. M., Beltman, J. B. & De Boer, R. J. MHC polymorphism under host-pathogen coevolution. *Immunogenetics* vol. 55 732–739 Preprint at https://doi.org/10.1007/s00251-003-0630-5 (2004).
20. Lenz, T. L., Spirin, V., Jordan, D. M. & Sunyaev, S. R. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. *Mol. Biol. Evol.* **33**, 2555–2564 (2016).
21. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. *Nat. Rev. Immunol.* **18**, 325–339 (2018).
22. Bitarello, B. D. *et al.* Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biol. Evol.* **10**, 939–955 (2018).
23. Wende, H., Colonna, M., Ziegler, A. & Volz, A. Organization of the leukocyte receptor cluster (LRC) on human chromosome 19q13.4. *Mamm. Genome* **10**, 154–160 (1999).
24. Trowsdale, J., Jones, D. C., Barrow, A. D. & Traherne, J. A. Surveillance of cell and tissue perturbation by receptors in the LRC. *Immunol. Rev.* **267**, 117–136 (2015).
25. Vivier, E. *et al.* Innate or adaptive immunity? The example of natural killer cells. *Science* **331**, 44–49 (2011).
26. Single, R. M., Martin, M. P., Meyer, D., Gao, X. & Carrington, M. Methods for assessing gene content diversity of KIR with examples from a global set of populations. *Immunogenetics* **60**, 711–725 (2008).
27. Dupont, B., Selvakumar, A. & Steffens, U. The killer cell inhibitory receptor genomic region on

human chromosome 19q13.4. *Tissue Antigens* **49**, 557–563 (1997).

28. Qi, Y. *et al.* KIR/HLA pleiotropism: protection against both HIV and opportunistic infections. *PLoS Pathog.* **2**, e79 (2006).
29. Jones, D. C. *et al.* Killer immunoglobulin-like receptor gene repertoire influences viral load of primary human cytomegalovirus infection in renal transplant patients. *Genes Immun.* **15**, 562–568 (2014).
30. Farias, T. D. J., Augusto, D. G., de Almeida, R. C., Malheiros, D. & Petzl-Erler, M. L. Screening the full leucocyte receptor complex genomic region revealed associations with pemphigus that might be explained by gene regulation. *Immunology* **156**, 86–93 (2019).
31. Kulkarni, S., Martin, M. P. & Carrington, M. The Yin and Yang of HLA and KIR in human disease. *Semin. Immunol.* **20**, 343–352 (2008).
32. Diaz-Peña, R. *et al.* Analysis of Killer Immunoglobulin-Like Receptor Genes in Colorectal Cancer. *Cells* **9**, (2020).
33. Besson, C. *et al.* Association of killer cell immunoglobulin-like receptor genes with Hodgkin’s lymphoma in a familial study. *PLoS One* **2**, e406 (2007).
34. Shaffer, B. C. & Hsu, K. C. How important is NK alloreactivity and KIR in allogeneic transplantation? *Best Pract. Res. Clin. Haematol.* **29**, 351–358 (2016).
35. Wroblewski, E. E., Parham, P. & Guethlein, L. A. Two to Tango: Co-evolution of Hominid Natural Killer Cell Receptors and MHC. *Front. Immunol.* **10**, 177 (2019).
36. Norman, P. J. *et al.* Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. *PLoS Genet.* **9**, e1003938 (2013).
37. Single, R. M. *et al.* Global diversity and evidence for coevolution of KIR and HLA. *Nat. Genet.* **39**, 1114–1119 (2007).
38. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. *Nature* vol. 538 161–164 (2016).
39. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. *Cell* **177**, 1080 (2019).
40. Error - Cookies Turned Off. https://onlinelibrary.wiley.com/doi/epdf/10.1111/tan.13723.

41. Boquett, J. A., Bisso-Machado, R., Zagonel-Oliveira, M., Schüler-Faccini, L. & Fagundes, N. J. R. HLA diversity in Brazil. *Hladnikia* **95**, 3–14 (2020).

42. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.* **12**, 443–451 (2011).

43. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326–3328 (2012).

44. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* **5**, 931–941 (2015).

45. Byrska-Bishop, M. *et al.* High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv* (2021) doi:10.1101/2021.02.06.430068.

46. Aguiar, V. R. C., César, J., Delaneau, O., Dermitzakis, E. T. & Meyer, D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. *PLoS Genet.* **15**, e1008091 (2019).

47. Aguiar, V. R. C., Masotti, C., Camargo, A. A. & Meyer, D. HLApers: HLA Typing and Quantification of Expression with Personalized Index. *Methods Mol. Biol.* **2120**, 101–112 (2020).

48. Castelli, E. C., Paz, M. A., Souza, A. S., Ramalho, J. & Mendes-Junior, C. T. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. *Hum. Immunol.* **79**, 678–684 (2018).

49. Souza, A. S. *et al.* Hla-C genetic diversity and evolutionary insights in two samples from Brazil and Benin. *Hladnikia* **96**, 468–486 (2020).

50. Lima, T. H. A. *et al.* HLA-A promoter, coding, and 3'UTR sequences in a Brazilian cohort, and their evolutionary aspects. *Hladnikia* **93**, 65–79 (2019).

51. Weiss, E. *et al.* KIR2DL4 genetic diversity in a Brazilian population sample: implications for transcription regulation and protein diversity in samples with different ancestry backgrounds. *Immunogenetics* **73**, 227–241 (2021).

52. Meyer, D. & Nunes, K. HLA imputation, what is it good for? *Hum. Immunol.* **78**, 239–241 (2017).

53. Nunes, K. *et al.* HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. *Hum. Immunol.* **77**, 307–312 (2016).

54. Vince, N. *et al.* SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting

MHC-centric analyses in genomics. *Genet. Epidemiol.* **44**, 733–740 (2020).

55. Castelli, E. C. *et al.* 'Immunogenetics of resistance to SARS-CoV-2 infection in discordant couples'. (2021) doi:10.1101/2021.04.21.21255872.

56. Luis B. Barreiro, L. Q.-M. Evolutionary and Population (Epi)Genetics of Immunity to Infection. *Hum. Genet.* **139**, 723 (2020).

57. Meyer, D., C Aguiar, V. R., Bitarello, B. D., C Brandt, D. Y. & Nunes, K. A genomic perspective on HLA evolution. *Immunogenetics* **70**, 5–27 (2018).

58. Staff, T. P. G. & The PLOS Genetics Staff. Correction: Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics* vol. 12 e1005983 Preprint at https://doi.org/10.1371/journal.pgen.1005983 (2016).

59. Zhou, Q., Zhao, L. & Guan, Y. Strong Selection at MHC in Mexicans since Admixture. *PLOS Genetics* vol. 12 e1005847 Preprint at https://doi.org/10.1371/journal.pgen.1005847 (2016).

60. Norris, E. T. *et al.* Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biol.* **21**, 1–12 (2020).

61. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* **31**, 2824–2827 (2014).

62. Templeton, A. R. The future of human evolution. *How evolution shapes our lives: essays on biology and society, Losos JB, Lenski RE (eds)* 362–379 (2016).

63. Boca, S. M., Huang, L. & Rosenberg, N. A. On the heterozygosity of an admixed population. *J. Math. Biol.* **81**, 1217–1250 (2020).

64. Brandt, D. Y. C., César, J., Goudet, J. & Meyer, D. The Effect of Balancing Selection on Population Differentiation: A Study with HLA Genes. *G3* **8**, 2805–2815 (2018).

65. Nunes, K. *et al.* How natural selection shapes genetic differentiation in the MHC region: A case study with Native Americans. *Human Immunology* Preprint at https://doi.org/10.1016/j.humimm.2021.03.005 (2021).

66. Weir, B. S. & Goudet, J. A Unified Characterization of Population Structure and Relatedness. *Genetics* **206**, 2085–2103 (2017).

67. Nunes, K. *et al.* How Ancestry Influences the Chances of Finding Unrelated Donors: An Investigation

in Admixed Brazilians. *Front. Immunol.* **11**, 584950 (2020).

68. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. *Nat. Rev. Genet.* **16**, 333–343 (2015).

69. Henn, B. M. *et al.* Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E440–E449 (2016).

70. Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. *Curr. Opin. Genet. Dev.* **29**, 139–146 (2014).

71. Gravel, S. When Is Selection Effective? *Genetics* **203**, 451–462 (2016).

72. Szpiech, Z. A. *et al.* Long runs of homozygosity are enriched for deleterious variation. *Am. J. Hum. Genet.* **93**, 90–102 (2013).

73. Lemes, R. B. *et al.* Inbreeding estimates in human populations: Applying new approaches to an admixed Brazilian isolate. *PLoS One* **13**, e0196360 (2018).

74. Szpiech, Z. A. *et al.* Ancestry-Dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity. *Am. J. Hum. Genet.* **105**, 747–762 (2019).

75. da Silva, A. K. L. S., Madrigal, L., da Silva, H. P. & Others. Relationships among genomic ancestry, clinical manifestations, socioeconomic status, and skin color of people with sickle cell disease in the State of Pará, Amazonia, Brazil. *Antropologia Portuguesa* 159–176 (2020).

76. Sebastiani, P. *et al.* A network model to predict the risk of death in sickle cell disease. *Blood* **110**, 2727–2735 (2007).

77. Menzel, S. & Thein, S. L. Genetic Modifiers of Fetal Haemoglobin in Sickle Cell Disease. *Mol. Diagn. Ther.* **23**, 235–244 (2019).

78. Lewontin, R. C. The Apportionment of Human Diversity. in *Evolutionary Biology: Volume 6* (eds. Dobzhansky, T., Hecht, M. K. & Steere, W. C.) 381–398 (Springer US, 1972).

79. Kamariza, M., Crawford, L., Jones, D. & Finucane, H. Misuse of the term ‘trans-ethnic’ in genomics research. *Nature Genetics* vol. 53 1520–1521 Preprint at https://doi.org/10.1038/s41588-021-00952-6 (2021).

80. Popejoy, A. B. Too many scientists still say Caucasian. *Nature* **596**, 463 (2021).

81. Ramos, B. R. de A. *et al.* Neither self-reported ethnicity nor declared family origin are reliable

indicators of genomic ancestry. *Genetica* **144**, 259–265 (2016).

82. Guo, Y. *et al.* Development and evaluation of a transfusion medicine genome wide genotyping array. *Transfusion* **59**, 101–111 (2019).

83. Matey-Hernandez, M. L., Danish Pan Genome Consortium, Brunak, S. & Izarzugaza, J. M. G. Benchmarking the HLA typing performance of Polysolver and Optitype in 50 Danish parental trios. *BMC Bioinformatics* **19**, 239 (2018).

84. Bai, Y., Wang, D. & Fury, W. PHLAT: Inference of High-Resolution HLA Types from RNA and Whole Exome Sequencing. *Methods Mol. Biol.* **1802**, 193–201 (2018).

85. Dilthey, A. T. *et al.* HLA*LA-HLA typing from linearly projected graph alignments. *Bioinformatics* **35**, 4394–4396 (2019).

86. Szolek, A. *et al.* OptiType: precision HLA typing from next-generation sequencing data. *Bioinformatics* **30**, 3310–3316 (2014).

87. Lee, H. & Kingsford, C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. *Genome Biol.* **19**, 16 (2018).

88. Huang, Y. *et al.* HLAreporter: a tool for HLA typing from next generation sequencing data. *Genome Med.* **7**, 25 (2015).

89. Xie, C. *et al.* Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8059–8064 (2017).

90. Norman, P. J. *et al.* Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. *Am. J. Hum. Genet.* **99**, 375–391 (2016).

91. Sonon, P. *et al.* Human leukocyte antigen (HLA)-F and -G gene polymorphisms and haplotypes are associated with malaria susceptibility in the Beninese Toffin children. *Infect. Genet. Evol.* **92**, 104828 (2021).

92. Sonon, P. *et al.* HLA-G, -E and -F regulatory and coding region variability and haplotypes in the Beninese Toffin population sample. *Mol. Immunol.* **104**, 108–127 (2018).

93. Guimarães de Oliveira, M. L. *et al.* Genetic diversity of the LILRB1 and LILRB2 coding regions in an admixed Brazilian population sample. *bioRxiv* (2021) doi:10.1101/2021.04.16.440206.

94. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis

Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).

95. Patterson, M. *et al.* WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. *J. Comput. Biol.* **22**, 498–509 (2015).

96. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–4 (2012).

97. Deveci, M., Catalyürek, U. V. & Toland, A. E. mrSNP: software to detect SNP effects on microRNA binding. *BMC Bioinformatics* **15**, 73 (2014).

98. Sticht, C., De La Torre, C., Parveen, A. & Gretz, N. miRWalk: An online resource for prediction of microRNA binding sites. *PLoS One* **13**, e0206239 (2018).

99. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).

100. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14**, 192–200 (2014).

101. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

102. Delaneau, O. *et al.* A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

103. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

104. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2011).

105. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics* **26**, 2064–2065 (2010).

106. Kelleher, J. & Lohse, K. Coalescent Simulation with msprime. *Methods Mol. Biol.* **2090**, 191–230 (2020).

107. Mendoza-Revilla, J. *et al.* Disentangling Signatures of Selection Before and After European Colonization in Latin Americans. *Mol. Biol. Evol.* **39**, (2022).

108. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows

into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).

109. Mastrangelo, S. *et al.* Runs of homozygosity reveal genome-wide autozygosity in Italian sheep breeds. *Animal Genetics* vol. 49 71–81 Preprint at https://doi.org/10.1111/age.12634 (2018).

110. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

111. Szpiech, Z. A., Blant, A. & Pemberton, T. J. GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. *Bioinformatics* **33**, 2059–2062 (2017).

112. Maróstica, A. S. *et al.* How HLA diversity is apportioned: influence of selection and relevance to transplantation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **377**, 20200420 (2022).

113. Meirmans, P. G. & Hedrick, P. W. Assessing population structure: F(ST) and related measures. *Mol. Ecol. Resour.* **11**, 5–18 (2011).

114. Horimoto, A. R. V. R. *et al.* Genome-Wide Admixture Mapping of Estimated Glomerular Filtration Rate and Chronic Kidney Disease Identifies European and African Ancestry-of-Origin Loci in Hispanic and Latino Individuals in the United States. *Journal of the American Society of Nephrology* vol. 33 77–87 Preprint at https://doi.org/10.1681/asn.2021050617 (2022).

115. Gogarten, S. M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).

116. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* **39**, 276–293 (2015).

117. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

1. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. Genetics 129, 513–523 (1991).

2. Gillespie, J. H. Population Genetics: A Concise Guide. (Johns Hopkins University Press, 2004).

3. Nielsen, R. et al. Tracing the peopling of the world through genomics. Nature 541, 302–310 (2017).

4. Gravel, S. et al. Demographic history and rare allele sharing among human populations. Proc. Natl.

Acad. Sci. U. S. A. 108, 11983–11988 (2011).

5. Hey, J. On the number of New World founders: a population genetic portrait of the peopling of the Americas. PLoS Biol. 3, e193 (2005).

6. Harris, E. E. & Meyer, D. The molecular signature of selection underlying human adaptations. Am. J. Phys. Anthropol. Suppl 43, 89–130 (2006).

7. Fu, W. et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216–220 (2012).

8. Simons, Y. B., Turchin, M. C., Pritchard, J. K. & Sella, G. The deleterious mutation load is insensitive to recent population history. Nat. Genet. 46, 220–224 (2014).

9. Kehdy, F. S. G. et al. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. Proc. Natl. Acad. Sci. U. S. A. 112, 8696–8701 (2015).

10. Chi, C. et al. Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry. PLoS Genet. 15, e1007808 (2019).

11. Winkler, C. A., Nelson, G. W. & Smith, M. W. Admixture mapping comes of age. Annu. Rev. Genomics Hum. Genet. 11, 65–89 (2010).

12. Secolin, R. et al. Distribution of local ancestry and evidence of adaptation in admixed populations. Sci. Rep. 9, 13900 (2019).

13. Pena, S. D. J., Santos, F. R. & Tarazona-Santos, E. Genetic admixture in Brazil. in American Journal of Medical Genetics Part C: Seminars in Medical Genetics vol. 184 928–938 (Wiley Online Library, 2020).

14. Naslavsky, M. S. et al. Whole-genome sequencing of 1,171 elderly admixed individuals from the largest Latin American metropolis (São Paulo, Brazil). Cold Spring Harbor Laboratory 2020.09.15.298026 (2020) doi:10.1101/2020.09.15.298026.

15. Patrinos, G. P. et al. Roadmap for Establishing Large-Scale Genomic Medicine Initiatives in Low- and Middle-Income Countries. Am. J. Hum. Genet. 107, 589–595 (2020).

16. Rocha, C. S., Secolin, R., Rodrigues, M. R., Carvalho, B. S. & Lopes-Cendes, I. The Brazilian Initiative on Precision Medicine (BIPMed): fostering genomic data-sharing of underrepresented populations. NPJ Genom Med 5, 42 (2020).

17. Klein, J. & Sato, A. The HLA system. First of two parts. N. Engl. J. Med. 343, (2000).

18. Hedrick, P. W. What is the evidence for heterozygote advantage selection? Trends Ecol. Evol. 27, 698–704 (2012).

19. Borghans, J. A. M., Beltman, J. B. & De Boer, R. J. MHC polymorphism under host-pathogen coevolution. Immunogenetics vol. 55 732–739 Preprint at https://doi.org/10.1007/s00251-003-0630-5 (2004).

20. Lenz, T. L., Spirin, V., Jordan, D. M. & Sunyaev, S. R. Excess of Deleterious Mutations around HLA Genes Reveals Evolutionary Cost of Balancing Selection. Mol. Biol. Evol. 33, 2555–2564 (2016).

21. Dendrou, C. A., Petersen, J., Rossjohn, J. & Fugger, L. HLA variation and disease. Nat. Rev. Immunol. 18, 325–339 (2018).

22. Bitarello, B. D. et al. Signatures of Long-Term Balancing Selection in Human Genomes. Genome Biol. Evol. 10, 939–955 (2018).

23. Wende, H., Colonna, M., Ziegler, A. & Volz, A. Organization of the leukocyte receptor cluster (LRC) on human chromosome 19q13.4. Mamm. Genome 10, 154–160 (1999).

24. Trowsdale, J., Jones, D. C., Barrow, A. D. & Traherne, J. A. Surveillance of cell and tissue perturbation by receptors in the LRC. Immunol. Rev. 267, 117–136 (2015).

25. Vivier, E. et al. Innate or adaptive immunity? The example of natural killer cells. Science 331, 44–49 (2011).

26. Single, R. M., Martin, M. P., Meyer, D., Gao, X. & Carrington, M. Methods for assessing gene content diversity of KIR with examples from a global set of populations. Immunogenetics 60, 711–725 (2008).

27. Dupont, B., Selvakumar, A. & Steffens, U. The killer cell inhibitory receptor genomic region on human chromosome 19q13.4. Tissue Antigens 49, 557–563 (1997).

28. Qi, Y. et al. KIR/HLA pleiotropism: protection against both HIV and opportunistic infections. PLoS Pathog. 2, e79 (2006).

29. Jones, D. C. et al. Killer immunoglobulin-like receptor gene repertoire influences viral load of primary human cytomegalovirus infection in renal transplant patients. Genes Immun. 15, 562–568 (2014).

30. Farias, T. D. J., Augusto, D. G., de Almeida, R. C., Malheiros, D. & Petzl-Erler, M. L. Screening the full leucocyte receptor complex genomic region revealed associations with pemphigus that might be explained by gene regulation. Immunology 156, 86–93 (2019).

31. Kulkarni, S., Martin, M. P. & Carrington, M. The Yin and Yang of HLA and KIR in human disease. Semin. Immunol. 20, 343–352 (2008).

32. Diaz-Peña, R. et al. Analysis of Killer Immunoglobulin-Like Receptor Genes in Colorectal Cancer. Cells 9, (2020).

33. Besson, C. et al. Association of killer cell immunoglobulin-like receptor genes with Hodgkin's lymphoma in a familial study. PLoS One 2, e406 (2007).

34. Shaffer, B. C. & Hsu, K. C. How important is NK alloreactivity and KIR in allogeneic transplantation? Best Pract. Res. Clin. Haematol. 29, 351–358 (2016).

35. Wroblewski, E. E., Parham, P. & Guethlein, L. A. Two to Tango: Co-evolution of Hominid Natural Killer Cell Receptors and MHC. Front. Immunol. 10, 177 (2019).

36. Norman, P. J. et al. Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. PLoS Genet. 9, e1003938 (2013).

37. Single, R. M. et al. Global diversity and evidence for coevolution of KIR and HLA. Nat. Genet. 39, 1114–1119 (2007).

38. Popejoy, A. B. & Fullerton, S. M. Genomics is failing on diversity. Nature vol. 538 161–164 (2016).

39. Sirugo, G., Williams, S. M. & Tishkoff, S. A. The Missing Diversity in Human Genetic Studies. Cell 177, 1080 (2019).

40. Error - Cookies Turned Off. https://onlinelibrary.wiley.com/doi/epdf/10.1111/tan.13723.

41. Boquett, J. A., Bisso-Machado, R., Zagonel-Oliveira, M., Schüler-Faccini, L. & Fagundes, N. J. R. HLA diversity in Brazil. Hladnikia 95, 3–14 (2020).

42. Nielsen, R., Paul, J. S., Albrechtsen, A. & Song, Y. S. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443–451 (2011).

43. Zheng, X. et al. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28, 3326–3328 (2012).

44. Brandt, D. Y. C. et al. Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. G3 5, 931–941 (2015).

45. Byrska-Bishop, M. et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. bioRxiv (2021) doi:10.1101/2021.02.06.430068.

46. Aguiar, V. R. C., César, J., Delaneau, O., Dermitzakis, E. T. & Meyer, D. Expression estimation and eQTL mapping for HLA genes with a personalized pipeline. PLoS Genet. 15, e1008091 (2019).

47. Aguiar, V. R. C., Masotti, C., Camargo, A. A. & Meyer, D. HLApers: HLA Typing and Quantification of Expression with Personalized Index. Methods Mol. Biol. 2120, 101–112 (2020).

48. Castelli, E. C., Paz, M. A., Souza, A. S., Ramalho, J. & Mendes-Junior, C. T. Hla-mapper: An application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. Hum. Immunol. 79, 678–684 (2018).

49. Souza, A. S. et al. Hla-C genetic diversity and evolutionary insights in two samples from Brazil and Benin. Hladnikia 96, 468–486 (2020).

50. Lima, T. H. A. et al. HLA-A promoter, coding, and 3'UTR sequences in a Brazilian cohort, and their evolutionary aspects. Hladnikia 93, 65–79 (2019).

51. Weiss, E. et al. KIR2DL4 genetic diversity in a Brazilian population sample: implications for transcription regulation and protein diversity in samples with different ancestry backgrounds. Immunogenetics 73, 227–241 (2021).

52. Meyer, D. & Nunes, K. HLA imputation, what is it good for? Hum. Immunol. 78, 239–241 (2017).

53. Nunes, K. et al. HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set. Hum. Immunol. 77, 307–312 (2016).

54. Vince, N. et al. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. Genet. Epidemiol. 44, 733–740 (2020).

55. Castelli, E. C. et al. 'Immunogenetics of resistance to SARS-CoV-2 infection in discordant couples'. (2021) doi:10.1101/2021.04.21.21255872.

56. Luis B. Barreiro, L. Q.-M. Evolutionary and Population (Epi)Genetics of Immunity to Infection. Hum. Genet. 139, 723 (2020).

57. Meyer, D., C Aguiar, V. R., Bitarello, B. D., C Brandt, D. Y. & Nunes, K. A genomic perspective on HLA

evolution. Immunogenetics 70, 5–27 (2018).

58. Staff, T. P. G. & The PLOS Genetics Staff. Correction: Strong Selection at MHC in Mexicans since Admixture. PLOS Genetics vol. 12 e1005983 Preprint at https://doi.org/10.1371/journal.pgen.1005983 (2016).

59. Zhou, Q., Zhao, L. & Guan, Y. Strong Selection at MHC in Mexicans since Admixture. PLOS Genetics vol. 12 e1005847 Preprint at https://doi.org/10.1371/journal.pgen.1005847 (2016).

60. Norris, E. T. et al. Admixture-enabled selection for rapid adaptive evolution in the Americas. Genome Biol. 21, 1–12 (2020).

61. Szpiech, Z. A. & Hernandez, R. D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol. Biol. Evol. 31, 2824–2827 (2014).

62. Templeton, A. R. The future of human evolution. How evolution shapes our lives: essays on biology and society, Losos JB, Lenski RE (eds) 362–379 (2016).

63. Boca, S. M., Huang, L. & Rosenberg, N. A. On the heterozygosity of an admixed population. J. Math. Biol. 81, 1217–1250 (2020).

64. Brandt, D. Y. C., César, J., Goudet, J. & Meyer, D. The Effect of Balancing Selection on Population Differentiation: A Study with HLA Genes. G3 8, 2805–2815 (2018).

65. Nunes, K. et al. How natural selection shapes genetic differentiation in the MHC region: A case study with Native Americans. Human Immunology Preprint at https://doi.org/10.1016/j.humimm.2021.03.005 (2021).

66. Weir, B. S. & Goudet, J. A Unified Characterization of Population Structure and Relatedness. Genetics 206, 2085–2103 (2017).

67. Nunes, K. et al. How Ancestry Influences the Chances of Finding Unrelated Donors: An Investigation in Admixed Brazilians. Front. Immunol. 11, 584950 (2020).

68. Henn, B. M., Botigué, L. R., Bustamante, C. D., Clark, A. G. & Gravel, S. Estimating the mutation load in human genomes. Nat. Rev. Genet. 16, 333–343 (2015).

69. Henn, B. M. et al. Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc. Natl. Acad. Sci. U. S. A. 113, E440–E449 (2016).

70. Lohmueller, K. E. The distribution of deleterious genetic variation in human populations. Curr. Opin.

Genet. Dev. 29, 139–146 (2014).

71. Gravel, S. When Is Selection Effective? Genetics 203, 451–462 (2016).

72. Szpiech, Z. A. et al. Long runs of homozygosity are enriched for deleterious variation. Am. J. Hum. Genet. 93, 90–102 (2013).

73. Lemes, R. B. et al. Inbreeding estimates in human populations: Applying new approaches to an admixed Brazilian isolate. PLoS One 13, e0196360 (2018).

74. Szpiech, Z. A. et al. Ancestry-Dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity. Am. J. Hum. Genet. 105, 747–762 (2019).

75. da Silva, A. K. L. S., Madrigal, L., da Silva, H. P. & Others. Relationships among genomic ancestry, clinical manifestations, socioeconomic status, and skin color of people with sickle cell disease in the State of Pará, Amazonia, Brazil. Antropologia Portuguesa 159–176 (2020).

76. Sebastiani, P. et al. A network model to predict the risk of death in sickle cell disease. Blood 110, 2727–2735 (2007).

77. Menzel, S. & Thein, S. L. Genetic Modifiers of Fetal Haemoglobin in Sickle Cell Disease. Mol. Diagn. Ther. 23, 235–244 (2019).

78. Lewontin, R. C. The Apportionment of Human Diversity. in Evolutionary Biology: Volume 6 (eds. Dobzhansky, T., Hecht, M. K. & Steere, W. C.) 381–398 (Springer US, 1972).

79. Kamariza, M., Crawford, L., Jones, D. & Finucane, H. Misuse of the term ‘trans-ethnic’ in genomics research. Nature Genetics vol. 53 1520–1521 Preprint at https://doi.org/10.1038/s41588-021-00952-6 (2021).

80. Popejoy, A. B. Too many scientists still say Caucasian. Nature 596, 463 (2021).

81. Ramos, B. R. de A. et al. Neither self-reported ethnicity nor declared family origin are reliable indicators of genomic ancestry. Genetica 144, 259–265 (2016).

82. Guo, Y. et al. Development and evaluation of a transfusion medicine genome wide genotyping array. Transfusion 59, 101–111 (2019).

83. Matey-Hernandez, M. L., Danish Pan Genome Consortium, Brunak, S. & Izarzugaza, J. M. G. Benchmarking the HLA typing performance of Polysolver and Optitype in 50 Danish parental trios. BMC Bioinformatics 19, 239 (2018).

84. Bai, Y., Wang, D. & Fury, W. PHLAT: Inference of High-Resolution HLA Types from RNA and Whole Exome Sequencing. Methods Mol. Biol. 1802, 193–201 (2018).

85. Dilthey, A. T. et al. HLA*LA-HLA typing from linearly projected graph alignments. Bioinformatics 35, 4394–4396 (2019).

86. Szolek, A. et al. OptiType: precision HLA typing from next-generation sequencing data. Bioinformatics 30, 3310–3316 (2014).

87. Lee, H. & Kingsford, C. Kourami: graph-guided assembly for novel human leukocyte antigen allele discovery. Genome Biol. 19, 16 (2018).

88. Huang, Y. et al. HLAreporter: a tool for HLA typing from next generation sequencing data. Genome Med. 7, 25 (2015).

89. Xie, C. et al. Fast and accurate HLA typing from short-read next-generation sequence data with xHLA. Proc. Natl. Acad. Sci. U. S. A. 114, 8059–8064 (2017).

90. Norman, P. J. et al. Defining KIR and HLA Class I Genotypes at Highest Resolution via High-Throughput Sequencing. Am. J. Hum. Genet. 99, 375–391 (2016).

91. Sonon, P. et al. Human leukocyte antigen (HLA)-F and -G gene polymorphisms and haplotypes are associated with malaria susceptibility in the Beninese Toffin children. Infect. Genet. Evol. 92, 104828 (2021).

92. Sonon, P. et al. HLA-G, -E and -F regulatory and coding region variability and haplotypes in the Beninese Toffin population sample. Mol. Immunol. 104, 108–127 (2018).

93. Guimarães de Oliveira, M. L. et al. Genetic diversity of the LILRB1 and LILRB2 coding regions in an admixed Brazilian population sample. bioRxiv (2021) doi:10.1101/2021.04.16.440206.

94. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr. Protoc. Bioinformatics 43, 11.10.1–11.10.33 (2013).

95. Patterson, M. et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads. J. Comput. Biol. 22, 498–509 (2015).

96. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19, 1655–1664 (2009).

97. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of

genomes. Nat. Methods 9, 179–181 (2011).

98. Ewing, G. & Hermisson, J. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26, 2064–2065 (2010).

99. Kelleher, J. & Lohse, K. Coalescent Simulation with msprime. Methods Mol. Biol. 2090, 191–230 (2020).

100. Mendoza-Revilla, J. et al. Disentangling Signatures of Selection Before and After European Colonization in Latin Americans. Mol. Biol. Evol. 39, (2022).

101. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of homozygosity: windows into population history and trait architecture. Nat. Rev. Genet. 19, 220–234 (2018).

102. Mastrangelo, S. et al. Runs of homozygosity reveal genome-wide autozygosity in Italian sheep breeds. Animal Genetics vol. 49 71–81 Preprint at https://doi.org/10.1111/age.12634 (2018).

103. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559–575 (2007).

104. Szpiech, Z. A., Blant, A. & Pemberton, T. J. GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. Bioinformatics 33, 2059–2062 (2017).

105. Maróstica, A. S. et al. How HLA diversity is apportioned: influence of selection and relevance to transplantation. Philos. Trans. R. Soc. Lond. B Biol. Sci. 377, 20200420 (2022).

106. Meirmans, P. G. & Hedrick, P. W. Assessing population structure: F(ST) and related measures. Mol. Ecol. Resour. 11, 5–18 (2011).

107. Horimoto, A. R. V. R. et al. Genome-Wide Admixture Mapping of Estimated Glomerular Filtration Rate and Chronic Kidney Disease Identifies European and African Ancestry-of-Origin Loci in Hispanic and Latino Individuals in the United States. Journal of the American Society of Nephrology vol. 33 77–87 Preprint at https://doi.org/10.1681/asn.2021050617 (2022).

108. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35, 5346–5348 (2019).

109. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. Genet. Epidemiol. 39, 276–293 (2015).

110. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894 (2019).