Received: 23 January 2024 | Revised: 3 May 2024 | Accepted: 13 May 2024

DOI: 10.1111/tan.15543

ORIGINAL ARTICLE

HLA Immune Response Genetics WILEY

A multi-ethnic reference panel to impute *HLA* classical and non-classical class I alleles in admixed samples: Testing imputation accuracy in an admixed sample from Brazil

Nayane S. B. Silva[1,2,3] | **Sonia Bourguiba-Hachemi**[1] | **Viviane A. O. Ciriaco**[2] | **Stefan H. Y. Knorst**[4] | **Ramon T. Carmo**[4] | **Cibele Masotti**[4] | **Diogo Meyer**[5] | **Michel S. Naslavsky**[5,6] | **Yeda A. O. Duarte**[5,7] | **Mayana Zatz**[5,6] | **Pierre-Antoine Gourraud**[1] | **Sophie Limou**[1] | **Erick C. Castelli**[2,3] | **Nicolas Vince**[1]

[1]Center for Research in Transplantation and Translational Immunology, Nantes Université, INSERM, Ecole Centrale Nantes, Nantes, France

[2]Molecular Genetics and Bioinformatics Laboratory, School of Medicine, São Paulo State University, Botucatu, State of São Paulo, Brazil

[3]Genetics Program, Institute of Biosciences of Botucatu, São Paulo State University, Botucatu, State of São Paulo, Brazil

[4]Department of Molecular Oncology, Hospital Sírio-Libanes, São Paulo, Brazil

[5]Department of Genetics and Evolutionary Biology, Biosciences Institute, University of São Paulo, São Paulo, State of São Paulo, Brazil

[6]Human Genome and Stem Cell Research Center, University of São Paulo, São Paulo, State of São Paulo, Brazil

[7]Medical-Surgical Nursing Department, School of Nursing, University of São Paulo, São Paulo, State of São Paulo, Brazil

Correspondence
Erick C. Castelli, Departamento de Patologia, Faculdade de Medicina de Botucatu, Unesp—Botucatu, State of São Paulo, 18618970, Brazil.
Email: erick.castelli@unesp.br

Nicolas Vince, CRTI UMR1064—ITUN, CHU Nantes Hôtel Dieu, 30 bld Jean Monnet, 44093 Nantes Cedex 01, France.
Email: nicolas.vince@univ-nantes.fr

Funding information
Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Numbers: 2021/02815-8, 2021/14851-9; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior; Programme of Investments for the Future Agence Nationale de la Recherche PIA-Investment; Comité Français d'Évaluation de la Coopération Universitaire et Scientifique avec le Brésil (COFECUB)

The *MHC* class I region contains crucial genes for the innate and adaptive immune response, playing a key role in susceptibility to many autoimmune and infectious diseases. Genome-wide association studies have identified numerous disease-associated SNPs within this region. However, these associations do not fully capture the immune-biological relevance of specific *HLA* alleles. *HLA* imputation techniques may leverage available SNP arrays by predicting allele genotypes based on the linkage disequilibrium between SNPs and specific *HLA* alleles. Successful imputation requires diverse and large reference panels, especially for admixed populations. This study employed a bioinformatics approach to call SNPs and *HLA* alleles in multi-ethnic samples from the 1000 genomes (1KG) dataset and admixed individuals from Brazil (SABE), utilising 30X whole-genome sequencing data. Using HIBAG, we created three reference panels: 1KG ($n = 2504$), SABE ($n = 1171$), and the full model ($n = 3675$) encompassing all samples. In extensive cross-validation of these reference panels, the multi-ethnic 1KG reference exhibited overall superior performance than the reference with only Brazilian samples. However, the best results were achieved with the full model. Additionally, we expanded the scope

Erick C. Castelli and Nicolas Vince have contributed equally to this work and share the last authorship and corresponding author.

of imputation by developing reference panels for non-classical, *MICA*, *MICB* and *HLA-H* genes, previously unavailable for multi-ethnic populations. Validation in an independent Brazilian dataset showcased the superiority of our reference panels over the Michigan Imputation Server, particularly in predicting HLA-B alleles among Brazilians. Our investigations underscored the need to enhance or adapt reference panels to encompass the target population's genetic diversity, emphasising the significance of multiethnic references for accurate imputation across different populations.

KEYWORDS

Admixed Population, HLA imputation, HLA class I, Reference Panel

1 | INTRODUCTION

The major histocompatibility complex (MHC) constitutes a gene-dense region on the short arm of chromosome 6, at 6p21.3. It spans a 4 Mb region with around 250 coding and non-coding genes.[1] Its structural organisation evolved gradually through several mutations, duplications, deletions and genomic rearrangement events.[2] The MHC region is the most variable region of the human genome, particularly at the *HLA* genes.[3] The *HLA* genes encode molecules involved in the antigen processing and presentation pathway and are therefore critical to the immune response to pathogens. The most polymorphic *HLA* genes encode transmembrane glycoproteins that bind to endogenous or exogenous peptides and present them to T lymphocytes selected to distinguish self-peptides from non-self-peptides (derived from viruses, bacteria, other pathogens, mutated endogenous proteins or proteins in the context of transplantation),[4] thus playing a critical role in adaptive immune responses and susceptibility to various pathological conditions.[1] Additional less polymorphic *HLA* genes, also called non-classical, are involved in different immune processes such as NK cells modulation (*HLA-E*, *HLA-F*, and *HLA-G*).[5]

HLA alleles have been extensively studied in several disorders,[6] and particularly associated with autoimmmune and infectious disease susceptibility, resistance, severity and particular clinical outcomes. For instance, the *HLA-DQB1*03:02* is associated with type 1 diabetes development,[7] *HLA-C*12:02* and *HLA-DQB1*06:01* with severe acute hepatitis protection,[8] and *HLA-B*27* strongly with ankylosing spondylitis.[9] Generally, the mechanism underlying these associations is related to differential antigen presentation.

In recent decades, genome-wide association studies (GWAS) have emerged as a powerful tool to identify associations between common genetic variants throughout the human genome and phenotypes.[10] GWAS, which involves SNP microarrays for the cost-effective genotyping of hundreds of thousands of variants,[11] has substantiated the pivotal role of *MHC* genes in several diseases and traits.[10] However, SNP associations alone do not convey the immune-biological relevance of specific *HLA* alleles, which represent a haplotype of SNPs, particularly considering that variants in high LD (linkage disequilibrium) are usually not included in such SNP arrays. Therefore, causal variants that follow the associated SNP in a given *HLA* allele may not be included in the assay.[12] To overcome this issue, we can use *HLA* imputation techniques to leverage this data and predict individuals' *HLA* alleles based on LD between GWAS-derived SNP data for the *MHC* region and specific *HLA* alleles.[13]

SNP to *HLA* imputation algorithms provide a fast, cost-effective and straightforward method for obtaining unsurveyed *HLA* alleles from widely available GWAS SNP genotyping data.[12] These methods ultimately rely on reference datasets containing *HLA* alleles and SNP genotypes for the same individuals. Their predictions have become increasingly accurate as new algorithms are developed.[13,14] However, the imputation accuracy is affected by several factors, such as reference panel size, polymorphism extent in the imputed locus, presence of rare or population-specific alleles (leading to less precise imputation due to underrepresentation in reference panels), and evolutionary proximity between the reference panel and the samples being imputed.[14,15]

In this context, we have built different reference panels for all class I genes, including classical and non-classical *HLA*, along with the *HLA-H* pseudogene and *MICA*/*MICB* genes. Imputation of these genes was previously performed only using a reference panel composed of Japanese[16] or mainly European[17] samples. We used samples from both Brazil and worldwide populations for our reference panels. Extensive cross-validation and subsequent imputation within an independent Brazilian population validated the accuracy of our models. This study is part of the SNP-HLA Reference Consortium (SHLARC)[12] international network, which aims to improve and share large reference panels with the immunogenetic community.

2 | METHODS

2.1 | Samples used to build reference panels

The *MHC* region was investigated using high-coverage (30X) whole-genome sequencing data from 3675 samples from different populations worldwide. Among these samples, there were 2504 individuals from 26 different populations of the latest release of the 1000 Genomes Project (1KG),[18,19] and 1171 samples from individuals over 60 years of age from the city of São Paulo, Brazil, enrolled at the longitudinal Health, Well-Being and Aging cohort (SABE—Saúde, Bem-estar e Envelhecimento).[20] We downloaded the original BAM files from both, with reads aligned to the hg38 reference genome. For all samples, we extracted all reads mapped to the *MHC* region and all unmapped reads using *samtools*,[21] producing a smaller BAM file for each sample.

2.2 | Genotyping *HLA* genes by NGS: alignment optimization, SNP calls and allele calls

The methodology applied to call SNPs across the *MHC* and *HLA* alleles in the training data are available at https://github.com/erickcastelli/HLA_genotyping. Briefly, we used *hla-mapper dna*, version 4.1,[22] to optimise the *HLA* alignments minimising cross-alignments and unmapped reads from *HLA* genes. The input files for the hla-mapper are the BAM files obtained in the previous step. The *hla-mapper* program is available at www.castelli-lab.net/apps/hla-mapper. To call genotypes, we used the *HaplotypeCaller* algorithm from the *Genome Analysis Toolkit* (*GATK*),[23] version 4.1, in the GVCF mode, filtering out artefacts and false positives using the *GATK VQSR* tool. After applying the *VQSR* and filtering out the variants that did not pass the test, we selected the remaining variants and processed them with the *checkpl* tool of the *vcfx package* (www.castelli-lab.net/apps/vcfx). Haplotypes were detected by combining physical and probabilistic inference using *WhatsHap*[24] and *Shapeit4*.[25] After, the complete sequences of each individual were reconstructed from the phased VCF file using the *vcfx fasta* to create complete genomic (exon + introns) and CDS (only exons) sequences for each sample, resulting in one for each chromosome. Due to *HLA-B* and *HLA-C* being encoded on the reverse strand of chromosome 6 in the hg38 reference genome, the sequences were reversed and complemented. This process allowed the detection of *HLA* alleles through direct comparison with known sequences deposited in the IPD-IMGT/HLA database[3] at both genomic (4-field alleles) and exonic (3-field alleles) levels. Additionally, EMBOSS transeq[26] was employed to translate the CDS sequences into protein sequences. These protein sequences were then used to define the allotypes, referring to the 2-field allele. The outcome of this methodology was to obtain data on both SNPs and *HLA* alleles of each individual included in the study. Please refer to https://github.com/erickcastelli/HLA_genotyping for a detailed description of the methodology. Additionally, to evaluate the accuracy of our method, we compared our results with Sanger Sequencing HLA typing by Gourraud et al.[27] in a subset of 965 samples from the 1KG dataset.[15] An additional step was required for *MICA* and *HLA-H* genes evaluation since they present copy number variations (CNV) in some individuals. To evaluate the presence of *MICA* and *HLA-H* gene deletions, we compared the ratio between the read depth observed in both genes and a reference gene (*TNF*) using *samtools*.[21] Then, we plotted these ratios to observe the samples with no copies of these genes, single copies, or individuals with two copies.

2.3 | *HLA* imputation and cross-validation

We used HIBAG (version 1.4) and its GPU extension (version 1.19.1) in R, to create reference panels and perform HLA imputation. Reference panels for *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E*, *HLA-F*, *HLA-G*, *HLA-H*, *MICA*, and *MICB* were built using 100 classifiers and SNPs within a 500-kilobase flanking region for each gene.[15] The parameters used for building the models were set according to the recommendations provided by the HIBAG authors.[28] The computations were executed on GPU nodes at the *Centre de calcul intensif des Pays de la Loire* (CCIPL) located at *Nantes Université* in France. We created three distinct imputation models: 1KG, SABE (Brazilians only), and full (1KG + SABE). For evaluation, a 10-fold cross-validation approach was undertaken, involving the random subsampling of individuals from each super-population (defined by their genetic ancestry group) in 1KG, and SABE, that is, we generated numerous pairs of reference panels and test sets by randomly subsampling individuals from each super-population in 1KG (AFR, AMR, EAS, SAS and EUR) and SABE (Figure S1). Individuals of the test set were excluded from the paired reference panel. SNP data were managed using PLINK,[29] which underwent quality control procedures such as removing SNPs exhibiting more than 2% missing genotypes or less than 1% minor allele frequency. Additionally, we removed SNPs with A/T or G/C ambiguities and selected those within the HSL (Hospital Sirio-Libanes) validation

dataset SNP array (Axiom_HGCoV2_1, Thermo Fisher Scientific). The resulting dataset contained 6292 SNPs within the *MHC* region (chromosome 6, from 29 to 34 Mb). Overall, within genes and their flanking regions extending 500 kb upstream and downstream of it, there were a total of 1527 SNPs for *HLA-A*, 2077 for *HLA-B*, 2166 for *HLA-C*, 1708 for *HLA-E*, 1099 for *HLA-F*, 1255 for *HLA-G*, 1371 for *HLA-H*, 2067 for *MICA*, and 1961 for *MICB*. HIBAG randomly selects a subset of these SNPs for each locus, and then determines which SNP haplotypes are the best *HLA* allele predictors, repeating this process 100 times. The selected SNPs are depicted in Figure S2 for *HLA-A*, *HLA-B* and *HLA-C*. All SNPs within the target gene were included in the model.

2.4 | Validation of the reference panels

We validated the reference panels by assessing the accuracy of the imputation models in predicting *HLA-A*, *HLA-B* and *HLA-C* genes using an independent Brazilian population cohort consisting of 192 samples from São Paulo city. For these individuals, we had access to *HLA* allele calls obtained through the AllTypeTM NGS 11-Loci Amplification kit (OneLambda, ThermoFisher Scientific), and the Axiom Human Genotyping SARS-CoV-2 Research Array (Axiom_HGCoV2_1, Affymetrix, ThermoFisher Scientific). Additionally, we had access to their global ancestry inference (Figure S3A), generated using ADMIXTURE[30] through supervised analysis ($K = 4$) based on multilocus SNP genotypes, with African, European, East Asian and Native American samples from the Human Genome Diversity Project[31] utilised as parental populations. This sample is part of a broader study conducted by the Department of Molecular Oncology at Hospital Sírio-Libanes (HSL) in Brazil, related to COVID-19 susceptibility. Furthermore, we compared the *HLA* imputation performance of our models with those employed by the Michigan Imputation Server, which utilised a reference panel containing approximately 20,000 multi-ethnic samples.[32]

2.5 | Statistical analyses

To evaluate the imputation accuracy, we measured the proportion of correct predictions under the number of alleles (i.e., the number of correct predictions divided by twice the number of individuals in the study, represented as $2n$). A prediction was considered correct when it was concordant with known genotypes. We also considered other metrics when evaluating the models' performance to predict specific alleles, such as precision, sensitivity and the F1 score. Precision represents the proportion of true positive predictions for a specific allele to the total number of positive predictions. It indicates the proportion of positive predictions that were correct for that particular allele. Sensitivity, on the other hand, is the proportion of true positive predictions for a specific allele to the total number of actual positive instances. It indicates the proportion of actual positive alleles that the model correctly identified. The F1 score is computed for each allele as $2 \times (\text{precision} \times \text{sensitivity})/(\text{precision} + \text{sensitivity})$, which ranges from 0 to 1. A score of 1 indicates perfect precision and sensibility, meaning that all positive predictions are correct and all positive instances are identified. A score of 0 indicates that the model's predictions are always wrong. Here, we computed the mean F1 score for each allele and its average for each model.

3 | RESULTS

3.1 | *HLA* class I diversity

Using the hla-mapper bioinformatics pipeline for genotyping at the SNP and *HLA* allele levels, we evaluated the *HLA* class I genetic diversity in 3674 reference samples from 27 worldwide populations. When comparing our results from a subset of 965 samples from the 1KG dataset for *HLA-A*, *HLA-B* and *HLA-C* at the 2-field resolution with Sanger Sequencing typing method, we found compatibility of 99.6%, 99.3% and 98.9%, respectively. This pipeline can assess the full *HLA* class I genetic diversity in all levels, SNPs, indels, haplotypes, and 2 to 4-field resolution alleles obtained from whole-genome sequencing. We used the 2-field alleles (allotypes) and the SNPs across the *MHC* to build the reference panel. Our results are based on 2-field alleles, and we will use 'alleles' to denote 2-field alleles (allotypes) throughout the text. This resolution (2-field alleles) was chosen because it is the most used by association studies. Furthermore, these sequences can be encoded by distinct variations across the gene, all encoding the same protein. Therefore, increasing the resolution would add even more complexity to the imputation process, potentially increasing the error rate. Newly identified alleles, that is not previously described, or alleles not fully characterised, were omitted. However, their summed frequencies are still reported in Table S1. As expected, classical *HLA* genes (*HLA-A*, *HLA-B* and *HLA-C*) showed higher diversity, with *HLA-B* having the highest diversity, consequently harbouring the greatest number of less frequent and rare alleles (Figure S4). Our data cover 81%, 86% and 85% of the alleles classified as 'common' at the G group level by CIWD version 3.0.0[33] for *HLA-A*, *HLA-B* and *HLA-C*,

TABLE 1 The number of different alleles per gene found in this study.

Gene	No of alleles	Gene	No of alleles	Gene	No of alleles
HLA-A	90	*HLA-E*	10	*HLA-H*	24
HLA-B	164	*HLA-F*	5	*MICA*	34
HLA-C	73	*HLA-G*	8	*MICB*	14

respectively. These alleles represent 95%, 92% and 98% of the cumulative allele frequencies identified in our dataset. Table 1 shows the number of different alleles we found per gene across the whole study population. Other genes, such as *HLA-H*, *MICA* and *MICB*, also presented significant diversity, although to a lesser extent. Furthermore, our analysis revealed copy number variations for *HLA-H* and *MICA*, with frequencies of 15% and 1.4% for the full deletion, respectively. The frequency of each allele in the entire dataset is presented in Table S1.

3.2 | *HLA* imputation in worldwide populations

We used three reference panels for HLA imputation—1KG ($n = 2504$), SABE (Brazilian population, $n = 1170$) and 1KG + SABE full reference (all samples, $n = 3674$)—to predict classical and non-classical *HLA* class I alleles. To assess the performance of each reference panel, we conducted 10 resampling iterations of 200 samples for each biogeographic region. Across all populations, the accuracy of all reference panels consistently exceeded 90% for most evaluated genes, with some (for instance *HLA-E*, *HLA-F* and *HLA-G*) reaching remarkably high accuracies ranging from 95% to 100% (Figure 1). Nonetheless, the accuracy was relatively lower for *HLA-A* and *HLA-B* (Figure 1) due to their higher allelic diversity in the general population (Table 1) and, consequently, the increased prevalence of less common and rare alleles, particularly within admixed populations (Figure S4).

In predicting non-classical alleles, the three reference panels exhibited robust performance, achieving an overall mean accuracy that surpassed 95% across all populations. However, lower accuracies were noted for *MICA* in AFR populations across all reference panels, ranging from 92% to 94%. Additionally, a slight reduction in performance was observed when predicting *MICB* in EAS populations, and *MICA*/*HLA-H* in SAS populations using the SABE reference (Table S2). Nevertheless, even in these instances, the average accuracy remained consistently above 90%, underscoring the reliability and effectiveness of the imputation process for non-classical *HLA* class I alleles.

Notably, the EUR population exhibited the highest accuracy of imputation. Conversely, EAS and SAS populations had particularly low accuracies when imputed using the SABE reference, which was expected, since there are just a few samples with Asian ancestry in the SABE cohort and EAS is not an ancestral population of Brazil. The full reference panel demonstrated the best performance in imputing all populations, followed closely by 1KG and SABE. Compared with SABE, the 1KG reference panel generally showed enhanced overall performance, with exceptions observed when predicting subsets from the SABE dataset (Figure 1 and Table S2).

The performance of imputation for specific alleles within the training set generally correlates with their frequencies in the reference panels (Table S3), with some exceptions. For instance, when predicting *HLA* alleles in AFR populations, all reference panels showed limited precision and sensitivity for imputing *HLA-B*51:01*, despite its relatively high frequency within the reference panels (ranging between 5% and 8%). Interestingly, the SABE reference panel displayed improved sensitivity for this allele (Figure S5). Low precision indicates that other alleles are inaccurately predicted as *HLA-B*51:01*, including less frequent alleles, such as *HLA-B*52:01* and *HLA-B*78:01*.

In the AMR population, all reference panels showed low precision (less than 0.5) when predicting *HLA-B*35:01*, despite it being one of the most common *HLA-B* alleles. This pattern was similar in EAS, where precision decreased (to less than 0.6) when predicting the most frequent *HLA-A* allele, *HLA-A*02:01*. Moreover, SAS has reduced precision specifically when utilising the SABE reference panel (with a precision of 0.24) for *HLA-A*02:01*. However, the SABE reference panel displayed perfect sensitivity (100%) for predicting *HLA-A*02:01* in this population, surpassing the performance of the other reference panels (see Table S3).

It is worth noting that population-specific and less frequent alleles are more difficult to predict due to the lack of these alleles in the reference panels. Nevertheless, the absence of some rare alleles within the reference panels, stemming from their exclusion for test dataset purposes, contributed to suboptimal sensitivity and precision in imputation for population-specific and less frequent alleles.

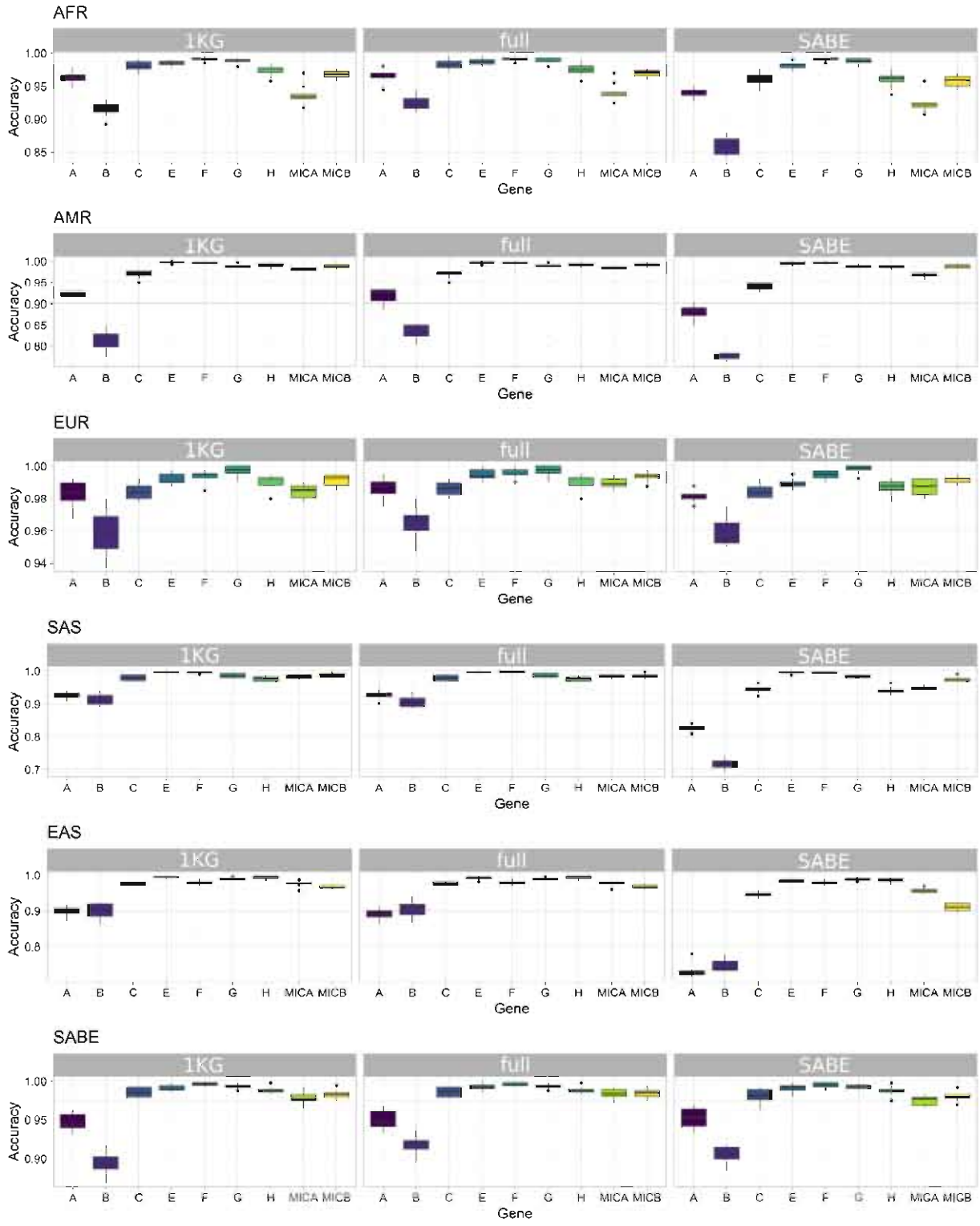

FIGURE 1 *HLA* imputation accuracy across different populations. The bar plots illustrate the distribution of imputation accuracy, represented as the percentage of correct predictions, based on 10 resampling iterations of 200 samples for each population. The populations are grouped into five categories: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). Additionally, there is a sixth category called SABE, which refers to the Brazilian population sample from the SABE cohort. The reference panels used for imputation include 1KG ($n = 2504$), SABE (Brazilian population, $n = 1170$) and the full model (all samples, $n = 3674$).

We also computed the mean F1 score for each allele (Table S3) and its average for each model (Figure 2 and Table S2). Our results indicated that *HLA-B* tended to exhibit lower F1 scores than other genes, except within Asian populations where *HLA-A* displayed lower F1 scores in some cases. Notably, among non-classical genes,

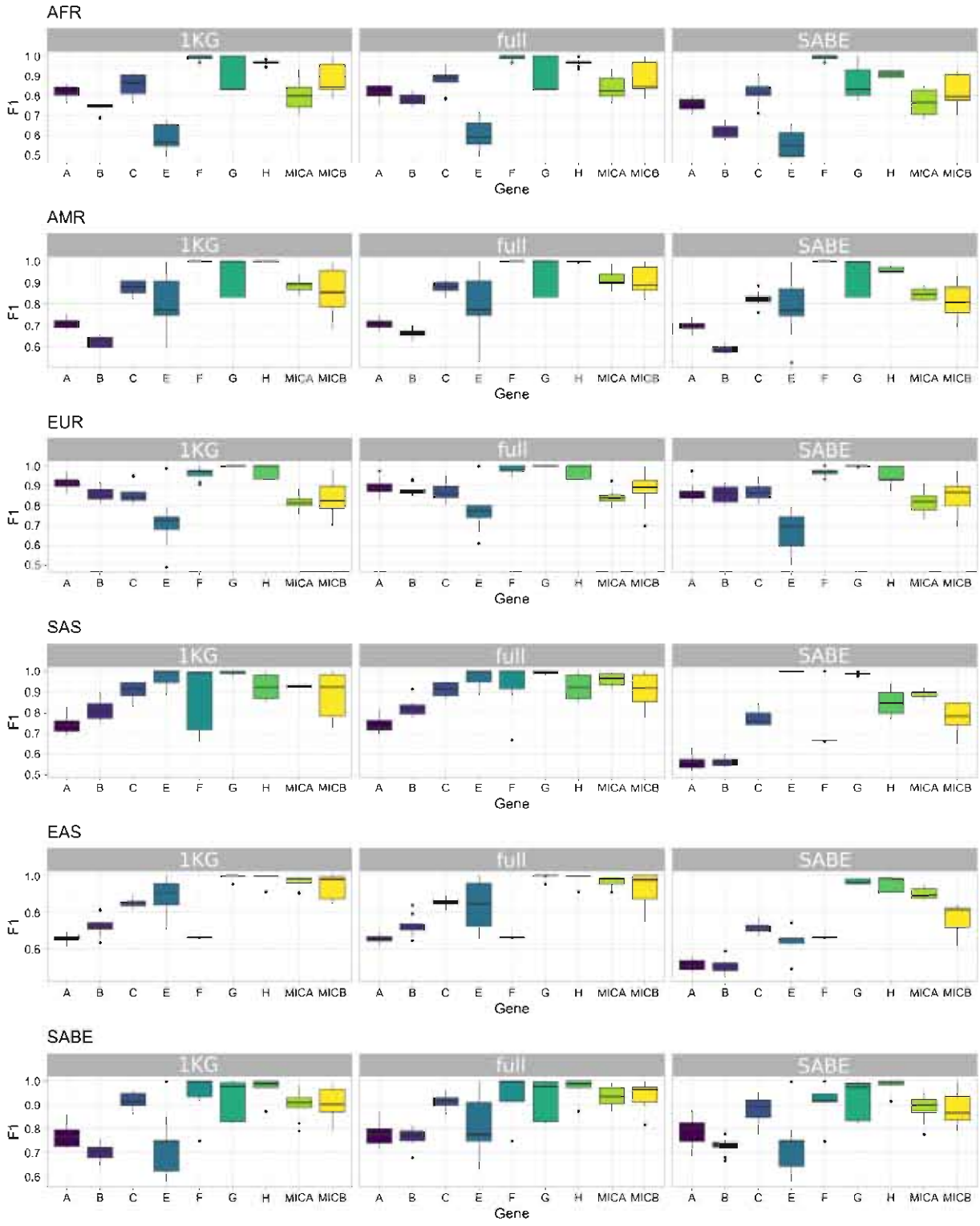

FIGURE 2 *HLA* imputation mean F1 score across different populations. The bar plots illustrate the distribution of the mean F1 score of the 10 resampling iterations of 200 samples for each gene in a specific population. The populations are grouped into five categories: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR) and South Asian (SAS). Additionally, there is a sixth category called SABE, which refers to the Brazilian population sample from the SABE cohort. The reference panels used for imputation include 1KG ($n = 2504$), SABE (Brazilian population, $n = 1170$) and the full model (all samples, $n = 3674$).

HLA-E showed reduced F1 scores, primarily due to the presence of some less frequent alleles. For instance, alleles such as *HLA-E*01:09* among AMR, EUR, AFR and SABE, and *HLA-E*01:11* between EUR and SABE (as shown in Table S3), contributed to this reduction. These alleles exhibit only one or two base substitutions

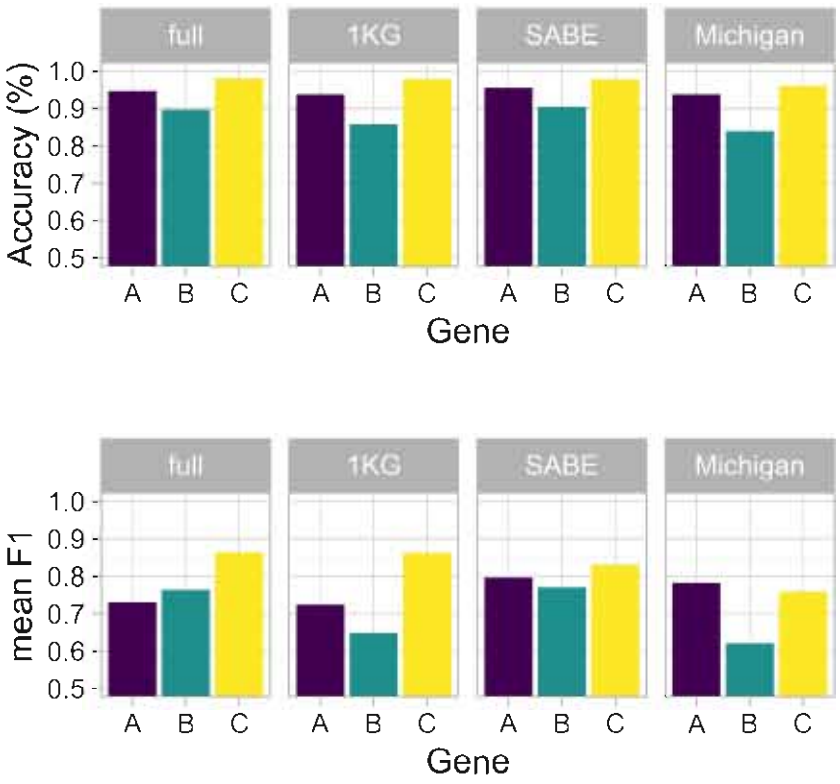

FIGURE 3 Accuracy and mean F1 score of each reference panel to predict *HLA-A*, *HLA-B* and *HLA-C* alleles in a Brazilian population. The reference panels used for imputation include 1KG ($n = 2504$), SABE (Brazilian population, $n = 1170$), and full model (1KG + SABE, $n = 3674$). 'Michigan' refers to the results obtained using the Michigan Imputation server and its associated reference panel. The upper panel shows the accuracy measured as the percentage of correct predictions, and the bottom panel shows the F1 score average of each gene.

compared to *HLA-E*01:01*, which is the most frequent allele worldwide. There are only six copies of *HLA-E*01:09* and *HLA-E*01:11* considering the entire dataset (Table S1). Furthermore, the *HLA-E*01:10* allele presents only one copy in the EAS population, therefore it could not be correctly predicted when selected among the test set. Additionally, among Africans, the presence of infrequent and population-specific alleles, such as *HLA-E*01:13* and *HLA-E*01:47*, which are exclusive to African populations in our dataset, further contributed to the observed reduction in F1 scores but not observed in weighted F1 scores (Figure S6).

3.3 | Testing *HLA* imputation in an independent Brazilian population

We assessed the performance of *HLA* imputation methods for *HLA-A*, *HLA-B* and *HLA-C* using an independent dataset from São Paulo, Brazil. Our findings showed that the SABE reference panel yielded the highest imputation accuracy for *HLA-A* and *HLA-B* in the Brazilian population cohort (Figure 3 upper panel). In addition, we tested *HLA* imputation using the Michigan Imputation Server, which employs a reference panel including over than 20,000 samples from worldwide populations. The SABE reference surpasses the Michigan Server despite being 15× smaller in sample size.

When examining the sensitivity and precision of each model in predicting individual alleles of the target gene, we noticed that the multiethnic reference panel used in the Michigan Server performed poorly, particularly for less common alleles (see Figure S7). This resulted in a lower mean F1 score for the Michigan Server than the other reference panels, as depicted in the bottom panel in Figure 3. In addition, the Michigan Server had the worst performance for *HLA-B* as compared with any other model.

The imputation errors for *HLA-A*, *HLA-B* and *HLA-C* when using the full model are primarily associated with samples exhibiting a higher degree of admixture, defined as those with less than 80% ancestry from any single parental population. The error proportion is higher among admixed samples for *HLA-A* and *HLA-B*, with all errors in *HLA-C* being attributed to admixed samples (Figure S3B).

4 | DISCUSSION

Our findings contribute to the SHLARC ongoing efforts to enhance the power of *HLA* association studies by providing the immunogenetics community with large and diverse reference panels for *HLA* imputation.[12,34] In this study, we assessed the accuracy of different reference panels, built with samples from different genetic

backgrounds, in predicting *HLA* alleles in randomly selected samples and an independent Brazilian dataset. Our results highlighted that imputation accuracy is higher when the model includes samples genetically close to the predicted population. Additionally, we expanded the scope of imputation by developing reference panels for predicting alleles from *MHC* genes typically not considered in imputation but relevant in several immunologic pathways and possibly linked to disease outcomes, including *HLA-H*, *MICA*, *MICB*, *HLA-G*, *HLA-E* and *HLA-F*.[1]

The *HLA* genes are the most polymorphic in the human genome, with over 36,000 alleles currently described in the IPD-IMGT/HLA database.[3] However, this number of alleles is expected to increase significantly, as studies estimate that millions of *HLA* alleles exist in the human population[35] and evidence suggests that de novo mutations will continue to introduce novel *HLA* alleles.[36,37] Due to the high degree of polymorphism observed in these genes, we adopt a methodology suitable to minimise *HLA* genotyping errors, which has been previously validated by other studies to obtain data on the *HLA* genes[38,39] and also here through comparison of our results with those obtained using Sanger sequencing methodologies.[27]

We used SNPs within a 500-kilobase flanking region of each *HLA* gene, thereby including some SNPs from neighbouring genes, to construct the reference panels. Genes located close to the beginning of the *MHC* region (29 Mb), such as *HLA-A*, *HLA-G* and *HLA-F*, had less than 500 kb considered in their upstream regions (Figure S2, bottom panel). However, it does not appear to jeopardise the accuracy of the models. The selection of SNPs for model building using the HIBAG method involves a systematic yet randomised process[28]; thereby, we lack sufficient data to rank SNPs based on their significance for the imputation result. While this systematic approach aims to capture informative SNPs, there remains a rather small chance that significant variants could be randomly omitted, especially for rare alleles. However, utilising 100 classifiers increases the likelihood that each SNP was selected at least once across them, ensuring comprehensive coverage. We assume that most frequently selected SNPs hold greater importance or informational value for *HLA* allele imputation, as suggested by the similarity between *HLA-B* and *HLA-C* overlapping selected SNPs (Figure S2, bottom panel). Conversely, less frequently selected SNPs may be less significant for imputation accuracy. Additionally, SNPs exhibiting strong linkage disequilibrium with the target *HLA* alleles are more likely to be important for imputation accuracy. However, a precise SNP ranking remains a challenge to identify the most informative SNPs for imputation with the randomness at play with HIBAG and the large number of SNP haplotypes involved.

Furthermore, the polymorphism of *HLA* genes differs across populations,[38] with differences in the prevalence of alleles and the way they are linked together into haplotypes within populations.[40] Therefore, specific alleles and haplotypes may be highly frequent or absent in certain populations.[41] This can affect imputation accuracy because the target alleles must be present in the reference for prediction. However, despite our data covering over 80% of the common alleles[33] for classical class I genes, it is essential to consider the haplotype structure into which the target allele is inserted. In our data, a notable example of different haplotypes within populations emerges within the AFR population, where we identified 26 instances of the *HLA-B*51:01* allele, with 20 of these linked to *HLA-C*16:01*. Interestingly, other populations exhibit distinct linkage patterns to different *HLA-C* alleles. Because of this complexity, the observed low precision (Figure S5) for *HLA-B*51:01* suggests the misclassification of other alleles as being *HLA-B*51:01*. In contrast, low sensitivity indicates instances where the model struggles to identify *HLA-B*51:01* itself accurately. haplotypes, specifically *HLA-B*51:01/HLA-C*16:01*, *HLA-B*51:01/HLA-C*18:02* and *HLA-B*51:01/HLA-C*02:10*. These haplotypes seem to be unique to the AFR populations, except *HLA-B*51:01/HLA-C*16:01*, which is also present in the SABE population and have two copies in the EUR samples. Furthermore, *HLA-B*51:01/HLA-C*02:10* is only found once in Brazil (SABE). This advocates for the addition of African ancestry samples in *HLA* imputation reference panels.

While common alleles are often easy to impute due to their great representation in reference panels, challenges arise when specific alleles are found in population-specific haplotype structures, as discussed in the context of *HLA-B*51:01* in AFR populations. Moreover, the presence of rare alleles closely related to common alleles, differing by only a few base substitutions, can drop the precision of common alleles. An example was observed with *HLA-B*35:01* in the AMR populations. Here, alleles exclusive to AMR, such as *HLA-B*35:12*, *HLA-B*35:10*, *HLA-B*35:09* and *HLA-B*35:17*, were often misclassified as *HLA-B*35:01*. Consequently, they reduce the precision of *HLA-B*35:01* due to the presence of false positive predictions. In contrast, we noted a good sensitivity of reference panels in predicting *HLA-B*35:01*, indicating accurate predictions for this allele in most instances. A similar pattern was observed with the common *HLA-A*02:01* allele in the SAS population.

To comprehensively assess the performance of our reference panels, we computed their mean F1 score (Figure 2)—a harmonic mean of sensitivity and precision.

The F1 score provides an overall assessment of how well the model identifies both rare and common alleles. A higher F1 score indicates that the model achieves a good balance between precision and recall, while a lower F1 score suggests that the model may lack accuracy in predicting certain alleles. The use of multiple metrics is crucial for a comprehensive evaluation of the model's performance, particularly when dealing with rare or less common alleles. The F1 score assumes equal importance for all alleles, which may not always be the case. Here, we calculated the mean F1 score, which can lead to a decrease in this metric for genes that have a limited number of alleles, such as *HLA-E*. For instance, *HLA-E* exhibited a mean F1 score lower than expected for non-classical genes while maintaining an accuracy higher than 98% in all cases (Table S2). We complemented this mean F1 score by showing the F1 scores weighted by the frequency of the alleles in Figure S6. Therefore, examining various metrics helps ensure the model accurately identifies all relevant alleles.

In the Brazilian population, which is characterised by high rates of admixture between Europeans, Africans and Amerindian ancestral populations, there is a greater diversity of *HLA* alleles as compared to other populations.[38] The *HLA* allele composition of a given population usually reflects its demographic history and admixture proportions[42] due to the unique demographic history and admixture patterns of each region.[43] For instance, in the northeastern state of Piauí, from Brazil, European and African influences are evident in the moderate frequencies of their corresponding *HLA* alleles, whereas Native American alleles are relatively rare.[44] In contrast, in the southern state of Rio Grande do Sul, the *HLA-B*35* allele is more common, while the *HLA-B*53* allele is prevalent among both admixed and Afro-Brazilians. Additionally, European-origin alleles, such as *HLA-B*35*, *HLA-B*44* and *HLA-B*51* show lower frequencies in admixed Afro-Brazilians.[45] These regional differences highlight the complex, varied origins of the Brazilian population, and underscore the importance of considering local demographics and admixture histories when studying *HLA* allele distributions.

When imputing an independent Brazilian dataset, we noticed a significant enhancement in imputation precision when the SABE reference (composed of Brazilian samples) or the full reference (which includes the SABE reference) is used, compared with the 1KG reference. Notably, we extended this investigation to impute the Brazilian dataset using the Michigan Imputation Server,[32] leveraging a reference panel of more than 20,000 samples. However, our reference panels exhibited superior performance, especially for *HLA-B*. It is worth emphasising that the similarity in genetic ancestry profiles and geographic origins between the Brazilian samples and those comprising the SABE reference likely contributed to this heightened efficacy. Furthermore, the Michigan Server employs a distinct methodology for *HLA* imputation, which may also influence the observed results. In light of this, while extensive multi-ethnic reference panels are certainly essential, the incorporation of sharing the same genetic heritage as the target data remains imperative for optimal performance.

Accurate prediction of most *HLA* alleles demands extensive training datasets, typically requiring approximately 10 copies of a specific allele within the training database to achieve a high level of sensitivity.[28] Despite this, predicting population-specific and less common alleles presents a challenge due to their underrepresentation in the reference panel. This underrepresentation leads to imputation errors, particularly for less frequent and rare alleles. The likelihood of encountering one of these haplotypes might be higher in admixed populations, especially in samples with a higher degree of admixture (Figure S3B). Therefore, training sets for admixed populations need samples from multiple ancestries to circumvent poor imputation accuracy attributed to underrepresentation.[46] Degenhardt et al.[47] demonstrated this by integrating multiple existing single-ancestry reference panels to construct a multiethnic reference panel covering ethnically heterogeneous populations. Given the ancestry-specific nature of the LD and haplotype structure in the *MHC* region, a multiethnic reference panel can maintain high accuracy across different ethnicities.[48]

In summary, our models achieved great accuracy when predicting *HLA* class I alleles across different populations. In general, the 1KG reference performed better than SABE alone. However, the SABE reference brings valuable information to the panel, improving the accuracy in predicting certain alleles. Consequently, the optimal performance was achieved with the full reference, a consensus of both SABE and 1KG datasets. In addition, it is important to emphasise that this is the first study to examine the imputation accuracy in non-classical alleles, which showed good results as expected for less polymorphic genes but also shows the same limitation as the classical alleles when predicting rare and population-specific alleles. In conclusion, our investigation underscores the paramount importance of enhancing reference panels or adapting existing ones to capture the genetic diversity inherent to the target population comprehensively.

ACKNOWLEDGEMENTS

The SHLARC project has received support from Nantes Métropole, the Pays de la Loire Region and the European Union (via the FEDER) under the Programme of Investments for the Future. This work was supported by ANR

PIA-Investment (NExT, SHLARC Project, Nantes Université), and the Fundação de Amparo à Pesquisa do Estado de São—FAPESP/Brazil (grants 2021/02815-8 and 2021/14851-9). The Brazilian HSL cohort was sequenced and genotyped with funding from The COVID-19 Research Campaign of the Hospital Sirio-Libanes donors. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) within the scope of the CAPES/COFECUB Program. This study was financed in part by the Comité Français d'Évaluation de la Coopération Universitaire et Scientifique avec le Brésil (COFECUB) within the scope of the CAPES/COFECUB Program (Me 1044/24).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The raw sequencing data used in this study are available in public repositories: 1000 Genomes at https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections, the Brazilian/SABE cohort at https://ega-archive.org/datasets/EGAD00001008640 and the Brazilian HSL cohort NGS and genotype array data are available under request to the authors. The reference panels generated in this study will be soon available on the SHLARC website.

ORCID

Nayane S. B. Silva https://orcid.org/0000-0001-5511-8426

Pierre-Antoine Gourraud https://orcid.org/0000-0003-1131-9554

Nicolas Vince https://orcid.org/0000-0002-3767-6210

REFERENCES

1. Klein J, Sato A. The HLA system—first of two parts. *N Engl J Med*. 2000;343:702-709.
2. Hughes AL. Natural selection and the evolutionary history of major histocompatibility complex loci. *Front Biosci*. 1998;3(4):A298-d516. doi:10.2741/A298
3. Robinson J, Barker DJ, Georgiou X, Cooper MA, Flicek P, Marsh SGE. IPD-IMGT/HLA database. *Nucleic Acids Res*. 2020;48(D1):D948-D955. doi:10.1093/nar/gkz950
4. Amigorena S. Antigen presentation: from cell biology to physiology. *Immunol Rev*. 2016;272(1):5-7. doi:10.1111/imr.12436
5. Wyatt RC, Lanzoni G, Russell MA, Gerling I, Richardson SJ. What the HLA-I!—classical and non-classical HLA class I and their potential roles in type 1 diabetes. *Curr Diab Rep*. 2019;19(12):159. doi:10.1007/s11892-019-1245-z
6. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. *Nat Rev Immunol*. 2018;18(5):325-339. doi:10.1038/nri.2017.143
7. Sticht J, Álvaro-Benito M, Konigorski S. Type 1 diabetes and the HLA region: genetic association besides classical HLA class II genes. *Front Genet*. 2021;12:683946. doi:10.3389/fgene.2021.683946
8. Okamoto T, Okajima H, Ogawa E, et al. The protective association of HLA-C*12:02 and HLA-DQB1*06:01 with severe acute hepatitis of unknown origin in the Japanese population. *HLA*. 2023;103:e15215. doi:10.1111/tan.15215
9. Bowness P. HLA-B27. *Annu Rev Immunol*. 2015;33(1):29-48. doi:10.1146/annurev-immunol-032414-112110
10. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Res*. 2017;45(D1):D896-D901. doi:10.1093/nar/gkw1133
11. Visscher PM, Wray NR, Zhang Q, et al. 10 years of GWAS discovery: biology, function, and translation. *Am J Hum Genet*. 2017;101(1):5-22. doi:10.1016/j.ajhg.2017.06.005
12. Vince N, Douillard V, Geffard E, et al. SNP-HLA reference consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. *Genet Epidemiol*. 2020;44(7):733-740. doi:10.1002/gepi.22334
13. Douillard V, Castelli EC, Mack SJ, et al. Approaching genetics through the MHC lens: tools and methods for HLA research. *Front Genet*. 2021;12:774916. doi:10.3389/fgene.2021.774916
14. Meyer D, Nunes K. HLA imputation, what is it good for? *Hum Immunol*. 2017;78(3):239-241. doi:10.1016/j.humimm.2017.02.007
15. Douillard V, dos Santos Brito Silva N, Bourguiba-Hachemi S, et al. Optimal population-specific HLA imputation with dimension reduction. *HLA*. 2024;103:e15282. doi:10.1111/tan.15282
16. Hirata J, Hosomichi K, Sakaue S, et al. Genetic and phenotypic landscape of the major histocompatibilty complex region in the Japanese population. *Nat Genet*. 2019;51(3):470-480. doi:10.1038/s41588-018-0336-0
17. Squire DM, Motyer A, Ahn R, Nititham J, Huang Z-M, Oksenberg JR. MHC*IMP—imputation of alleles for genes in the major histocompatibility complex. bioRxiv. Preprint posted online January 26, 2020. 2020. doi:10.1101/2020.01.24.919191
18. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
19. Byrska-Bishop M, Evani US, Zhao X, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*. 2022;185(18):3426-3440.e19. doi:10.1016/j.cell.2022.08.004
20. Naslavsky MS, Yamamoto GL, de Almeida TF, et al. Exomic variants of an elderly cohort of Brazilians in the ABraOM database. *Hum Mutat*. 2017;38(7):751-763. doi:10.1002/humu.23220
21. Li H, Handsaker B, Wysoker A, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
22. Castelli EC, Paz MA, Souza AS, Ramalho J, Mendes-Junior CT. Hla-mapper: an application to optimize the mapping of HLA sequences produced by massively parallel sequencing procedures. *Hum Immunol*. 2018;79(9):678-684. doi:10.1016/j.humimm.2018.06.010
23. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110
24. Patterson MD, Marschall T, Pisanti N, et al. WhatsHap: weighted haplotype assembly for future-generation sequencing

reads. *J Comput Biol*. 2015;22(6):498-509. doi:10.1089/cmb.2014.0157
25. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods*. 2012;9(2):179-181. doi:10.1038/nmeth.1785
26. Rice P, Longden L, Bleasby A. EMBOSS: the European molecular biology open software suite. *Trends Genet*. 2000;16(6):276-277. doi:10.1016/S0168-9525(00)02024-2
27. Gourraud P-A, Khankhanian P, Cereb N, et al. HLA diversity in the 1000 genomes dataset. *PLoS One*. 2014;9(7):e97282. doi:10.1371/journal.pone.0097282
28. Zheng X, Shen J, Cox C, et al. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J*. 2014;14(2):192-200. doi:10.1038/tpj.2013.18
29. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575. doi:10.1086/519795
30. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109
31. Cann HM, Toma CD, Cazes L, et al. A human genome diversity cell line panel. *Science*. 2016;296:261-262.
32. Luo Y, Kanai M, Choi W, et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nat Genet*. 2021;53(10):1504-1516. doi:10.1038/s41588-021-00935-7
33. Hurley CK, Kempenich J, Wadsworth K, et al. Common, intermediate and well-documented HLA alleles in world populations: CIWD version 3.0.0. *HLA*. 2020;95(6):516-531. doi:10.1111/tan.13811
34. Silva NSB, Bourguiba-Hachemi S, Douillard V, et al. 18th international HLA and immunogenetics workshop: report on the SNP-HLA reference consortium (SHLARC) component. *HLA*. 2023;103:e15293. doi:10.1111/tan.15293
35. Robinson J, Guethlein LA, Cereb N, et al. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet*. 2017;13(6):e1006862. doi:10.1371/journal.pgen.1006862
36. Klitz W, Hedrick P, Louis EJ. New reservoirs of HLA alleles: pools of rare variants enhance immune defense. *Trends Genet*. 2012;28(10):480-486. doi:10.1016/j.tig.2012.06.007
37. Baxter-Lowe LA. The changing landscape of HLA typing: understanding how and when HLA typing data can be used with confidence from bench to bedside. *Hum Immunol*. 2021;82(7):466-477. doi:10.1016/j.humimm.2021.04.011
38. dos Santos Brito SN, Souza AS, Andrade HS, et al. Immunogenetics of HLA-B: SNP, allele, and haplotype diversity in populations from different continents and ancestry backgrounds. *HLA*. 2023;101:634-646. doi:10.1111/tan.15043
39. Souza AS, Sonon P, Paz MA, et al. HLA-C genetic diversity and evolutionary insights in two samples from Brazil and Benin. *HLA*. 2020;96(4):468-486. doi:10.1111/tan.13996
40. Gragert L, Madbouly A, Freeman J, Maiers M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol*. 2013;74(10):1313-1320. doi:10.1016/j.humimm.2013.06.025
41. Gonzalez-Galarza FF, Christmas S, Middleton D, Jones AR. Allele frequency net: a database and online repository for immune gene frequencies in worldwide populations. *Nucleic Acids Res*. 2011;39(Suppl. 1):D913-D919. doi:10.1093/nar/gkq1128
42. Boquett JA, Bisso-Machado R, Zagonel-Oliveira M, Schüler-Faccini L, Fagundes NJR. HLA diversity in Brazil. *HLA*. 2020;95(1):3-14. doi:10.1111/tan.13723
43. Salzano FM, Sans M. Interethnic admixture and the evolution of Latin American populations. *Genet Mol Biol*. 2014;37(1 suppl 1):151-170. doi:10.1590/S1415-47572014000200003
44. Carvalho MG, Tsuneto LT, Moita Neto JM, et al. HLA-A, HLA-B and HLA-DRB1 haplotype frequencies in Piauí's volunteer bone marrow donors enrolled at the Brazilian registry. *Hum Immunol*. 2013;74(12):1598-1602. doi:10.1016/j.humimm.2013.08.283
45. Bortolotto AS, Petry MG, da Silveira JG, et al. HLA-A, -B, and -DRB1 allelic and haplotypic diversity in a sample of bone marrow volunteer donors from Rio Grande do Sul state, Brazil. *Hum Immunol*. 2012;73(2):180-185. doi:10.1016/j.humimm.2011.11.009
46. Nunes K, Zheng X, Torres M, et al. HLA imputation in an admixed population: an assessment of the 1000 genomes data as a training set. *Hum Immunol*. 2016;77(3):307-312. doi:10.1016/j.humimm.2015.11.004
47. Degenhardt F, Wendorff M, Wittig M, et al. Construction and benchmarking of a multi-ethnic reference panel for the imputation of HLA class I and II alleles. *Hum Mol Genet*. 2019;28(12):2078-2092. doi:10.1093/hmg/ddy443
48. Naito T, Okada Y. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. *Semin Immunopathol*. 2022;44(1):15-28. doi:10.1007/s00281-021-00901-9

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Silva NSB, Bourguiba-Hachemi S, Ciriaco VAO, et al. A multi-ethnic reference panel to impute *HLA* classical and non-classical class I alleles in admixed samples: Testing imputation accuracy in an admixed sample from Brazil. *HLA*. 2024;103(6):e15543. doi:10.1111/tan.15543