Title

SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics

Journal

Genetic Epidemiology, 44(7)

ISSN

0741-0395

Authors

Vince, Nicolas
Douillard, Venceslas
Geffard, Estelle
et al.

Publication Date

2020-10-01

DOI

10.1002/gepi.22334

Peer reviewed

Received: 6 January 2020 | Revised: 19 June 2020 | Accepted: 3 July 2020

DOI: 10.1002/gepi.22334

RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

WILEY

SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics

Nicolas Vince[1] | **Venceslas Douillard**[1] | **Estelle Geffard**[1] | **Diogo Meyer**[2] | **Erick C. Castelli**[3] | **Steven J. Mack**[4] | **Sophie Limou**[1,5] | **Pierre-Antoine Gourraud**[1]

[1]Centre de Recherche en Transplantation et Immunologie, ITUN, UMR 1064, Université de Nantes, CHU Nantes, Inserm, Nantes, France

[2]University of São Paulo, São Paulo, Brazil

[3]UNESP—Universidade Estadual Paulista, Botucatu, São Paulo, Brazil

[4]Department of Pediatrics, University of California, San Francisco, UCSF Benioff Children's Hospital Oakland, Oakland, California

[5]Ecole Centrale de Nantes, Nantes, France

Correspondence
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
H2020 Marie Sklodowska-Curie Actions, Grant/Award Number: 846520

Abstract

Genome-wide associations studies have repeatedly identified the major histocompatibility complex genomic region (6p21.3) as key in immune pathologies. Researchers have also aimed to extend the biological interpretation of associations by focusing directly on human leukocyte antigen (*HLA*) polymorphisms and their combination as haplotypes. To circumvent the effort and high costs of HLA typing, statistical solutions have been developed to infer *HLA* alleles from single-nucleotide polymorphism (SNP) genotyping data. Though *HLA* imputation methods have been developed, no unified effort has yet been undertaken to share large and diverse imputation models, or to improve methods. By training the HIBAG software on SNP + HLA data generated by the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) to create reference panels, we highlighted the importance of (a) the number of individuals in reference panels, with a twofold increase in accuracy (from 10 to 100 individuals) and (b) the number of SNPs, with a 1.5-fold increase in accuracy (from 500 to 24,504 SNPs). Results showed improved accuracy with CAAPA compared to the African American models available in HIBAG, highlighting the need for precise population-matching. The SNP-HLA Reference Consortium is an international endeavor to gather data, enhance *HLA* imputation and broaden access to highly accurate imputation models for the immunogenomics community.

KEYWORDS

consortium, HLA, imputation, SNP

Nicolas Vince and Venceslas Douillard contributed equally to this work.

1 | INTRODUCTION

Beginning with the discovery of the HLA system in the 1950s, the characterization of *HLA* polymorphism and *HLA* disease associations have been performed in parallel (Dausset, 1999; Trowsdale & Knight, 2013). In the genome-wide association study (GWAS) era, the focus was shifted on single-nucleotide polymorphisms (SNP) with little to no biological relevance. Even when located in the major histocompatibility complex (MHC) region (6p21.3), these SNP associations have largely supplanted the traditional study of *HLA* allele associations. GWASs have however confirmed the crucial role of the *HLA* loci for the genetic epidemiology of nearly a quarter of all diseases and traits (MacArthur et al., 2017; Trowsdale & Knight, 2013), but SNP associations do not convey the immune-biological relevance that specific *HLA* alleles have. For example, GWASs of HIV disease identified the rs2395029 SNP near the *HCP5* gene on chromosome 6 as being the strongest associated with viral control (Fellay et al., 2007; Limou & Zagury, 2013). This SNP, which is located 100 kb from *HLA-B,* is in nearly complete linkage disequilibrium with the *HLA-B*57:01*, which can present HIV peptides crucial for HIV detection by the immune system (Chen et al., 2012; Limou & Zagury, 2013). Using novel bioinformatic approaches, we now have the ability to statistically infer *HLA* alleles from genotypic SNP data (imputation), returning HLA molecular functions to the forefront of disease-associated research (Meyer & Nunes, 2017; Pappas et al., 2018). Imputations are statistical methods that infer or predict missing information based on haplotypes. Haplotypes are a combination of genetic variants on one chromosome, they can be SNP haplotype (e.g., 011010, referring as the presence or absence of SNPs), gene haplotype (e.g., *HLA-A*01:01~HLA-B*08:01~HLA-C*07:01~HLA-DRB1*03:01~HLA-DQB1*02:01*) or a combination of different genetic variants (SNP, indels, substitution) haplotype (e.g., *HLA* alleles). In genomics, SNP imputation can infer the identity of missing SNPs that were not genotyped on GWAS arrays (Delaneau, Zagury, & Marchini, 2013; McCarthy et al., 2016) by comparing whole-genome SNP genotypes to a large reference panel of SNP haplotypes (Delaneau et al., 2013). Filling the genotyping gaps, SNP imputation performance and accuracy increased significantly when new large reference haplotype panels became available (McCarthy et al., 2016), which has contributed to a large number of discoveries over the past decade (Visscher et al., 2017).

In parallel to SNP, imputation also applies to *HLA* polymorphisms themselves, alone or in combination. It has revealed key associations in numerous diseases (Fellay et al., 2007; Limou & Zagury, 2013; MacArthur et al., 2017; Trowsdale & Knight, 2013; Vince et al., 2020) and can, as such, lead to the development of new drugs or patient-care guidance. Efforts to impute *HLA* alleles from these GWAS should be pursued to empower the community to go beyond simple SNP associations and to discover new disease associations (Khor et al., 2015; Meyer & Nunes, 2017; Shen et al., 2018); as an example, *HLA* alleles can bring new functional immunogenomics data such as prediction of amino acid, haplotypes (five genes: *A~B~C~DRB1~DQB1*) or imputed HLA-C expression easily implemented with Easy-HLA (Geffard et al., 2019; Vince et al., 2016). *HLA* allele imputation appears as a time and cost-effective alternative to the laborious *HLA* typing of all GWAS subjects. However, to rely on *HLA* imputation we must consider its accuracy, which depends on the reference panel quality (e.g., matching ancestry background, matching SNPs composition; Khor et al., 2015) and size (e.g., number of individuals with both SNP as well as *HLA* typing data, referred as SNP + HLA data; Pappas et al., 2018; Zheng et al., 2014). Successful *HLA* imputation, therefore, depends on the availability of large and diverse reference panels, which warrants a major collective effort in organizing community resources. Here, we advocate for the development of the SNP-HLA Reference Consortium (SHLARC), a new international network focused on collecting a large collection of high-quality HLA and SNP data, especially from an ethnically diverse population, with the goal to develop and share large reference panels and help worldwide researchers exploring *HLA* allelic information from their cohorts.

2 | RESULTS

We had access to the CAAPA (Consortium on Asthma among African-ancestry Populations in the Americas) data set (Daya et al., 2019; Vince et al., 2020) that consists of 880 whole-genome sequenced African American subjects with associated SNP GWAS data and typed *HLA* alleles at a two-field resolution (corresponding to the protein level). We chose the *HLA* Genotype Imputation with Attribute Bagging (HIBAG) R package (Zheng et al., 2014) to test the impact of the number of subjects and SNPs on *HLA* imputation accuracy. HIBAG demonstrates improved imputation accuracy over other available methods (Pappas et al., 2018) and allows the creation of custom reference panels, using the machine-learning technique of attribute bagging. Building reference panels requires heavy computing power which is related to the number of subjects and number of SNPs in an almost linear correlation (Zheng et al., 2014). The development of machine-learning algorithms heavily

relies on the evolution of computational power. We used graphics processing units (GPUs) as they are architecturally better suited to handle the computationally intensive tasks. For this project, we took advantage of the upgraded HIBAG version (HIBAG v1.15.3, HIBAG.gpu v0.9.1; Zheng, 2018) and used GPUs to build and compare multiple reference panels with a fivefold reduction in computation time relative to central processing units).

Starting with the complete data set ($n = 880$ individuals), we simulated scenarios of reference panel building by creating a collection of training and test sets. Each of the condition was replicated 10 times to assess the variability in the frequency of SNPs and HLA types and display confidence intervals for each prediction: (a) from a set of 100 samples ($n_{\text{training}} = 100$), we created 40 different reference panels with either increasing numbers of individuals (10/20/500/1,000) or increasing numbers of SNPs (500/1,000/5,000/10,000/24,504; see Supporting Information Methods) and (b) a test set ($n_{\text{test}} = 780$) used to assess the accuracy of *HLA* imputation from the 40 different reference panels (5 *HLA* genes × [4 different number of individuals + 4 different number of SNPs]; Figure 1). Accuracy is defined by the percentage of correct HLA allele prediction.

We observed that increasing the number of individuals in the reference panel increased *HLA* imputation accuracy (two-field resolution) for all loci (Figure 1a). As an example, accuracy rose from 60% with 10 individuals to 93% with 100 individuals for *HLA-DQB1*, and from 27% with 10 individuals to 71% with 100 individuals for *HLA-B* on average. We then compared the *HLA* imputation accuracies obtained from our CAAPA-based test set with pre-existing reference panels available on the HIBAG website (http://www.biostat.washington.edu/~bsweir/HIBAG/). These precomputed reference panels were all created with more than 100 individuals of African American ancestry (from 137 for *HLA-DQB1* to 171 for *HLA-B*) from the HLARES data and the HapMap Yoruba population. The accuracies using the precomputed HIBAG reference panels (represented as horizontal lines in Figure 1a) ranged from 70% (*HLA-DRB1*) to 87% (*HLA-A*) and were lower than those obtained using the CAAPA-based reference panels using a smaller number of individuals. This illustrates the importance of close matching of ancestry between the reference panel and the genotyped subjects, even within a single ancestry group (here African ancestry).

In addition, we reduced the number of SNPs in the training data set (500, 1,000, 5,000 and 10,000 out of the 24,504 available chromosome-6 SNPs) and observed that increasing the number of SNPs in the reference panel increased the *HLA* imputation accuracy for all genes (Figure 1b). For example, accuracy rose from 86% with 500 SNPs to 91% with the full set of 24,504 SNPs for *HLA-A*, and from 65% with 500 SNPs to 77% accuracy with the full set of SNPs for *HLA-B*. The number of SNPs in the training data set differs from the number of SNPs in the statistical model (or bag) as HIBAG does not use all SNPs provided in the input to create the reference panels (see Tables S1.1 and 1.2 for exact numbers). Indeed, HIBAG only includes SNPs within a 500-kb window around the gene of interest, and only keeps those improving the model after random selection (see Supporting Information Methods). For in-depth analysis of *HLA* imputation, we have also plotted the sensitivity and frequency of each allele to predict in the validation data set, to identify alleles decreasing the overall accuracy (see Figures S1–S5 and Table S2).

3 | DISCUSSION

Our results illustrate the importance of matching large reference panels with high SNP coverage to the input data set for efficient and accurate *HLA* allele imputation (Dilthey et al., 2016; Jia et al., 2013; Khor et al., 2015; Pappas et al., 2018). The goal of the SHLARC is to combine international expertise with data and computational resources. It will bring data to a level of interpretation that is key to solving questions on immune-related pathologies through innovative algorithms and powerful computation tool development. To achieve this goal, we determined three main objectives (Figure 2):

1. *Data*. By bringing together scientists from around the world, we will collectively increase the amount of SNP + HLA data available, both in terms of quantity and genetic diversity. Building new reference panels from these data will improve the performance of *HLA* allele imputation from SNPs as large, diverse, well-defined genomic data are the *prima materia* of successful collaborations and machine-learning applications for dissecting the genetic determinants of disease association.
2. *Applied mathematical and computer sciences*. We will further optimize SNP-HLA imputation methods using the HIBAG tool, and particularly for genetically diverse and admixed populations as (a) the higher complexity of their *MHC* region is a challenge for imputation and (b) these populations are still underrepresented in genomic studies (Sirugo, Williams, & Tishkoff, 2019). In addition, we will explore new machine-learning approaches such as deep learning to develop new, more efficient methods of *HLA* imputation.
3. *Accessibility and service to the scientific community*. Following the Haplotype Reference Consortium

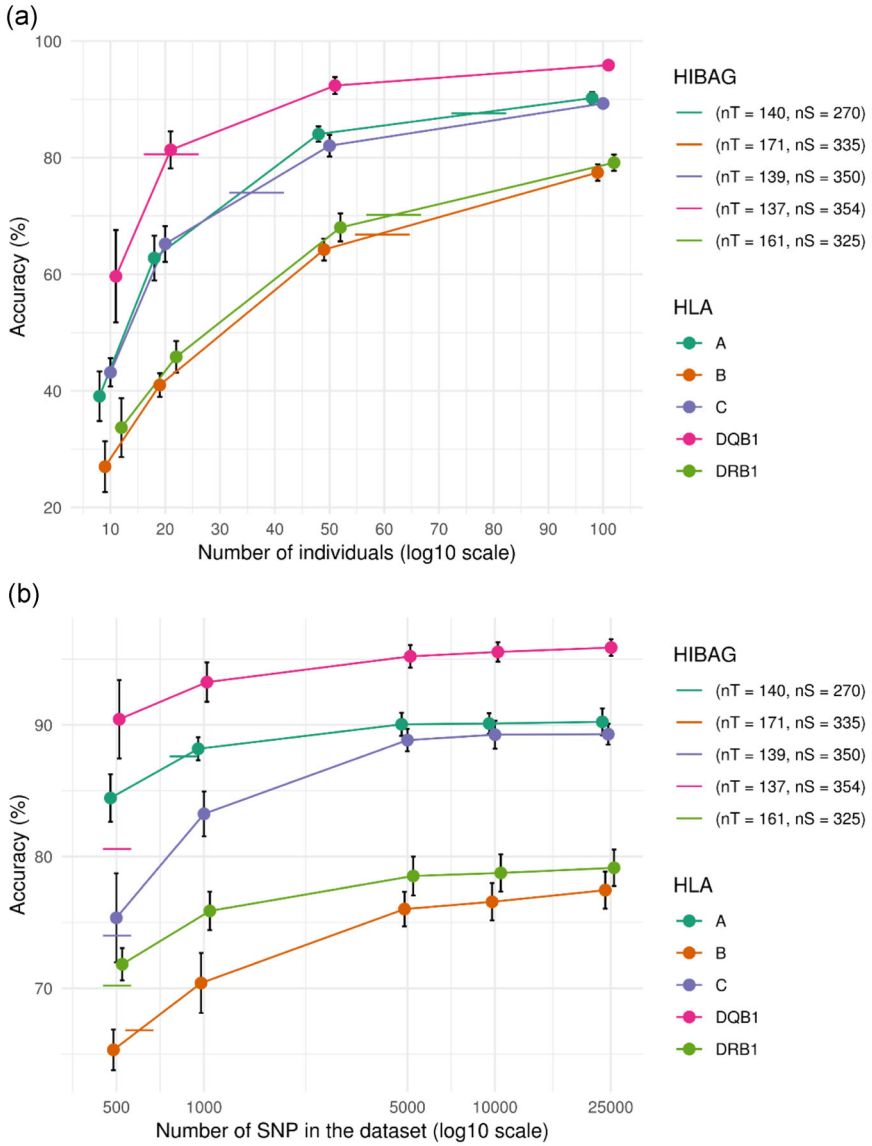

FIGURE 1 Influence of the number of individuals (a) and SNPs (b) in the HIBAG reference panel building on the accuracy of *HLA* alleles prediction. From the CAAPA data set ($N = 880$ and SNPs = 24,504), we produced a set of 10 training subsets ($n_{\text{training}} = 100$) and test ($n_{\text{test}} = 780$) sets to assess *HLA* imputation accuracy in different scenarios. Each model was validated by comparing the typed *HLA* alleles to the model-predicted *HLA* alleles across all individuals to provide an accuracy percentage (postprobability call threshold = 0). (a) By randomly selecting individuals in the training data set, we created sub-datasets containing 10, 20, and 50 individuals. Custom HIBAG models were computed for these subsets as well as for the whole 100 training individuals, using every available SNP. (b) Subsets of the training data set with 500, 1,000, 5,000, 10,000 randomly selected SNPs (out of the 24,504 available SNPs) were created and the corresponding models computed. The number of SNPs on the *x*-axis is indicative of the number of SNPs in the data set. The number of SNPs kept to create the model, which varies depending on the gene studied and the subset, is five times lower on average (see Tables S1.1 and S1.2). Note that the horizontal marks on each *HLA* gene curve indicate the accuracies obtained with the default African American HIBAG models. HIBAG, *HLA* Genotype Imputation with Attribute Bagging; HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism; nS, number of SNPs in the model; nT, number of individuals in the model

initiative (McCarthy et al., 2016), our network envisions building a free, user-friendly webserver where researchers can access our improved imputation protocols by simply uploading their data and obtaining the most accurate possible *HLA* imputation for their data set. This service will offer several solutions (a) ready-to-use anonymized reference panels for researchers wishing to impute the *HLA* themselves, (b) allow the on-demand creation and sharing of tailored (customized) reference panels based on data available

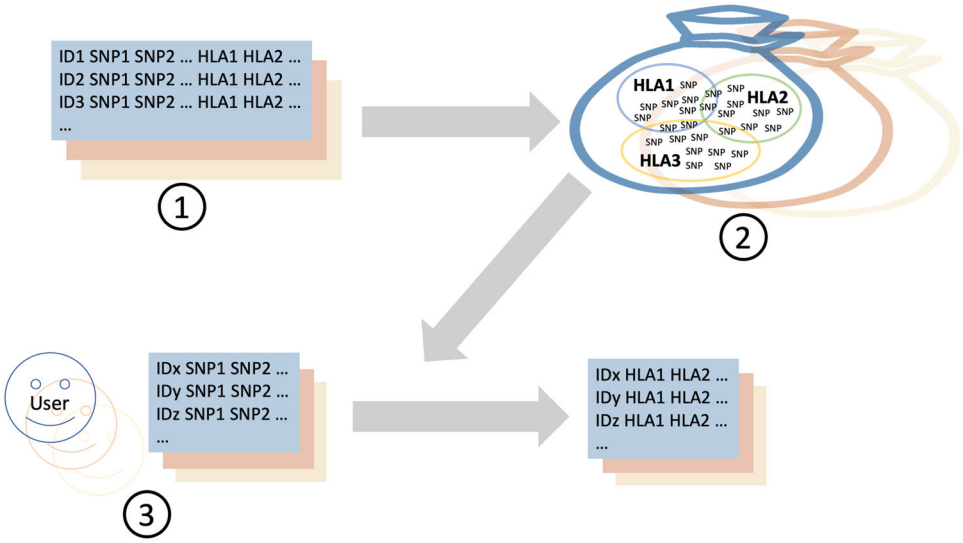

FIGURE 2 The SNP-HLA Reference Consortium (SHLARC) design. Aim 1: Increase the amount of SNP + HLA data available both in terms of quantity and diversity. Aim 2: Optimize SNP-HLA imputation methods. Aim 3: The SHLARC website will allow users from the scientific community to benefit from the data and knowledge accumulated by the consortium on SNP-to-HLA allele imputation. From a list of SNPs and a selected ethnicity of interest, or alternatively from uploading SNP genotype data sets, the best custom reference panel for *HLA* allele imputation will be built in our servers. HLA, human leukocyte antigen; SNP, single-nucleotide polymorphism

in our database, or (c) provide a full SNP-to-HLA imputation service from uploaded raw SNP genotypes. We will also explore how to create the reference panel with the best fit for ancestry and genotyping platforms, given the queried samples, without the need for the centralization of individual data. Indeed, distributed calculation techniques may allow to create reference panels from data hosted on different servers without collecting all the information in a single place.

Our objectives require access to the extensive computation power that is readily available through several GPU servers within the Université de Nantes. For each submission, we aim to design custom reference panels, for which SNPs, *HLA*, and reference panel data will be securely stored on University's servers. Importantly, reference panels represent statistical models that do not allow individual re-identification. The current SHLARC partners share complementary expertise including but not limited to bioinformatics, population genetics, and immunogenetics. Importantly, our network is designed around data sharing to facilitate open research as we believe research can be accelerated by freely sharing knowledge and data. With this in mind, we have added this consortium as a component of the 18th International HLA and Immunogenetics Workshop (https://www.ihiw18.org/).

HLA imputation is primarily intended for research applications, as clinical applications such as hematopoietic stem cell transplantation (HSCT) cannot tolerate statistical uncertainty, even though it might be used to accelerate pre-selection of HSCT patients as well (Meyer & Nunes, 2017; Pappas et al., 2018). The 1000 Genomes project (1000 Genomes Project Consortium et al., 2015) generated a large collection of polymorphisms from 2,504 individuals of diverse ancestry (SNPs, indels, and copy number variants), along with *HLA* allele typings (Gourraud et al., 2014), providing an informative overview of genetic diversity among human populations. However, a recent study by Abi-Rached et al. (2018) highlighted the absence of several common *HLA* alleles (>1% allele frequency) from the 1000 Genomes project which shows how *HLA* imputation results could be biased by an insufficient reference panel. With the proper sampling and a shared effort in gathering diverse data, *HLA* imputation could bridge the gap between *HLA* allele diversity and the understanding of its impact on phenotypes by harnessing the latent information stored in GWAS data sets to upgrade genetic epidemiological knowledge of immune-related diseases. As shown previously (Okada et al., 2015), predicting *HLA* alleles from population-matching reference panels not only increases the confidence in the predicted *HLA* but above all, allows prediction of specific *HLA* alleles that could not be imputed otherwise. Therefore, the informed choice of the applied model would strengthen the relation between *HLA*, ancestry, and disease risk factor. By applying this customization at a general level, we would assess ancestry with SNP relatedness, a

consistent marker of population, rather than using self-reported ancestry which can be often misleading (Sanchez-Mazas et al., 2012).

To develop this ambitious project, we encourage willing participants with available two-fields *HLA* alleles + SNPs data sets to join 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.

ACKNOWLEDGEMENTS

The authors thank Labex IGO (ANR-11-LABX-0016-01) and IHU CESTI for their support. 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. This study is supported by the ATIP-Avenir Inserm program, the Region Pays de Loire ConnectTalent, the ANR PIA-Investment (NExT), and the 18th International HLA and Immunogenetics Workshop. SNP-HLA Reference Consortium (SHLARC) Partners: Pierre-Antoine Gourraud, Nicolas Vince, Sophie Limou, Estelle Geffard, and Venceslas Douillard, Nantes Université, Centrale Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, F-44000 Nantes, France; Mario Südholt, Damien Eveillard, and Fatima-Zahra Boujdad, LS2N, UMR6004 CNRS, Université de Nantes, Centrale Nantes, IMTA, Nantes, France; Luisa Rocha Da Silva, Hugues Digonnet, and Domenico Borzacchiello, Ecole Centrale de Nantes, Nantes, France; Diogo Meyer, Vitor Aguiar, Kelly Nunes, University of São Paulo, São Paulo, Brazil; Erick C. Castelli, Unesp—Universidade Estadual Paulista, Botucatu-SP, Brazil; Surakameth Mahasirimongkol, Nuanjun Wichukchinda, Nusara Satproedprai, Sukanya Wattanapokayakit, Sacarin Bunbanjerdsuk, Punna Kunhapan, Thanyapat Wanitchanon, Penpitcha Thawong, and Pundharika Piboonsiri, Medical Genetics Center, Medical Life Sciences Insitute, Department of Medical Sciences, Ministry of Public Health; Soranun Chantarangsu, Chulalongkorn University, Department of Oral Pathology, Bangkok, Thailand; Sasithorn Chotewutmontri, Faculty of Medicine and Public Health, HRH Princess Chulabhorn College of Medical Science, Bangkok, Thailand; Supichaya Boonvisut, Environmental Toxicology, Chulabhorn Graduate Institute, Chulabhorn Royal Academy, Bangkok, Thailand; Derek Middleton, University of Liverpool, Liverpool, UK; Faviel Gonzalez, University of Liverpool, Liverpool, UK and Autonomous University of Coahuila, Mexico; James Traherne and Vitalina Kirgizova, University of Cambridge, Cambridge, UK; Andre Franke, Frauke Degenhardt, David Ellinghaus, and Mareike Wendorff, Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Kiel, Germany; Mehmet Dorak, Kingston University London, London, UK; Xiuwen Zheng, Department of Biostatistics, University of Washington, Seattle, WA, USA; Benedicte A. Lie, Marte Kathrine Viken, and Riad Hajdarevic, Department of Medical Genetics University of Oslo and Oslo University Hospital, Oslo, Norway; Department of Immunology, Rikshospitalet, University of Oslo and Oslo University Hospital, Oslo, Norway; Veron Ramsuran, University of KwaZulu-Natal, Durban, South Africa; Dara Torgerson and Ryan Hernandez, McGill University, Montreal, Canada; Zachary Szpiech, Auburn University, Auburn, AB, USA; Jill Hollenbach and Melissa Spear, University of California, San Francisco, CA, USA; Steven J. Mack, Department of Pediatrics, University of California, San Francisco and UCSF Benioff Children's Hospital Oakland, Oakland, CA, USA; Martin Maiers, Bioinformatics Research, Center for International Blood and Marrow Transplant Research, Minneapolis, MN, USA; Satu Koskela, Finnish Red Cross Blood Service, Helsinki, Finland; Anders Albrechtsen and Torben Hansen, The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark; Zorana Grubic, Katarina Stingl Jankovic, and Marija Maskalan, University Hospital Center Zagreb, Zagreb, Croatia; Martin Petrek and Katerina Sikorova, Faculty of Medicine and Dentistry, Palacky University, Olomouc, Czechia; Fatma Oguz, Instanbul University, Instanbul, Turkey; Jeremie Decouchant, Marcus Volp, Maria Fernandes, University of Luxembourg, Luxembourg, Luxembourg; Piotr Kusnierczyk, Hirszfeld Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, Wroclaw, Poland; Blanka Vidan-Jeras and Sendi Montanic, Blood Transfusion Center of Slovenia, Ljubljana, Slovenia.

CONFLICT OF INTERESTS

The authors declare that there are no conflict of interests.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at dbGAP (CAAPA, dbGaP Study Accession: phs001123.v1.p1) and from the 1000 Genomes Project website, using the latest SNP (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/release/20181203_biallelic_SNV/) and *HLA* data at the time (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/).

ORCID

Nicolas Vince http://orcid.org/0000-0002-3767-6210
Venceslas Douillard http://orcid.org/0000-0002-6762-4083

REFERENCES

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., ... 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Abi-Rached, L., Gouret, P., Yeh, J.-H., Di Cristofaro, J., Pontarotti, P., Picard, C., & Paganini, J. (2018). Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLOS One*, *13*(10), e0206512. https://doi.org/10.1371/journal.pone.0206512

Chen, H., Ndhlovu, Z. M., Liu, D., Porter, L. C., Fang, J. W., Darko, S., ... Walker, B. D. (2012). TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. *Nature Immunology*, *13*(7), 691–700. https://doi.org/10.1038/ni.2342

Dausset, J. (1999). The HLA adventure. *Transplantation Proceedings*, *31*(1–2), 22–24.

Daya, M., Rafaels, N., Brunetti, T. M., Chavan, S., Levin, A. M., Shetty, A., ... CAAPA. (2019). Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. *Nature Communications*, *10*(1), 880. https://doi.org/10.1038/s41467-019-08469-7

Delaneau, O., Zagury, J.-F., & Marchini, J. (2013). Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, *10*(1), 5–6.

Dilthey, A. T., Gourraud, P.-A., Mentzer, A. J., Cereb, N., Iqbal, Z., & McVean, G. (2016). High-accuracy HLA type inference from whole-genome sequencing data using population reference graphs. *PLOS Computational Biology*, *12*(10), e1005151. https://doi.org/10.1371/journal.pcbi.1005151

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., ... Goldstein, D. B. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, *317*(5840), 944–947.

Geffard, E., Limou, S., Walencik, A., Daya, M., Watson, H., Torgerson, D., ... Vince, N. (2019). Easy-HLA, a validated web application suite to reveal the full details of HLA typing. *Bioinformatics*, *36*(7), https://doi.org/10.1093/bioinformatics/btz875

Gourraud, P.-A., Khankhanian, P., Cereb, N., Yang, S. Y., Feolo, M., Maiers, M., ... Oksenberg, J. (2014). HLA diversity in the 1000 genomes dataset. *PLOS One*, *9*(7), e97282.

Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.-M., Concannon, P. J., Rich, S. S., ... de Bakker, P. I. W. (2013). Imputing amino acid polymorphisms in human leukocyte antigens. *PLOS One*, *8*(6), e64683. https://doi.org/10.1371/journal.pone.0064683

Khor, S.-S., Yang, W., Kawashima, M., Kamitsuji, S., Zheng, X., Nishida, N., ... Tokunaga, K. (2015). High-accuracy imputation for HLA class I and II genes based on high-resolution SNP data of population-specific references. *The Pharmacogenomics Journal*, *15*(6), 530–537. https://doi.org/10.1038/tpj.2015.4

Limou, S., & Zagury, J.-F. (2013). Immunogenetics: Genome-wide association of non-progressive HIV and viral load control: HLA genes and beyond. *Frontiers in Immunology*, *4*, 118. https://doi.org/10.3389/fimmu.2013.00118

MacArthur, J., Bowler, E., Cerezo, M., Gil, L., Hall, P., Hastings, E., ... Parkinson, H. (2017). The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, *45*(D1), D896–D901. https://doi.org/10.1093/nar/gkw1133

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., ... Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, *48*(10), 1279–1283. https://doi.org/10.1038/ng.3643

Meyer, D., & Nunes, K. (2017). HLA imputation, what is it good for? *Human Immunology*, *78*(3), 239–241. https://doi.org/10.1016/j.humimm.2017.02.007

Okada, Y., Momozawa, Y., Ashikawa, K., Kanai, M., Matsuda, K., Kamatani, Y., ... Kubo, M. (2015). Construction of a population-specific HLA imputation reference panel and its application to Graves' disease risk in Japanese. *Nature Genetics*, *47*(7), 798–802. https://doi.org/10.1038/ng.3310

Pappas, D. J., Lizee, A., Paunic, V., Beutner, K. R., Motyer, A., Vukcevic, D., ... Maiers, M. (2018). Significant variation between SNP-based HLA imputations in diverse populations: The last mile is the hardest. *The Pharmacogenomics Journal*, *18*(3), 367–376. https://doi.org/10.1038/tpj.2017.7

Sanchez-Mazas, A., Vidan-Jeras, B., Nunes, J. M., Fischer, G., Little, A.-M., Bekmane, U., ... Tiercy, J.-M. (2012). Strategies to work with HLA data in human populations for histocompatibility, clinical transplantation, epidemiology and population genetics: HLA-NET methodological recommendations. *International Journal of Immunogenetics*, *39*(6), 459–472. https://doi.org/10.1111/j.1744-313X.2012.01113.x. quiz 473–476.

Shen, J. J., Yang, C., Wang, Y.-F., Wang, T.-Y., Guo, M., Lau, Y. L., ... Sheng, Y. (2018). HLA-IMPUTER: An easy to use web application for HLA imputation and association analysis using population-specific reference panels. *Bioinformatics*, *37*(7), https://doi.org/10.1093/bioinformatics/bty730

Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The missing diversity in human genetic studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048

Trowsdale, J., & Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. *Annual Review of Genomics and Human Genetics*, *14*, 301–323. https://doi.org/10.1146/annurev-genom-091212-153455

Vince, N., Li, H., Ramsuran, V., Naranbhai, V., Duh, F.-M., Fairfax, B. P., ... Carrington, M. (2016). HLA-C level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. *American Journal of Human Genetics*, *99*(6), 1353–1358. https://doi.org/10.1016/j.ajhg.2016.09.023

Vince, N., Limou, S., Daya, M., Morii, W., Rafaels, N., Geffard, E., ... CAAPA. (2020). Association of HLA-DRB1*09:01 with tIgE levels among African ancestry individuals with asthma. *The Journal of Allergy and Clinical Immunology*, https://doi.org/10.1016/j.jaci.2020.01.011

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 years of GWAS discovery: Biology, function, and translation. *American Journal of Human Genetics*, *101*(1), 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005

Zheng, X. (2018). Imputation-based HLA typing with SNPs in GWAS studies. *Methods in Molecular Biology (Clifton, NJ)*, *1802*, 163–176. https://doi.org/10.1007/978-1-4939-8546-3_11

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., & Weir, B. S. (2014). HIBAG—HLA genotype imputation with attribute bagging. *The Pharmacogenomics Journal*, *14*(2), 192–200. https://doi.org/10.1038/tpj.2013.18

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: 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. *Genetic Epidemiology*. 2020;44:733–740. https://doi.org/10.1002/gepi.22334