UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Helda

18th International HLA and Immunogenetics Workshop : Report on the SNP-HLA Reference Consortium (SHLARC) component

Silva, Nayane S. B.

Wiley Blackwell
2024-01

Silva, N S B, Bourguiba-Hachemi, S, Douillard, V, Koskela, S, Degenhardt, F, Clancy, J, Limou, S, Meyer, D, Masotti, C, Knorst, S, Naslavsky, M S, Franke, A, Castelli, E C, Gourraud, P-A & Vince, N 2024, '18th International HLA and Immunogenetics Workshop : Report on the SNP-HLA Reference Consortium (SHLARC) component', HLA, vol. 103, no. 1, e15293. https://doi.org/10.1111/tan.15293

http://hdl.handle.net/10138/592802
10.1111/tan.15293

cc_by
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Check for updates

Received: 14 June 2023 | Revised: 24 October 2023 | Accepted: 1 November 2023

DOI: 10.1111/tan.15293

MEETING REPORT

HLA Immune Response Genetics WILEY

18th International HLA and Immunogenetics Workshop: Report on the SNP-HLA Reference Consortium (SHLARC) component

Nayane S. B. Silva[1,2] | **Sonia Bourguiba-Hachemi**[1] | **Venceslas Douillard**[1] | **Satu Koskela**[3] | **Frauke Degenhardt**[4] | **Jonna Clancy**[3] | **Sophie Limou**[1] | **Diogo Meyer**[5] | **Cibele Masotti**[6] | **Stefan Knorst**[6] | **Michel Satya Naslavsky**[5] | **Andre Franke**[4] | **Erick C. Castelli**[2] | **Pierre-Antoine Gourraud**[1] | **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 – Unesp, Botucatu, Brazil

[3]Finnish Red Cross Blood Service Biobank, Helsinki, Finland

[4]Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, University Hospital Schleswig Holstein - Campus Kiel, Kiel, Germany

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

[6]Molecular Oncology Center, Hospital Sírio-Libanês, São Paulo, Brazil

Correspondence
Nicolas Vince, Nantes Université, CR2TI UMR1064 – ITUN, CHU Nantes Hôtel Dieu, 30 bld Jean Monnet, 44093 Nantes Cedex 01, France.
Email: nicolas.vince@univ-nantes.fr

Funding information
European Union (via the FEDER) under the Programme of Investments for the Future; European Union's Horizon 2020 research and innovation program; INSERM; Nantes Métropole, the Pays de la Loire Region

The SNP-HLA Reference Consortium (SHLARC), a component of the 18th International HLA and Immunogenetics Workshop, is aimed at collecting diverse and extensive human leukocyte antigen (HLA) data to create custom reference panels and enhance HLA imputation techniques. Genome-wide association studies (GWAS) have significantly contributed to identifying genetic associations with various diseases. The HLA genomic region has emerged as the top locus in GWAS, particularly in immune-related disorders. However, the limited information provided by single nucleotide polymorphisms (SNPs), the hallmark of GWAS, poses challenges, especially in the HLA region, where strong linkage disequilibrium (LD) spans several megabases. HLA imputation techniques have been developed using statistical inference in response to these challenges. These techniques enable the prediction of HLA alleles from genotyped GWAS SNPs. Here we present the SHLARC activities, a collaborative effort to create extensive, and multi-ethnic reference panels to enhance HLA imputation accuracy.

KEYWORDS
18th International HLA and Immunogenetics Workshop, HLA, imputation, reference panel, SHLARC, SNP

Nayane S. B. Silva and Sonia Bourguiba-Hachemi these authors contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. HLA: Immune Response Genetics published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Over the past decade, genome-wide association studies (GWASs) have identified more than 10,000 SNP-disease associations.[1] The HLA genomic region within the human MHC has consistently demonstrated the highest number of SNP associations, mainly with immune-related diseases.[2] However, SNPs as genetic markers provide limited information, especially for HLA, as this region is the most polymorphic in the human genome, and linkage disequilibrium (LD) is strong and covers large areas. Therefore, a SNP associated with a particular pathology is only a marker of a genomic region. To better understand the functional mechanisms at play in HLA-related diseases and develop therapeutic targets, it is necessary to look beyond this simple association and, rather than analyze SNPs, move to the analysis of HLA alleles.[3] Although crucial for HLA studies, HLA typing techniques based on sequencing are expensive, require specialized laboratory infrastructure, and are subject to constant evolution. HLA imputation represents a fast and cost-effective way to explore genetic associations in diseases already studied by GWASs. HLA imputation was developed to predict HLA genotypes from genotyped GWAS SNPs through statistical inference. SNP to HLA imputation relies on reference panels of individuals with known SNPs and HLA genotypes to generate links between SNPs, haplotypes, and HLA alleles using machine learning algorithms.[4] We led the creation of an international consortium, the SNP-HLA Reference Consortium (SHLARC), aiming to improve HLA imputation.[3]

The primary goal of the SHLARC project is to create reference panels to improve HLA imputation from GWAS datasets.[4] Our goals include the following:

1. Collect a large number of samples with both HLA and SNP genotyping data, as diverse as possible, to improve our imputation models.
2. Apply mathematical and computer sciences to improve existing HLA imputation techniques and potentially develop new ones.
3. Provide accessibility and service to the scientific community: supercomputer power is made available through a freely accessible website (hla.univ-nantes.fr).

By combining international expertise as well as data and computational resources, the SHLARC will bring HLA imputation data to a level of interpretation that is key to solving questions on immune-related pathologies.[3] Concretely, SHLARC aims to investigate the performance of reference panels, their applicability in diverse populations, and the advancement of imputation methodologies.

The SHLARC component activities started in 2019. The successful implementation and planning of this collaborative endeavor necessitated the application of diverse scientific, administrative, and technological approaches. Since its initiation, the SHLARC component has achieved significant milestones and made a notable impact on the field of HLA imputation. Here, we present the 18th International HLA and Immunogenetics Workshop report for the SHLARC component.

2 | METHODS

2.1 | Dataset construction

The SNP and HLA genotyping data used in SHLARC came from different sources. We have joint data from whole-genome sequencing (WGS), genome-wide SNP genotyping arrays, and direct HLA typing, including 1KG (1000 Genomes Project, $n = 3197$),[5] HGDP (Human Genome Diversity Project, $n = 828$),[6] CAAPA (Consortium on Asthma among African-ancestry Populations in the Americas, $n = 880$, dbGAP phs001123.v2.p1),[7] and SABE (Saúde, Bem-estar e Envelhecimento, $n = 1322$).[8] Individuals from all cohorts used in this work gave informed consent to their primary promoter for research use of their data. The SNP genotypes within the MHC region on chromosome 6, ranging from 29 to 34 Mb, from multiple sources to construct a multiethnic panel, are stored in PLINK file format. We have applied standard quality control measures to ensure that only high-quality data was used in the subsequent analyses. Palindromic SNPs (A/T or G/C) were removed to prevent potential ambiguities due to misclassification. Additionally, SNPs with a minor allele frequency (MAF) below 1% were excluded, as well as those with a genotype missing call rate exceeding 2%. SNP positions are based on the GRCh38/hg38 genome assembly. We did not apply any Hardy–Weinberg Equilibrium filter. For WGS data (1KG, HGDP, and SABE), HLA genotypes were called using the hla-mapper[9] workflow (https://github.com/erickcastelli/HLA_genotyping/), our own HLA calling tool, which was validated previously.[10,11] The hla-mapper was chosen for its ability to provide comprehensive SNPs and HLA alleles data from MHC genomic region, a significant advantage over other methods (for example PolyPheMe and Optitype) that typically provide only HLA types. hla-mapper includes the detection of null alleles, which were considered for the imputation process. On the other hand, the detection of "new" alleles, not yet been documented by IMGT, was not taken into consideration in our analysis. While we cannot quantify the amount of noise attributed to the WGS-to-HLA method versus the

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SNP-to-HLA method, it is worth noting that SNP-to-HLA relies on WGS-to-HLA results. Thus, we begin with the assumption that the WGS-to-HLA method is accurate, and the SNP-to-HLA method is considered correct when it yields the same results as WGS-to-HLA.

To evaluate the genetic diversity of our data, we conducted a principal component analysis (PCA) using PLINK. For the PCA analysis, we used the SNPs of the combined datasets from chromosome 6, excluding the MHC region. PLINK format is required for HIBAG reference panel construction and HIBAG-based imputation.

2.2 | Generation of HLA imputation reference panels using the HIBAG package

For each HLA gene, namely HLA-A, HLA-B, HLA-C (class I), and HLA-DQB1, HLA-DRB1, HLA-DPA1, HLA-DPB1 (class II), we created different reference panels for HLA imputation using our different available dataset. The reference panels were constructed based on SNP and HLA genotype information. We did not have HLA-DRB3/HLA-DRB4/HLA-DRB5 types in the reference panels as our genotyping methodology is still under test for these genes due to the complexity of the HLA-DRB locus: haplotypes structure and copy number variations. The underlying principle of HLA imputation involves inferring HLA alleles from SNP genotypes using these reference panel models.

Among various HLA imputation software options, we selected the HIBAG R package due to its superior performance compared to other methods[12] and its simplicity of use. Indeed, HIBAG offers the flexibility to create custom reference panels easily. The statistical models were computed using R 3.5.1, utilizing the HIBAG package (v1.4)[13] and its extension HIBAG.gpu (v0.9.1). All calculations were performed on GPU nodes available in our local Nantes Université servers. During the model-building process, SNPs within a 500 kb window of the target gene were retained. The proportion of SNPs integrated into each classifier followed the default value of HIBAG. Indeed, the HIBAG algorithm randomly picks a subset of SNPs at each round of bagging, before selecting from them the most capable of accurately predicting HLA alleles.[13]

2.3 | Validation of HLA imputation reference panels using the HIBAG package

The effects of models composition on HLA imputation were validated by applying this methodology to independent datasets of admixed individuals (CAAPA and SABE).[14]

2.4 | Statistical analyses

To evaluate the performance of the reference panels, we used several scores. The HIBAG output provided accuracy statistics. No specific call threshold was set, and we estimated accuracy as the number of correctly predicted HLA alleles out of all predicted alleles. The results were aggregated and visualized using R 3.5.1 with the ggplot2 package. In addition, we used the F1-score. This metric is a harmonic mean of precision and sensitivity, which provides a balanced assessment of both measures. The F1-score is especially important when assessing the accuracy of rare allele predictions because it emphasizes the coverage of a specific allele prediction.[15] However, F1-score can be influenced by the presence of rare alleles, also depending on other factors such as locus diversity (highly polymorphic locus means high number of rare alleles) and the number of different alleles present for a given locus. For example: if we consider a locus with 10 different alleles, 4 rare alleles are present only once in the data, and have an F1-score of 0, while the other 6 common and frequent alleles have an F1-score of 1, the average F1-score for that locus will be 0.6. This is influenced by the contribution of low F1-score from rare alleles.

3 | RESULTS

The SHLARC component session during the 18th International HLA and Immunogenetics Workshop was held on May 12, 2022. SHLARC partners were invited, as well as anyone interested in HLA imputation. We had four presentations about data and last results concerning HLA imputation: Venceslas Douillard, with "Improving HLA imputation in an admixed population with dimension reduction"; Nayane dos Santos Brito Silva, with "HLA-DPA1 and HLA-DPB1 diversity and imputation in a Brazilian population sample"; Satu Koskela, with "HLA variation in Finland—population specific HLA imputation tool and its applications"; and Frauke Degenhardt, with "Trans-ancestry, haplotype-aware HLA analysis in ulcerative colitis".

3.1 | Data collection

At the time of the 18th International HLA and Immunogenetics Workshop in May 2022, we had four datasets

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

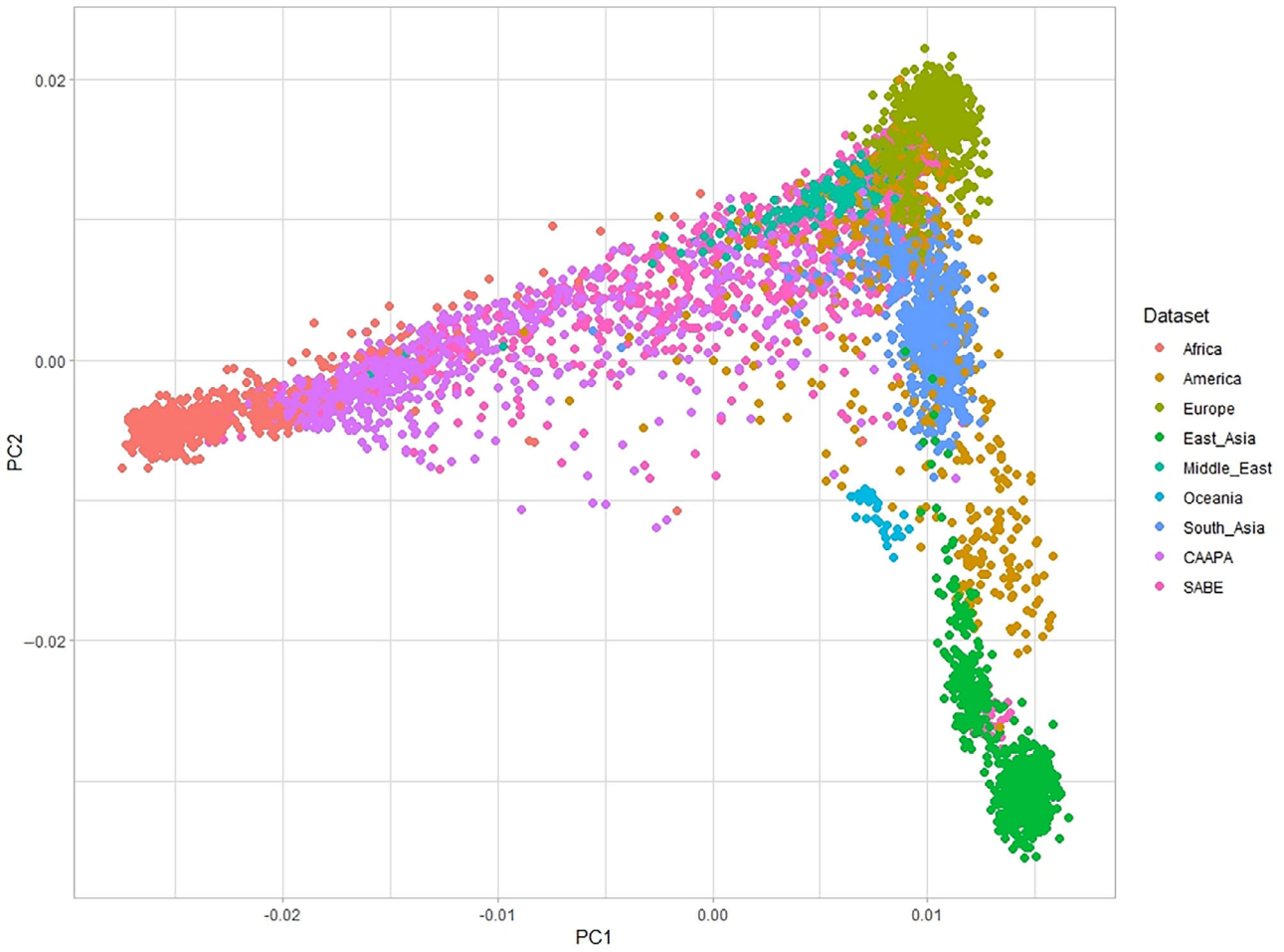

FIGURE 1 Analysis of genetic diversity in different datasets used in SHLARC. PCA representation of merged genotypes of chromosome 6 (excluding the MHC region) from the 1KG, HGDP, CAAPA, and SABE datasets. The 1KG and HGDP datasets were used as references and classified into 7 super-populations: African, American, European, East Asian, Middle Eastern, Oceanian, and South Asian.

already available (Figure 1, Table S1): 1KG ($n = 3197$),[5] HGDP ($n = 828$),[6] CAAPA ($n = 880$, dbGAP phs001123.v2.p1),[7] and SABE ($n = 1322$).[8] Currently, the SHLARC network includes 20 different laboratories in 16 countries. Our effort during the workshop led to multiple contacts to bring additional large datasets to the SHLARC.

3.2 | HLA imputation performance in an admixed population (Venceslas Douillard)

We obtained genotyping data of 1KG (diverse ancestry) and CAAPA (African American) from the whole genome sequencing,[5,7,16] storing two-field HLA alleles for HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 in a CSV file. HLA imputation models were computed using HIBAG.[17] First, we explored the HLA diversity of 1KG by creating imputation models leaving one population out at a time. For example, creating an imputation model with the whole 1KG dataset but without the Yoruba samples, and testing imputation accuracy on the Yoruba samples. Overall, we observed high accuracies in most tests (Figure 2) but low accuracy in specific situations, such as HLA-B and HLA-DRB1 among Peruvians due to the population-specific alleles HLA-B*15:04 and HLA-DRB1*09:06, or HLA-B among Mexicans due to HLA-B*15:15 (Figure 2).

Further, we investigated the possibility of creating specific models, that is, imputation models with reference samples that closely match the ancestry of the samples to impute. Models were built using the 1KG dataset as training and CAAPA as a test dataset. These models consisted of the full dataset (all 1KG samples) and smaller ones with subsets of 200 individuals, selected from the same super-population (African, American, European, East Asian, and South Asian), or with individuals genetically close to the test sample selected by using dimension reduction (PCA or UMAP) changing SNP and dimension parameters (custom models). The full dataset reference panel always outperformed other models, with an

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

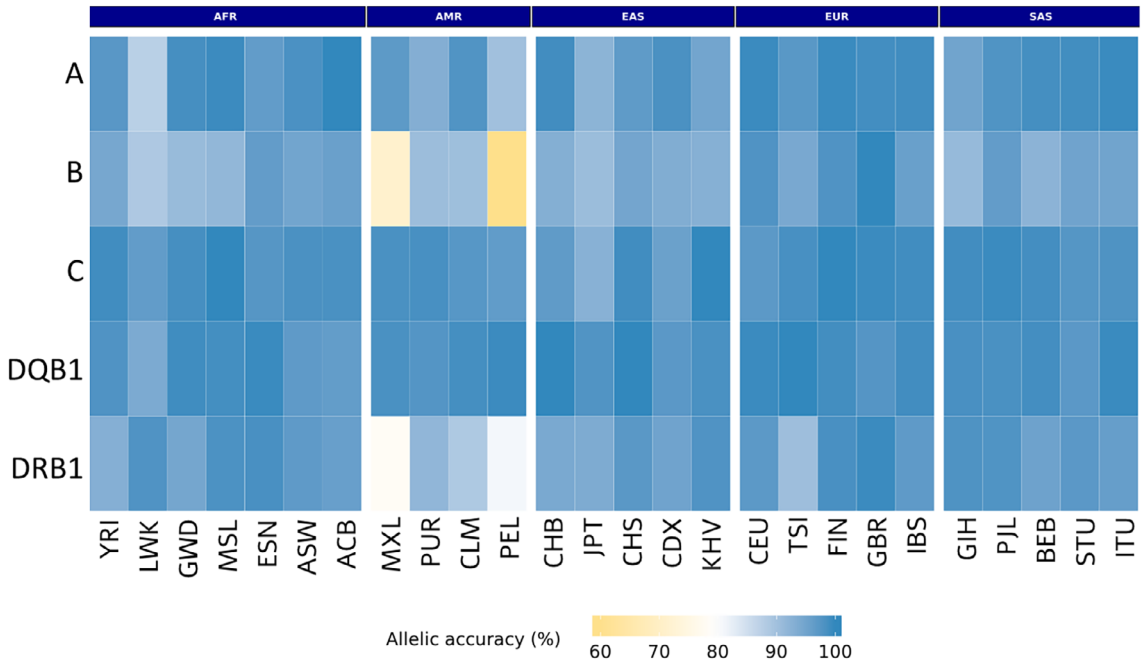

FIGURE 2 Accuracy depends of the population origin, which reflects HLA allele diversity. For each population, we used all the others to impute the HLA alleles. HLA-A, HLA-B, and HLA-DRB1 accuracies are >90% for most population samples, but some carry specific alleles (e.g., PEL: HLA-B*15:04, MXL: HLA-B*51:26), absent from the reference panels, hence impossible to impute. Overall accuracy is dependent of the gene allelic diversity and specificity of the population. AFR, Africans. AMR, Americans. EAS, East Asians. EUR, Europeans. SAS, South Asians. Please refer to the Table S2 for a full list of population samples.

F1-score of 0.66 for HLA-B. However, custom models consistently outperformed multiethnic or superpop models of similar size, with F1-scores up to 0.53, against F1-scores up to 0.42. These last data are now published as a preprint.[14]

3.3 | HLA-DPA1 and HLA-DPB1 diversity and imputation in a Brazilian population sample (Nayane dos Santos Brito Silva)

The reference panels for HLA-DPA1 and HLA-DPB1 genes were generated using HIBAG.[13] The process involved analyzing high-coverage (30X) whole-genome sequencing data from 5347 samples collected from worldwide populations. These samples comprised 3197 samples from 1KG,[5,17] 828 samples from HGDP,[6] and 1322 samples from SABE/Brazil.[8] The HLA genotyping methodology applied to evaluate the samples used in the reference panel is available at https://github.com/erickcastelli/HLA_genotyping/.[9]

Five HLA imputation reference panels were computed for both genes. These panels included the following: 1KG; HGDP; SABE; 1KG + HGDP; 1KG + HGDP +SABE (full model). The reference panels were validated by predicting HLA-DPA1 and HLA-DPB1 alleles in 192 samples from São Paulo with available HLA allele calls and SNP-array data. The best imputation model varied depending on the locus. For HLA-DPA1, the full model performed better (F1-score of 0.78) than others. For HLA-DPB1, the SABE (Brazilian) model alone provided better results, with an F1-score of 0.83, than the full model (Brazilian data included) with an F1-score of 0.79.

3.4 | HLA variation in Finland (Satu Koskela)

Finnish HLA landscape has been shaped by population history and the geopolitical location at the crossroads of Eastern and Western civilization. The small founder population underwent several bottlenecks and received limited gene flow from Western and Eastern European populations.[18,19] As a result, today's population in Finland is a genetic patchwork and a genetic outlier in Europe.[20,21] The Finnish MHC is characterized by a reduced allele pool and enrichment of alleles and haplotypes regarding HLA and MIC genes.[22–24] Several HLA haplotypes, common in Finland but rare or missing in other European populations, have been identified and named as Finnish enriched rare (FER) haplotypes.[23]

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

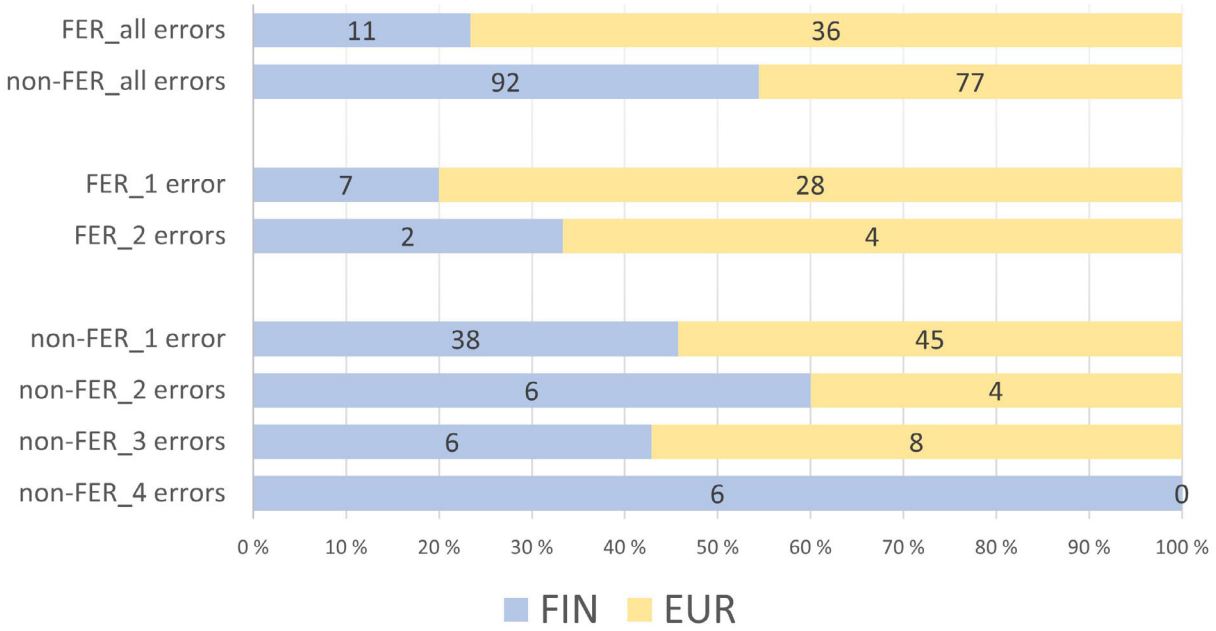

FIGURE 3 Numbers and relative proportions of imputation errors in the FER and non-FER haplotypes. Imputation errors observed in the Finnish samples using the European reference are concentrated among the FER haplotypes while errors with the Finnish reference are concentrated among the non-FER haplotypes. Errors in the non-FER haplotypes are explained by a small sample size of the Finnish reference dataset and by uncommon haplotypes not included in the reference. FER, Finnish enriched rare HLA haplotype; non-FER, HLA haplotype not enriched in the Finnish population; FIN, the Finnish reference dataset ($n = 1150$); EUR, the European reference dataset ($n = 2668$).

Due to the distinctive HLA repertoire, we collected a Finnish population-specific reference dataset for imputing samples of Finnish origin. Seven classical HLA genes (HLA-A, HLA-B, HLA-C, HLA-DRB1, HLA-DQA1, and HLA-DPB1) were imputed at two field resolution level with the HIBAG algorithm.[25] Comparison of Finnish and European reference panels regarding imputation accuracy underlines the importance of matching the reference population with the target population; errors with the European reference are specifically concentrated among the FER haplotypes due to the inability of the European panel to fully capture the target population's LD structure despite the informative SNP markers (Figure 3).

The aim of establishing a population-based reference panel was to design a high-precision HLA imputation tool to generate in silico HLA typing for large sample collections containing SNP-based genotyping data, such as Finnish biobanks. Also, our imputation model has been made available to the national research initiative, the FinnGen project, a data collection of patient cohorts and healthy controls of 500,000 data subjects with thousands of phenotypes ($n = 4600$, data freeze 11), https://www.finngen.fi/en. Extensive national biobanking and research cooperation in FinnGen enable predicted high-accuracy HLA type for about 10% of Finland's population of 5.6 million. The FinnGen dataset has already been used to extract new HLA disease associations, both allelic and pleiotropic.[26]

3.5 | Trans-ancestry, haplotype-aware HLA analysis in ulcerative colitis (Frauke Degenhardt)

Besides the analyses mentioned above, additional topics in this Workshop included discussing a previously published trans-ethnic analysis of the HLA in the immune-related disease Ulcerative Colitis.[27] The study by the IBDGenetics consortium and international partners used a custom-built multi-ethnic and extensively validated HIBAG HLA reference panel[28] specifically designed for this HLA fine-mapping study to cover the diversity of ancestral backgrounds investigated.

Additional focus in the presentation was set on some of the general drawbacks of HLA genotyping using SNP genotypes derived from different GWAS genotyping arrays that may result in some variability across studies for specific HLA alleles. Besides the evident problem of typing rare HLA alleles, some alleles, with a focus here on HLA-DRB1, such as HLA-DRB1*11:01/11:04 or HLA-DRB1*04:01/04:03/04:04 are difficult to distinguish using the current methodologies, because of similarities

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

in their SNP haplotypes.[28] Leveraging the advantages of their multi-ethnic reference panel and the large multi-ethnic GWAS dataset, Degenhardt et al. identified previously unreported HLA associations, like at the rare HLA-DRB1*10:01 allele, confirming the genetic associations observed in predominantly investigated Caucasian Ulcerative colitis of HLA-DQB1*06 ~ HLA-DRB1*15 haplotypes across different ancestries. Beyond that, they reported ancestry-specific associations that align with HLA allele frequencies observed in the respective ancestries. This study also showed the importance of haplotype-based analysis in the context of HLA fine-mapping studies, rather than the investigation of associations at single HLA alleles and presented a method for HLA haplotyping, based on SNP haplotypes generated within the HIBAG models and external SNP haplotyping tools such as SHAPEIT2.[27,29]

4 | DISCUSSION

GWAS studies have primarily focused on the cohorts of patients and subjects of European descent, resulting in limited exploration of the genetic makeup of complex traits in non-European populations.[30,31] Genotype imputation is a crucial component of GWAS. It relies on a large reference panel of sequenced individuals with similar ancestry to the cohort being studied.[32] However, determining the best strategy for imputation and subsequent association analysis is challenging due to various trade-offs, such as panel size, imputable variant types, and population specificity across different reference panels. Multiple methods for HLA imputation have been developed, each possessing distinct benefits, and more recent approaches have made advancements in both accuracy and computational efficiency.[4] The results presented here underline various aspects of HLA imputation and analysis in different populations and disease contexts.

To comprehensively evaluate the performance of our reference panels, we utilized the mean F1 score as a metric, a harmonic mean of sensitivity and precision. A higher F1 score signifies a well-balanced trade-off between precision and recall, while a lower score suggests potential inaccuracies in predicting specific alleles. It is worth noting, however, that the mean F1 score assumes equal importance for all alleles, which might not always hold true. Thus, in this context, the F1 score offers an overall assessment of how well the model was in identifying both common and rare alleles.

Our findings show that HLA allele imputation accuracy in admixed populations demands a large multiethnic reference panel or, at least, a matched reference panel. When using models with similar sizes, the models with ancestral backgrounds close to the target population usually outperform multiethnic models for African-American (CAAPA) imputation. In addition, smaller population-specific models can also outperform larger multiethnic models, such as what we observed for HLA-DPB1 and Brazilians. The evolutionary proximity between the reference panel and the samples being imputed is an important factor.[33] Choosing an appropriate reference panel for HLA imputation depends on how well it matches the set of samples to be imputed.[4]

The findings from our Finnish collaborator, comparing the Finnish and European reference panels, highlighted the significance of matching the reference population with the target population. This suggests that even relatively close populations, such as Central Europeans and the Finns, exhibit notable differences in LD structure and allele frequencies that impact the imputation performance. The presence of rare or specific alleles underrepresented in the reference panels can be overcome by collecting a large and comprehensive set of samples with both HLA and SNP genotyping data from different populations for the reference panel, as in the SHLARC. This is expected to improve the imputation accuracy of any given sample from a similar ethnic origin by covering diverse HLA alleles, haplotypes, and MHC structures.

Another important outcome is that the accuracy of the HLA imputation methods depends on the locus being imputed. Using the Brazilian cohort, the best imputation model varies depending on the locus, for HLA-DPA1 the multiethnic model performed better than others. However, the specific population model alone provided better results for HLA-DPB1 than the multiethnic model. Moreover, HLA-B imputation was jeopardized for some populations when these populations were not included in the reference panel (Figure 1), because of the presence of population-specific alleles. The limitation of HLA imputation is most likely due to the underrepresentation of specific HLA alleles in the reference panels. The polymorphism of HLA genes varies across populations, resulting in the presence or absence of specific allelic variants in different populations, and variations in their frequencies.[34] This again highlights the importance of using reference panels that match the ancestry of the target dataset when imputing specific HLA alleles. To improve accuracy, especially in admixed populations like Brazilians, it is crucial to include underrepresented populations in the reference panels.

To avoid poor imputation accuracy caused by underrepresentation,[35] admixed populations require training sets that consist of samples from multiple ancestries. Degenhardt et al. demonstrated this by integrating

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

multiple existing single-ancestry reference panels to create a multiethnic reference panel that encompasses ethnically diverse populations.[28] Since the LD and haplotype structure of the MHC region are highly specific to the genetic background, a multiethnic reference panel can maintain a high level of accuracy across different ethnicities.[36] Therefore, the success of HLA imputation depends on the quality and availability of large and diverse reference panels.

The evolution of imputation tools will also consequently improve HLA imputation, for instance, HLA-IMP*03[37] and CookHLA[15] showed improved results compared to their predecessor algorithms. Additionally, DeepHLA[38] also demonstrated high accuracy, with a specific focus on rare HLA alleles. However, these algorithms are yet to be tested by an independent team and formally compared to other HLA imputation tools. In our opinion, it is important to acknowledge that these efforts will eventually reach a limit. To overcome these limitations and further enhance HLA imputation, we advocate for a global research focus on gathering diverse data from around the world, ensuring the inclusion of a broad range of populations to enrich the reference panels and overcome the limitations posed by underrepresented HLA alleles. Such global efforts will ultimately contribute to the continuous advancement of HLA imputation and its applications in various fields of immunogenetics research.

The HIBAG algorithm has exhibited promising outcomes and has demonstrated its superiority based on previous research.[12] One of its notable strengths lies in its user-friendliness and its ability to create reference panels that distinguish it from other existing programs. Nonetheless, a shared limitation in HIBAG and other imputation tools is the accurate identification of less frequent and rare alleles. Successfully addressing this challenge is predominantly contingent on the composition of the reference panel utilized.

Our findings emphasize the paramount importance of expanding reference panels to improve the accuracy of HLA imputation. While customized reference panels may enhance HLA imputation accuracy for specific genes and population groups, they should not substitute the preference for extensive multi-ethnic panels. Confidence in the quality of HLA imputation is notably high when the study population shares genome intervals with the reference sequences. Conversely, confidence in HLA imputation accuracy decreases when the population lacks these shared intervals, particularly for genetically distant individuals, which can lead to elevated error rates. A similar dynamic applies to SNP-SNP imputation, although the available diversity in the reference population for SNP-SNP imputation significantly surpasses SNP-HLA imputation. For instance, resources such as the TopMed imputation server (imputation.biodatacatalyst.nhlbi.nih.gov) offer an amplitude of options. When performing GWAS analyses, it is important to consider the genetic diversity present in the study population. This diversity can arise from differences in ancestry or geographic origin, known as population substructure. To handle this, a common practice is to use "principal component analysis" (PCA). These components allow researchers to account for the genetic differences between individuals that might be due to their different backgrounds. Essentially, creating principal components is a way to control for the potential influence of population substructure on the GWAS results. In cases where the population under study lacks a reference panel of genetically closely related individuals for HLA imputation, the option of creating a custom reference panel may be considered to enhance imputation quality.

Again, our investigation underscores the preference for large, multi-ethnic reference panels that encompass a wide spectrum of HLA alleles and haplotypes. It is equally crucial that these panels include samples from the same genetic background as the target data, particularly when dealing with admixed populations that may exhibit unique allele and haplotype structures. Furthermore, our findings consistently demonstrate improved imputation performance when utilizing population-specific or large multiethnic models. This performance enhancement can be attributed to the increased representativity of these alleles and haplotype structures within the reference panels. Consequently, building up larger and more diverse reference panels emerges as the optimal strategy, and we encourage researchers worldwide to participate in the SHLARC initiative, contributing valuable data to advance the field.

Overall, studies presented in the SHLARC component highlight the importance of enhancing the inclusion of non-European ancestry samples in genomic datasets to mitigate the potential of exacerbating health disparities and to identify disease-associated variants that are specific to non-European populations. The SHLARC advocates for the coordination of these endeavors to establish comprehensive and diverse panels of substantial size, facilitating HLA imputation for researchers through an easy-to-use website to explore HLA imputation.

4.1 | SHLARC perspectives

As we conclude this article, it is important to acknowledge that our ongoing efforts extend beyond the scope of the discussed objectives. We are actively engaged in the pursuit of advancing HLA imputation techniques

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

through the application of mathematical and computer sciences. Furthermore, we are continuously seeking to enhance our reference panels by expanding our sample collection. To develop this ambitious project, we invite individuals with accessible two-field HLA alleles + SNPs datasets to actively participate in the SNP-HLA reference consortium (https://www.ihiw18.org/component-bioinformatics/snp-hla-reference/) to contribute empowering the immunogenetic community to move into the era of immunogenomic association.

In addition, our commitment to facilitating accessibility and service to the scientific community remains steadfast. To this end, we will soon be introducing a dedicated server specifically designed for imputation purposes.

The webserver is set to go live shortly, providing an easily accessible gateway to our reference data and the imputation process. Our comprehensive imputation pipeline will efficiently process user's uploaded SNP data, leveraging a reference panel with over 10,000 samples representing a diverse array of ancestries. The imputation process will predict the corresponding HLA alleles based on the information extracted from the user's SNP data and our reference panel. The SHLARC webserver will not propose SNP-to-SNP imputation which will not compromise the integrity of the imputation results. The key factor here is the overlap of SNPs between the reference panel and the data to impute. It is worth noting that even with a relatively modest number of SNPs within the HLA region, a high-quality HLA imputation can be achieved if a perfect matching between the reference panel and the imputed data exists. It is relevant to note that our supercomputers play a crucial role in intensive model training, demanding significant computational power to achieve accurate results. This initiative aims to provide the scientific community with an efficient and accessible resource, enabling them to perform accurate HLA imputations and contribute to advancements in the field of immunogenetics.

AUTHOR CONTRIBUTIONS

Nayane S. B. Silva and Sonia Bourguiba-Hachemi analyzed the data and took the lead in writing the article. Venceslas Douillard, Satu Koskela, Frauke Degenhardt, and Jonna Clancy contributed to the analysis of the results and to the writing of parts of the article. Nicolas Vince supervised the work, analyzed the data, and wrote and edited the article. Sophie Limou, Pierre-Antoine Gourraud, Andre Franke, and Erick C. Castelli provided critical feedback and helped shape the analysis and the article. Michel Satya Naslavsky and Diogo Meyer provided SABE cohort genomic data. Cibele Masotti and Stefan Knorst provided genomic data from São Paulo samples. All authors revised and commented on the manuscript.

ACKNOWLEDGMENTS

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. The authors wish to thank all SHLARC partners and the organizers of the 18th International HLA & Immunogenetics Workshop. Venceslas Douillard has received funding from the Inserm and Région Pays de la Loire. Nicolas Vince has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 846520.

CONFLICT OF INTEREST STATEMENT

All authors have declared no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ETHICS STATEMENT

All individual data were obtained through third-party access. The SHLARC project obtained approval from the ethics committee for noninterventional research. Nantes Université IRB number: 14092022.

ORCID

Nayane S. B. Silva https://orcid.org/0000-0001-5511-8426
Sonia Bourguiba-Hachemi https://orcid.org/0000-0002-2452-2861
Venceslas Douillard https://orcid.org/0000-0002-6762-4083
Satu Koskela https://orcid.org/0000-0001-9258-9163
Frauke Degenhardt https://orcid.org/0000-0001-7516-3179
Jonna Clancy https://orcid.org/0000-0002-0568-6676
Sophie Limou https://orcid.org/0000-0002-7702-8234
Diogo Meyer https://orcid.org/0000-0002-7155-5674
Cibele Masotti https://orcid.org/0000-0003-4462-0941
Michel Satya Naslavsky https://orcid.org/0000-0002-9068-1713
Erick C. Castelli https://orcid.org/0000-0003-2142-7196
Pierre-Antoine Gourraud https://orcid.org/0000-0003-1131-9554
Nicolas Vince https://orcid.org/0000-0002-3767-6210

REFERENCES

1. 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
2. MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

catalog). *Nucleic Acids Res*. 2017;45(D1):D896-D901. doi:10.1093/nar/gkw1133
3. 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
4. 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
5. 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. doi:10.1016/j.cell.2022.08.004
6. Cann HM, de Toma C, Cazes L, et al. A human genome diversity cell line panel. *Science*. 2002;296(5566):261-262. doi:10.1126/science.296.5566.261b
7. Vince N, Limou S, Daya M, et al. Association of HLA-DRB1*09:01 with tIgE levels among African-ancestry individuals with asthma. *J Allergy Clin Immunol*. 2020;146(1):147-155. doi:10.1016/j.jaci.2020.01.011
8. Naslavsky MS, Scliar MO, Yamamoto GL, et al. Whole-genome sequencing of 1,171 elderly admixed individuals from São Paulo, Brazil. *Nat Commun*. 2022;13(1):1004. doi:10.1038/s41467-022-28648-3
9. 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
10. Silva NDSB, Souza A d S, Andrade H d S, et al. Immunogenetics of HLA-B: SNP, allele, and haplotype diversity in populations from different continents and ancestry backgrounds. *HLA*. 2023;101(6):634-646. doi:10.1111/tan.15043
11. 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
12. Pappas DJ, Lizee A, Paunic V, et al. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. *Pharmacogenomics J*. 2018;18(3):367-376. doi:10.1038/tpj.2017.7
13. 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
14. Douillard V, Santos Brito Silva N d, Bourguiba-Hachemi S, et al. Optimal HLA imputation of admixed population with dimension reduction. *bioRxiv*. 2023. doi:10.1111/tan.15282
15. Cook S, Choi W, Lim H, et al. Accurate imputation of human leukocyte antigens with CookHLA. *Nat Commun*. 2021;12(1):1264. doi:10.1038/s41467-021-21541-5
16. Abi-Rached L, Gouret P, Yeh JH, et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PloS One*. 2018;13(10):e0206512. doi:10.1371/journal.pone.0206512
17. Auton A, Brooks LD, Durbin RM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
18. Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C, Pääbo S. Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc Natl Acad Sci U S A*. 1996;93(21):12035-12039. doi:10.1073/pnas.93.21.12035
19. Palo JU, Ulmanen I, Lukka M, Ellonen P, Sajantila A. Genetic markers and population history: Finland revisited. *Eur J Hum Genet EJHG*. 2009;17(10):1336-1346. doi:10.1038/ejhg.2009.53
20. Kerminen S, Havulinna AS, Hellenthal G, et al. Fine-scale genetic structure in Finland. *G3 Bethesda Md*. 2017;7(10):3459-3468. doi:10.1534/g3.117.300217
21. Lamnidis TC, Majander K, Jeong C, et al. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nat Commun*. 2018;9(1):5018. doi:10.1038/s41467-018-07483-5
22. Polvi A, Peräsaari J, Linjama T, et al. Description of four new HLA alleles in the Finnish population: a*03:283N, a*68:167, C*03:327, C*03:361. *HLA*. 2018;91(1):61-62. doi:10.1111/tan.13158
23. Linjama T, Eberhard HP, Peräsaari J, Müller C, Korhonen M. A European HLA isolate and its implications for hematopoietic stem cell transplant donor procurement. *Biol Blood Marrow Transplant J Am Soc Blood Marrow Transplant*. 2018;24(3):587-593. doi:10.1016/j.bbmt.2017.10.010
24. Koskela S, Tammi S, Clancy J, et al. MICA and MICB allele assortment in Finland. *HLA*. 2023;102:52-61. doi:10.1111/tan.15023
25. Ritari J, Hyvärinen K, Clancy J, Partanen J, Koskela S. Increasing accuracy of HLA imputation by a population-specific reference panel in a FinnGen biobank cohort. *NAR Genomics Bioinforma*. 2020;2(2):lqaa030. doi:10.1093/nargab/lqaa030
26. Ritari J, Koskela S, Hyvärinen K, FinnGen PJ. HLA-disease association and pleiotropy landscape in over 235,000 Finns. *Hum Immunol*. 2022;83(5):391-398. doi:10.1016/j.humimm.2022.02.003
27. Degenhardt F, Mayr G, Wendorff M, et al. Transethnic analysis of the human leukocyte antigen region for ulcerative colitis reveals not only shared but also ethnicity-specific disease associations. *Hum Mol Genet*. 2021;30(5):356-369. doi:10.1093/hmg/ddab017
28. 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
29. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods*. 2013;10(1):5-6. doi:10.1038/nmeth.2307
30. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x
31. Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. *Cell*. 2019;177(4):1080. doi:10.1016/j.cell.2019.04.032
32. Marchini J, Howie B. Genotype imputation for genome-wide association studies. *Nat Rev Genet*. 2010;11(7):499-511. doi:10.1038/nrg2796
33. 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
34. 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:D913-D919. doi:10.1093/nar/gkq1128
35. Nunes K, Zheng X, Torres M, et al. HLA imputation in an admixed population: an assessment of the 1000 genomes data

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

as a training set. *Hum Immunol*. 2016;77(3):307-312. doi:10.1016/j.humimm.2015.11.004

36. 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
37. Motyer A, Vukcevic D, Dilthey A, Donnelly P, McVean G, Leslie S. Practical use of methods for imputation of HLA alleles from SNP genotype data. *bioRxiv*. 2016:091009. doi:10.1101/091009
38. Naito T, Suzuki K, Hirata J, et al. A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes. *Nat Commun*. 2021;12(1):1639. doi:10.1038/s41467-021-21975-x

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, Douillard V, et al. 18th International HLA and Immunogenetics Workshop: Report on the SNP-HLA Reference Consortium (SHLARC) component. *HLA*. 2024; 103(1):e15293. doi:10.1111/tan.15293

20592310, 2024, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tan.15293 by University Of Helsinki, Wiley Online Library on [20/02/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License