Immunogenetic study of human immune-related diseases

Nicolas Vince

Habilitation à Diriger des Recherches

Soutenue le 25 août 2022

Immunogenetic study of human immune-related diseases

Nicolas VINCE

CRCN Inserm

Nantes Université, CHU Nantes, Inserm, Centre de Recherche Translationnelle en Transplantation et Immunologie (CR2TI), UMR 1064, ITUN, Nantes, France

Présidente du jury : Elise Launay

Rapporteure : Cheryl Winkler

Rapporteure : An Goris

Rapporteur : Anavaj Sakuntabhai

Examinateur : Alexis Elbaz

Examinateur : Jean-Luc Taupin

Acknowledgments - remerciements

I sincerely thank Alexis Elbaz and Jean-Luc Taupin, whom pre-reviewed this work and accepted to be part of my jury. I also am grateful for Cheryl Winkler, An Goris and Anavaj Sakuntabhai for reviewing this HDR thesis. I would like to add particular thanks to Elise Launay to have accepted to be part of my jury and for the many years of collaborations ahead of us.

My research peregrinations have taken me on path and locations I had not anticipated. My first encounter with a research lab was where I finally ended up, but in a very different field. Working with rats and neurobiology (Delphine Michel, Philippe Naveilhan, 2005) finished to convince myself to move from Nantes and embrace immune-related disease genetics. The Pasteur institute was my first meeting with capillary Sanger sequencing, studying Dengue (Anavaj Sakuntabhai, 2005). Then, I threw myself towards CVID and a multitude of genes within a young and budding team (Claire Fieschi, Marion Malphettes, Eric Oksenhendler, David Boutboul, Jean-Christophe Bories, 2006-2010). I did not know at that time, by joining the St-Louis Hospital in the old lab of Jean Dausset, I would dedicate a large part of my science to HLA but only after leaving Paris. Indeed, after a giant step across the ocean, in this small Frederick town, within the Fort Detrick military campus, I found myself again Sanger sequencing genes, only this time it was *HLA* (Mary Carrington, Arman Bashirova, Richard Apps, Veron Ramsuran, Stephen Anderson, 2011-2016). And it is with enthusiasm but not without a bit of apprehension that I moved back to Nantes where it had all started, after 11 years (Pierre-Antoine Gourraud). Here, I want to thank all mentors, colleagues, students whom accompany this science journey.

I want to add a special thanks to all the past and present ATIP-Avenir team and future team 3 for the wonderful science we make together: Ven, Estelle, Axelle, Rokhaya, Abel, Léo, Martin, Irène, Morgane, Elise, Olivia, Vincent, Pierre-Antoine, Alexandre, Gilles, Fabienne, Stéphanie, Christine, Sarah, Lucile…

Looking forward to pursue our promising collaboration and thanks to the whole NEMO team: David, Laureline, Arnaud, Alex, Emilie, Sita…

Thanks to the whole CR2TI and especially: Régis, Sophie, Nico, Laurence, Laurent, Amédée, Magali…

And of course, thanks to collaborators from all over the world: Erick, Florent, Laure, Satu, Frauke, Martin, Diogo, Alexis, Agnès, Emmanuelle, Cande, Ana, the SHLARC colleagues…

Merci à toute ma famille qui sont toujours là même quand ils sont loin : maman, papa, Télio, Taïna, Flo, Lili, Thanase, Yoyo, Roger, Marité, Dominique, Marie-Hélène…

Je veux terminer par la personne la plus importante de cette équipée scientifique, celle qui m'accompagne sur tous les chemins, celle sans qui rien de tout cela ne serait possible, merci Sophie.

I. Curriculum Vitae

A. Civil status

Nicolas VINCE, born December 4th 1982 in Saint-Nazaire (44), France.
French nationality.

B. Contact information

email: nicolas.vince@univ-nantes.fr
Tel.: +33240087424
Address: 30 bd Jean Monnet, 44093 Nantes cedex 1

C. Current status

Research scientist (**Chargé de recherche classe normale,** INSERM competitive recruitment 2021).

D. Host laboratory

Centre de Recherche Translationnelle en Transplantation et Immunologie (CR2TI), UMR1064, Inserm, Nantes Université | Institut de Transplantation en Urologie-Néphrologie (ITUN), CHU de Nantes, France.

E. Education

- 2010: **Ph.D** "Genetics and immunologic basis of Common Variable Immuno-deficiency" (Université Paris Diderot, IUH, Paris, France)
- 2006: **Master's degree** of Genetics (Université Paris XI, Orsay, France)
- 2005: **Pasteur Course** on Genome Analysis (Pasteur Institute of Paris, France)
- 2004: **Bachelor's degree** of Physiology and Cellular Biology (Université de Nantes, Nantes, France)

F. Work experience

- 2016-now: **Research scientist**, team 3 iTHINK, CR2TI UMR1064 - ITUN, Nantes, France. Genomic and genetic studies in HLA/KIR, Multiple Sclerosis, Kidney Transplantation.

- 2016-2018: **Guest Researcher** in Mary Carrington's lab, Leidos, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. Genomic and genetic studies in HLA expression and trans-eQTL.
- 2011-2016: **Post-Doc** in Mary Carrington's lab, Leidos, Frederick National Laboratory for Cancer Research, Frederick, MD, and Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. Genomic and genetic studies in HLA/KIR and candidate genes in HIV disease.
- 2006-2010: **Ph.D student** in EA3963 laboratory, IUH, St-Louis Hospital, Paris, France. Identification and functional description of genes involved in CVID.
- 2007-2008: **OpenOffice teacher** in Diderot University, Paris, France. Teaching basics concepts of Word, Excel and PowerPoint.
- 2005: 2 months **internship** in Pasteur Institute of Paris, France. Genotyping Malaria and Dengue fever patients.
- 2005: 2 months **internship** in INSERM U643, Nantes, France. Cell line implantation in rat brain and study of their differentiation by immunohistochemistry.

G. Awards

- 2019: 2 years Individual Fellowship - Marie Skłodowska-Curie Actions
- 2018: EFI personal bursary
- 2006: 3 years Ph.D fellowship from LNCC (Ligue Nationale Contre le Cancer)

H. Funding of research projects

In the past 5 years, I obtained 880 k€ funding as lead investigator for my research projects. I also participated in several other funding for large projects (total: 65M€).

- **PI**. 2022-2023: INSERM starting package after competitive recruitment. INSERM. 30k€. 2 years.
- Partner. 2022-2026: PRIMUS. RHU PIA5. 30M€. 5 years. (PI: Gilles Edan).
- **PI**. 2021-2022: Elucidating the genetic architecture of Neuro-Myelitis Optica Spectrum Disorders patients using the Multiple Sclerosis genetic burden score. ARSEP. 100k€. 2 years.
- **PI**. 2021-2022: HLA-3Diff. SFHI. 3.5k€. 1 year.
- **PI**. 2021-2024: SHLARC network. I-site NExT. 400k€. 3 years.

- **PI**. 2020-2021: Causal genetic variants in pneumococcal sepsis: exome analysis in children with extreme phenotype. ESPID. 10k€. 1 year. (PIs: Elise Launay, Nicolas Vince).
- **PI**. 2020-2021: KiT-DSApredict. Labex-IGO. 10k€. 1 year.
- **PI**. 2019-2022: PhD grant for Venceslas Douillard. INSERM/Région Pays de la Loire. 95k€. 3 years.
- Partner. 2019-2022: DELPHI. I-site NExT. Partner. 235k€. 3 years. (PIs: Frédéric Benhamou, Pierre-Antoine Gourraud).
- Partner. 2019-2022: International Research Network "Pathogen-Host Interaction and the Search for Biomarkers in infections". Universidade Estadual Paulista, Sao Paulo, Brazil. Partner. (PI: Erick Castelli).
- **PI**. 2019-2020: KiT-TEMRA. ABM. 35k€. 1 year.
- **PI**. 2019-2021: KiT-FIG. Individual Fellowships - Marie Skłodowska-Curie Actions. EU. 200k€. 2 years.
- Partner. 2019-2020: MSGB and GWAS in NMO. ARSEP. Partner. 50k€. 1 year. (PI: Pierre-Antoine Gourraud).
- Partner. 2019-2023: KTD Innov. RHU PIA3. Partner. 26,45M€. 5 years. (PIs: Sophie Brouard, Alexandre Loupy).
- Partner. 2019-2024: EU-TRAIN. EU. Partner. 6,6M€. 5 years. (PIs: Sophie Brouard, Alexandre Loupy).
- Partner. 2018-2019: KiT-GeniD. ABM. Partner. 30k€.1 year. (PI: Sophie Limou).
- Partner. 2018-2019: Déterminants génétiques de la fonction rénale chez les donneurs vivants. SFNDT. Partner. 30k€. 1 year. (PI: Sophie Limou).
- Partner. 2019-2022: SysMics. I-site NExT. Partner. 235k€. 3 years. (PI: Richard Redon).
- Partner. 2017-2018: AURORA. PHC Campus France. Partner. 10k€. 1 year. (PI: Pierre-Antoine Gourraud).
- **PI**. 2007-2009: Financement de thèse. LNCC. 90k€. 3 years.
- Partner. 2006-2009: CVID. ANR MRAR-06. Partner. 220k€. 3 years. (PI: Claire Fieschi).

I. Scientific networks

1. Scientific society membership

- European Federation of Immunogenetics (EFI): 2017-now
- American Society of Human Genetics (ASHG): 2013-now
- Société Française de Génétique Humaine (SFGH): 2018-now

- Association for Artificial Intelligence in Nantes (NaonedIA): 2018-now
- International AIDS Society (IAS): 2012-2013
- European Society of Immunodeficiency (ESID): 2008-2011

2. Consortia

- SHLARC: SNP-HLA Reference Consortium, **PI**. 2019-now.
- COVID-19 HLA and Immunogenetics Consortium (hlacovid19.org): 2020-now.

J. Scientific evaluation

1. Grant review

- ARSEP: 1 grant

2. Paper review

- Allergy: 2 papers
- American Journal of Transplantation: 1 paper
- AIDS: 3 papers
- Briefings in Bioinformatics: 1 paper
- Clinics and Research in Hepatology and Gastroenterology: 2 papers
- Frontiers in Genetics: 1 paper
- Frontiers in Immunology: 1 paper
- HLA: 1 paper
- Human Immunology: 11 papers
- International Journal of Immunogenetics: 5 papers
- Infection, Genetics and Evolution: 2 papers
- Journal of Virology: 1 paper
- Journal of AIDS: 1 paper
- PLoS One: 2 papers
- Scientific Report: 3 papers

K. Scientific animation

1. Conference chairing

- Abstract Session: Bioinformatics. Chairpersons: Nicolas Vince, Martin Maiers. EFI 2022, Amsterdam, Netherlands.

2. Conference organization

- LOC member for the EFI (European Federation of Immunogenetics) 2023.

- LOC member for Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) 2019.
- LOC member for Nantes Actualités Transplantation (NAT) 2018.

3. Committees

- European Federation of Immunogenetics (EFI) IT & Bioinformatics committee: 2021-now
- Technical committee of the mesocenter CNSC users (Centrale Nantes Supercomputing Centre): 2019-now

4. Jury

- MD thesis reviewer (1)
- PhD follow up committee (2)
- Master thesis reviewer (1)

5. Association

- NaonedIA, Association for the promotion of artificial intelligence in Nantes: 2018-now.

6. Miscellaneous

- Escape Game Transplant'Action. Partnership between CR2TI team 3 bioinfo with INSERM and Nantes Université to create an escape game about the genetic of kidney transplantation.

L. Teaching

- 2022: 2h course. Easy-HLA: at the crossroad of Immunology, Genetics, Bioinformatics and Precision medicine. UE scientific English, Nantes Université, Nantes, France.
- 2020: 2h course. HLA imputation as a stepping stone for HLA analyses in complex diseases: Example of asthma in African American individuals. UE initiation à la recherche, IMT-Atlantique, Nantes, France
- 2019: 2h course. Le challenge des données en biologie: exemple de l'imputation HLA. UE initiation à la recherche, IMT-Atlantique, Nantes, France
- 2019: 2h course. Greffe de cellules souches hématopoïétiques. UE médecine, Université de Nantes, France
- 2016: 2h course. Immunogénétique et auto-immunité. M2 BBRT, Université de Nantes, France
- 2007-2008: OpenOffice teacher for C2I (certificat informatique et internet) in first year of bachelor (60 hours; Université Paris VII Diderot, Paris).

II. Scientific management

In the course of my PhD, I had the chance of mentoring several students from diverse level: MD, Master, Engineer student, PhD student. However, during my postdoc, I could mentor very few students. Indeed, the NIH are not a university and do not foster many student programs. Since my arrival in Nantes, I could resume to welcoming students. Particularly, I mentored Estelle Geffard up to her PhD defense in December 2020 and am currently mentoring Venceslas Douillard and Nayane Brito for their PhD. In the next paragraphs, "Mentor" means that I was the main person involved in the student's training and "Co-mentor" means that I participated in the mentoring.

A. Bachelor

1. Mentor (N=4)

- 2020: Léo Hermet. 3 months. HLA, Genetics.
- 2015: Edward Yoon. 2 months. HLA, Genetics.
- 2009: Cécile Parmentier. 2 months. CVID, Genetics.
- 2008: Laura Simon. 2 months. CVID, Genetics.

2. Co-mentor (N=2)

- 2019: Arnaud Landry. 6 months. Biostatistics.
- 2017: Marinna Gaudin. 2 months. Biostatistics.

B. Master 1

1. Mentor (N=7)

- 2022: Kétsia Mortant. 3 months. Data analysis of KIR populations.
- 2021: Antoine Tanguy. 6 months. Hematopoeitic Stem Cell Transplantation. 1 presentation (SFHI 2021).
- 2021: Nabilah Ouro-Agouda. 6 months. Gene expression prediction.
- 2020: Lina Benabid. 6 months. HLAfix (gitlab.univ-nantes.fr/Nico_V/hlafix).
- 2020: Elgeta Hysaj. 6 months. HLA imputation.
- 2019: Léo Boussamet. 6 months. HLA-Epi (hla.univ-nantes.fr). 2 presentations (JOBIM 2019 as presenter, EFI 2019). 1 paper [1].
- 2019: Ayan Ianniello. 6 months. HLAfix (gitlab.univ-nantes.fr/Nico_V/hlafix). 1 presentation (EFI 2021). 1 poster (first author).

2. Co-mentor (N=5)

- 2020: Maëlle Guillout. 6 months. Nephrogenomic pipeline.
- 2020: Camille Picquet. 6 months. Surface protein pipeline.
- 2020: Antoine Tréhello. 3 months. Nephrogenomics.
- 2017: Kevin Hoang. 6 months. EWAS, HIV.
- 2017: Axelle Durand. 2 months. GWAS kidney disease in kids.

C. Master 2

1. Mentor (N=15)

- 2022: Justine Baron. 6 months. HLA genetic association in kidney transplantation.
- 2022: Augustin Moreau. 6 months. HLA-3Diff (hla.univ-nantes.fr).
- 2021: Morgane Gélin. 1 year (année recherche). Sepsis, Pediatric, Genetics. 2 presentations (Assises de génétique 2022, ESPID 2022), 2 posters.
- 2021: Irène Charles. 6 months. Multiple Sclerosis, Polygenic Risk Score. 1 presentation (Labex IGO days 2022), 4 posters.
- 2021: Flavie Durand-Perdriel. 6 months. Multiple Sclerosis, Database.
- 2020: Chloé Boulard. 6 months. Neuromyelitis Optica, Polygenic Risk Score, GWAS. 1 presentation (EFI 2021), 2 posters.
- 2019: Hadrien Règue. 6 months. Multiple Sclerosis, Polygenic Risk Score. 1 presentation (EFI 2021), 3 posters.
- 2019: Rémi Guimon. 6 months. Population genetics. 1 paper [2], 1 presentation (EFI 2019), 1 poster.
- 2019: Amandine Lecerf-Defer. 6 months. Transplantation, compatibility score. 1 poster.
- 2018: Venceslas Douillard. 6 months. HLA imputation. See PhD students.
- 2017: Corentin Hervé. 6 months. Precision medicine application. 2 posters.
- 2017: Estelle Geffard. 6 months. Easy-HLA website (hla.univ-nantes.fr). See PhD students.
- 2009: Jean-Marie Michot. 6 months. CVID, Genetics.
- 2007: Sophie Georgin. 6 months. CVID, Genetics.
- 2007: David Boutboul. 6 months. CVID, Genetics. See PhD students.

2. Co-mentor (N=8)

- 2021: Clémence Petit. 6 months. FSGS.
- 2021: Vincent Mauduit. 6 months. GWAS. Transplantation.
- 2020: Léo Boussamet. 6 months. Multiple Sclerosis, microbiota.

- 2019: Raphaël Gaisne. 6 months. Hereditary Hypophosphatemic Rickets.
- 2019: Claire Leman. 6 months. Autosomal dominant polycystic kidney disease.
- 2018: Rokhaya Ba. 6 months. Ferret (limousophie35.github.io/Ferret/).
- 2018: Axelle Durand. 6 months. GWAS kidney disease in kids.
- 2018: Abel Garnier. 6 months. EWAS, HIV.

D. Engineer students

1. Mentor (N=3)

- 2020: Valentin Redon. 3 months. HLA, Hematopoeitic Stem Cell Transplantation. 2 presentations (Labex IGO days 2022, EFI 2021).
- 2019: Marie Le Bougeant. 2 months. Multiple Sclerosis, data management.
- 2006: Muriel Berget. 2 months. CVID.

2. Co-mentor (N=2)

- 2019: Vincent Mauduit. 2 months. Transplantation, Machine learning.
- 2019: Romain Guédon. 5 months. Transplantation, Data management.

E. MD students (N=6)

- 2021: Morgane Gélin. Thesis director with Elise Launay. Defended on December 10th 2021.
- 2019: Camille Cany. 1 month. Transplantation.
- 2019: Morgan Jamot--Dubois. 1 month. Multiple Sclerosis.
- 2018: Hélène Beaucoudray. 1 month. Transplantation.
- 2018: Hélène Stork. 1 month. Transplantation.
- 2007: Zita Travnickova. 2 months. CVID, Genetics.

F. Technicians (N=2)

- 2009: Mariem Raho. CVID.
- 2007-2008: Angélique Guignet. CVID.

G. Engineer (N=6)

- 2021-now: Irène Charles, research engineer. Multiple Sclerosis, Neuromyelitis Optica.
- 2020-now: Olivia Rousseau, research engineer. Avatar, Transplantation.
- 2018-2020: Axelle Durand, research engineer. Project management.
- 2018-2020: Abel Garnier. Bioinformatic support.

- 2018-2019: Venceslas Douillard. Bioinformatic support.
- 2017-2018: Pauline Scherdel. Precision medicine application.

H. Postdoc

1. Co-mentor (N=1)

- 2021-now: Martin Morin. EWAS, HIV. 2 posters.

I. PhD students

I mentored Estelle Geffard during her PhD, I signed her 2 PhD papers as last author and am also last author on 1 other submitted paper. I am currently mentoring Venceslas Douillard PhD, we signed a paper as co-first author in 2020, we published 2 reviews where he is first author and I am co-last author with Sophie Limou and we have a paper in preparation.

1. Mentor (N=3)

- 2021-now: Nayane dos Santos Brito Silva (50% along with Sophie Limou). HLA imputation. Nayane is a Brazilian student, she is starting her PhD in cotutelle with Erick Castelli's lab, UNESP, Botucatu, Sao Paulo, Brazil. She is funded by the SHLARC project (NExT). She is included in **2** abstracts selected for **oral presentation** in international conferences, including **1 as presenter** (IHIW 2022). She participated in **2 posters** for conferences including **1 as first author**.
- 2019-now: Venceslas Douillard (50% along with Pierre-Antoine Gourraud). HLA imputation. Venceslas published **6 original articles** in peer review journals (**one as first author**) [2–7]. He published **3 review articles** including 2 in Frontier in genetics as first author [8–10]. He is included in **9** abstracts selected for **oral presentation** in international conferences, including **3 as presenter** (IHIW 2022, EFI 2021, IHIW satellite meeting 2021). He participated in **18 posters** for conferences including **6 as first author**.
- 2017-2020: Estelle Geffard (50% along with Pierre-Antoine Gourraud). Easy-HLA website (hla.univ-nantes.fr), precision medicine application. PhD obtained December 14th 2020. Estelle published **4 original articles** in peer review journals (**2 as first author**) [1,3,4,11]. She also published **2 proceeding papers** [12,13]. She has **1 submitted paper** as first author [14]. She is included in **18** abstracts selected for **oral presentation** in international conferences, including **6 as presenter** (EFI 2021, JOBIM 2019, EFI 2019, GDR 2018, NAT 2018). She participated in **21 posters** for conferences including **9 as first author**.

2. Co-mentor (N=2)

- 2019-now: Sirine Sayadi. Database and calculation distribution. 2 proceeding papers [12,13]. 2 oral presentations.
- 2008-2010: David Boutboul. CVID, Genetics. 6 papers [15–20]. 3 posters.

III. Scientific production

A. Bibliometric summary

In the past 13 years, I have contributed to the publication of 43 papers in peer review journals (38), online archive (3) or proceeding (2), including 7 as first author and 11 as last author (Fig. 1). In addition, we recently submitted one paper where I hold a last author position.

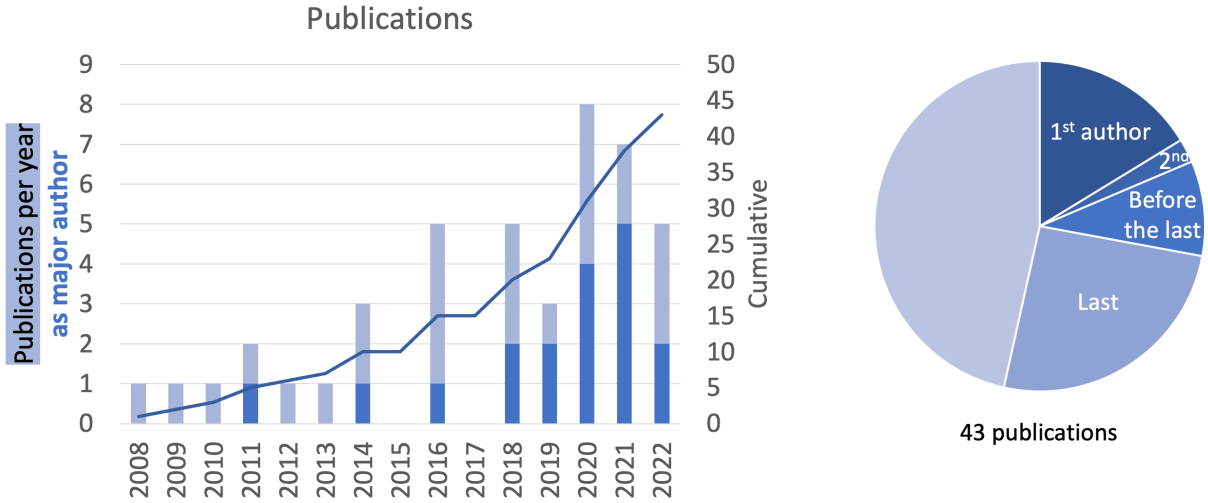

Figure 1: Evolution of the number of published papers throughout the years and distribution according to position hold.

As of June 8th 2022, Google scholar reported 1,575 citations of my work with an h-index of 17 (17/43 publications with ≥17 citations).

B. Publications

1. Original articles in peer review journals (35)

a) First or last author (14):

2022

1. Walencik A, Geffard E, Gautier AC, Delbos F, Chevallier P, Rialland F, Limou S, Gourraud PA, **Vince N***. EasyMatch-R software facilitates identification of compatible unrelated bone marrow donors, saves time-to-decision and money. HLA. Submitted (IF=4.5) *Corresponding author.
2. Geffard E, Boussamet L, Walencik A, Delbos F, Limou S, Gourraud PA, **Vince N***. HLA-EPI: A new EPIsode in exploring Donor/Recipient epitopic compatibilities. HLA. 2022 Feb;99(2):79-92. doi: 10.1111/tan.14505. Epub 2021 Dec 16. PMID: 34862850 (IF=4.5) *Corresponding author.

2021

3. Goodin DS, Oksenberg JR, Douillard V, Gourraud PA, **Vince N**. Genetic Susceptibility to Multiple Sclerosis in African Americans. PLoS One. 2021 Aug 9;16(8):e0254945. doi: 10.1371/journal.pone.0254945. eCollection 2021. (IF=3.6)
4. Goodin DS, Khankhanian P, Gourraud PA, **Vince N**. Genetic Susceptibility to Multiple Sclerosis: Interactions between Conserved Extended Haplotypes of the MHC and other Susceptibility Regions. BMC Med Genomics. 2021 Jul 10;14(1):183. doi: 10.1186/s12920-021-01018-6. PMID: 34246256 (IF=3.1, 6 citations)
5. Goodin DS, Khankhanian P, Gourraud PA, **Vince N**. The Nature of Genetic Susceptibility to Multiple Sclerosis. PLoS One. 2021 Mar 22;16(3):e0246157. doi: 10.1371/journal.pone.0246157. eCollection 2021. PMID: 33750973 (IF=3.6, 7 citations)

2020

6. **Vince N***, Douillard V, Geffard E, Meyer D, Castelli EC, Mack SJ, SHLARC investigators, Limou S, Gourraud PA. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics. Genet Epidemiol. 2020 Jul 18. doi: 10.1002/gepi.22334. Online ahead of print. PMID: 32681667 (IF=2.5, 5 citations) Front cover. *Corresponding author.
7. **Vince N**, Limou S, Daya M, Morii W, Rafaels N, Geffard E, Douillard V, Walencik A, Boorgula MP, Chavan S, Vergara C, Ortega VE, Wilson JG, Lange LA, Watson H, Nicolae DL, Meyers DA, Hansel NN, Ford JG, Faruque MU, Bleecker ER, Campbell M, Beaty TH, Ruczinski I, Mathias RA, Taub MA, Ober C, Noguchi E, Barnes KC on behalf of CAAPA, Torgerson D, Gourraud PA. Association of HLA-DRB1*09:01 with tIgE levels among African ancestry individuals with asthma. J Allergy Clin Immunol. 2020 Jan 22. pii: S0091-6749(20)30098-1. doi: 10.1016/j.jaci.2020.01.011. Online ahead of print. PMID: 31981624 (IF=14.1, 8 citations)
8. Geffard E, Limou S, Walencik A, Daya M, Watson H, Torgerson D, Barnes KC, Gautier AC, Gourraud PA, **Vince N**. Easy-HLA, a validated web application suite to reveal the full details of HLA typing. Bioinformatics. 2020 Apr 1;36(7):2157-2164. doi: 10.1093/bioinformatics/btz875. PMID: 31750874 (IF=4.5, 9 citations)

2019

9. **Vince N**, Poschmann J, Josien R, Anegon I, Limou S, Gourraud PA. 23rd Nantes Actualités Transplantation: "Genomics and Immunogenetics of Kidney and Inflammatory Diseases – Lessons for Transplantation". Transplantation. 2019 May;103(5):857-861. doi: 10.1097/TP.0000000000002517. [Epub ahead of print] PMID: 30399125 (IF=4.6, 1 citation)

2018

10. **Vince N***, Mouillot G, Malphettes M, Limou S, Boutboul D, Guignet A, Bertrand V, Pellet P, Gourraud PA, Debré P, Oksenhendler E, Theodorou I, Fieschi C, and the DEFI Study Group. Genetic screening of male patients with primary hypogammaglobulinemia can guide diagnosis and clinical management. Hum Immunol. 2018 Jul;79(7):571-577. doi: 10.1016/j.humimm.2018.04.014. Epub 2018 Apr 27. PMID: 29709555 (IF=2.0, 7 citations) *Corresponding author.
11. Goodin DS, Khankhanian P, Gourraud PA, **Vince N**. Highly Conserved Extended Haplotypes of the Major Histocompatibility Complex and their Relationship to Multiple Sclerosis Susceptibility. PLoS One. 2018 Feb 13;13(2):e0190043. doi: 10.1371/journal.pone.0190043. eCollection 2018. PMID: 29438392 (IF=2.8, 24 citations)

2016

12. **Vince N**, Li H, Ramsuran V, Naranbhai V, Duh FM, Fairfax BP, Saleh B, Knight JC, Anderson S, Carrington M. HLA-C level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. Am J Hum Genet. 2016 Dec 1;99(6):1353-1358. doi: 10.1016/j.ajhg.2016.09.023. Epub 2016 Nov 3. (IF=10.5, 44 citations)

2014

13. **Vince N**, Bashirova AA, Lied A, Gao X, Dorrell L, McLaren PJ, Fellay J, Carrington M. HLA class I and KIR genes do not protect against HIV-1 infection in highly exposed uninfected individuals with hemophilia A. J Infect Dis. 2014 Oct 1;210(7):1047-51. doi: 10.1093/infdis/jiu214. Epub 2014 Apr 8. (IF=6.3, 17 citations)

2011

14. **Vince N**, Boutboul D, Mouillot G, Just N, Peralta M, Casanova JL, Conley ME, Bories JC, Oksenhendler E, Malphettes M, Fieschi C, and the DEFI Study Group. Defects in the CD19 complex predispose to glomerulonephritis as well as IgG1 subclass deficiency. J Allergy Clin Immunol. 2011 Feb;127(2):538-541. (IF=13.1, 33 citations)

b) <u>Collaborations (21):</u>

2022

15. Domenighetti C, Douillard V, Sugier PE, Sreelatha AAK, Schulte C, Grover S, May P, Bobbili DR, Radivojkov-Blagojevic M, Lichtner P, Singleton AB, Hernandez DG, Edsall C, Gourraud PA, Mellick GD, Zimprich A, Pirker W, Rogaeva E, Lang AE, Koks S, Taba P, Lesage S, Brice A, Corvol JC, Chartier-Harlin MC, Mutez E, Brockmann K, Deutschländer AB, Hadjigeorgiou GM, Dardiotis E, Stefanis L, Simitsi AM, Valente EM, Petrucci S, Duga S, Straniero L, Zecchinelli A, Pezzoli G, Brighina L, Ferrarese C, Annesi

G, Quattrone A, Gagliardi M, Matsuo H, Nakayama A, Hattori N, Nishioka K, Chung SJ, Kim YJ, Kolber P, van de Warrenburg BPC, Bloem BR, Aasly J, Toft M, Pihlstrøm L, Correia Guedes L, Ferreira JJ, Bardien S, Carr J, Tolosa E, Ezquerra M, Pastor P, Diez-Fairen M, Wirdefeldt K, Pedersen NL, Ran C, Belin AC, Puschmann A, Rödström EY, Clarke CE, Morrison KE, Tan M, Krainc D, Burbulla LF, Farrer MJ, Krüger R, Gasser T, Sharma M, **Vince N**, Elbaz A, on behalf of the Comprehensive Unbiased Risk Factor Assessment for Genetics and Environment in Parkinson's Disease (Courage-PD) consortium. The interaction between HLA-DRB1 and smoking in Parkinson's disease revisited. Movement Disorders. Accepted. (IF=10.3)

16. Valencia A, Vergara C, Thio CL, **Vince N**, Douillard V, Grifoni A, Cox AL, Johnson E, Kral AH, Goedert JJ, Mangia A, Piazzolla V, Mehta SH, Kirk GD, Kim AY, Lauer GM, Chung RT, Price JC, Khakoo SI, Alric L, Cramp ME, Donfield SM, Edlin BR, Busch MP, Alexander G, Rosen HR, Murphy EL, Wojcik GL, Carrington M, Gourraud PA, Sette A, Thomas DL, Duggal P. Trans-Ancestral Fine-Mapping of MHC Reveals Key Amino Acids Associated with Spontaneous Clearance of Hepatitis C in HLA-DQβ1. Am J Hum Genet. 2022 Feb 3;109(2):299-310. doi: 10.1016/j.ajhg.2022.01.001. Epub 2022 Jan 31.. PMID: 35090584 (IF=11.0)

2021

17. Bashirova A, Zheng W, Akdag M, Augusto D, **Vince N**, Dong K, O'hUigin C, Carrington M. Population-specific diversity of the immunoglobulin constant heavy G chain (IGHG) genes. Genes Immun. 2021 Dec;22(7-8):327-334. doi: 10.1038/s41435-021-00156-2. Epub 2021 Dec 4. PMID: 34864821 (IF=2.7, citations)

2020

18. Gauttier V, Pengam S, Durand J, Biteau K, Mary C, Morello A, Néel M, Porto G, Teppaz G, Thepenier V, Danger R, **Vince N**, Wilhelm E, Girault I, Abes R, Ruiz C, Trilleaud C, Ralph KL, Trombetta ES, Garcia A, Vignard V, Martinet B, Glémain A, Bruneau S, Haspot F, Dehmani S, Duplouye P, Miyasaka M, Labarrière N, Laplaud D, Le Bas-Bernardet S, Blanquart C, Catros V, Gouraud PA, Archambeaud I, Aublé H, Metairie S, Mosnier JF, Costantini D, Blancho G, Conchon S, Vanhove B, Poirier N. Selective SIRPα blockade reverses tumor T cell exclusion and overcomes cancer immunotherapy resistance. J Clin Invest. 2020 Oct 19:135528. doi: 10.1172/JCI135528. Online ahead of print. PMID: 33074246 (IF=11.9, 18 citations)
19. Savage SA, Viard M, O'hUigin C, Zhou W, Yeager M, Li SA, Wang T, Ramsuran V, **Vince N**, Vogt A, Hicks B, Burdett L, Chung C, Dean M, de Andrade KC, Freedman ND, Berndt

S, Rothman N, Lan Q, Cerhan JR, Slager SL, Zhang Y, Teras LR, Haagenson M, Chanock SJ, Spellman SR, Wang Y, Willis A, Askar M, Lee SJ, Carrington M, Gadalla SM. Genome-wide association study identifies HLA-DPB1 as a significant risk factor for severe aplastic anemia. Am J Hum Genet. 2020 Feb 6;106(2):264-271. doi: 10.1016/j.ajhg.2020.01.004. Epub 2020 Jan 30. PMID: 32004448 (IF=10.5, 15 citations)

20. Montassier E, Al-Ghalith GA, Le Bastard Q, Douillard V, Garnier A, Guimon R, Raimondeau B, **Vince N**, Limou S, Gourraud PA, Laplaud DA, Nicot AB, Soulillou JP, Berthelot L. Enterobacteriaceae are the main providers of α1,3-Gal antigen in the human gut microbiome. Front Immunol. 2020 Jan 13;10:3000. doi: 10.3389/fimmu.2019.03000. eCollection 2019. PMID: 31998300 (IF=6.4, 28 citations)

2019

21. Daya M, Rafaels N, Brunetti TM, Chavan S, Levin AM, Shetty A, Gignoux CR, Boorgula MP, Wojcik G, Campbell M, Vergara C, Torgerson DG, Ortega VE, Doumatey A, Johnston HR, Acevedo N, Araujo MI, Avila PC, Belbin G, Bleecker E, Bustamante C, Caraballo L, Cruz A, Dunston GM, Eng C, Faruque MU, Ferguson TS, Figueiredo C, Ford JG, Gan W, Gourraud PA, Hansel NN, Hernandez RD, Herrera-Paz EF, Jiménez S, Kenny EE, Knight-Madden J, Kumar R, Lange LA, Lange EM, Lizee A, Maul P, Maul T, Mayorga A, Meyers D, Nicolae DL, O'Connor TD, Oliveira RR, Olopade CO, Olopade O, Qin ZS, Rotimi C, **Vince N**, Watson H, Wilks RJ, Wilson JG, Salzberg S, Ober C, Burchard EG, Williams LK, Beaty TH, Taub MA, Ruczinski I, Mathias RA, Barnes KC; CAAPA. Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. Nat Commun. 2019 Feb 20;10(1):880. doi: 10.1038/s41467-019-08469-7. PMID: 30787307 (IF=12.4, 48 citations)

2018

22. Renand A, Habes S, Mosnier JF, Aublé H, Judor JP, **Vince N**, Hulin P, Nedellec S, Métairie S, Archambaud I, Brouard S, Gournay J, Conchon S. Immune Alterations in Patients With Type 1 Autoimmune Hepatitis Persist Upon Standard Immunosuppressive Treatment. Hepatol Commun. 2018 Aug 6;2(8):968-981. doi: 10.1002/hep4.1202. eCollection 2018 Aug. PMID: 30094407 (IF=5.1, 29 citations)
23. Martin MP, Naranbhai V, Shea PR, Qi Y, Ramsuran V, **Vince N**, Gao X, Thomas R, Brumme ZL, Carlson JM, Wolinsky S, Goedert JJ, Walker B, Segal FP, Deeks S, Haas D, Migueles S, Connors M, Michael N, Fellay J, Price DA, Lafont BA, Pymm P, Saunders PM, Widjaja J, Wong SC, Vivian JP, Rossjohn J, Brooks AG, Carrington M. Killer cell immunoglobulin-like receptor 3DL1 variation modifies HLA-B*57 protection against HIV-

1. J Clin Invest. 2018 May 1;128(5):1903-1912. doi: 10.1172/JCI98463. Epub 2018 Apr 3. PMID: 29461980 (IF=13.3, 40 citations)

2016

24. Wang Y, Hwangpo T, Martin MP, **Vince N**, Qi Y, Reynolds RJ 4th, Absher D, Gao X, Ballinger CA, Burrows PD, Atkinson TP, Brown EE, Elgavish A, Liu C, Carrington M, Schroeder HW. Killer cell immunoglobulin-like receptors are associated with common variable immune deficiency pathogenesis. J Allergy Clin Immunol. 2016 Nov;138(5):1495-1498. doi: 10.1016/j.jaci.2016.07.011. Epub 2016 Sep 21. (IF=13.1, 4 citations)
25. Le Gall C, Laurent J, **Vince N**, Lizee A, Agrawal A, Walencik A, Rettman P, Gagne K, Retiere C, Hollenbach J, Cesbron A, Limou S, Gourraud PA. Multidimensional Reduction of Multicentric Cohort Heterogeneity: An Alternative Method to Increase Statistical Power and Robustness. Hum Immunol. 2016 Nov;77(11):1024-1029. doi: 10.1016/j.humimm.2016.05.013. Epub 2016 Jun 1. (IF=2.3)
26. Ebbo M, Gérard L, Carpentier S, Vély F, Cypowyj S, Farnarier C, **Vince N**, Malphettes M, Fieschi C, Oksenhendler E, Schleinitz N, Vivier E, for the DEFI Study Group. Low Circulating Natural Killer Cell Counts are Associated With Severe Disease in Patients With Common Variable Immunodeficiency. EBioMedicine. 2016 Apr;6:222-30. doi: 10.1016/j.ebiom.2016.02.025. Epub 2016 Mar 2. (IF=2.0, 51 citations)
27. Boutboul D, **Vince N**, Mahevas M, Bories JC, Fieschi C; DEFI Study Group. TNFA, ANXA11 and BTNL2 polymorphisms in CVID patients with granulomatous disease. J Clin Immunol. 2016 Feb;36(2):110-2. doi: 10.1007/s10875-016-0234-0. Epub 2016 Jan 18. (IF=3.3, 7 citations)

2014

28. Ericsen AJ, Starrett GJ, Greene JM, Lauck M, Raveendran M, Deiros DR, Mohns MS, **Vince N**, Cain BT, Pham NH, Weinfurter JT, Bailey AL, Budde ML, Wiseman RW, Gibbs R, Muzny D, Friedrich TC, Rogers J, O'Connor DH. Whole genome sequencing of SIV-infected macaques identifies candidate loci that may contribute to host control of virus replication. Genome Biol. 2014 Nov 7;15(11):478. doi: 10.1186/s13059-014-0478-z. (IF=11.9, 30 citations)
29. Bashirova AA, Apps R, **Vince N**, Mochalova Y, Yu XG, Carrington M. Diversity of the human LILRB3/A6 locus encoding a myeloid inhibitory and activating receptor pair. Immunogenetics. 2014 Jan;66(1):1-8. doi: 10.1007/s00251-013-0730-9. Epub 2013 Oct 6. (IF=2.1, 35 citations)

2013

30. Gouilleux-Gruart V, Chapel H, Chevret S, Lucas M, Malphettes M, Fieschi C, Patel S, Boutboul D, Marson MN, Gérard L, Lee M, Watier H, Oksenhendler E; **DEFI study group**. Efficiency of immunoglobulin G replacement therapy in common variable immunodeficiency: correlations with clinical phenotype and polymorphism of the neonatal Fc receptor. Clin Exp Immunol. 2013 Feb;171(2):186-94. doi: 10.1111/cei.12002. (IF=3.3, 57 citations)

2011

31. Rivoisy C, Gérard L, Boutboul D, Malphettes M, Fieschi C, Durieu I, Tron F, Masseau A, Bordigoni P, Alric L, Haroche J, Hoarau C, Bérézné A, Carmagnat M, Mouillot G, Oksenhendler E; **DEFI study group**. Parental consanguinity is associated with a severe phenotype in common variable immunodeficiency. J Clin Immunol. 2012 Feb;32(1):98-105. doi: 10.1007/s10875-011-9604-9. Epub 2011 Oct 15. (IF=3.4, 36 citations)
32. Boileau J, Mouillot G, Gérard L, Carmagnat M, Rabian C, Oksenhendler E, Pasquali JL, Korganow AS; **DEFI Study Group**. Autoimmunity in common variable immunodeficiency: correlation with lymphocyte phenotype in the French DEFI study. J Autoimmun. 2011 Feb;36(1):25-32. doi: 10.1016/j.jaut.2010.10.002. Epub 2010 Nov 13. (IF=7.6, 148 citations)

2010

33. Mouillot G, Carmagnat M, Gérard L, Garnier JL, Fieschi C, **Vince N**, Karlin L, Viallard JF, Jaussaud R, Boileau J, Donadieu J, Gardembas M, Schleinitz N, Suarez F, Hachulla E, Delavigne K, Morisset M, Jacquot S, Just N, Galicier L, Charron D, Debré P, Oksenhendler E, Rabian C; DEFI Study Group. B-cell and T-cell phenotypes in CVID patients correlate with the clinical phenotype of the disease. J Clin Immunol. 2010 Sep;30(5):746-55. doi: 10.1007/s10875-010-9424-3. Epub 2010 May 1. (IF=3.3, 158 citations)

2009

34. Malphettes M, Gérard L, Carmagnat M, Mouillot G, **Vince N**, Boutboul D, Bérezné A, Nove-Josserand R, Lemoing V, Tetu L, Viallard JF, Bonnotte B, Pavic M, Haroche J, Larroche C, Brouet JC, Fermand JP, Rabian C, Fieschi C, Oksenhendler E; DEFI Study Group. Late-onset combined immune deficiency: a subset of common variable immunodeficiency with severe T cell defect. Clin Infect Dis. 2009 Nov 1;49(9):1329-38. doi: 10.1086/606059. (IF=8.2, 205 citations)

2008

35. Oksenhendler E, Gérard L, Fieschi C, Malphettes M, Mouillot G, Jaussaud R, Viallard JF, Gardembas M, Galicier L, Schleinitz N, Suarez F, Soulas-Sprauel P, Hachulla E, Jaccard

A, Gardeur A, Théodorou I, Rabian C, Debré P; **DEFI Study Group**. Infections in 252 patients with common variable immunodeficiency. Clin Infect Dis. 2008 May 15;46(10):1547-54. doi: 10.1086/587669. (IF=8.2, 467 citations)

2. Review articles in peer review journals (4)

2022

36. Ba R, Geffard E, Douillard V, Simon F, Mesnard L, **Vince N**, Gourraud PA, Limou S. Surfing the big data wave: omics data challenges in transplantation. Transplantation. 2022 Feb 1;106(2):e114-e125. doi: 10.1097/TP.0000000000003992. PMID: 34889882 (IF=4.9)

2021

37. Douillard V, Castelli E, Mack SJ, Hollenbach J, Gourraud PA, **Vince N***, Limou S*, for the Covid-19|HLA & Immunogenetics Consortium and the SNP-HLA Reference Consortium. Approaching genetics through the MHC lens: tools and methods for HLA research. Front Genet. 2021 Dec 2;12:774916. doi: 10.3389/fgene.2021.774916. eCollection 2021. PMID: 34925459 (IF=4.4, 2 citations) * Joint last and corresponding author.
38. Douillard V, Castelli E, Mack SJ, Hollenbach J, Gourraud PA, **Vince N***, Limou S*, for the Covid-19|HLA & Immunogenetics Consortium and the SNP-HLA Reference Consortium. Current HLA investigations on SARS-CoV-2 and perspectives. Front Genet. 2021 Nov 29;12:774922. doi: 10.3389/fgene.2021.774922. eCollection 2021. PMID: 34912378 (IF=4.4, 7 citations) * Joint last and corresponding author.

2017

39. Limou S, **Vince N**, Parsa A. Lessons from CKD-related genetic association studies – moving forward. Clin J Am Soc Nephrol. 2017 Dec 14. pii: CJN.09030817. doi: 10.2215/CJN.09030817. [Epub ahead of print]. (IF=4.8, 15 citations)

3. Original articles in online archives (3)

2022

40. Goodin DS, Khankhanian P, Gourraud PA, **Vince N**. Multiple Sclerosis: Exploring the Limits of Genetic and Environmental Susceptibility. medRxiv 2022.03.09.22272129; doi: https://doi.org/10.1101/2022.03.09.22272129

2020

41. Goodin DS, Khankhanian P, Gourraud PA, **Vince N**. The Nature of Genetic Susceptibility to Multiple Sclerosis. bioRxiv 2020.08.13.249920; doi: https://doi.org/10.1101/2020.08.13.249920 (1 citation)

2019

42. Goodin DS, Khankhanian P, Gourraud PA, **Vince N**. Genetic Susceptibility to Multiple Sclerosis: Interactions between Conserved Extended Haplotypes of the MHC and other Susceptibility Regions. bioRxiv 603878; doi: https://doi.org/10.1101/603878 (5 citations)

4. Proceeding papers (2)

2021

43. Sayadi S, Geffard E, Südholt M, **Vince N**, Gourraud PA. (2021) Secure Distribution of Factor Analysis of Mixed Data (FAMD) and Its Application to Personalized Medicine of Transplanted Patients. In: Barolli L., Woungang I., Enokido T. (eds) Advanced Information Networking and Applications. AINA 2021. Lecture Notes in Networks and Systems, vol 225. Springer, Cham. https://doi.org/10.1007/978-3-030-75100-5_44 (1 citation)

2020

44. Sayadi S, Geffard E, Südholt M, **Vince N**, Gourraud PA. Distributed Contextualization of Biomedical Data: a case study in precision medicine. ACS/IEEE International Conference on Computer Systems and Applications, 2020. https://doi.org/10.1109/AICCSA50499.2020.9316502 (2 citations)

C. Oral presentations

63 presentations (Speaker is underlined)

1. Oral presentations in conferences (41)

a) International invited conference (8):

5 as speaker

1. **Vince N**. The SNP-HLA Reference Consortium (SHLARC), wrap up session. 18th International HLA & Immunogenetics Workshop, Noordwijkerhout, Netherlands.
2. Silva NSB, **Vince N**. Presentation: Vince N. *HLA-DPA1* and *HLA-DPB1* diversity and imputation in a Brazilian population. 18th International HLA & Immunogenetics Workshop, Noordwijkerhout, Netherlands.
3. Douillard V, **Vince N**. Improving HLA imputation in an admixed population with dimension reduction. 18th International HLA & Immunogenetics Workshop, Noordwijkerhout, Netherlands.
4. **Vince N**. The SNP-HLA Reference Consortium (SHLARC), introduction session. 18th International HLA & Immunogenetics Workshop, Noordwijkerhout, Netherlands.

5. **Vince N**. HLA-Epi: at the crossroad of Immunology, Genetics, Bioinformatics and Precision medicine. Invited conference. TIGED 2022, Antalya, Turkey.
6. **Vince N**. The SNP-HLA Reference Consortium: where we stand, where we go. International HLA & Immunogenetics Workshop satellite meeting 2021, Virtual meeting.
7. Douillard V, **Vince N**. The SNP-HLA Reference Consortium: promising results in HLA imputation. International HLA & Immunogenetics Workshop satellite meeting 2021, Virtual meeting.
8. **Vince N**, Gourraud PA. SNP-HLA reference consortium. IHIWS 2017, Pacific Grove, CA, USA.

b) International after selection (20):

7 as speaker

9. Douillard V, Silva NSB, Limou S, Gourraud PA, Castelli EC, **Vince N**. Improving HLA imputation in admixed population with UMAP dimension reduction. EFI 2022, Amsterdam, Netherlands.
10. Gelin M, Limou S, Gourraud PA, Durand A, Rousseau O, Gras-Leguen C, Lorton F, Launay E, **Vince N**. Identification of rare causal genetic variants in invasive pneumococcal diseases by exome analysis in children. ESPID 2022, Athens, Greece.
11. Petit C, Dantal J, Durand A, Ba R, Rousseau O, Mauduit V, Morin M, Gatault P, Garrouste C, Couzi L, Mesnard L, **Vince N**, Limou S. First genome-wide association study of post-transplantation FSGS recurrence. SFT 2021, Genève, Switzerland.
12. Sayadi S, Geffard E, Südholt M, **Vince N**, Gourraud PA. Secure distribution of Factor Analysis of Mixed Data (FAMD) and its application to personalized medicine of transplanted patients. AINA-2021: The 35-th International Conference on Advanced Information Networking and Applications, Virtual meeting.
13. Delbos F, Redon V, Douillard V, Limou S, Geffard E, Walencik A, Gourraud PA, **Vince N**. Resampling and Bayesian regression model to predict the graft-specific risk of donor specific HLA antibodies development for a given recipient awaiting solid organ transplantation. EFI 2021, virtual meeting.
14. Boulard C, Regue H, Limou S, Laplaud DA, Gourraud PA, **Vince N**. Polygenic risk score in Multiple Sclerosis: understanding the challenges of the post-GWAS era for complex diseases. EFI 2021, virtual meeting.
15. Douillard V, Limou S, Gourraud PA, **Vince N**. Sailing towards the new horizon of immunogenomics along with the SHLARC (SNP-HLA Reference Consortium). EFI 2021, virtual meeting.

16. Geffard E, Ianniello A, Limou S, Gourraud PA, **Vince N**. Development of a complete HLA analysis pipeline: HLA-Functional Immunogenomic eXploration (HLA*fix*). EFI 2021, virtual meeting.
17. Geffard E, Goronflot T, Limou S, **Vince N**, Wargny M, Gourraud PA. Data-driven methods generating synthetic data in genomics: the HLA "avatars" are shifting paradigms in data sharing. CHIST-ERA Conference 2020, Virtual meeting.
18. Sayadi S, Südholt M, Geffard E, **Vince N**. Distributed Contextualization of Biomedical Data: a case study in precision medicine. 17th ACS/IEEE International Conference on Computer Systems and Applications AICCSA 2020, Virtual meeting.
19. **Vince N**. RefGenSEP, an MS French cohort to reveal new potentials in MS genetic burden. Plenary. IMSGC face to face meeting 2019, Stockholm, Sweden.
20. **Vince N**, Douillard V, Limou S, Gourraud PA. Navigating the treacherous waters of HLA imputation with the SHLARC (SNP-HLA Reference Consortium). Best abstract plenary session. EFI 2019, Lisbon, Portugal.
21. Guimon R, Haspot F, Poirier N, Blancho G, Limou S, Gourraud PA, **Vince N**. The SIRP gene family: widespread conservation in animals, haplotypic polymorphisms in humans and its therapeutic consequences for monoclonal antibody reactivity. EFI 2019, Lisbon, Portugal.
22. Geffard E, Boussamet L, Walencik A, Limou S, **Vince N**, Gourraud PA. HLA eplet in Easy-HLA: prediction and compatibility level assessment available in a complete webtool suite. EFI 2019, Lisbon, Portugal.
23. Ba R, **Vince N**, Geffard E, Malguid D, Lanza M, Gourraud PA, Limou S. Burrowing functional and immunogenetic information through the 1000 Genomes Project with Ferret v3.0. EFI 2019, Lisbon, Portugal.
24. Daya M, Rafaels N, Brunetti T, Chavan S, Levin AM, Shetty A, Gignoux CR, Boorgula MP, Wojcik G, Campbell M, Vergara C, Torgerson DG, Ortega VE, Doumatey A, Johnston HR, Acevedo N, Araujo MA, Avila PC, Belbin G, Bleecker E, Bustamante C, Caraballo L, Cruz AA, Dunston GM, Eng C, Faruque MU, Ferguson TS, Figueiredo C, Ford JG, Gan W, Gourraud PA, Hansel N, Hernandez RD, Herrera-Paz EP, Jimenez S, Kenny EE, Knight-Madden J, Kumar R, Lange LE, Lange EM, Lizee A, Maul P, Maul T, Mayorga Sirera AJ, Meyers DA, Nicolae DL, O'Connor TD, Oliveira RR, Olopade CO, Olopade O, Qin ZS, Rotimi C, **Vince N**, Watson H, Wilks RJ, Wilson JG, Salzberg S, Ober C, Burchard EG, Williams LK, Beaty TH, Taub MA, Ruczinski I, Mathias RA, Barnes KC. Association

study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. AAAAI 2019, San Francisco, CA, USA.

25. **Vince N**, Limou S, Daya M, Rafaels, N, Hollenbach J, Lizée A, Geffard E, Walencik A, Pino-Yanes M, Salzberg S, Kim D, Watson H, Lange L, Wilson J, Beaty T, Taub M, Ruczinski I, Mathias R, CAAPA, Barnes K, Torgerson D, Gourraud PA. HLA-DRB1 09:01 is associated with a severity asthma outcome in the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). EFI 2018, Venice, Italy.
26. Hoang K, Nelson G, Binns-Roemer E, **Vince N**, Gourraud PA, Goedert JJ, Winkler C, Limou S. Epigenome-wide association study reveals immunogenetic targets of DNA methylation modification by HIV-1. Best abstract plenary session. EFI 2018, Venice, Italy.
27. **Vince N**, Daya M, Hollenbach J, Lizée A, Pino-Yanes M, Salzberg S, Kim D, Beaty T, Taub M, Ruczinski I, Mathias R, Rafaels N, CAAPA, Barnes K, Torgerson D, Gourraud PA. HLA and asthma in the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). EFI 2017, Mannheim, Germany.
28. Walencik A, Geffard E, **Vince N**, Garnier F, Cesbron A, Gourraud PA. Easy-HLA: a package of applications to reveal the full details of HLA typing. EFI 2017, Mannheim, Germany.

c) National invited conference (1):

29. **Vince N**. Big Data et médecine de précision. Invited closing plenary session. SFTS 2019, Nantes, France.

d) National after selection (12):

2 as speaker

30. Gelin M, Limou S, Gourraud PA, Durand A, Rousseau O, Gras-Leguen C, Lorton F, Launay E, **Vince N**. Identification of rare causal genetic variants in invasive pneumococcal diseases by exome analysis in children. Assises de génétique humaine et médicale 2022, Rennes, France.
31. Boussamet L, Geffard E, Walencik A, Limou S, Gourraud PA, **Vince N**. Easy-HLA web application: new tools for HLA genotypes studies. Plenary. JOBIM 2019, Nantes, France.
32. Geffard E, Scherdel P, Limou S, Brouard S, Giral M, **Vince N**, Gourraud PA. A precision medicine application: personalized contextualization of patients after kidney transplantation. JOBIM 2019, Nantes, France.

33. **Vince N**, Douillard V, Geffard E, Limou S, Gourraud PA. SNP-HLA Reference Consortium: partage de données HLA et SNP pour la promotion des études immunogénomiques. Plenary. GDR science et santé 2018, Nantes, France.
34. Geffard E, Walencik A, Limou S, **Vince N**, Gourraud PA. Easy-HLA, outils de mise à niveau des génotypes HLA par imputation statistique et computationnelle. Plenary. GDR science et santé 2018, Nantes, France.
35. **Vince N**, Geffard E, Douillard V, Limou S, Gourraud PA. Harnessing the power of functional immunogenomics parameters to discover new associations with diseases. Plenary. NAT 2018, Nantes, France.
36. Geffard E, Scherdel P, Limou S, Brouard S, Tissot A, COLT investigators, DIVAT investigators, Magnan A, Giral M, Blancho G, **Vince N**, Gourraud PA. A precision medicine application: personalized contextualization of patients after solid organ transplantation. Plenary. NAT 2018, Nantes, France.
37. Walencik A, Geffard E, Limou S, Cesbron A, **Vince N**, Gourraud PA. EasyMatch-R: a web application to facilitate donor query in Hematopoietic Stem Cell Transplantation (HSCT). Plenary. NAT 2018, Nantes, France.
38. Limou S, Durand A, Ng D, **Vince N**, Reidy K, Woroniecki R, Furth S, Warady B, Wong CS, Gourraud PA, Kopp JB, Winkler C, Kaskel FJ. New Candidate Genetic Loci Associated with Pediatric Proteinuria in the CKiD Cohort. Plenary. NAT 2018, Nantes, France.
39. Limou S, Hoang K, Garnier A, Nelson G, **Vince N**, Gourraud PA, Winkler C. Epigenome-wide association study (EWAS) with HIV infection status. Plenary. NAT 2018, Nantes, France.
40. Limou S, Taverner A, Ba R, **Vince N**, Gourraud PA, Servières M, Winkler C. Ferret, a user-friendly Java tool to extract data from the 1000 Genomes Project. Plenary. NAT 2018, Nantes, France.
41. Hoang K, Nelson G, **Vince N**, Gourraud PA, Winkler C, Limou S. Epigenome-wide association study (EWAS) with HIV infection status. Assises de génétique humaine et médicale 2018, Nantes, France.

2. Oral presentations in other contexts (22)

a) Talks in institutes of international reputation (6)

42. **Vince N**. Deciphering the control of HLA-C expression using the 1000 genomes dataset. Invited seminar at Université de Luminy, TAGC, INSERM U1090 - Marseille, France, 2015.

43. **Vince N**. Deciphering the control of HLA-C expression using the 1000 genomes dataset. Invited seminar at Université de Nantes, INSERM UMR 915 - Nantes, France, 2015.
44. **Vince N**. HLA in hemophiliac highly exposed HIV seronegative. CIP seminar - Frederick, MD, USA, 2013.
45. **Vince N**. HLA-E and HLA class I signal peptide during HIV progression. CIP seminar - Frederick, MD, USA, 2012.
46. **Vince N**. LILRA4 in HIV. CIP seminar - Frederick, MD, USA, 2012.
47. **Vince N**. Genetics and immunologic basis of CVID. BCGC seminar series - Frederick, MD, USA, 2011.

b) Local oral interventions (9)

8 as speaker

48. Delbos F, Redon V, Douillard V, Limou S, Geffard E, Walencik A, Gourraud PA, **Vince N**. Artificial intelligence algorithms to predict donor specific antibody development in kidney transplanted patients (KiT-DSApredict). LabEx IGO days 2022, Nantes, France.
49. Charles I, Limou S, Laplaud DA, Gourraud PA, **Vince N**. Elucidating the genetic architecture of Neuromyelitis Optica disease using Genome wide association approach. LabEx IGO days 2022, Nantes, France.
50. **Vince N**. Easy-HLA: at the crossroad of Immunology, Genetics, Bioinformatics and Precision medicine. Invited conference. Congrès Licence Science de la Vie 2022, Nantes, France.
51. **Vince N**. Round table. Afterwork: "Et si le calcul intensif était aussi pour vous ?". ICI, Ecole Centrale de Nantes 2018, Nantes, France.
52. **Vince N**. Computer intensive HLA imputation in the context of HLA association with asthma in African American population - User experience. Plan de Mutualisation des Moyens de Calcul (MMC) de l'INSERM - Demi-journée d'information, Nantes, France, 2017.
53. **Vince N**. HLA-E signal peptide - protection during CMV infection? LSOCA investigator meeting - Frederick, MD, USA, 2012.
54. **Vince N**. Mutations de CD19 dans le DICV. Rencontre clinique DEFI - Paris, France, 2008.
55. **Vince N**. Bases génétique et immunologique du DICV. Rencontre doctorale B2T - Paris, France, 2008.
56. **Vince N**. Etude de TACI dans le DICV. RJS 2007 - Paris, France, 2007.

c) Oral interventions – dissemination to the general public (6)

57. **Vince N**. Les enjeux de la génétique. Nuit blanche des chercheurs 2022, Nantes, France.
58. **Vince N**. Speed searching. Nuit blanche des chercheurs 2022, Nantes, France.
59. **Vince N**, Douillard V, Garnier A, Geffard E, Durand A, Ba R, Limou S. Escape Game: il faut sauver le soldat Barack. Fête de la science 2019, Nantes, France.
60. **Vince N**. Genetic round table. Invited conference. Utopiales 2019, Nantes, France.
61. **Vince N**. Speed searching. Nuit des chercheurs 2019, Angers, France.
62. **Vince N**. De Guérande aux USA - A la recherche du gène "malade". Lycée Galilée – Guérande, France, 2013.
63. **Vince N**. Etude de TACI dans le DICV. Ligue nationale contre le cancer - Blois, France, 2007.

D. Posters (63)

1. Poster: Douillard V, Silva NSB, Limou S, Gourraud PA, Castelli EC, **Vince N**. Improving HLA imputation in admixed population with UMAP dimension reduction. HLA Symposium - in health and disease 2022, Zurich, Switzerland.
2. Poster: Charles I, Limou S, Rousseau O, Douillard V, Mauduit V, Laplaud DA, Gourraud PA, **Vince N**. Polygenic risk score in Multiple Sclerosis: understanding the challenges of the post-GWAS era for complex diseases. ARSEP 2022, Paris, France.
3. Poster: Gelin M, Limou S, Durand A, Rousseau O, Gourraud PA, Gras-Leguen C, Lorton F, Toubiana J, Launay E, **Vince N**. Identification of rare causal genetic variants in invasive pneumococcal diseases by exome analysis in children. EFI 2022, Amsterdam, Netherlands.
4. Poster: **Vince N**, Haspot F, Le Bas-Bernardet S, Poirier N, Gourraud PA, Blancho G, Limou S. SIRPA V1/V2 haplotypes were selected by balancing selection impacting therapy. EFI 2022, Amsterdam, Netherlands.
5. Poster: Silva NSB, Knorst SHY, Carmo RT, Masotti C, Naslavsky MS, Duarte YAO, Zatz M, Douillard V, Limou S, **Vince N**, Castelli EC. Improving HLA imputation in admixed population with UMAP dimension reduction. EFI 2022, Amsterdam, Netherlands.
6. Poster: Domenighetti C, Douillard V, Sugier PE, consortium Courage-PD, **Vince N**, Elbaz A. Une valine 11 codée par le gène HLA-DRB1 est associée inversement au risque de maladie de Parkinson et interagit avec l'initiation du tabagisme. JNLF 2022, Strasbourg, France.

7. Poster: **Vince N**, Haspot F, Le Bas-Bernardet S, Poirier N, Gourraud PA, Blancho G, Limou S. SIRPA V1/V2 haplotypes were selected by balancing selection. Assises de génétique humaine et médicale 2022, Rennes, France.
8. Poster: Charles I, Limou S, Rousseau O, Douillard V, Mauduit V, Gourraud PA, **Vince N**. Polygenic risk score in Multiple Sclerosis: understanding the challenges of the post-GWAS era for complex diseases. Assises de génétique humaine et médicale 2022, Rennes, France.
9. Poster: Mauduit V, Durand A, Ba R, Morin M, Petit C, Gourraud PA, **Vince N**, Limou S. Integrating large-scale genetic data to study kidney graft survival. SFT 2021, Genève, Suisse.
10. Poster: **Vince N**, Haspot F, Le Bas-Bernardet S, Poirier N, Gourraud PA, Blancho G, Limou S. SIRPA V1/V2 haplotypes were selected by balancing selection. ASHG 2021, Virtual meeting.
11. Poster: Charles I, Limou S, Rousseau O, Douillard V, Mauduit V, Gourraud PA, **Vince N**. Polygenic risk score in Multiple Sclerosis: understanding the challenges of the post-GWAS era for complex diseases. ASHG 2021, Virtual meeting.
12. Poster: Gelin M, Limou S, Gourraud PA, Durand A, Rousseau O, Gras-Leguen C, Lorton F, Launay E, **Vince N**. Identification of causal genetic variants in invasive pneumococcal diseases by exome analysis in children. ASHG 2021, Virtual meeting.
13. Charles I, Boulard C, Regue H, Shah S, Garcia A, Limou S, Gourraud PA, Laplaud D, **Vince N**. Polygenic risk score in Multiple Sclerosis: understanding the challenges ofthe post-GWAS era for complex diseases. NAT & IGO joint meeting 2021, virtual meeting.
14. Douillard V, Limou S, Gourraud PA, **Vince N**. Sailing towards the new horizon of immunogenomics along with the SHLARC (SNP-HLA Reference Consortium). NAT & IGO joint meeting 2021, virtual meeting.
15. Geffard E, Ba R, Durand A, Limou S, Sayadi S, Sudholt M, Brouard S, Loupy A, **Vince N**, Gourraud PA. De-centralized database: new challenges to design innovative contextualization algorithms. NAT & IGO joint meeting 2021, virtual meeting.
16. Douillard V, Limou S, Gourraud PA, **Vince N**. Sailing towards the new horizon of immunogenomics along with the SHLARC (SNP-HLA Reference Consortium). MCAA 2021, virtual meeting.
17. **Vince N**, Boulard C, Regue H, Shah S, Garcia A, Limou S, Gourraud PA, Laplaud DA. Elucidating the genetic architecture of Neuromyelitis Optica Spectrum Disorders patients using the Multiple Sclerosis genetic burden score. ASHG 2020, Virtual meeting.

18. Douillard V, Limou S, SHLARC investigators, Gourraud PA, **Vince N**. Sailing towards new immunogenomics horizons along with the SHLARC (SNP-HLA Reference Consortium). ASHG 2020, Virtual meeting.
19. Douillard V, **Vince N**, Limou S, Gourraud PA. Navigating the treacherous waters of HLA imputation with the SHLARC (SNP-HLA Reference Consortium). IPW 2019, New Orleans, LA, USA.
20. **Vince N**, Douillard V, Limou S, Gourraud PA. Navigating the treacherous waters of HLA imputation with the SHLARC (SNP-HLA Reference Consortium). ASHG 2019, Houston, TX, USA.
21. Gaisne R, Figueres L, Houillier P, Vargas-Poussou R, Lemoine S, Leman C, **Vince N**, Gourraud PA, Limou S. Exome sequencing in Hereditary Hypophosphatemic Rickets with Hypercalciuria. JOBIM 2019, Nantes, France.
22. Leman C, Gaisne R, Durand A, Karakachoff M, Bourcier R, Figueres L, **Vince N**, Redon R, Desal H, Gourraud PA, Hourmant M, Limou S. Genetic determinants of intracranial aneurism in autosomal dominant polycystic kidney disease. JOBIM 2019, Nantes, France.
23. Regue H, **Vince N**, Gourraud PA, Laplaud D. Exploring relationship between to neuro-inflammatory diseases. JOBIM 2019, Nantes, France.
24. Ba R, **Vince N**, Geffard E, Malguid D, Lanza M, Gourraud PA, Limou S. Burrowing functional and immunogenetic information through the 1000 Genomes Project with Ferret v3.0. JOBIM 2019, Nantes, France.
25. Guimon R, Haspot F, Poirier N, Blancho G, Limou S, Gourraud PA, **Vince N**. The SIRP gene family: widespread conservation in animals, haplotypic polymorphisms in humans and its therapeutic consequences for monoclonal antibody reactivity. JOBIM 2019, Nantes, France.
26. Garnier A, **Vince N**, Nelson G, Binns-Roemer E, David V, Hoang K, Gourraud PA, Goedert JJ, Winkler C, Limou S. Epigenome-wide association study reveals immunogenetic targets of DNA methylation modification by HIV-1. JOBIM 2019, Nantes, France.
27. Douillard V, **Vince N**, Limou S, Gourraud PA. Navigating the treacherous waters of HLA imputation with the SHLARC (SNP-HLA Reference Consortium). JOBIM 2019, Nantes, France.
28. Durand A, Winkler C, **Vince N**, Douillard V, Geffard E, Ng D, Gourraud PA, Warady B, Furth S, Kopp JB, Kaskel FJ, Limou S. Statistical inference of immunogenetic parameters reveals an HLA allele associated with pediatric Focal Segmental Glomerulosclerosis. JOBIM 2019, Nantes, France.

29. Durand A, Geffard E, Ba R, Limou S, Brouard S, Loupy A, **Vince N**, Gourraud PA. De-centralized database: new challenges to design innovative contextualization algorithms. JOBIM 2019, Nantes, France.
30. Geffard E, Goronflot T, Limou S, **Vince N**, Wargny M, Gourraud PA. Pioneer data-driven methods generating synthetic data: the HLA "avatars" are shifting paradigms in data sharing. JOBIM 2019, Nantes, France.
31. Ianniello A, Geffard E, Douillard V, Limou S, Gourraud PA, **Vince N**. Development of a complete HLA analysis pipeline: HLA-Functional Immunogenomic eXploration (HLA-FIX). JOBIM 2019, Nantes, France.
32. Lecerf-Defer A, Mesnard L, Gourraud PA, Limou S, **Vince N**. Development and validation of an alloscore in kidney transplantation. JOBIM 2019, Nantes, France.
33. Geffard E, Durand A, Ba R, Limou S, Brouard S, Loupy A, **Vince N**, Gourraud PA. De-centralized database: new challenges to design innovative contextualization algorithms. EFI 2019, Lisbon, Portugal.
34. Geffard E, Goronflot T, Limou S, **Vince N**, Wargny M, Gourraud PA. Pioneer data-driven methods generating synthetic data: the HLA "avatars" are shifting paradigms in data sharing. EFI 2019, Lisbon, Portugal.
35. Durand A, Winkler C, **Vince N**, Douillard V, Geffard E, Ng D, Gourraud PA, Warady B, Furth S, Kopp JB, Kaskel FJ, Limou S. Statistical inference of immunogenetic parameters reveals an HLA allele associated with pediatric Focal Segmental Glomerulosclerosis. EFI 2019, Lisbon, Portugal.
36. **Vince N**, Douillard V, Geffard E, Limou S, Gourraud PA. SNP-HLA Reference Consortium: HLA and SNP data sharing for promoting HLA-centric analyses in genomics. DHU2020 2018, Nantes, France.
37. Geffard E, Scherdel P, Limou S, Brouard S, Tissot A, Magnan A, Giral M, Blancho G, **Vince N**, Gourraud PA. A precision medicine application: personalized contextualization of patients after solid organ transplantation. DHU2020 2018, Nantes, France.
38. **Vince N**, Douillard V, Geffard E, Limou S, Gourraud PA. SNP-HLA Reference Consortium: HLA and SNP data sharing for promoting HLA-centric analyses in genomics. EFI 2018, Venise, Italy.
39. Geffard E, Scherdel P, Limou S, Brouard S, Tissot A, COLT investigators, DIVAT investigators, Magnan A, Giral M, Blancho G, **Vince N**, Gourraud PA. A precision medicine application: personalized contextualization of patients after solid organ transplantation. EFI 2018, Venise, Italy.

40. Geffard E, Walencik A, Limou S, Scherdel P, Tissot A, DIVAT investigators, COLT investigators, Blancho G, Giral M, Magnan A, Cesbron A, **Vince N**, Gourraud PA. Application of Easy-HLA to 2,260 solid organ transplant donor-recipient pairs from 2 cohorts: statistically upgraded HLA typing for research use. EFI 2018, Venise, Italy.
41. Walencik A, Geffard E, Limou S, Cesbron A, **Vince N**, Gourraud PA. EasyMatch-R: the validation of a web application for donor query in Hematopoietic Stem Cell Transplantation (HSCT). EFI 2018, Venise, Italy.
42. **Vince N**, Geffard E, Douillard V, Limou S, Gourraud PA. Harnessing the power of functional immunogenomics parameters to discover new associations with diseases. Labex IGO meeting 2018.
43. Geffard E, Walencik A, Limou S, Cesbron A, **Vince N**, Gourraud PA. EasyMatch-R: a web application to facilitate donor query in Hematopoietic Stem Cell Transplantation (HSCT). Labex IGO meeting 2018, Nantes, France.
44. Garnier A, Hoang K, Nelson G, **Vince N**, Gourraud PA, Winkler C, Limou S. Epigenome-wide association study (EWAS) with HIV infection status. Labex IGO meeting 2018, Nantes, France.
45. Ba R, Taverner A, **Vince N**, Gourraud PA, Winkler C, Servieres M, Limou S. Ferret, a user-friendly Java tool to extract data from the 1000 Genomes Project. Labex IGO meeting 2018, Nantes, France.
46. **Vince N**, Limou S, Daya M, Rafaels N, Hollenbach J, Lizee A, Geffard E, Walencik A, Pino-Yanes M, Salzberg S, Kim D, Watson H, Lange L, Wilson J, Beaty T, Taub M, Ruczinski I, Mathias R, CAAPA, Barnes KC, Torgerson D, Gourraud PA. HLA-DRB1*09:01 is associated with a severity outcome of asthma in the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). Assises de génétique humaine et médicale 2018, Nantes, France.
47. Geffard E, Walencik A, Limou S, Garnier F, Cesbron A, **Vince N**, Gourraud PA. Easy-HLA : logiciel en ligne d'haplotypage HLA accessible à tous délivrant un panel complet destiné aux analyses génétiques. Assises de génétique humaine et médicale 2018.
48. Taverner A, Belkacem O, Lam E, **Vince N**, Gourraud PA, Winkler C, Limou S. Ferret, a user-friendly Java tool to extract data from the 1000 Genomes Project. Assises de génétique humaine et médicale 2018, Nantes, France.
49. **Vince N**, Walencik A, Geffard E, Limou S, Cesbron A, Gourraud PA. SNP-HLA Reference Consortium: HLA and SNP data sharing for promoting HLA centric analyses in genomics. ASHI 2017, San Francisco, CA, USA.

50. **Vince N**, Daya M, Hollenbach J, Lizée A, CAAPA, Barnes K, Torgerson D, Gourraud PA. HLA component in the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA). ASHI 2017, San Francisco, CA, USA.
51. Hervé C, **Vince N**, Brouard S, Giral M, Blancho G, Gourraud PA. The Kidney transplantation application (KiTapp): a visualization and contextualization tool in a kidney graft patients' cohort. ASHI 2017, San Francisco, CA, USA.
52. Walencik A, **Vince N**, Garnier F, Cesbron A, Gourraud PA. Easy-HLA: a suite of software using haplotypes to handle HLA typing data. ASHI 2017, San Francisco, CA, USA.
53. **Vince N**, Limou S, Walencik A, Cesbron A, Geffard E, Gourraud PA. SNP-HLA Reference Consortium: HLA and SNP data sharing for promoting HLA centric analyses in genomics. EFI 2017, Mannheim, Germany.
54. Limou S, Taverner A, **Vince N**, Gourraud PA, Winkler C. Ferret, a User-Friendly Tool to Extract Data from the 1000 Genomes Project. EFI 2017, Mannheim, Germany.
55. Hervé C, **Vince N**, Brouard S, Giral M, Limou S, Blancho G, Gourraud PA. The Kidney transplantation application (KiTapp): a visualization and contextualization tool in a kidney graft patients' cohort. EFI 2017, Mannheim, Germany.
56. **Vince N**, Li H, Anderson S, Carrington M. Deciphering the control of HLA-C expression using the 1000 genomes dataset. ASHG 2015, Baltimore, MD, USA.
57. **Vince N**, Bashirova AA, Nelson G, Carrington M. Revealing the detailed MHC implication in seven common diseases from the WTCCC by HLA imputation. ASHG 2014, San Diego, CA, USA.
58. **Vince N**, Bashirova AA, Apps R, Mochalova Y, Yu XG, Carrington M. Diversity of the human LILRB3/A6 locus encoding a myeloid inhibitory and activating receptor pair. ASHG 2013, Boston, MA, USA.
59. Boutboul D, **Vince N**, Just N, Oksenhendler E, Malphettes M, Fieschi C. Extending spectrum of phenotype in 2 new patients with CD19 deficiency. ESID 2010, Istanbul, Turkey.
60. **Vince N**, Boutboul D, Just N, Oksenhendler E, Malphettes M, Fieschi C. CD19 deficiency in 2 patients with CVID [French]. ICB 2009, Paris, France.
61. **Vince N**, Malphettes M, Guignet A, Rabian C, Oksenhendler E, Fieschi C. TACI mutations in DEFI, a french update. ESID 2008, 's-Hertogenbosch, Netherland.
62. **Vince N**, Boutboul D, Just N, Oksenhendler E, Malphettes M, Fieschi C. Complete CD19 Deficiency in 2 patients previously diagnosed as CVID. ESID 2008, 's-Hertogenbosch, Netherland.

63. Malphettes M, **Vince N**, Theodorou I, Bories JC, Oksenhendler E, Fieschi C. TNFRSF13B mutation screening in the French national cohort “DEFI”. ESID 2006, Budapest, Hungary.

IV. Previous work and project

A. Previous work

1. PhD work

EA3963, Université Paris Diderot, Paris, France, 2006-2010

Director: Claire Fieschi

Genetic and immunological bases of the common variable immunodeficiency

Publications: 10 publications including 2 as first author [19,20]

International context. Common variable immunodeficiency (CVID) is a rare disease (Prevalence ~1/25,000) whose pathophysiology is still largely unknown. This syndrome is characterized by hypogammaglobulinemia. Approximately 20% of family cases are reported in CVID cohorts [21], suggesting a potential genetic origin; however, a genetic etiology was identified in only 10% of cases [22]. The objective of my PhD was to improve the **molecular understanding of CVID**, by correlating the **clinical and biological data** of the DEFI (DEFicit Immunitaire) registry with the **genetic and functional data** obtained from each patient.

Methodologies applied. The **DEFI registry** identifies adults with primary **hypogammaglobulinemia** at the national level [21]. Screening of genes of interest was done by **sequencing** patients' DNA using the Sanger method. I used classical immunology techniques (Western blot, cytometry, immunofluorescence...) to demonstrate the **functional effects** of these genes. This work was funded by the LNCC (N. Vince) and the ANR (C. Fieschi).

CD19. We selected 18 DEFI patients lacking CD19 expression (classical B-cell surface marker) and sequenced their *CD19* gene. I could then identify 2 unrelated patients carrying novel ***CD19* mutations** [19]. I demonstrated that the **absence of CD19 expression** (Fig. 2) was **responsible** for both patients' clinical and biological **phenotypes**, even if the clinical presentation was slightly different for one of them with kidney injuries. This work has improved the understanding of this immune deficiency by placing CD19 as a major player in the effective response to infections with B-cells involvement. The discovery of new *CD19* mutations clarified the phenotypic spectrum of CD19 deficiency (residual Ig

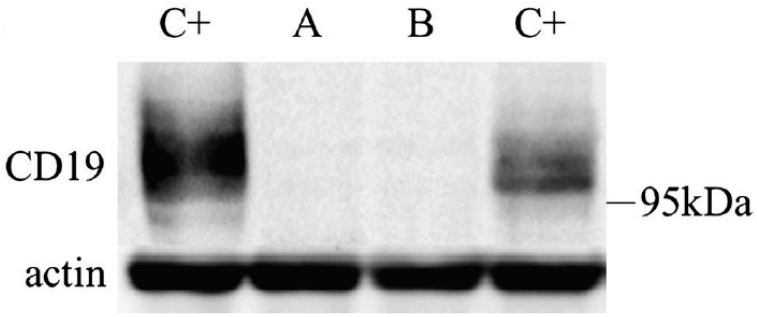

Figure 2: Both A and B patients carrying CD19 mutations show a lack of gene expression. Western blot made from B-EBV cell line protein extract. C+: control. A: patient A. B: patient B.

production, inconsistence in memory B-cells deficiency) [19], and defined a new immune deficiency [23].

Genetic screening of male patients. We screened more than 150 male patients with hypogammaglobulinemia for mutations in three genes involved in pediatric X-linked primary immune deficiency: *CD40LG*, *SH2D1A* and *BTK* (Fig. 3). We described 5 patients labeled CVID and 1 IgG subclass deficiency carrying a *BTK* mutation; as well as one CVID patient carrying a *SH2D1A* mutation [20]. The DEFI cohort is one of the largest worldwide in terms of size and homogeneity. The discovery of mutations in *BTK* and *SH2D1A* calls for the inclusion of these genes in newly developed next-generation sequencing (NGS) panels to screen patients with immune deficiency. Genetic characterization of patients with primary hypogammaglobulinemia remains essential, especially with the help of a large cohort, in order to provide the most appropriate medical treatment and informed genetic counseling.

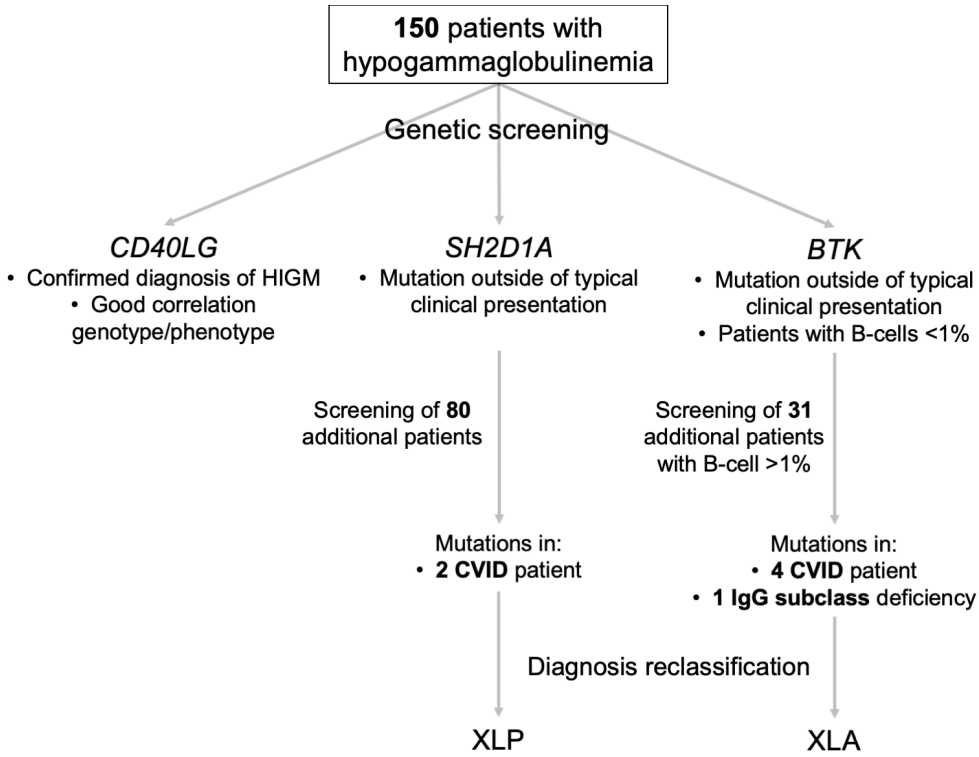

Figure 3: To identify disease-causing mutations and their potential to support a diagnosis, we explored a large cohort of French hypogammaglobulinemic patients. We therefore systematically screened this population for mutations in the SH2D1A and BTK genes, as well as for mutations in CD40LG.

These genetic studies are the precursors of genetically-informed **precision medicine**, the precise characterization of the pathology allowing a **personalized patient's management**.

2. NIH postdoctoral work

NCI, NIH, Frederick MD and Ragon Institute of MGH, MIT and Harvard, Cambridge MA, USA, 2011-2016

Lab head: Mary Carrington

Functional immunogenomics of complex pathologies

Publications: 7 publications including 2 as first author [24,25]

As my PhD projects focused on monogenic approaches about CVID (immunogenetics), it seemed important to enlarge my knowledge and to learn how to use the modern tools available in **immunogenomics**. I joined Mary Carrington's laboratory to study **complex pathologies**, such as HIV infection, from a broader perspective. There, I oriented my research towards **bioinformatic** tools to study hypervariable regions of the genome (*HLA*, *KIR*).

International context. The ***HLA*** (Human Leukocyte Antigen) system plays a central role in the immune response for presentation of self (from the individual) and non-self (foreign to the individual) peptides to T-cells, and as ligands for Killer-cell immunoglobulin-like receptors (**KIR**s). KIRs are NK-cell (natural killer) receptors. Studies about **progression from HIV infection to AIDS** showed several associations within the *HLA* region, however, no convincing association with HIV acquisition involving *HLA* or *KIR* were found [26]. HLA-C surface expression is differentially regulated according to alleles and this has a direct impact on HIV disease progression and Crohn's disease [27]. The control of this expression is still not fully elucidated. My investigations during this post-doctoral fellowship were twofold: 1) What are the ***HLA*** **alleles** or **KIR-HLA combinations** involved in **HIV infection**? 2) What are the **mechanisms controlling HLA-C expression**?

HLA and *KIR* polymorphisms in HIV infection. Access to large cohorts where the viral exposure probability is near 100% is mandatory to study HIV infection. I had access to a unique dataset of 325 hemophilia A patients transfused with contaminated clotting factor concentrates between 1979 and 1984 [25]. I performed the ***HLA*** **and** ***KIR*** **typing** of these patients. The *HLA* alleles frequencies are different across populations even within individuals of European ancestry, it is therefore essential to correct for these differences (**population stratification**, Fig. 4). Genomic correction techniques are very effective in addressing these differences [28]. This study revealed the complexity of HIV infection. While the deletion of the HIV coreceptor CCR5 explains resistance to infection for some individuals, persons without CCR5 deletion still lack explanation for their resistance to infection. I showed that the **HLA-KIR system does not seem to have any influence on the potential for HIV infection** through bloodstream. This study of

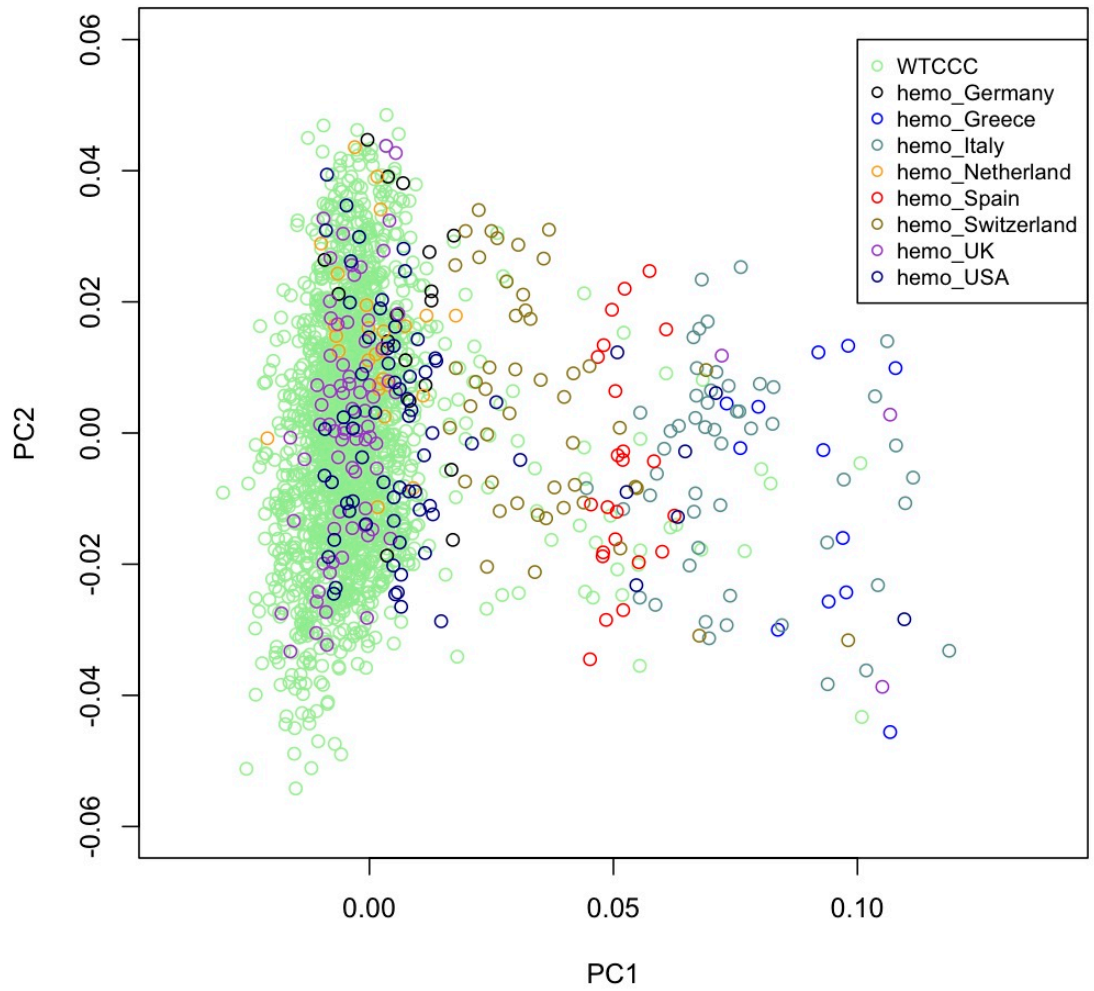

Figure 4: Principal component analysis (PCA) comparing HIV infected hemophilia A patients to WTCCC controls. Hemophilia A patients are from different European genetic background while controls are all from UK. PC1: first principal component. PC2: second principal component.

highly exposed seronegative individuals revealed that adaptive immune responses or NK-cells do not seem to be responsible for resistance to infection. The proportion of individuals resistant to blood-borne HIV infection being extremely low (<10%), this would tend to predict that if intrinsic immunogenetic factors were involved, they are probably **rare or even private variants**.

Functional genomics of HLA-C expression. To identify single nucleotide **polymorphisms** (SNPs) involved in the **differential expression of HLA-C**, I have developed a novel methodology based on statistical inference within a large public dataset. 1) The **mean surface expression of each HLA-C allele** is known and published [24,27]. 2) ***HLA* typing** of individuals from the 1000 Genomes project (1KG) is publicly available as well as **SNPs** from the *HLA* region [29,30]. 3) From these data, I **imputed HLA-C expression**, i.e. statistically predicted HLA-C surface expression for each 1KG individual from the available expression means. 4) I tested the association between the 68,726 SNPs from the *HLA* region and the HLA-C imputed expression (continuous variable) with linear regression to identify **impeQTL**

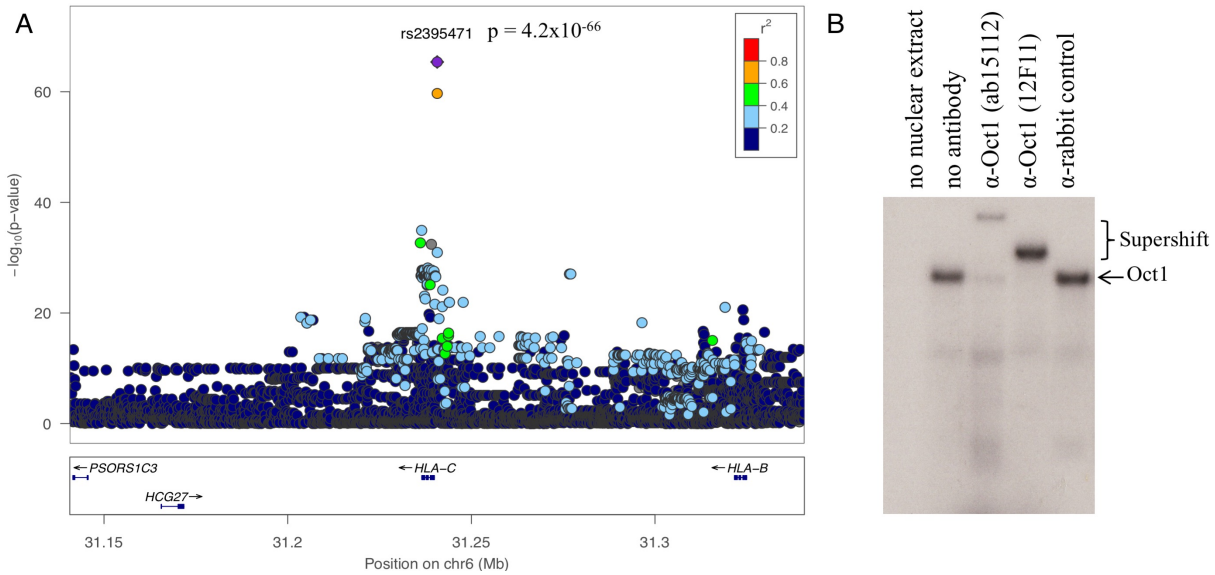

Figure 5: The rs2395471 SNP is an impeQTL (imputed expressed quantitative trait loci) for HLA-C. A: Manhattan plot representing SNPs associations in a 200kb window around HLA-C with the HLA-C imputed expression. B: Electrophoretic mobility shift assay showed the Oct-1 transcription factor binding to the oligonucleotide containing rs2395471.

(imputed expressed quantitative trait loci, Fig. 5A). 5) I have proven by a **functional study** that the most strongly associated impeQTL with HLA-C imputed expression (rs2395471, $p=4.2x10^{-66}$) modifies the Oct-1 (*POU2F1*) transcription factor binding site (Fig. 5B): the rs2395471_A allele showed a better Oct-1 binding, therefore a higher expression, than the G allele. Overall, I have developed a **new methodology** and described the **first impeQTL**, a SNP associated with the predicted (imputed) differential expression of a gene. I confirmed this association using measured surface expression. The use of predicted expression to identify involved genetic factors is unprecedented. This allows to better characterize functionally relevant genetic polymorphisms and helps understand HLA effect on diseases.

3. Nantes research work

Team 3 iTHINK, CR2TI, UMR1064, Inserm, Nantes Université, Nantes, France, 2016-now

Team leaders: Sophie Limou and Gilles Blancho

Immunogenomics of inflammatory diseases

Publications: 25 publications including 3 first author [3,4,31] and 11 last author [1,6,8,9,11,32–37]

The synergy of my research projects at the **crossroads** of immunology, genetics and bioinformatics put me in position to face a new challenge by joining a new team. This Inserm ATIP-Avenir (Pierre-Antoine Gourraud) team at the CRTI was built to bring **bioinformatics, genomics, HLA and large-scale** expertise to a recognized Inserm lab in the field of immunology. There, I use **bioinformatics and statistics** to explore relevant immunogenetic determinants in inflammatory diseases on large population datasets. Since January 2022, we merged 2 former CRTI teams into the CR2TI team 3 iTHINK (Integrative transplantation, HLA, Immunology and genomics of kidney injury) to pursue our work on bioinformatics and genomics and integrate more functional immunology.

International context. In recent years, genetic associations have expanded at a high rate thanks to GWASs (genome-wide association studies), particularly in the major histocompatibility complex (*MHC*) region, host of the *HLA*, origin of more than 25% of all associations listed in the GWAS catalog (ebi.ac.uk/gwas, Fig. 6) [38,39].

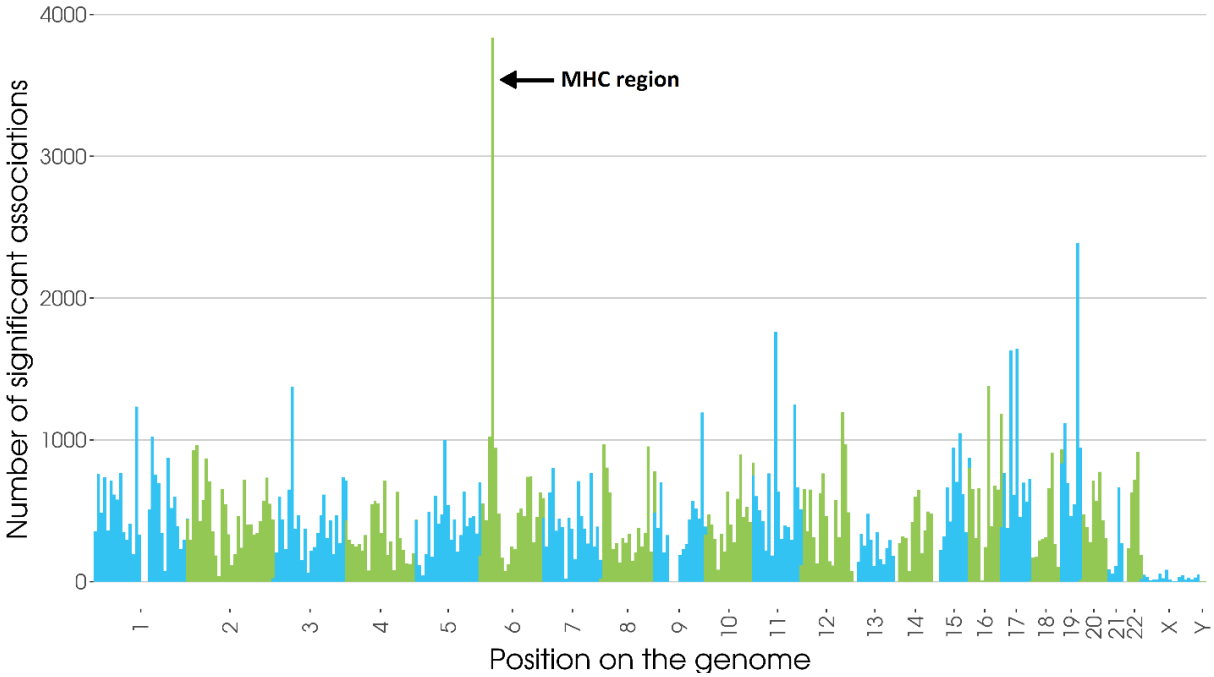

Figure 6: Number of significant SNP associations from the GWAS catalog within the whole genome divided in 350 bins (Ref. 7). 4,080 associations fall within the extended MHC region (chromosome 6, 25Mb-34Mb). Updated on January 14th, 2021. Figure was built with a total of 155,456 associations with p<5x10-8.

However, the information of a SNP is very limited, especially in the *HLA* where linkage disequilibrium is high. Indeed, the information of a SNP associated with a pathology is rather the marker of a region of the genome; it is necessary to go beyond this simple association to improve our understanding of functional mechanisms and eventually develop therapeutic targets. *HLA* typing techniques are expensive and require a specialized laboratory infrastructure. With the help of statistical inference, it is now possible to **impute *HLA* alleles from GWAS genotyped SNPs** [40], requiring adequate **reference panels** [41], and also **impute *HLA* haplotypes from genotypes**. Haplotypes are a combination of genetic variants on one chromosome, they can be SNP haplotype, gene haplotype or a combination of different genetic variants (SNP, indels, substitution) haplotype (such as *HLA* alleles). ***HLA* haplotypes** are a combination of *HLA* alleles typically considering the 5 main studied genes: *A~B~C~DRB1~DQB1*. These statistical methods allow the generation of **new data** and the discovery of **new genetic associations**. We developed a **web suite of tools** dedicated to study the *HLA*: Easy-HLA (hla.univ-nantes.fr) [11]. We launched the first international network dedicated to HLA imputation: the SHLARC (SNP-HLA Reference Consortium) [3]. In addition, I studied 2 immune-mediated pathologies: 1) *HLA* haplotypic associations with MS [32], 2) *HLA* allelic associations with asthma in a large dataset of African origin individuals [4].

Easy-HLA, HLA-oriented online tools. [1,11] **Easy-HLA** (hla.univ-nantes.fr, >300 users so far) implements a computerized and statistical method of *HLA* haplotype inference based on a large dataset of more than 600,000 haplotype frequencies provided by the National Marrow Donor Program (USA, 6.59 million donors). Easy-HLA includes several tools (Fig. 7): 1) **EasyMatch-R** effectively quantifies probabilities to find Hematopoietic Stem Cell

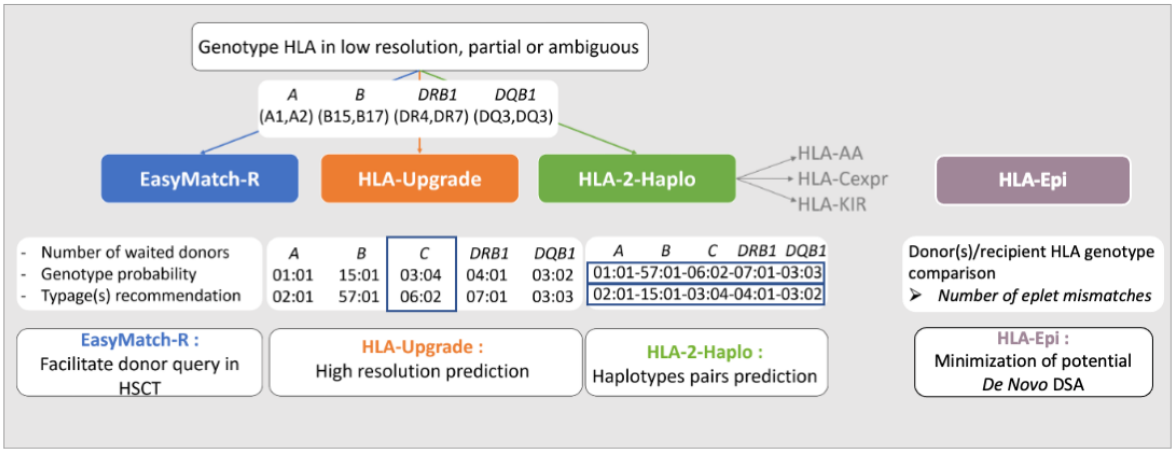

Figure 7: Easy-HLA includes several tools: (1) EasyMatch-R (in blue) is designed to help donor search decision-making in the context of HSCT; (2) HLA-Upgrade (in orange) provides high resolution HLA typing inference; (3) HLA-2-Haplo (in green) allows imputation of haplotype pairs and provides additional functional annotations (amino acids, HLA-C expression or epitopes); (4) HLA-Epi (in purple) gives the number of epitope differences between a patient (or several) and a potential donor (or several).

Transplantation (HSCT) candidates (Walencik et al., submitted); 2) **HLA-Upgrade** infers high resolution *HLA* typing [11]; 3) **HLA-2-Haplo** imputes haplotype pairs [11]; 4) **HLA-Epi** explores donor to recipient matching at the epitopic level to facilitate epitopes contribution study during transplantation [1].

HLA-Upgrade has a 92% predictive accuracy when upgrading low-resolution genotypes in a population of European origin. We observed a 96% call rate and 76% accuracy when predicting haplotype pairs with HLA-2-Haplo. Our tool is a perfect example of how **computer and statistical modelling can relay and upgrade high-value experimental data** with direct research usefulness.

HLA-Epi [1] provides the number of HLA epitopes mismatches between donor and recipient in transplantation. HLA-Epi is relying on the largest and more up-to-date epitope database (Epregistry.com.br [42,43]) comprising correspondences of 560 epitopes carried by major *HLA* genes (representing >99% of the total observed allele frequency). The freely accessible web tool HLA-Epi calculates an epitopic mismatch load between different sets of potential recipient-donor pairs at different resolution levels. We have characterized the epitopic mismatches distribution in a cohort of more than 10,000 kidney transplanted pairs from European ancestry (Fig. 8), which showed low number of epitopic mismatches, 57 incompatibilities on average; while a simulated random sampling genotype pairs of 5,000 hypothetical donor-recipient showed 70 to 74 incompatibilities on average (3 populations:

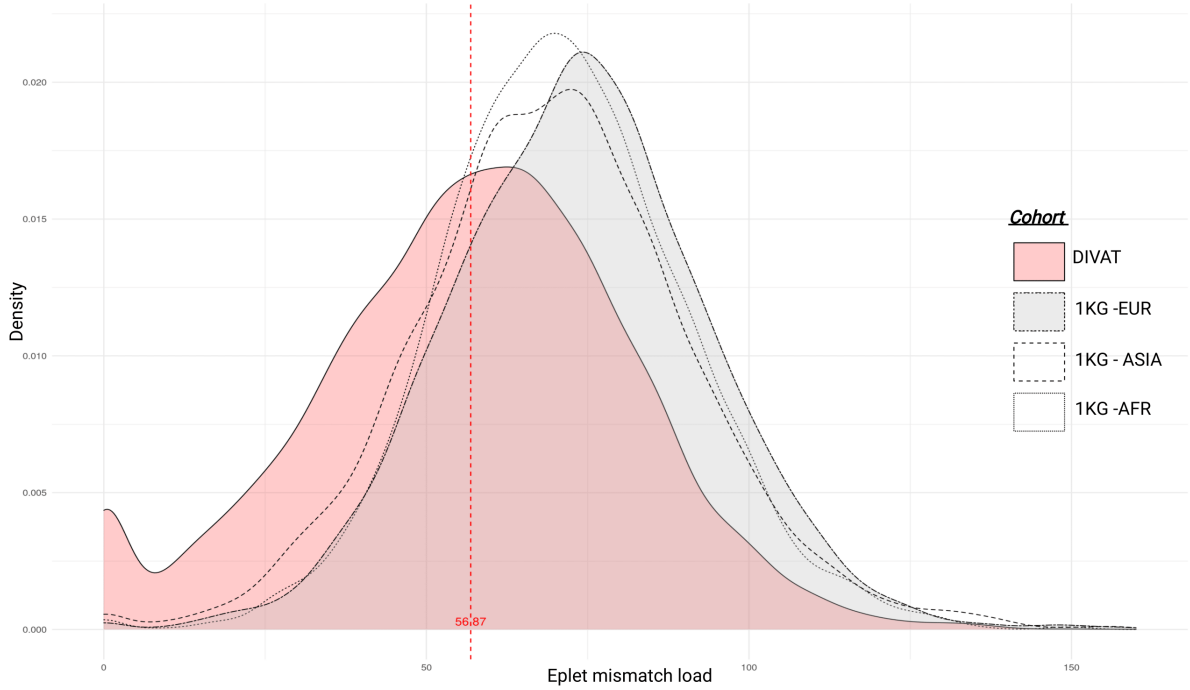

Figure 8: Eplet mismatch load distribution (Ref. 1). In grey, 1KG simulations mismatch load distribution from 3 tested populations (European, Asian, and African), 5,000 donor-recipient pairs were simulated and mismatch load calculation was performed based on their high-definition HLA genotypes. In red, real eplet mismatch load distribution (from 10,667 DIVAT cohort kidney transplanted individuals, red curve).

European, Asian and African). HLA-Epi allows the exploration of epitope pairing matching to better understand epitopes contribution to immune responses regulation, particularly during transplantation. This free and ready-to-use bioinformatics tool not only addresses limitations of other related tools, such as the heavy and slow spreadsheet-based HLAMatchmaker interface [42], but also offers a cost-efficient and reproducible strategy to analyze HLA epitopes as an alternative to *HLA* allele compatibility. In the future, this could improve sensitization prevention for allograft allocation decisions and reduce the risk of alloreactivity.

SHLARC. [3] HLA imputation methods were developed to circumvent the effort and high costs of HLA typing; however, no unified effort has yet been undertaken to **share large and diverse imputation models**, or to improve methods. By training the HIBAG software [40] on an African ancestry SNP+HLA dataset to create reference panels, we highlighted the importance of having more data, notably the importance of the number of individuals in reference panels with a fourfold increase in accuracy (from 10 to 100 individuals) [3]. Results showed improved accuracy with our dataset compared to the African American models available in HIBAG, emphasizing the **need for precise population-matching**. This work [3] is the first step of an international endeavor to gather data, **enhance HLA imputation** quality and broaden access to highly accurate imputation models for the immunogenomics community (Fig. 9).

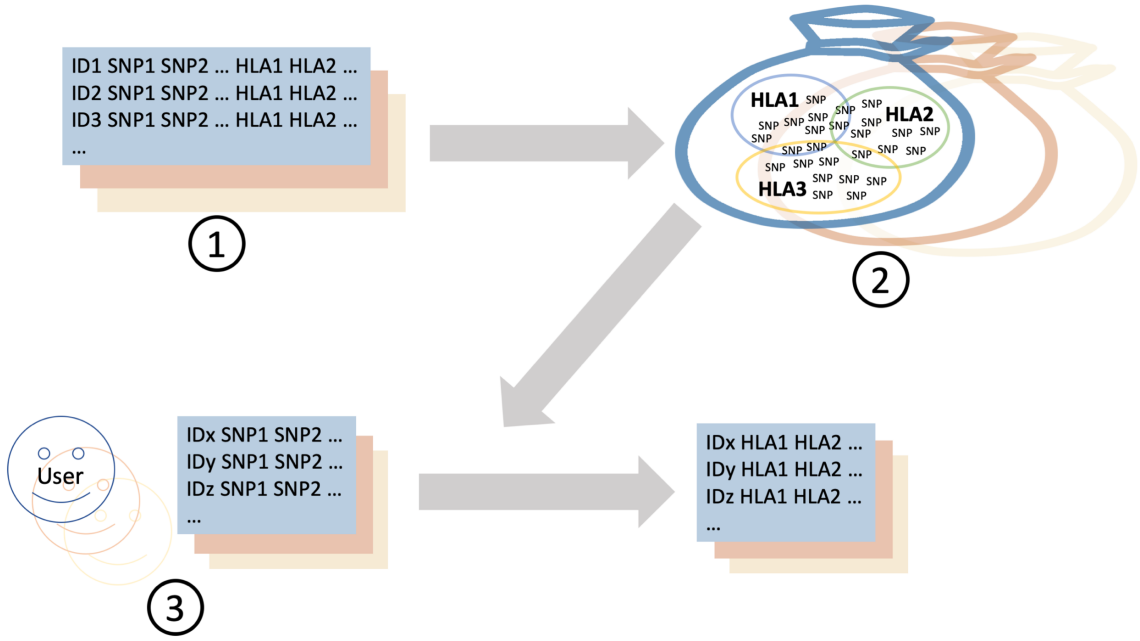

Figure 9: The SNP-HLA Reference Consortium (SHLARC) design (Ref. 3). Aim 1. Increase the amount of SNP+HLA data available both in terms of quantity and diversity. Aim 2. Optimize SNP-HLA imputation methods. Aim 3. The SHLARC website will allow users from the scientific community to benefit from the data and knowledge accumulated by the consortium on SNP-to-HLA allele imputation.

MS (multiple sclerosis): *HLA* haplotypes association. [6,32,35,36] Few *HLA* haplotypes are highly conserved during evolution and are transmitted in blocks [32], suggesting a **selective advantage** for these haplotypes when the immune system is under selection pressure. The

exploration of *HLA* haplotypes allows to consider the evolutionary dimension and can also reflect a combined effect of their alleles, which can reveal new associations and contribute to a better understanding of diseases. A good example of this paradigm shift is the study of MS. The strongest genetic association identified with MS is the *HLA-DRB1*15:01* allele (OR~3). In 3 studies [6,32,36], we wanted to assess the MS risk effect of HLA haplotypes and their interaction in European and African American ancestry individuals.

We phased (haplotyped) extended haplotypes (*HLA-A*, *HLA-B*, *HLA-C*, *HLA-DRB1*, *HLA-DQB1* and SNP haplotypes) from 29,636 individuals of European ancestry [32] and 2,460 African American individuals [6] and obtained 10,078 and 2,744 unique haplotypes, respectively. This difference only reflects the considered number of individuals from each ancestry. Khankhanian et al. previously described 174 SNP haplotypes (*a1* to *a174*), combination of 11 SNPs from *HLA-DRB1* regulatory regions [44]. We then compared these haplotype frequencies between MS patients and controls: European, 11,144 vs. 18,492; African American, 1,305 vs 1,155. We showed that some haplotypes carrying *HLA-DRB1*15:01* were not associated with MS. On the contrary, all haplotypes with *a1*, even without *HLA-DRB1*15:01*, were associated with MS, suggesting that *HLA-DRB1*15:01* may only be a marker of SNPs influencing **HLA class II genes expression regulation.** Differences in expression could lead to higher risk of developing MS by increasing inflammation. The haplotypic approach to MS susceptibility has revealed the need for large patient cohorts and specific bioinformatic tools adapted to low frequencies. The study of *HLA* haplotypes allows us to consider the evolutionary complexity of HLA and to dichotomize (alleles vs. regulation) functional immunological mechanisms.

Association of *HLA-DRB1*09:01* with total serum IgE levels among African ancestry individuals with asthma. [4] So far, asthma GWASs have focused on populations of European and East Asian ancestry. Furthermore, these GWASs only provided SNP associations, which are poorly informative biologically, particularly in the *HLA* region. To overcome these, I conducted the **first immunogenomic study of a large dataset of subjects with asthma from African ancestry** to explore the role of *HLA* in this disease. Through a collaboration with **CAAPA** (Consortium on Asthma among African-ancestry Populations in the Americas), we obtained **genome-wide genotyping data** for 10 cohorts with a total of 2,608 asthmatics (including 1,182 with total serum IgE [tIgE] level data) and 4,056 controls [45]. I **imputed the *HLA*** alleles from SNPs with a customized reference panel, as this is key to obtain high quality imputation. Indeed, I was able to create a **CAAPA-specific SNP-HLA reference panel** from a subgroup of 917 individuals with whole-genome sequencing data (providing both SNPs and *HLA* typing data) to impute the whole dataset.

Case/control association tests did not reveal any *HLA* allele significantly associated with asthma. At the opposite, I identified an **association between *HLA-DRB1*09:01* allele and tIgE levels in CAAPA subjects with asthma** ($p=8.5x10^{-4}$, weighted effect size 0.51 [0.15-0.87]; Fig. 10). The tIgE level in asthmatics is related to the severity of the phenotype, hence individuals with the *HLA-DRB1*09:01* allele may present a more severe phenotype. This allele appears to have an important role in inflammation and was previously associated with an increased risk of developing rheumatoid arthritis or lupus in East Asian populations [46]. The allele could be responsible of an increased activation of CD4+ T-cells compared to other alleles, by presentation of specific dust mite peptides, which would trigger an amplified inflammation environment and result in a higher IgE production by B-cells. Further experiments are needed to confirm the association and define the functional relationship between *HLA-DRB1*09:01* allele and asthma severity. By studying this exceptional dataset (the largest cohort of genotyped African ancestry asthma patients), I could both **create a unique reference panel for *HLA* imputation**, and **identify an *HLA* allele significantly associated with tIgE level in asthmatics**. The **generation of new data by innovative bioinformatic and biostatistic methods** therefore makes it possible to reveal new immunogenetic associations and to better understand the functional mechanisms of immunological pathologies.

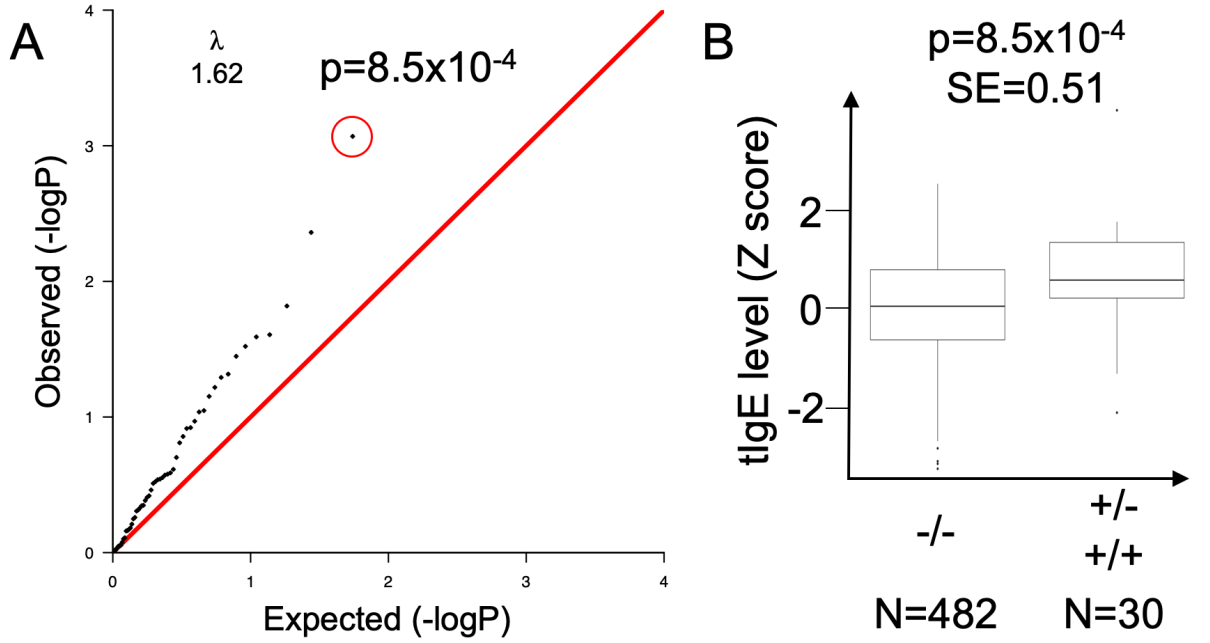

*Figure 10: HLA-DRB1*09:01 is associated with tIgE level in CAAPA subjects with asthma. A: QQplot of p-values obtained by comparing HLA alleles and tIgE level (continuous variable) with a linear regression model. B: tIgE levels are higher in patients carrying HLA-DRB1*09:01 allele. SE: size effect.*

B. Research project

Immunogenetic study of human immune-related diseases

My project is a continuation of Easy-HLA and SHLARC developments as well as the *HLA* association studies of MS and asthma. It has 2 main objectives: first, develop cutting-edge tools for the community which can be used for investigating multiple HLA-related questions (association genetics, population genetics, transcriptomics…); second, apply these tools to explore 2 specific pathological conditions: **kidney transplantation (KiT) and neuro-inflammatory** diseases, such as MS and Neuromyelitis Optica Syndrom Disorder (NMOSD). *Medical need*. Chronic kidney failure affects **~10% of the world population** and can progressively lead to end-stage kidney disease requiring replacement therapy (dialysis or transplantation). **KiT is the best treatment for end-stage kidney disease.** The one-year survival of kidney transplant is >90% and the graft half-life is >10 years (Fig. 11) [47]. ESKD is

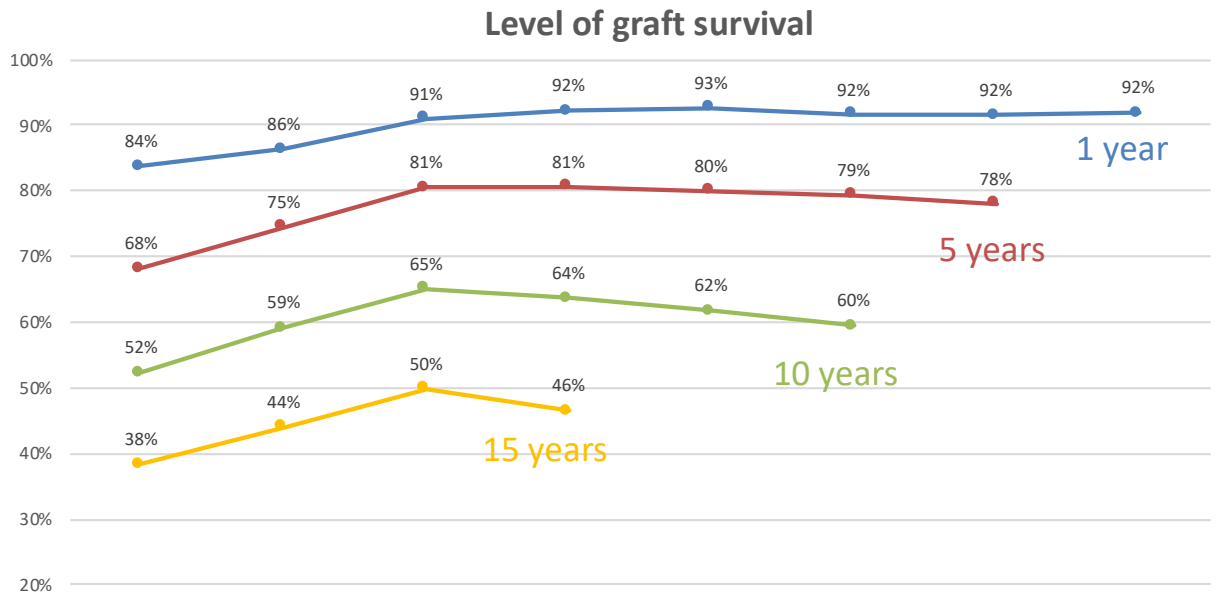

Figure 11: Level of graft survival after kidney transplantation for different time period. The figure depicts level of graft survival post-transplantation after 1 year (blue), 5 years (red), 10 years (green), 15 years (yellow). The level is grouped for 8 time period of 5 or 3 years from 1986 to 2018. These data are from the ABM (agence de la biomédecine), they consider only French patients.

a critical **public health issue** since organs are a rare resource. Overall, the triggers leading to rejection are not fully understood and need to be further studied through innovative big data immune strategies. The MS is a severe neuro-inflammatory disease of the central nervous system and the first cause of neurological disability in young adults **affecting ~1‰ person in France**, and NMOSD is ten times less frequent. While clinical research continues to identify better treatment strategies, immunogenomic studies in cohorts could drive HLA and a **polygenic score**, such as MSGB, to better define and categorize neuro-inflammatory disease

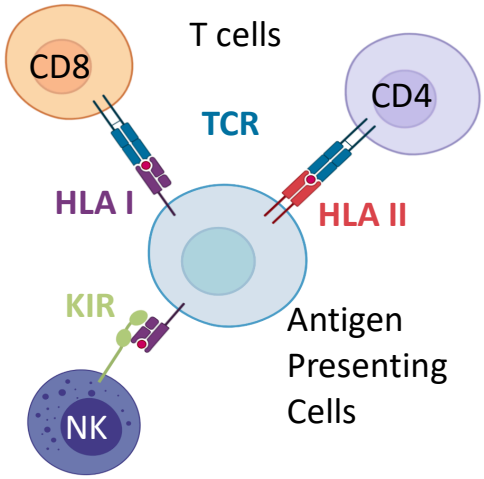

Figure 12: HLA system plays a central role in immunity.

patients in a personalized risk prediction and management. *Pathological challenge*. HLA bears a **central role in the immune system** for presentation of self (from the individual) and non-self (foreign to the individual) peptides to T-cells, and as ligands for **KIRs** expressed on NK-cell surface (Fig. 12). **My goal is to genetically dissect HLA implication in 2 immune-related conditions where HLA is well-known to have a strong impact**. In KiT, **HLA** molecules produce immunodominant antigens for both humoral and cellular alloreactive responses [48]. The number of *HLA* allelic differences between donor and recipient (**mismatches**) is strongly linked to graft survival, hence ***HLA* compatibility** between donor and recipient **is essential** [49,50]. Nevertheless, **mismatches do not explain all rejections** and the intrinsic effect of *HLA* alleles, beyond these mismatches, remains elusive. **I hypothesize that HLA** (alleles and/or expression levels) **plays a major role in post-kidney transplant complications and graft dysfunction, and not only HLA mismatches**. In neuroinflammatory disease, the **strongest genetic association** of MS consistently maps to the ***HLA-DRB1*15:01*** allele (odds ratio ~3) in the class II *HLA* region of the *MHC*. Beyond this signal, 200 genetic risk factors show a small susceptibility effect on MS (with odds ratios from 1.1 to 1.2) [51]. These risk factors can be summarized into a **polygenic score, which can be used** to predict disease occurrence or severity [52]. The MS genetic burden (**MSGB**) [53] was implemented few years ago and will gain in precision with the new genetic knowledge available. Nevertheless, ***HLA* stays the main contributor in MS** risks and I am convinced that the development of specific in silico HLA e-tools will allow to **dissect its implication in MS development**. *Technological challenge*. Genetic data inside *HLA* genomic regions are available through 2 forms: 1) **SNPs** represent a difference of one nucleotide at a specific position with an allelic frequency >1% in a population; 2) ***HLA* alleles** are determined with typing technologies (from serology to NGS). A great paradox in *HLA* genetic studies is the imbalance between the extensive GWAS reported associations originating from this locus [38] (see Fig. 6) and the fragmented exploitation of typed *HLA* alleles data available in some cohorts (e.g. transplanted patients). In fact, **the question of HLA functional consequences on disease etiology**, even if central**, is still under-exploited because of its analysis and interpretation complexity**. **Easy-to-use tools are necessary to solve this issue**.

1. HLA data transformation as a bioinformatic stepping stone for immunogenomic studies

a) The SNP-HLA Reference Consortium (SHLARC)

To overcome the paradox of numerous *MHC* SNP associations in GWASs but few HLA allelic associations, ***HLA* allele imputation** from SNP data offers a **simple and efficient statistical alternative** to current *HLA* typing, simultaneously reducing costs and delays (Fig. 13) [9,41]. The goal is to better understand HLA function by exploring *HLA* alleles associations. In addition, as in SNP genomic studies [54], *HLA* imputation is more effective for populations from European ancestry because of their data over-representation, pointing on an essential **need to include under-represented populations to improve their *HLA* imputation**.

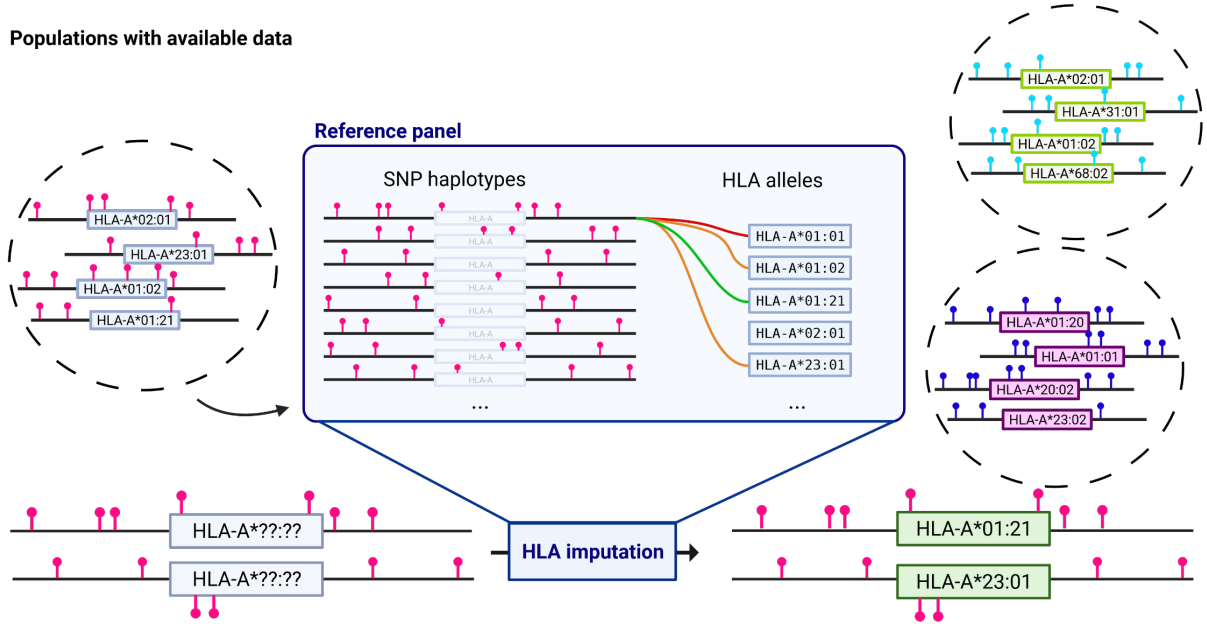

Figure 13: HLA imputation from GWAS data (Ref. 8). Reference panels are created from individuals with known SNP and HLA data. Depending on the method, an algorithm will deduce the probability of a specific HLA allele in the population given a SNP haplotype. These new found links are stored for that reference panel and applied to new SNP data to infer HLA genotypes. HLA-A is given as an example with a truncated list of alleles; other MHC genes are imputed using the same method. Different populations are represented in different circles and imply different allele frequencies. Pinpoints represent SNPs and are only indicative. HLA imputation results are highly dependent on the population chosen for the reference panel.

The ambition of the SNP-HLA Reference Consortium (**SHLARC**) is **to improve and facilitate the statistical inference of *HLA* alleles to discover new genetic associations**. There are currently **no gold standards for SNP-HLA imputation**. We will **improve** the **accuracy** of *HLA* alleles imputation using existing machine learning software: the HIBAG R package [40], based on attribute bagging, and DEEP*HLA [55], based on deep learning. The **essential prerequisite** to achieve this goal is large, diverse and well-defined genomic **data**. By bringing together scientists from around the world, we will increase the amount of **SNP+*HLA* data** available both in terms of **quantity** and **diversity**. I obtained a **400k€ funding** to create and

animate the SHLARC international network; **two PhD students** will be instrumental in the realization of this project (Venceslas Douillard and Nayane Brito). Building on this network, we also plan to create a free and easy-to-use web server where researchers can obtain optimized *HLA* imputation of their data. Furthermore, a key impact of this network is to facilitate **open science** through the free sharing of knowledge and data.

Our paper describing the SHLARC consortium was published in ***Genetic Epidemiology*** in October 2020 [3]. In this report, we also showed that rising data size and matching ancestry can lead to increased accuracy.

SHLARC's expected results are threefold: 1) a reliable and secure ***HLA* imputation facility**, 2) facilitating the creation of *HLA* alleles data from genotyped patients will **reveal new *HLA* associations** in immune-related diseases, 3) an **enhanced and validated statistical imputation tool** could accelerate the access of *HLA* alleles information in health care. SHLARC's 25 partners (Fig. 14) come from various backgrounds and **share complementary expertise**: immunology, bioinformatics, artificial intelligence, population genetics, genomics... The unification of these complementary skills will have multiple impacts on my own research. The **scientific impact** will accelerate all immunogenomic studies in existing Nantes cohorts (kidney transplantation: DIVAT, lung transplantation: COLT, multiple sclerosis: REFGENSEP) to discover new genetic associations, with the ultimate goal to **impact the clinic** (e.g. by optimizing the selection of KiT donors). Finally, by facilitating access to high-confident *HLA* allele imputation, SHLARC may allow multiple immunogenetic studies beyond my own interest on genetic association such as the HLA impact on human evolution (selective pressure and arm race with pathogens) or a better understanding of HLA expression regulation (conditioning strength of immune responses).

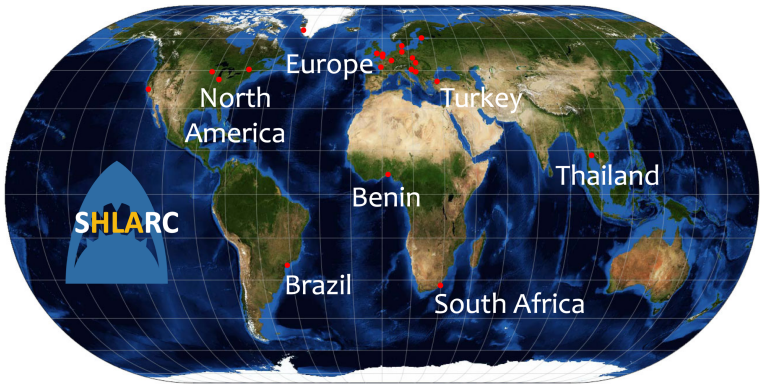

Figure 14: SNP-HLA Reference Consortium (SHLARC) partners.

The success of improving *HLA* imputation is dependent on data access. Reference panels only contain aggregated data, which allows *HLA* alleles statistical inference without the possibility to re-identify individuals, thus **ensuring data anonymity**. This will help to overcome potential difficulties in the diversity of legal frameworks regarding individual data sharing and protection.

I set up and took over the coordination of the SHLARC to collect large and diverse data (crucial for reference panels) and improve methods (e.g. machine learning) with a large international immunogenomic task force. The objective is to reveal new *HLA* association with diseases.

b) Harness HLA capacities to go beyond simple allelic association and better functionally define *HLA* associations discovered in immune-related pathologies

***HLA* studies need specific tools** because of its analysis and interpretation complexity, some tools already exist but are not fully satisfactory. For instance, BIGDAWG [56] is an R package designed to handle *HLA* and *KIR* data association studies but is still lacking high-standard statistical analyses such as regression models with covariates. My goal here is to build on the tools and knowledge we have developed to design new high-quality applications which will convey additional parameters to better define *HLA* associations.

Easy-HLA (hla.univ-nantes.fr) was published in April 2020 [11]. Easy-HLA is available to all as a free and easy-to-use web suite. We are now pursuing Easy-HLA development by adding new functionalities for facilitating clinical care and immunogenetic association studies. The **EasyMatch-R** tool estimates the number of unrelated donors present in a population in regards to a patient's *HLA* genotype within hematopoietic stem cell transplantation context. The goal is to estimate the chances of a patient to find a suitable donor. In a retrospective real day-

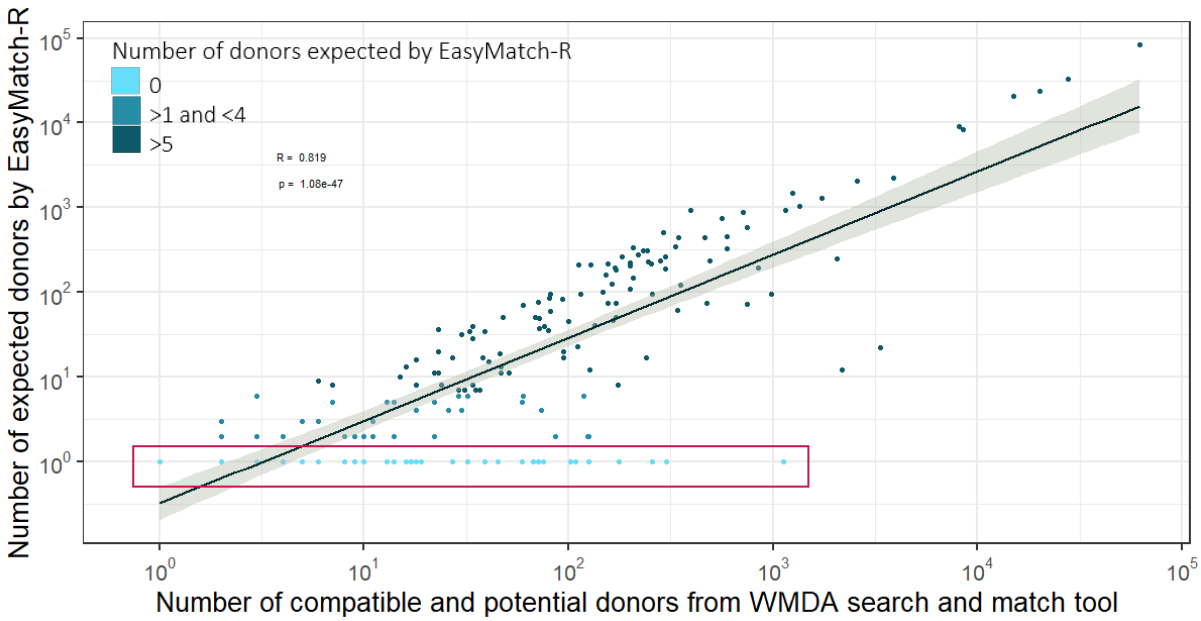

Figure 15: Correlation between the number of expected donors from EasyMatch-R and total number of 10/10 match potential donor from the WMDA search and match tool (n=202). The dot plot represents the correlation between the number potential 10/10 donors found by the WMDA search and match tool and the expected donors given by EasyMatch-R after a log transformation log(x+1) for better visualization. The histogram allows visualizing the dispersion of individuals on the dot plot. The blue scale represents the number of donors expected by EasyMatch-R: 0, n=53; >1 and <4, n=26; >5, n=113. The red frame includes patients for which WMDA search and match tool gives potential donors but EasyMatch-R does not.

to-day dataset from the EFS HLA lab in Nantes of 202 patients in need of a compatible donor, we found a significant positive correlation between the number of expected donors obtained from EasyMatch-R and real data from the World Marrow Donor Association search and match tool (Fig. 15). The World Marrow Donor Association search and match tool is an international collaborative registry of 40 million potential hematopoietic stem cell donor from 53 countries (wmda.info). We retrospectively compared the impact of the algorithm recommendation with the number of additional typings originally requested by the laboratory. Whatever the setting of EasyMatch-R, our algorithm reduced the number of extended *HLA* locus genotyping requested by the recipient center (n=696) to refine or complete donor's *HLA* typing (-279, -400 and even -513 requested loci for a 2%, 5% and 10% post-probability threshold, respectively). EasyMatch-R aims to facilitate compatible donor search by delivering statistical yet actionable information supporting early adoption of alternative options when suitable potential donors may lack. It simultaneously improves the efficiency and diminishes the additional *HLA* typing related costs.

To bring together essential functional immunogenomics parameters from Easy-HLA and high-standard statistical analysis, we are developing a **complete HLA analysis pipeline R package**: HLA-Functional Immunogenomic eXploration (**HLA*fix***). HLA*fix* was released as an alpha test R package (gitlab.univ-nantes.fr/Nico_V/hlafix) as further performance tests are still necessary. HLA*fix* is part of the **COVID-19 HLA and Immunogenetics Consortium** analyses task force strategy to ease the exploration of *HLA* alleles in immune-related diseases such as the SARS-CoV-2 infection. HLA*fix* can receive diverse input data: raw GWAS data, post-imputation GWAS data, imputed *HLA* alleles or even typed *HLA* alleles. HLA*fix* handles GWAS quality control and *HLA* imputation. HLA*fix* performs **regression model analyses to test *HLA* alleles for association** with given traits and can include covariates such as ancestry principal components or sex. HLA*fix* will help to **accelerate research** and highlight an interesting path for further exploration of immune-related diseases.

These 2 tools are still under development and validation before publication, but are already available to all as an open science willingness within our team.

The alliance of *HLA* bioinformatic imputation tools from SNPs with Easy-HLA and HLA*fix* aims to generate new functional immunogenomic parameters and facilitate large-scale disease association analyses. These new parameters will make it possible to go beyond the simple marker association and to better functionally define the *HLA* associations in immune-related pathologies.

2. Unravel innovative immunogenomic associations to describe the fundamental and functional role of HLA in immune-related diseases (KiT: Kidney Transplantation; MS: Multiple Sclerosis).

a) Identify *HLA*/*KIR* genetic factors associated with KiT immunological complications

Chronic kidney disease affects 1 out of 10 people worldwide and can progress to end-stage kidney disease requiring dialysis or kidney transplantation. KiT is the best treatment for end-stage kidney disease; compared to dialysis, it simultaneously improves survival and quality of life. In 2019, 41,374 patients were living with a functional kidney transplant in France [47]. The one-year survival of kidney transplant is now 90% and the graft half-life is about 10 years. However, despite the enormous progress of immunosuppressive treatments, the host immune response against the graft **still leads to rejection in many cases** [57]. This **public health concern** is critical considering that **organs are a scarce resource**. Rejections are grouped based on kidney biopsies and the Banff classification in 2 categories [58]: 1) **cell-mediated rejection** which involves graft infiltration by recipient T-cells and 2) **humoral rejection** involving antibodies against the graft (i.e. donor-specific antibody, DSA). Moreover, immunosuppressive treatments are not without side effects (e.g. cancers). Overall, the **triggers** leading to rejection are still largely unknown and **need to be further studied**.

HLA molecules produce immunodominant antigens target for both humoral and cellular alloreactivity [48]. ***HLA* compatibility** between donor and recipient **is essential** as the number of *HLA* allelic differences between donor and recipient (**mismatches**) is strongly linked to graft survival [49]. In addition, HLA bears a **central role in the immune system** for peptides presentation to T-cells, and as **KIRs** ligands (NK cells receptors). In another hand, HLA expression is partially regulated by SNPs (single nucleotide polymorphism) within their regulatory regions. As an example, 2 studies explored the effect of rs9277534, a SNP (single nucleotide polymorphism) associated with differential expression of HLA-DPB1, and showed opposite conclusions regarding the effect on de novo DSA development in solid organ transplantation [59,60]. Beyond these contradictory results, the overall impact of SNPs regulating HLA expression has not been extensively studied in solid organ transplantation, including in KiT. Thus, the role of SNPs regulating HLA expression still needs further investigation in KiT. The few immunogenetic studies carried out in KiT were mainly about *HLA* and *KIR* mismatches between donor and recipient [49,50,61]. Nevertheless, **mismatches do not explain all rejections** and the intrinsic effect of *HLA* alleles, beyond these mismatches, remains elusive. Altogether, I **hypothesize that HLA** (alleles and/or expression levels) **plays a major role in post-kidney**

transplant complications and graft dysfunction, not only HLA mismatches. Exploring the *HLA* diversity and complexity (alleles, haplotypes), but also aspects of regulation/expression and KIR receptors is definitely a **present-day research project** and will allow us to better understand cellular and humoral rejections and provide new insights in **transplanted patients management and graft allocation** for example by selecting potentially less immunogenic *HLA* alleles.

CRTI is part of the Institute of Transplantation Urology Nephrology (ITUN) at CHU Nantes and as such benefits from an **access to the DIVAT network**, which **collects biological, clinical and epidemiological data for kidney transplanted patients** from 8 centers in France, representing more than **10,400 transplanted patients since 2000**. In addition, we have collected **DNA for 2,200 donor/recipient pairs from the DIVAT Nantes cohort**. The genotyping was recently performed for all samples using the high-density Affymetrix® Axiom Precision Medicine Research Array (PMRA, 96 samples/chip) that targets more than 900k SNPs and is enriched for immune- and transplant-related variants (e.g. within *HLA* genes) and for functional variants (KiT-Genie, Sophie Limou).

Upgrade HLA in the DIVAT cohort. The *HLA* genotypes of DIVAT subjects are for the most part in low resolution; **Easy-HLA will provide new parameters** from these genotypes: upgrade to high-resolution genotypes, 5-gene haplotyping, number of epitope mismatches, HLA amino acids, KIR ligand groups and HLA-C imputed expression level. *SNP genotyping quality control.* Quality controls and SNPs imputation of this genome-wide genotyping arrays are underway to get a high SNP coverage within the *MHC* region. The SNPs will allow us to: 1) **correct for possible population structure** in our analyses (even if >90% of the individuals are from European origin); 2) study **SNPs associated with HLA regulation** such as HLA-C or HLA-DPB1 expression [62]. *KIR typing of the DIVAT cohort.* In parallel, in collaboration with Katia Gagne (researcher at EFS, Nantes), ***KIR* genes typing** was performed in a subset of **247 DIVAT patients** from Nantes (ABM funding). A very preliminary analysis showed an increased frequency of the telomeric B haplotype in KiT patients compared to controls (16% vs. 7%, p=0.024). Further analyses are needed to confirm this result. *HLA association study in kidney transplanted patients from the DIVAT cohort. HLA* is well known to contribute to rejection because of mismatches between donor and recipient pairs [49]. Here, we hypothesize that *HLA* and its related parameters are linked to transplantation adverse outcomes beyond compatibility (Fig. 16). The computationally generated immunogenomic parameters will allow us to **study in depth the *HLA* role in KiT dysfunctions**. The DIVAT cohort provides clearly defined biological and immunological parameters. We plan to test for associations with cellular

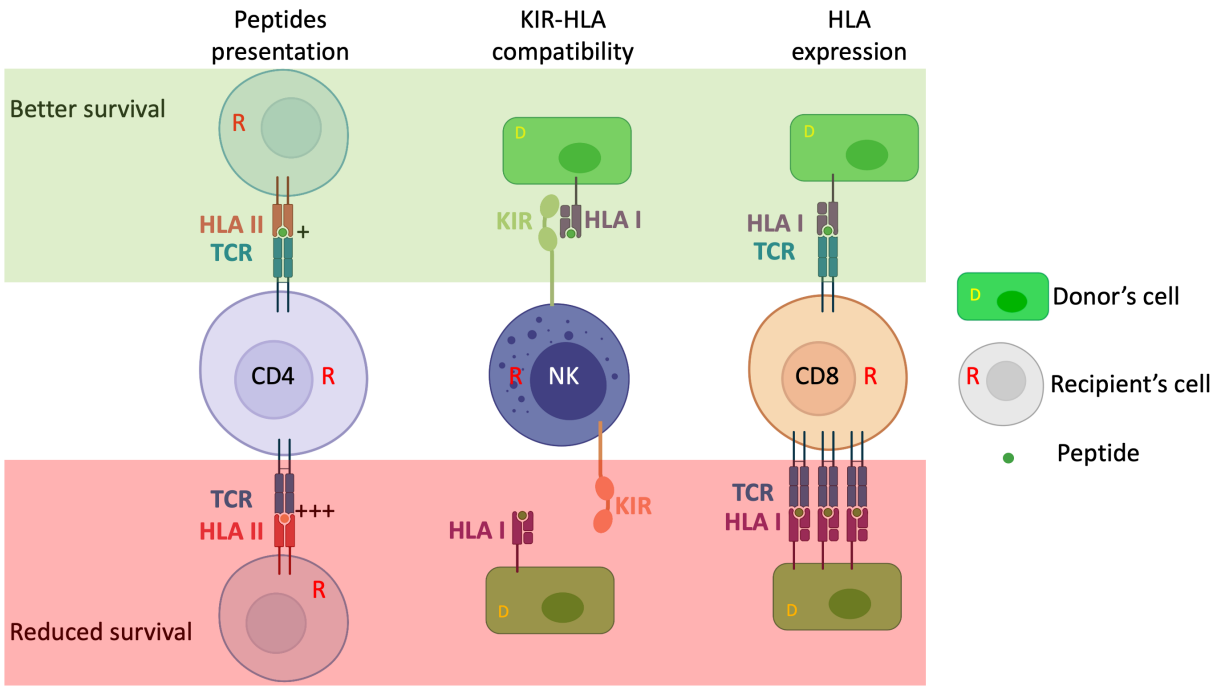

Figure 16: HLA functional role in KiT survival. Beyond HLA mismatches between donor and recipient, HLA can impact different aspect of graft immunity. Peptides presentation: the capacity of HLA to present peptides is dependent of the alleles carried by the host and this can lead to humoral rejection. KIR-HLA compatibility: KIR genes bind specific HLA alleles, KIR are mainly inhibitory receptors, if recipient's KIR do not recognize donor's HLA alleles, this can lead to NK-cells activation. HLA expression: HLA alleles are differentially expressed and this can have an impact on CD8 T-cells activation against the graft leading to cellular rejection. Top green background represents better situation for the graft while bottom red situation represents higher risk of immune activation. TCR: T-cell receptor. CD8: CD8 T-cell. CD4: CD4 T-cell. D: donor. R: recipient.

rejection and humoral rejection transplanted patients in the 4,400 genotyped individuals (2,200 donor/recipient pairs). The statistical methodologies implemented will be **logistic regression** models for case/control analyses (e.g. rejection vs. controls) or **Cox models** for time to phenotype analyses. These models will be corrected with potential confounding factors such as age, gender, population stratification and particularly the number of *HLA* mismatches (alleles and epitopes). The *HLA* and *KIR* alleles will be tested according to **allelic and dominant models**. Phenotypes can be tested using data of the donor or the recipient as an explaining factor. The free R platform will be used to perform these analyses. We estimated that we can expect a **statistical power of 90% to detect *HLA* associations** with an allele frequency of 5% and an effect size of 1.2. **KIR data** will be **combined with *HLA* data**. Indeed, HLA class I and especially HLA-C and some HLA-B alleles are ligands for KIR. For example, specific *HLA-C* alleles such as *HLA-C* group C2 (including *HLA-C*02*) bind to specific *KIR* genes such as *KIR2DL1* [25]. This functional grouping is well known and will be implemented in the analyses. The same models and statistics will be used as for *HLA* alone. *Impact of HLA alleles differential expression in the occurrence of chronic humoral rejection in renal transplanted patients of the DIVAT cohort.* My hypothesis is that **high HLA expression level in the graft is more**

immunogenic whatever the recipient *HLA* type. As with the HLA-C expression, the HLA class II expression level is differentially expressed between alleles (Fig. 17). We will use different strategies to impute this expression level to donor and recipient: first, with the help of a reference HLA class II mean expression cohort (Fig. 17); second, with PrediXcan [63], a software using SNP and Elastic net to predict expression level. We will implement this imputation strategy to both *HLA-DRB1* and *HLA-DQB1* genes. The difference in expression will be considered in a regression model comparing patients with humoral rejection versus others. SNPs associated with HLA-DRB1 and HLA-DPB1 expression will also be considered in the analyses.

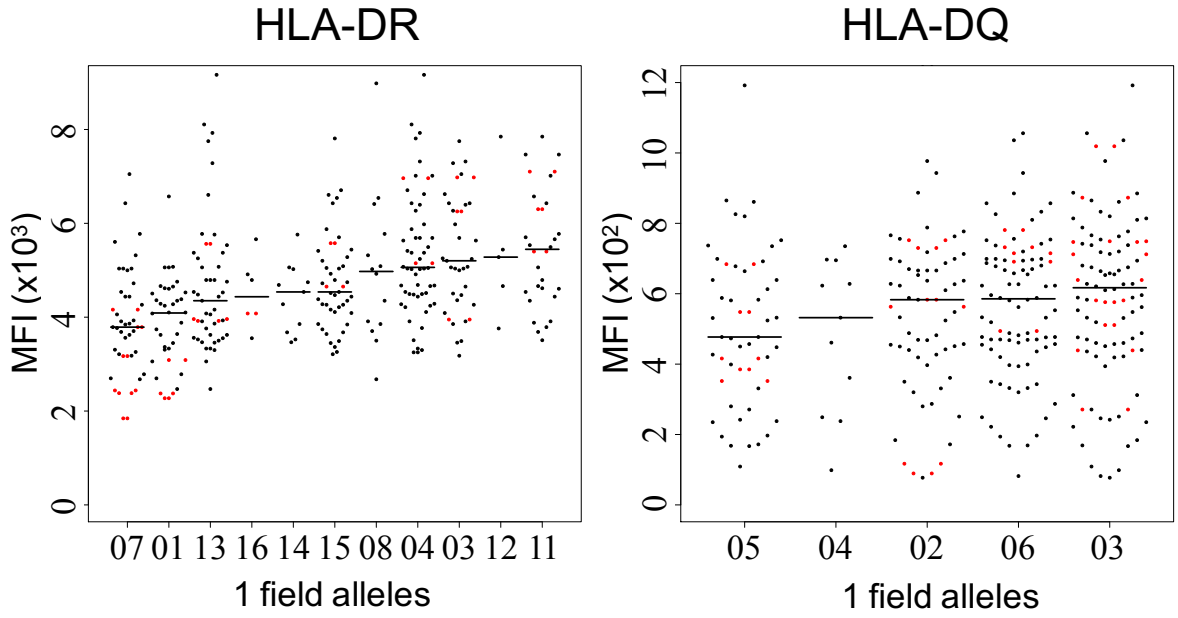

Figure 17: HLA class II gradient of expression across alleles from European ancestry blood donors. HLA-DR expression levels correlate significantly with HLA-DR alleles and are continuously distributed ($p=9.3x10^{-12}$; ANOVA). Vince et al. Unpublished data.

HLA and KIR are at the heart of innate and acquired immune responses. The *in silico* HLA functional involvement study in the occurrence of KiT complications is unprecedented, especially at the DIVAT scale. Even in the absence of a significant association, the generated immunogenetic data will be used to evaluate other phenotypes associated with KiT such as cancers, infections or response to immunosuppressive treatments.

The main objective is to identify *HLA* or *KIR* genetic factors associated with KiT immunological complications (e.g. humoral or cellular rejection, see Fig. 16). The functional immunogenomic parameters provided by our bioinformatics tools will be instrumental in the biological characterization of these associations.

b) Extract new knowledge from the MSGB polygenic score and further characterized the MS-related NMOSD disease for new genetic and *HLA* associations

MS is a debilitating and complex autoimmune, genetic, neuro-inflammatory and neurodegenerative chronic condition that manifests in young adults, typically around ages 25-35 (bit.ly/35P1lQU). In France, more than **110,000 individuals suffer from MS** and it is the **second cause of handicap in young adults** (after trauma); **each year 5,000 new cases** are diagnosed with around 75% females. Progress against MS has been remarkable these past years with approval of new medications [64].

Within the past 10 years, several **MS GWASs identified more than 200 genomic regions** [51,65]. The strongest signal consistently maps to the *HLA-DRB1* gene in the class II *HLA* region. Unlike *HLA-DRB1*15:01* (odds ratio ~3.0), most of the 200 genetic risk factors identified so far only have a slight effect on susceptibility to MS (odds ratios around 1.1 to 1.2); however, the risk alleles in these loci are common in people of European ancestry, with allele frequencies >10% [51]. This complex genetic architecture could be used to profile individuals through a **polygenic risk score** to summarize their susceptibility to disease [66]. In MS, such score was developed few years ago (MS Genetic Burden, MSGB) [53] and now needs refinement with the new accumulated genetic knowledge.

NMOSD was for a long time considered a variant form of MS, although with distinctive clinical, radiological and pathological features [67]. The identification of **2 specific auto-antibodies**, against aquaporin-4 (AQP4) and myelin oligodendrocyte glycoprotein (MOG), confirmed the distinction of NMOSD in regards of MS [68]. Only one limited study explored the genetics of NMOSD and revealed an *HLA* implication in the pathology without overlap with MS [69].

In the lab, we have access to several cohorts: GWAS data from the WTCCC MS cohort (18,872 controls and 11,376 cases), DNA from 372 NMOSD patients which were genotyped on the PMRA Affymetrix® Axiom chip. An additional 302 NMOSD patients were genotyped and the results are yet to be analyzed. Early results show that the mean MSGB score of NMOSD patients is higher than control individuals but lower than MS patients revealing a potential partial sharing of genetic risk factor between NMOSD and MS (Fig. 18).

MSGB. The MSGB score is computed based on a weighted scoring algorithm using independent MS-SNPs, typically one in each genomic region of interest. The MSGB computation follows a log-additive model: $MSGB = \sum_1^n [number\ of\ risk\ allele_i * log(OR)_i]$. I will explore the **evolution of the score** throughout

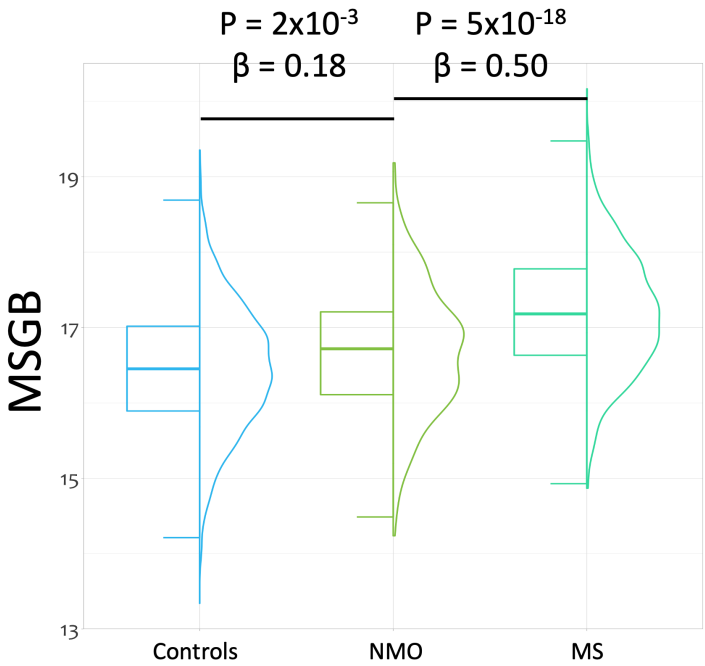

Figure 18: NMOSD patients MSGB score are on average higher than controls but lower than MS patients. MSGB was calculated with 174 SNPs extracted from the last published GWAS in 2019 [43]. P-values and beta were calculated using linear regression (MSGB as continuous variable) corrected by ancestry (4 PCs) and sex. Controls: N=5,205. NMOSD: N=295. MS: N=931.

the discovery of new SNP associations for the past 10 years (N>200 now). I will also study the updated MSGB in different situations: impact of *HLA* SNPs, association with severity, association with specific ancestry, male-female difference, association with age, etc. This stratification strategy will allow to better **understand the distribution and impact of genetic determinants on MS etiology**. Furthermore, I will explore the **implementation of new machine learning methods** to increase the polygenic score predictability.

NMOSD. NMOSD is now considered as a different entity from MS. The genetics of NMOSD is not clearly demonstrated yet, I have access to 372 genotyped NMOSD patients including 161 AQP4+, 82 MOG+, 69 seronegative and 60 to be determined; plus an additional 302 yet to be analyzed. With this large French cohort, I will **perform several GWASs:** first, on the whole NMO cohort, then, stratified for each subgroup to explore genome-wide associations. In addition, I will **impute and test *HLA* alleles** and other immunogenomic parameters for association with NMOSD phenotypes. The goal here is to **genetically characterized NMOSD compared to MS** but also within the different NMOSD subgroups. This cohort could be used to feed larger collaborative studies on the genetics of NMOSD.

The purpose here is to better define and improve the MSGB score for MS patients, to further characterize *HLA* and non-*HLA* genetic risk factors for the MS-related NMOSD disease.

V. References

1. Geffard E, Boussamet L, Walencik A, et al. HLA-EPI: A new EPIsode in exploring donor/recipient epitopic compatibilities. *HLA*. 2022;99(2):79-92. doi:10.1111/tan.14505

2. Montassier E, Al-Ghalith GA, Mathé C, et al. Distribution of Bacterial α1,3-Galactosyltransferase Genes in the Human Gut Microbiome. *Front Immunol*. 2019;10:3000. doi:10.3389/fimmu.2019.03000

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. 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

5. Valencia A, Vergara C, Thio CL, et al. Trans-ancestral fine-mapping of MHC reveals key amino acids associated with spontaneous clearance of hepatitis C in HLA-DQβ1. *Am J Hum Genet*. 2022;109(2):299-310. doi:10.1016/j.ajhg.2022.01.001

6. Goodin DS, Oksenberg JR, Douillard V, Gourraud PA, Vince N. Genetic susceptibility to multiple sclerosis in African Americans. *PLoS One*. 2021;16(8):e0254945. doi:10.1371/journal.pone.0254945

7. Domenighetti C, Douillard V, Sugier PE, Vince N, Elbaz A. The interaction between HLA-DRB1 and smoking in Parkinson's disease revisited. *Movement Disorders*. Published online Accepted.

8. Douillard V, Castelli EC, Mack SJ, et al. Current HLA Investigations on SARS-CoV-2 and Perspectives. *Front Genet*. 2021;12:774922. doi:10.3389/fgene.2021.774922

9. 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

10. Ba R, Geffard E, Douillard V, et al. Surfing the Big Data Wave: Omics Data Challenges in Transplantation. *Transplantation*. 2022;106(2):e114-e125. doi:10.1097/TP.0000000000003992

11. Geffard E, Limou S, Walencik A, et al. Easy-HLA: a validated web application suite to reveal the full details of HLA typing. *Bioinformatics*. 2020;36(7):2157-2164. doi:10.1093/bioinformatics/btz875

12. Sayadi S, Geffard E, Sudholt M, Vince N, Gourraud PA. Distributed Contextualization of Biomedical Data: A Case Study in Precision Medicine. In: *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. IEEE; 2020:1-6. doi:10.1109/AICCSA50499.2020.9316502

13. Sayadi S, Geffard E, Südholt M, Vince N, Gourraud PA. Secure Distribution of Factor Analysis of Mixed Data (FAMD) and Its Application to Personalized Medicine of Transplanted Patients. In: Barolli L, Woungang I, Enokido T, eds. *Advanced Information Networking and Applications*. Vol 225. Lecture Notes in Networks and Systems. Springer International Publishing; 2021:507-518. doi:10.1007/978-3-030-75100-5_44

14. Walencik A, Geffard E, Cesbron Gautier A, et al. EasyMatch-R software facilitates identification of compatible unrelated bone marrow donors, saves time-to-decision and money. *Transplant Cell Ther*. Published online 2021. doi:Submitted

15. Boutboul D, Vince N, Mahevas M, Bories JC, Fieschi C, DefI Study Group. TNFA, ANXA11 and BTNL2 Polymorphisms in CVID Patients with Granulomatous Disease. *J Clin Immunol*. 2016;36(2):110-112. doi:10.1007/s10875-016-0234-0

16. Gouilleux-Gruart V, Chapel H, Chevret S, et al. Efficiency of immunoglobulin G replacement therapy in common variable immunodeficiency: correlations with clinical phenotype and polymorphism of the neonatal Fc receptor. *Clin Exp Immunol*. 2013;171(2):186-194. doi:10.1111/cei.12002

17. Malphettes M, Gérard L, Carmagnat M, et al. Late-Onset Combined Immune Deficiency: A Subset of Common Variable Immunodeficiency with Severe T Cell Defect. *Clinical Infectious Diseases*. 2009;49(9):1329-1338. doi:10.1086/606059

18. Rivoisy C, Gérard L, Boutboul D, et al. Parental consanguinity is associated with a severe phenotype in common variable immunodeficiency. *J Clin Immunol*. 2012;32(1):98-105. doi:10.1007/s10875-011-9604-9

19. Vince N, Boutboul D, Mouillot G, et al. Defects in the CD19 complex predispose to glomerulonephritis, as well as IgG1 subclass deficiency. *J Allergy Clin Immunol*. 2011;127(2):538-541.e1-5. doi:10.1016/j.jaci.2010.10.019

20. Vince N, Mouillot G, Malphettes M, et al. Genetic screening of male patients with primary hypogammaglobulinemia can guide diagnosis and clinical management. *Hum Immunol*. Published online April 27, 2018. doi:10.1016/j.humimm.2018.04.014

21. Oksenhendler E, Gérard L, Fieschi C, et al. Infections in 252 Patients with Common Variable Immunodeficiency. *Clinical Infectious Diseases*. 2008;46(10):1547-1554. doi:10.1086/587669

22. Thaventhiran JED, Lango Allen H, Burren OS, et al. Whole-genome sequencing of a sporadic primary immunodeficiency cohort. *Nature*. 2020;583(7814):90-95. doi:10.1038/s41586-020-2265-1

23. van Zelm MC, Reisli I, van der Burg M, et al. An antibody-deficiency syndrome due to mutations in the CD19 gene. *New England Journal of Medicine*. 2006;354(18):1901-1912.

24. Vince N, Li H, Ramsuran V, et al. HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. *Am J Hum Genet*. 2016;99(6):1353-1358. doi:10.1016/j.ajhg.2016.09.023

25. Vince N, Bashirova AA, Lied A, et al. HLA class I and KIR genes do not protect against HIV type 1 infection in highly exposed uninfected individuals with hemophilia A. *J Infect Dis*. 2014;210(7):1047-1051. doi:10.1093/infdis/jiu214

26. Limou S, Zagury JF. Immunogenetics: Genome-Wide Association of Non-Progressive HIV and Viral Load Control: HLA Genes and Beyond. *Front Immunol*. 2013;4:118. doi:10.3389/fimmu.2013.00118

27. Apps R, Qi Y, Carlson JM, et al. Influence of HLA-C Expression Level on HIV Control. *Science*. 2013;340(6128):87-91. doi:10.1126/science.1232685

28. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38(8):904-909. doi:10.1038/ng1847

29. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

30. Gourraud PA, Khankhanian P, Cereb N, et al. HLA diversity in the 1000 genomes dataset. *PloS one*. 2014;9(7):e97282.

31. Vince N, Poschmann J, Josien R, Anegon I, Limou S, Gourraud PA. 23rd Nantes Actualités Transplantation: "Genomics and Immunogenetics of Kidney and Inflammatory Diseases - Lessons for Transplantation." *Transplantation*. Published online November 5, 2018. doi:10.1097/TP.0000000000002517

32. Goodin DS, Khankhanian P, Gourraud PA, Vince N. Highly conserved extended haplotypes of the major histocompatibility complex and their relationship to multiple sclerosis susceptibility. *PLoS ONE*. 2018;13(2):e0190043. doi:10.1371/journal.pone.0190043

33. Goodin D, Khankhanian P, Gourraud P, Vince N. *Genetic Susceptibility to Multiple Sclerosis: Interactions between Conserved Extended Haplotypes of the MHC and*

Other Susceptibility Regions. Epidemiology; 2019. doi:10.1101/603878

34. Goodin D, Khankhanian P, Gourraud P, Vince N. *The Nature of Genetic Susceptibility to Multiple Sclerosis*. Genetics; 2020. doi:10.1101/2020.08.13.249920

35. Goodin DS, Khankhanian P, Gourraud PA, Vince N. The nature of genetic and environmental susceptibility to multiple sclerosis. *PLoS One*. 2021;16(3):e0246157. doi:10.1371/journal.pone.0246157

36. Goodin DS, Khankhanian P, Gourraud PA, Vince N. Genetic susceptibility to multiple sclerosis: interactions between conserved extended haplotypes of the MHC and other susceptibility regions. *BMC Med Genomics*. 2021;14(1):183. doi:10.1186/s12920-021-01018-6

37. Goodin D, Khankhanian P, Gourraud P, Vince N. *Multiple Sclerosis: Exploring the Limits of Genetic and Environmental Susceptibility*. Genetic and Genomic Medicine; 2022. doi:10.1101/2022.03.09.22272129

38. 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

39. 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

40. 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

41. 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

42. Duquesnoy RJ, Marrari M, Tambur AR, et al. First report on the antibody verification of HLA-DR, HLA-DQ and HLA-DP epitopes recorded in the HLA Epitope Registry. *Hum Immunol*. 2014;75(11):1097-1103. doi:10.1016/j.humimm.2014.09.012

43. Duquesnoy RJ, Marrari M, Marroquim MS, et al. Second update of the International Registry of HLA Epitopes. I. The HLA-ABC Epitope Database. *Hum Immunol*. 2019;80(2):103-106. doi:10.1016/j.humimm.2018.11.007

44. Khankhanian P, Gourraud PA, Lizee A, Goodin DS. Haplotype-based approach to known MS-associated regions increases the amount of explained risk. *J Med Genet*. 2015;52(9):587-594. doi:10.1136/jmedgenet-2015-103071

45. Daya M, Rafaels N, Brunetti TM, et al. Association study in African-admixed

populations across the Americas recapitulates asthma risk loci in non-African populations. *Nat Commun*. 2019;10(1):880. doi:10.1038/s41467-019-08469-7

46. Shimane K, Kochi Y, Suzuki A, et al. An association analysis of HLA-DRB1 with systemic lupus erythematosus and rheumatoid arthritis in a Japanese population: effects of *09:01 allele on disease phenotypes. *Rheumatology*. 2013;52(7):1172-1182. doi:10.1093/rheumatology/kes427

47. Agence de la Biomédecine. *REIN annual report 2019 -Réseau Epidémiologie et Information en Néphrologie.* Agence de la Biomédecine; 2019.

48. Montgomery RA, Tatapudi VS, Leffell MS, Zachary AA. HLA in transplantation. *Nat Rev Nephrol*. 2018;14(9):558-570. doi:10.1038/s41581-018-0039-x

49. Held PJ, Kahan BD, Hunsicker LG, et al. The impact of HLA mismatches on the survival of first cadaveric kidney transplants. *N Engl J Med*. 1994;331(12):765-770. doi:10.1056/NEJM199409223311203

50. Heidt S, Haasnoot GW, van Rood JJ, Witvliet MD, Claas FHJ. Kidney allocation based on proven acceptable antigens results in superior graft survival in highly sensitized patients. *Kidney Int*. 2018;93(2):491-500. doi:10.1016/j.kint.2017.07.018

51. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*. 2019;365(6460). doi:10.1126/science.aav7188

52. Duncan L, Shen H, Gelaye B, et al. Analysis of polygenic risk score usage and performance in diverse human populations. *Nat Commun*. 2019;10(1):3328. doi:10.1038/s41467-019-11112-0

53. Gourraud PA, McElroy JP, Caillier SJ, et al. Aggregation of multiple sclerosis genetic risk variants in multiple and single case families. *Ann Neurol*. 2011;69(1):65-74. doi:10.1002/ana.22323

54. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell*. 2019;177(1):26-31. doi:10.1016/j.cell.2019.02.048

55. Naito T, Suzuki K, Hirata J, et al. *A Multi-Task Convolutional Deep Learning Method for HLA Allelic Imputation and Its Application to Trans-Ethnic MHC Fine-Mapping of Type 1 Diabetes*. Genetic and Genomic Medicine; 2020. doi:10.1101/2020.08.10.20170522

56. Pappas DJ, Marin W, Hollenbach JA, Mack SJ. Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An integrated case-control analysis pipeline. *Hum Immunol*. 2016;77(3):283-287. doi:10.1016/j.humimm.2015.12.006

57. Danger R, Sawitzki B, Brouard S. Immune monitoring in renal transplantation:

The search for biomarkers. *Eur J Immunol*. 2016;46(12):2695-2704. doi:10.1002/eji.201545963

58. Roufosse C, Simmonds N, Clahsen-van Groningen M, et al. A 2018 Reference Guide to the Banff Classification of Renal Allograft Pathology. *Transplantation*. 2018;102(11):1795-1814. doi:10.1097/TP.0000000000002366

59. Shieh M, Hayeck TJ, Dinh A, et al. Complex Linkage Disequilibrium Effects in HLA-DPB1 Expression and Molecular Mismatch Analyses of Transplantation Outcomes. *Transplantation*. 2021;105(3):637-647. doi:10.1097/TP.0000000000003272

60. Tang C, Unterrainer C, Fink A, et al. Analysis of de novo donor-specific HLA-DPB1 antibodies in kidney transplantation. *HLA*. 2021;98(5):423-430. doi:10.1111/tan.14422

61. Koenig A, Chen CC, Marçais A, et al. Missing self triggers NK cell-mediated chronic vascular rejection of solid organ transplants. *Nat Commun*. 2019;10(1):5350. doi:10.1038/s41467-019-13113-5

62. Petersdorf EW, Malkki M, O'hUigin C, et al. High HLA-DP Expression and Graft-versus-Host Disease. *N Engl J Med*. 2015;373(7):599-609. doi:10.1056/NEJMoa1500140

63. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091-1098. doi:10.1038/ng.3367

64. Bloomgren G, Richman S, Hotermans C, et al. Risk of natalizumab-associated progressive multifocal leukoencephalopathy. *New England Journal of Medicine*. 2012;366(20):1870-1880.

65. Baranzini SE, Oksenberg JR. The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. *Trends Genet*. Published online October 5, 2017. doi:10.1016/j.tig.2017.09.004

66. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet*. 2018;19(9):581-590. doi:10.1038/s41576-018-0018-x

67. Wingerchuk DM, Hogancamp WF, O'Brien PC, Weinshenker BG. The clinical course of neuromyelitis optica (Devic's syndrome). *Neurology*. 1999;53(5):1107-1114. doi:10.1212/wnl.53.5.1107

68. McCreary M, Mealy MA, Wingerchuk DM, Levy M, DeSena A, Greenberg BM. Updated diagnostic criteria for neuromyelitis optica spectrum disorder: Similar outcomes of previously separate cohorts. *Mult Scler J Exp Transl Clin*. 2018;4(4):2055217318815925. doi:10.1177/2055217318815925

69. Estrada K, Whelan CW, Zhao F, et al. A whole-genome sequence study

identifies genetic risk factors for neuromyelitis optica. *Nat Commun*. 2018;9(1):1929. doi:10.1038/s41467-018-04332-3

VI. Selection of five publications

A. HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region (AJHG, 2016)

B. Association of HLA-DRB1*09:01 with tIgE levels among African ancestry individuals with asthma (JACI, 2020)

C. Easy-HLA: a validated web application suite to reveal the full details of HLA typing (Bioinformatics, 2020)

D. SNP-HLA Reference Consortium (SHLARC): HLA and SNP data sharing for promoting MHC-centric analyses in genomics (Genetic Epidemiology, 2020)

E. Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research (Frontiers in Genetics, 2022)

HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the *HLA-C* Promoter Region

Nicolas Vince,[1,2] Hongchuan Li,[1] Veron Ramsuran,[1,2] Vivek Naranbhai,[2,3,4] Fuh-Mei Duh,[1] Benjamin P. Fairfax,[3] Bahara Saleh,[1] Julian C. Knight,[3] Stephen K. Anderson,[1] and Mary Carrington[1,2,*]

Differential HLA-C levels influence several human diseases, but the mechanisms responsible are incompletely characterized. Using a validated prediction algorithm, we imputed HLA-C cell surface levels in 228 individuals from the 1000 Genomes dataset. We tested 68,726 SNPs within the *MHC* for association with HLA-C level. The *HLA-C* promoter region variant, rs2395471, 800 bp upstream of the transcription start site, gave the most significant association with *HLA-C* levels ($p = 4.2 \times 10^{-66}$). This imputed expression quantitative trait locus, termed impeQTL, was also shown to associate with *HLA-C* expression in a genome-wide association study of 273 donors in which *HLA-C* mRNA expression levels were determined by quantitative PCR (qPCR) ($p = 1.8 \times 10^{-20}$) and in two cohorts where HLA-C cell surface levels were determined directly by flow cytometry ($n = 369$ combined, $p < 10^{-15}$). rs2395471 is located in an Oct1 transcription factor consensus binding site motif where the A allele is predicted to have higher affinity for Oct1 than the G allele. Mobility shift electrophoresis demonstrated that Oct1 binds to both alleles in vitro, but decreased *HLA-C* promoter activity was observed in a luciferase reporter assay for rs2395471_G relative to rs2395471_A on a fixed promoter background. The rs2395471 variant accounts for up to 36% of the explained variation of HLA-C level. These data strengthen our understanding of HLA-C transcriptional regulation and provide a basis for understanding the potential consequences of manipulating HLA-C levels therapeutically.

Variation in the *MHC* associates with an extensive number of human diseases and traits, with about 500 (30%) reported in the human genome-wide association study (GWAS) catalog,[1] particularly that occurring within the *HLA* class I and II genes. Extreme polymorphism of the *HLA* loci[2] along with their central importance in both the acquired and innate immune response accounts for this over-representation relative to the rest of the genome. HLA class I and II molecules bind and present an extensive array of antigenic peptides to cytotoxic T lymphocytes and CD4 T cells, respectively, in order to initiate the acquired immune response.[3,4] The class I molecules also serve as ligands for killer cell immunoglobulin-like receptors (KIR) expressed on natural killer cells.[5] Relative to *HLA-A* (MIM: 142800) and *-B* (MIM: 142830), *HLA-C* (MIM: 142840) exhibits limited diversity, lower cell surface level,[6,7] and a more widely distributed role as ligands for KIR. *HLA-C* alleles are associated with many disease traits,[1] primarily with regard to autoimmune diseases.[8,9] High HLA-C level associates with better HIV control (MIM: 609423), but also with increased risk of developing Crohn disease (MIM: 266600).[10] Hence, understanding the mechanisms that determine HLA-C level could provide important insights into the management of complex human disease.

An insertion/deletion polymorphism in the 3′ UTR of *HLA-C* determines the binding and inhibition of *HLA-C* expression by the microRNA miR-148a (*MIR148A* [MIM: 613786]), contributing to differential HLA-C levels.[11,12] However, this polymorphism accounts for only 9% of the explained variation in HLA-C level, indicating that additional mechanisms participate in determining allele-specific expression levels at this locus. In the current study, we identified variants within and near *HLA-C* that significantly associate with imputed HLA-C levels among individuals from the 1000 Genomes dataset (1KG).[10,13,14] The most significant variant was then tested for its association with both RNA expression and cell surface HLA-C levels measured directly, validating the imputation approach, and the mechanism underlying this association was shown to involve the transcription factor (TF) Oct1.

HLA-C levels vary in an allele-specific manner over a range of 7-fold in a pattern that is consistent between African and European Americans and highly reproducible across study groups.[10] Based on the level value characteristic for each given HLA-C allotype as determined previously (Table S1), we imputed HLA-C levels for 228 European 1KG individuals who have previously been typed for *HLA-C*.[14] We restricted our analysis to Europeans with homogeneous ancestry background (Figure S1), which included 52 CEU (Utah residents with Northern and Western European ancestry), 87 GBR (British in England and Scotland), and 89 TSI (Toscani in Italy).

Imputed HLA-C levels were tested as a continuous variable for association with 68,726 SNPs within the *MHC* using linear regression in order to identify *cis*-acting variants that may cause (or mark) differential level of HLA-C. The peak association was centered in the *HLA-C* promoter region (Figure 1A), and correction for population structure did not alter the results.[17] The top signal identified was rs2395471 ($p = 4.2 \times 10^{-66}$), which is 800 bp upstream of the transcription start site (Figure 1B), and

[1]Cancer and Inflammation Program, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA; [2]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139, USA; [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK; [4]Center for the AIDS Programme of Research in South Africa, University of KwaZuluNatal, Durban 4091, South Africa
*Correspondence: carringm@mail.nih.gov
http://dx.doi.org/10.1016/j.ajhg.2016.09.023.

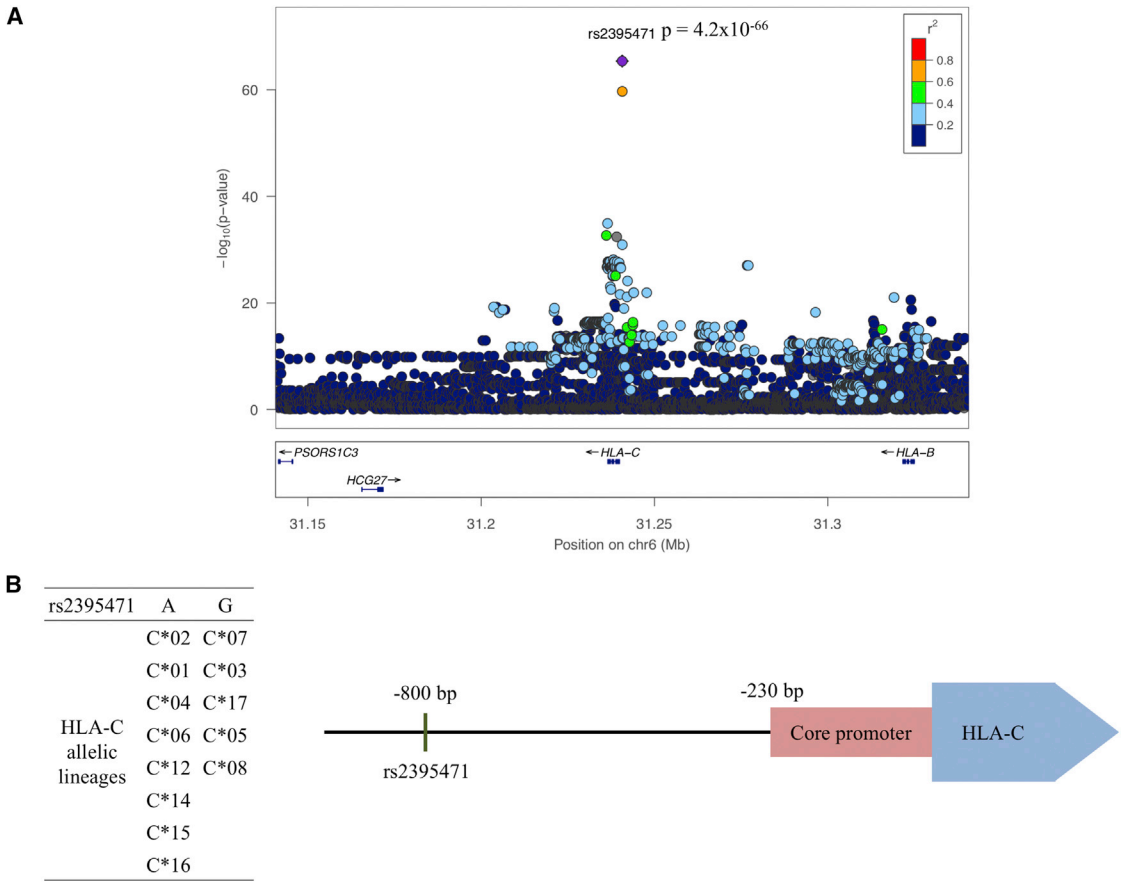

Figure 1. HLA-C Level Is Associated with rs2395471 in the Promoter Region of *HLA-C*

(A) Association between SNPs located in a 200 kb window around *HLA-C* and imputed HLA-C level. The Manhattan plot was created with LocusZoom.[15] Each dot represents a SNP for which the color depicts the linkage disequilibrium (LD) score (r^2) computed with PLINK[16] in the European 1KG dataset (n = 228). A complete list of LD scores between rs2395471 and the surrounding 1 Mb SNPs is presented in Table S2. Logistic regression tests for association between the 68,726 SNPs in the *HLA* locus and HLA-C level was performed in R (version 3.2.4) using the imputed HLA-C level values as a continuous variable.

(B) A list of *HLA-C* lineages that carry either the A or G at the rs2395471 SNP and a map of the *HLA-C* promoter region are shown.

only one other neighboring SNP, rs2249741, showed a similar level of significance ($p = 2.0 \times 10^{-60}$). The A versus G frequency at rs2395471 was fairly evenly distributed across *HLA-C* alleles (Figure 1B). We term these variants "imputed expression quantitative trait loci" (impeQTL) to distinguish them from those associating with expression levels of genes that were measured directly. impeQTL can only be used to identify candidates of expression modifiers in *cis* of the gene when its expression or protein level is imputed.

Genotyping of rs2395471 in two independent cohorts where HLA-C levels were measured directly by flow cytometry confirmed the association between this SNP and cell surface levels of HLA-C on CD3-positive cells: $p = 9.3 \times 10^{-16}$ in a cohort of 195 African Americans for whom levels were determined previously[10] (Figure 2A) and $p = 2.9 \times 10^{-17}$ in a cohort of 174 European Americans (Figure 2B). The explained variation based on the R^2 indicated that rs2395471 accounts for 28% and 36% of the HLA-C level variation in the African American and European American, respectively.[18] The rs2395471_A allele, which marks high level of HLA-C and is the ancestral allele, has a global frequency of 53% in the 1KG dataset ranging from 46% in East Asia to 62% in South Asia.

The association between rs2395471 and *HLA-C* expression was independently corroborated by a genome-wide expression quantitative trait loci (eQTL) study of *HLA-C* expression measured by qPCR in peripheral blood mononuclear cells (PBMCs) from 273 donors of European descent from Great Britain.[19] A locus containing two variants spaced 20 bp apart and in moderate linkage disequilibrium ($r^2 = 0.6$), namely rs2249741 (effect allele frequency [EAF] 48%, $p = 1.8 \times 10^{-24}$) and rs2395471 (EAF 36%, $p = 1.8 \times 10^{-20}$), was the most significantly associated locus genome wide, demonstrating that this locus probably represents the strongest eQTL for *HLA-C* expression across the genome (Figure S2A). We note that the differences in allele frequency may account for the relative difference in degree of statistical significance, but the effect sizes are comparable (rs2395471: beta = 0.72; SE = 0.074 versus rs2249741: beta = 0.74; SE = 0.071). The mean expression level of individual *HLA-C* alleles in

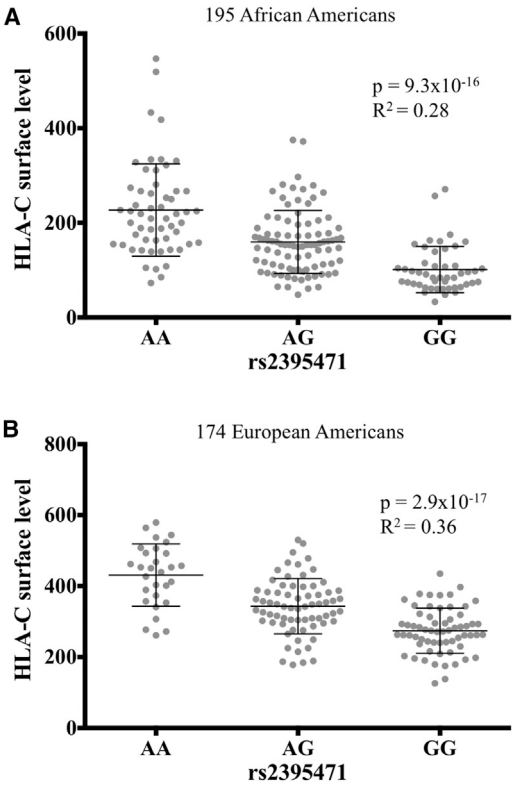

Figure 2. Association between rs2395471 and HLA-C Cell Surface Level

Cell surface level was measured directly in two independent cohorts of (A) 195 African Americans and (B) 174 European Americans. HLA-C cell surface level was measured on $CD3^+$ cells from peripheral blood by flow cytometry. For each plot, the p value from linear regression and R^2 (the level of variation explained by the SNP) are provided. The rs2395471 genotypes were determined by Sanger sequencing. The mean with standard deviation is represented. The National Health Institutes office of human subjects research protection approved this study.

this eQTL study (Figure S2B) was highly correlated with that published previously[12] ($p = 0.001$, $R = 0.95$), underscoring the reproducibility of the expression level measurements.

Given the location of the rs2395471 and rs2249741 variants in the promoter region of *HLA-C*, the possibility that one or both may alter TF binding was considered using the AliBaba online prediction tool.[20] A potential TF binding site for POU, a family of TFs containing well-conserved homeodomains,[21] was predicted to overlap with the rs2395471 variant. No TF binding site was predicted for rs2249741. The presence of a TF binding site within a sequence containing the rs2395471_A/G variant was confirmed by an electrophoretic mobility shift assay (EMSA) using rs2395471_A versus _G oligonucleotides and HeLa cell nuclear extracts (Figure 3A). POU-specific antibodies showed that Oct1 (*POU2F1* [MIM: 164175]), but not Oct2 (*POU2F2* [MIM: 164176]) or Oct3/4 (*POU5F1* [MIM: 164177]), can bind to the promoter region containing rs2395471 (Figure 3A). Further, the intensity of Oct1 binding was lower in the presence of oligonucleotide containing the G allele relative to that containing the A allele (Figures 3A and 3B). These experiments were reproduced using Jurkat cell line nuclear extract (Figure S3). Nuclear extracts of PBMCs from two healthy donors also demonstrated Oct1 binding within the region containing the rs2395471 variant (Figure 3C).

Given the limitations in terms of sensitivity and in vitro application, we proceeded to design a luciferase reporter assay to evaluate whether the rs2395471 alleles differentially impact *HLA-C* promoter activity. The promoter regions of two high expression *HLA-C* alleles that carry the rs2395471_A allele, *C*0102* and *C*0401*, were cloned into a pGL3 plasmid and alternates containing G at this position were generated by site-directed mutagenesis. In addition, the lower expression alleles *C*03:04* and *C*08:02*, both of which carry rs2395471_G, were used to generate a pGL3 construct together with an alternate that contains A at this position. Luciferase activity was assayed 48 hr after transfection of HeLa cells, revealing a significant decrease in promoter activity upon the single base pair change from A to G in a *C*01:02*, as well as a *C*04:01* background (Figure 4). These results strongly suggest that rs2395471 has a direct effect on *HLA-C* expression by impacting the binding affinity of the Oct1 TF. The promoter activities of the wild-type *C*03:04* and *C*08:02* alleles (i.e., rs2395471_G) were similar to the activity of *C*01:02* or *C*04:01* after the A residue in the Oct1 site was changed to G. However, replacement of the G residue with A in the Oct1 binding site of *C*03:04* and *C*08:02* did not increase the promoter activity of either allele (Figure 4), indicating that negative regulatory factors can dominate over the enhancer activity of the Oct1 binding site. Of note, wild-type *C*08:02* promoter activity was lower than that of *C*04:01* (Figure 4), in spite of their virtually identical levels observed previously on $CD3^+$ cells.[10] This discrepancy may be due to escape of *C*08:02* from miR-148a downregulation due to a deletion polymorphism in the 3′ UTR of this allele,[11] whereas *C*04:01* binds and is inhibited by miR-148a. These data demonstrate the complexity of *HLA-C* regulation, and this complexity further illustrates the remarkable observation that the Oct1 binding site variant has the most significant association with cell surface levels across *HLA-C* alleles overall.

Here we have shown that imputation of HLA-C level data allowed the identification of the *cis* variant rs2395471, located in the *HLA-C* promoter region, that contributes to determining allele-specific *HLA-C* mRNA expression and HLA-C cell surface levels. The association involving this impeQTL was confirmed in three independent cohorts (cumulative $n = 694$) with HLA-C expression and cell surface levels measured directly by either qPCR or flow cytometry. Importantly, this locus has the most significant effect on *HLA-C* expression levels genome wide based on qPCR data in which GWAS data were also available, indicating that it is likely the leading regulator of HLA-C levels.

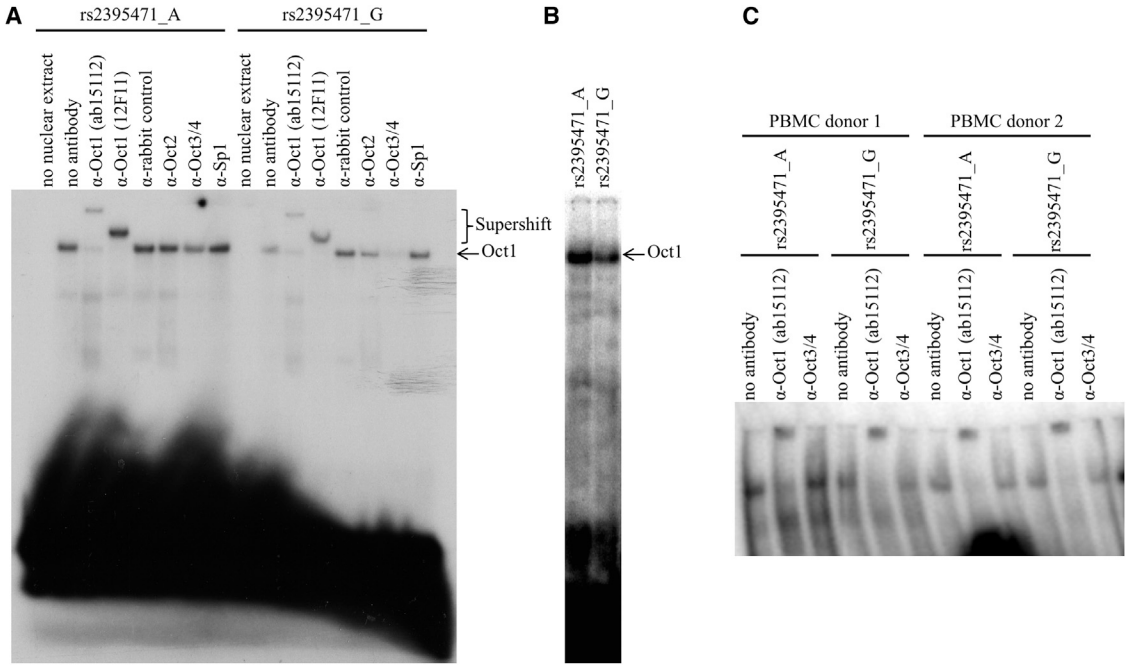

Figure 3. The Oct1 Transcription Factor Binds to the Genomic Region Containing rs2395471
Electrophoretic mobility shift assay (EMSA) was performed using nuclear protein extract from HeLa cell line (A and B) and PBMCs (C). Two oligonucleotides were designed, one containing the rs2395471_G allele and one containing the rs2395471_A allele (detailed methods can be found in Li et al.,[22] primers are available upon request).
(A) The presence of a shift (arrow) indicates that the genomic region containing rs2395471 binds a protein from the HeLa nuclear extract. A supershift is observed when anti-Oct1 antibodies (2 different clones) are added, but not anti-Oct2, anti-Oct3/4, or anti-Sp1 antibodies, indicating that Oct1 is a transcription factor that binds to the rs2395471 genomic region.
(B) HeLa nuclear extract binds to oligonucleotides containing rs2395471_A more strongly than it does to rs2395471_G.
(C) Nuclear proteins derived from PBMCs of two donors were extracted and incubated with the oligonucleotides and antibodies to anti-Oct1 and anti-Oct3/4. A supershift was detected only in the presence of the anti-Oct1 antibody, confirming the specificity of this TF binding site for Oct1.

POU2F1 (Oct1) is ubiquitously expressed across cell types[23–25] and recognizes the octamer DNA element ATGCAAAT and variations of it.[26] Its activity results in exceptionally diverse biological outcomes. Consistent with its ubiquitous expression, Oct1 is implicated in basal transcription regulation,[24,27,28] and it also appears to have a crucial role during embryogenesis, as Oct1 knockout mice are not viable.[29,30] Increased activity of Oct1 has been implicated in tumorigenesis, particularly epithelial tumors such as gastric (MIM: 613659) and breast (MIM: 114480) cancers,[31–35] and importantly in the context of HLA-C levels, Oct1 is linked to immune regulation of B cells, macrophages, T cells, and NK cells[31] by targeting production of cytokines (e.g., IL2),[36,37] pro-inflammatory mediators (e.g., NOS2),[38] and immunoglobulins.[39] Oct1 also has a role in signal response by serving as an adaptor for other TFs such as NF-κB.[28,40] An NF-κB binding site is predicted 633 bp downstream of the Oct1 binding site in the HLA-C core promoter region, raising the possibility that Oct1 may have a dual function in controlling *HLA-C* expression levels by regulating its basal expression and enhancing NF-κB-mediated effects on *HLA-C* expression levels upon cell activation.

Imputing HLA-C levels has led us to identify a locus, rs2395471, that accounts for up to 36% of the explained variation in HLA-C level. This SNP is not in significant LD with the insertion/deletion polymorphism in the 3′ UTR of *HLA-C* ($r^2 = 0.25$), which was previously shown to account in part for differential HLA-C levels as well.[11] The rs2395471 variant in combination with the *HLA-C* 3′ UTR variant explains up to 40% of the observed variability in measured HLA-C levels in European Americans. The approach described herein could be extended to other imputed expression or protein level data (e.g., additional *HLA* genes) in order to further our understanding of human gene regulation and their impact on disease.

Supplemental Data

Supplemental Data include three figures and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2016.09.023.

Acknowledgments

This project has been funded in whole or in part with federal funds from the Frederick National Laboratory for Cancer Research, under contract no. HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement

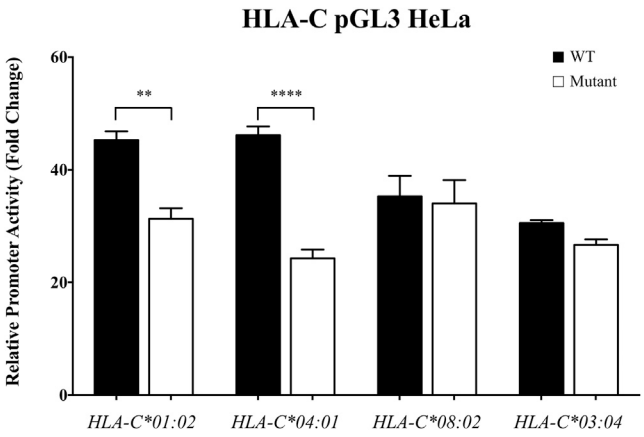

Figure 4. Impact of the rs2395471 on the *HLA-C* Promoter Activity Estimated by Luciferase Assay

The *HLA-C* promoters of two alleles carrying rs2395471_A, *C*01:02* and *C*04:01*, and two alleles carrying rs2395471_G, *C*03:04* and *C*08:02*, were cloned in a pGL3 plasmid. We mutated this position by site-directed mutagenesis to obtain new plasmids with a G nucleotide at the rs2395471 position for *C*01:02* and *C*04:01* and an A at this position for *C*03:04* and *C*08:02* (detailed methods can be found in Li et al.,[22] primers are available upon request). Both *HLA-C*01:02* and *C*04:01* promoters in which rs2395471_A was mutated to rs2395471_G show a significant decrease in promoter activity. Both *HLA-C*03:04* and *C*08:02* promoters in which rs2395471_G was mutated to rs2395471_A show no significant difference in promoter activity. Two-way ANOVA. $^{**}p < 0.01$, $^{****}p < 0.0001$. WT: wild-type. The mean of four replicated experiments with standard error is represented.

by the U.S. Government. This research was supported in part by the Intramural Research Program of the NIH, Frederick National Lab, Center for Cancer Research. J.C.K. is supported by European Research Council funding (grant agreement no. 281824) and the NIHR Oxford Biomedical Research Centre and Wellcome Trust (core facilities grant 090532/Z/09/Z). The Oxford cohort was supported by the Wellcome Trust (grants 074318 [J.C.K.], 088891 [B.P.F.], and 090532/Z/09/Z [core facilities Wellcome Trust Centre for Human Genetics including High-Throughput Genomics Group]), the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement no. 281824 (J.C.K.), the Medical Research Council (98082 [J.C.K]), and the NIHR Oxford Biomedical Research Centre. V.N. was supported by the Rhodes Trust.

Received: July 15, 2016
Accepted: September 29, 2016
Published: November 3, 2016

Web Resources

1000 Genomes, http://browser.1000genomes.org/index.html
AliBaba, http://www.gene-regulation.com/pub/programs/alibaba2/index.html
EIGENSOFT, https://www.hsph.harvard.edu/alkes-price/software/
GWAS Catalog, http://www.ebi.ac.uk/gwas/
LocusZoom, http://locuszoom.sph.umich.edu/locuszoom/
OMIM, http://www.omim.org/
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/
R statistical software, http://www.r-project.org/

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Robinson, J., Halliwell, J.A., Hayhurst, J.D., Flicek, P., Parham, P., and Marsh, S.G. (2015). The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res. *43*, D423–D431.
3. Hansen, T.H., and Bouvier, M. (2009). MHC class I antigen presentation: learning from viral evasion strategies. Nat. Rev. Immunol. *9*, 503–513.
4. Jones, E.Y., Fugger, L., Strominger, J.L., and Siebold, C. (2006). MHC class II proteins and disease: a structural perspective. Nat. Rev. Immunol. *6*, 271–282.
5. Bashirova, A.A., Martin, M.P., McVicar, D.W., and Carrington, M. (2006). The killer immunoglobulin-like receptor gene cluster: tuning the genome for defense. Annu. Rev. Genomics Hum. Genet. *7*, 277–300.
6. Apps, R., Meng, Z., Del Prete, G.Q., Lifson, J.D., Zhou, M., and Carrington, M. (2015). Relative expression levels of the HLA class-I proteins in normal and HIV-infected cells. J. Immunol. *194*, 3594–3600.
7. Parham, P., Lomen, C.E., Lawlor, D.A., Ways, J.P., Holmes, N., Coppin, H.L., Salter, R.D., Wan, A.M., and Ennis, P.D. (1988). Nature of polymorphism in HLA-A, -B, and -C molecules. Proc. Natl. Acad. Sci. USA *85*, 4005–4009.
8. Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., et al. (2007). A whole-genome association study of major determinants for host control of HIV-1. Science *317*, 944–947.
9. Wellcome Trust Case Control, C.; Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature *447*, 661–678.
10. Apps, R., Qi, Y., Carlson, J.M., Chen, H., Gao, X., Thomas, R., Yuki, Y., Del Prete, G.Q., Goulder, P., Brumme, Z.L., et al. (2013). Influence of HLA-C expression level on HIV control. Science *340*, 87–91.
11. Kulkarni, S., Savan, R., Qi, Y., Gao, X., Yuki, Y., Bass, S.E., Martin, M.P., Hunt, P., Deeks, S.G., Telenti, A., et al. (2011). Differential microRNA regulation of HLA-C expression and its association with HIV control. Nature *472*, 495–498.
12. Kulkarni, S., Qi, Y., O'hUigin, C., Pereyra, F., Ramsuran, V., McLaren, P., Fellay, J., Nelson, G., Chen, H., Liao, W., et al. (2013). Genetic interplay between HLA-C and MIR148A in HIV control and Crohn disease. Proc. Natl. Acad. Sci. USA *110*, 20705–20710.
13. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A.; 1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.
14. Gourraud, P.A., Khankhanian, P., Cereb, N., Yang, S.Y., Feolo, M., Maiers, M., Rioux, J.D., Hauser, S., and Oksenberg, J. (2014). HLA diversity in the 1000 genomes dataset. PLoS ONE *9*, e97282.
15. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R., and Willer, C.J. (2010). LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics *26*, 2336–2337.

16. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
17. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. *38*, 904–909.
18. Draper, N.R., and Smith, H. (1998). Applied Regression Analysis (New York: Wiley).
19. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., and Knight, J.C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. Nat. Genet. *44*, 502–510.
20. Grabe, N. (2002). AliBaba2: context specific identification of transcription factor binding sites. In Silico Biol. (Gedrukt) *2*, S1–S15.
21. Phillips, K., and Luisi, B. (2000). The virtuoso of versatility: POU proteins that flex to fit. J. Mol. Biol. *302*, 1023–1039.
22. Li, H., Wright, P.W., McCullen, M., and Anderson, S.K. (2016). Characterization of KIR intermediate promoters reveals four promoter types associated with distinct expression patterns of KIR subtypes. Genes Immun. *17*, 66–74.
23. Ferraris, L., Stewart, A.P., Kang, J., DeSimone, A.M., Gemberling, M., Tantin, D., and Fairbrother, W.G. (2011). Combinatorial binding of transcription factors in the pluripotency control regions of the genome. Genome Res. *21*, 1055–1064.
24. Sive, H.L., and Roeder, R.G. (1986). Interaction of a common factor with conserved promoter and enhancer sequences in histone H2B, immunoglobulin, and U2 small nuclear RNA (snRNA) genes. Proc. Natl. Acad. Sci. USA *83*, 6382–6386.
25. Staudt, L.M., Singh, H., Sen, R., Wirth, T., Sharp, P.A., and Baltimore, D. (1986). A lymphoid-specific protein binding to the octamer motif of immunoglobulin genes. Nature *323*, 640–643.
26. Reményi, A., Tomilin, A., Pohl, E., Lins, K., Philippsen, A., Reinbold, R., Schöler, H.R., and Wilmanns, M. (2001). Differential dimer activities of the transcription factor Oct-1 by DNA-induced interface swapping. Mol. Cell *8*, 569–580.
27. Kim, M.H., and Peterson, D.O. (1995). Stimulation of basal transcription from the mouse mammary tumor virus promoter by Oct proteins. J. Virol. *69*, 4717–4726.
28. Pance, A. (2016). Oct-1, to go or not to go? That is the PolII question. Biochim. Biophys. Acta *1859*, 820–824.
29. Hwang, S.S., Kim, L.K., Lee, G.R., and Flavell, R.A. (2016). Role of OCT-1 and partner proteins in T cell differentiation. Biochim. Biophys. Acta *1859*, 825–831.
30. Wang, V.E.H., Schmidt, T., Chen, J., Sharp, P.A., and Tantin, D. (2004). Embryonic lethality, decreased erythropoiesis, and defective octamer-dependent promoter activation in Oct-1-deficient mice. Mol. Cell. Biol. *24*, 1022–1032.
31. Vázquez-Arreguín, K., and Tantin, D. (2016). The Oct1 transcription factor and epithelial malignancies: Old protein learns new tricks. Biochim. Biophys. Acta *1859*, 792–804.
32. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al.; METABRIC Group (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.
33. Hernández, P., Solé, X., Valls, J., Moreno, V., Capellá, G., Urruticoechea, A., and Pujana, M.A. (2007). Integrative analysis of a cancer somatic mutome. Mol. Cancer *6*, 13.
34. Jeong, S.H., Lee, Y.J., Cho, B.I., Ha, W.S., Choi, S.K., Jung, E.J., Ju, Y.T., Jeong, C.Y., Ko, G.H., Yoo, J., and Hong, S.C. (2014). OCT-1 overexpression is associated with poor prognosis in patients with well-differentiated gastric cancer. Tumour Biol. *35*, 5501–5509.
35. Wang, J., Yang, Y.H., Wang, A.Q., Yao, B., Xie, G., Feng, G., Zhang, Y., Cheng, Z.S., Hui, L., Dai, T.Z., et al. (2010). Immunohistochemical detection of the Raf kinase inhibitor protein in nonneoplastic gastric tissue and gastric cancer tissue. Med. Oncol. *27*, 219–223.
36. Pfeuffer, I., Klein-Hessling, S., Heinfling, A., Chuvpilo, S., Escher, C., Brabletz, T., Hentsch, B., Schwarzenbach, H., Matthias, P., and Serfling, E. (1994). Octamer factors exert a dual effect on the IL-2 and IL-4 promoters. J. Immunol. *153*, 5572–5585.
37. Ullman, K.S., Flanagan, W.M., Edwards, C.A., and Crabtree, G.R. (1991). Activation of early gene expression in T lymphocytes by Oct-1 and an inducible protein, OAP40. Science *254*, 558–562.
38. Kim, Y.M., Ko, C.B., Park, Y.P., Kim, Y.J., and Paik, S.G. (1999). Octamer motif is required for the NF-kappaB-mediated induction of the inducible nitric oxide synthase gene expression in RAW 264.7 macrophages. Mol. Cells *9*, 99–109.
39. Ballard, D.W., and Bothwell, A. (1986). Mutational analysis of the immunoglobulin heavy chain promoter region. Proc. Natl. Acad. Sci. USA *83*, 9626–9630.
40. Xie, Q. (1997). A novel lipopolysaccharide-response element contributes to induction of nitric oxide synthase. J. Biol. Chem. *272*, 14867–14872.

Association of *HLA-DRB1*09:01* with tIgE levels among African-ancestry individuals with asthma

Nicolas Vince, PhD,[a] Sophie Limou, PhD,[a,p] Michelle Daya, PhD,[b] Wataru Morii, PhD,[c] Nicholas Rafaels, MS,[b] Estelle Geffard, MS,[a] Venceslas Douillard, MS,[a] Alexandre Walencik, PharmD,[a] Meher Preethi Boorgula, MS,[b] Sameer Chavan, MS,[b] Candelaria Vergara, MD, PhD,[d] Victor E. Ortega, MD, PhD,[e] James G. Wilson, MD,[f] Leslie A. Lange, PhD,[b] Harold Watson, MD,[g] Dan L. Nicolae, PhD,[h] Deborah A. Meyers, PhD,[i] Nadia N. Hansel, MD, MPH,[d] Jean G. Ford, MD,[j] Mezbah U. Faruque, MD, PhD,[k] Eugene R. Bleecker, MD,[i] Monica Campbell, MS,[b] Terri H. Beaty, PhD,[l] Ingo Ruczinski, PhD,[m] Rasika A. Mathias, ScD,[d,l] Margaret A. Taub, PhD,[m] Carole Ober, PhD,[n] Emiko Noguchi, MD, PhD,[c] Kathleen C. Barnes, PhD,[b] on behalf of CAAPA, Dara Torgerson, PhD,[o] and Pierre-Antoine Gourraud, PhD, MPH[a]

Nantes, France; Aurora, Colo; Ibaraki, Japan; Baltimore, Md; Winston-Salem, NC; Jackson, Miss; Bridgetown, Barbados; Chicago, Ill; Tucson, Ariz; Philadelphia, Pa; Washington, DC; and Montreal, Quebec, Canada

Background: Asthma is a complex chronic inflammatory disease of the airways. Association studies between *HLA* and asthma were first reported in the 1970s, and yet, the precise role of *HLA* alleles in asthma is not fully understood. Numerous genome-wide association studies were recently conducted on asthma, but were always limited to simple genetic markers (single nucleotide polymorphisms) and not complex *HLA* gene polymorphisms (alleles/haplotypes), therefore not capturing the biological relevance of this complex locus for asthma pathogenesis.

Objective: To run the first *HLA*-centric association study with asthma and specific asthma-related phenotypes in a large cohort of African-ancestry individuals.

Methods: We collected high-density genomics data for the Consortium on Asthma among African-ancestry Populations in the Americas (N = 4993) participants. Using computer-intensive machine-learning attribute bagging methods to infer *HLA* alleles, and Easy-HLA to infer *HLA* 5-gene haplotypes, we conducted a high-throughput HLA-centric association study of asthma susceptibility and total serum IgE (tIgE) levels in subjects with and without asthma.

Results: Among the 1607 individuals with asthma, 972 had available tIgE levels, with a mean tIgE level of 198.7 IU/mL. We could not identify any association with asthma susceptibility. However, we showed that *HLA-DRB1*09:01* was associated with increased tIgE levels ($P = 8.5 \times 10^{-4}$; weighted effect size, 0.51 [0.15-0.87]).

Conclusions: We identified for the first time an *HLA* allele associated with tIgE levels in African-ancestry individuals with asthma. Our report emphasizes that by leveraging powerful computational machine-learning methods, specific/extreme phenotypes, and population diversity, we can explore *HLA* gene polymorphisms in depth and reveal the full extent of complex disease associations. (J Allergy Clin Immunol 2020;146:147-55.)

Key words: *Asthma, HLA, tIgE levels, atopy, CAAPA, imputation, admixture*

Asthma is a complex chronic inflammatory disease of the airways presenting a large variety of phenotypes that can be divided between T_H2-high (eg, allergic asthma) and non-T_H2 (very late-onset).[1,2] The HLA molecules play a central role in the initiation and regulation of innate and acquired immune responses,[3,4] and, as such, have been extensively studied for their potential links with numerous diseases. According to the genome-wide association study (GWAS) catalog, discoveries inside the HLA genomic region account for 22% of all diseases and traits,[5] emphasizing the crucial role played by HLA in a large number of immune-related pathologies.[6] The study of potential associations between *HLA* and asthma was first reported in the

From [a]Université de Nantes, Centrale Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, Nantes; [b]the Department of Medicine, University of Colorado Denver, Aurora; [c]the Department of Medical Genetics, Faculty of Medicine, University of Tsukuba, Ibaraki; [d]the Department of Medicine, Johns Hopkins University, Baltimore; [e]the Department of Internal Medicine, Section on Pulmonary, Critical Care, Allergy and Immunologic Diseases, Center for Precision Medicine, Wake Forest School of Medicine, Winston-Salem; [f]the Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson; [g]the Faculty of Medical Sciences Cave Hill Campus, The University of the West Indies, Bridgetown; [h]the Department of Medicine, University of Chicago, Chicago; [i]the Department of Medicine, University of Arizona College of Medicine, Tucson; [j]the Department of Medicine, Einstein Medical Center, Philadelphia; [k]the National Human Genome Center, Howard University College of Medicine, Washington; the Departments of [l]Epidemiology and [m]Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore; [n]the Department of Human Genetics, University of Chicago, Chicago; [o]the McGill University and Genome Quebec Innovation Centre, Montreal, Quebec; and [p]Ecole Centrale de Nantes, Nantes.

N.V. has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement number 846520. Funding for this study was provided in part by the National Institutes of Health (grant nos. R01-HL129239 and R01HL104608 to K.C.B.).

Disclosure of potential conflict of interest: The authors declare that they have no relevant conflicts of interest.

Received for publication May 5, 2019; revised December 5, 2019; accepted for publication January 8, 2020.

Available online January 22, 2020.

Corresponding author: Pierre-Antoine Gourraud, PhD, MPH, ATIP-Avenir Team 5, CRTI UMR1064 - ITUN, CHU Nantes Hôtel Dieu, 30 Bld Jean Monnet, 44093 Nantes Cedex 01, France. E-mail: pierre-antoine.gourraud@univ-nantes.fr.

0091-6749/$36.00

https://doi.org/10.1016/j.jaci.2020.01.011

Abbreviations used
CAAPA: Consortium on Asthma among African-ancestry Populations in the Americas
GWAS: Genome-wide association study
tIgE: Total serum IgE

1970s,[7,8] and yet, the precise role of *HLA* alleles in asthma is not fully understood.[1] Difficulties for replication of associations include variability in asthma etiologies and biological characteristics,[2] small sample sizes,[1,9] population heterogeneity, the cost of *HLA* molecular typing, and the challenge of interpreting associations due to high allelic and structural variations and complex linkage disequilibrium patterns.[10] Overall, class II *HLA* alleles appear to be involved in late-onset and allergic asthma in European and Hispanic populations.[1,11-15]

With the generalization of genomic studies, an important number of GWASs were conducted on asthma in the past few years,[11,12,16-26] and notably reported several associations for single nucleotide polymorphisms (SNPs) within the major histocompatibility complex (MHC) region that includes the *HLA* genes. In particular, SNPs in class II *HLA* genes were associated with asthma in European and Japanese populations.[11,12,20,21,23-25] Variations in *HLA-DQ* appear to be the main asthma contributors identified through GWASs, because signals near this gene were found in European American, African American, and Latino populations, but with different SNPs identified in each ethnic group.[22] In addition, genetic ancestry at the *HLA* locus has been associated with both asthma and allergen-specific IgE levels in Latinos,[17,21] and other *HLA* variations also seem to play a role (eg, *HLA-G*[27] and *HLA-DRA*[18]). However, all these studies only focused on investigations at the SNP level, which do not convey HLA biological complexity and functional relevance. To fully capture the complex information related to antigen presentation and interactions with other immune-related molecules in the context of asthma, it is necessary to examine *HLA* alleles, defined by *HLA* gene sequence (www.ebi.ac.uk/ipd/imgt/hla/)[10] or SNP haplotypes, and *HLA* 5-gene haplotypes (*HLA-A*∼*HLA-B*∼*HLA-C*∼*HLA-DQB1*∼*HLA-DRB1*).

Because of the high degree of diversity in the *HLA* region, *HLA* alleles exhibit relatively low frequencies and only the top 10 most frequent 5-gene haplotypes are in the 1% frequency range[28]; therefore, a very large sample size or a large effect size is necessary for adequate statistical power to detect an association between an *HLA* allele and asthma. However, collecting a very large cohort with asthma often implies a mix of diverse phenotypes, which may introduce heterogeneity of genetic causes and lead to false-negative signals. Focusing on specific or extreme phenotypes may then contribute to find specific associations,[29,30] even though the working sample size is reduced.

Here, we leveraged genomic data generated by CAAPA (the Consortium on Asthma among African-ancestry Populations in the Americas) to explore for the first time the role of *HLA* alleles in asthma and specific asthma-related phenotypes in a large cohort of African-ancestry individuals (N = 4993).[31] CAAPA recently published a GWAS that found strong evidence for association at 4 previously reported asthma loci (whose discovery was driven largely by non-African populations), and also identified 2 novel loci that may be specific to asthma risk in African-ancestry populations.[31] Here, our aims were to deliver the first high-throughput HLA-centric study of asthma outcomes by performing a case/control analysis of asthma susceptibility and a total serum IgE (tIgE)-level quantitative analysis in subjects with asthma to explore atopy. Our hypothesis was that *HLA* alleles are associated with asthma phenotypes. We also hypothesized that a classical GWAS design cannot identify these associations because GWASs focus on simple genetic markers (SNPs) that do not fully recapitulate the complexity of the HLA genomic region.

METHODS

CAAPA

The CAAPA multicenter participants were previously described.[31] A total of 1607 cases with asthma and 3365 controls of African ancestry (the United States and Barbados) were recruited (for a full description, see Table E1 in this article's Online Repository at www.jacionline.org and Daya et al[31]). Briefly, 8 of the 17 CAAPA investigators contributed data to this study. The distributions of age, sex, and age of asthma onset for each cohort are summarized in Table E1. Participants in these studies were unrelated except for the Barbados Asthma Genetics Study (BAGS) and the Howard University Family Study (HUFS), which included families; however, we only included the founder's individuals to perform our statistical analyses on unrelated subjects. Childhood-onset asthma is defined as having been diagnosed with asthma before age 16 years. Studentized residuals of $\log_{10}$-transformed tIgE (adjusted for age and sex as previously described[32]) were available from 4 of 8 CAAPA studies (for mean and standard error of tIgE before transformation, see Table E1, and for population distribution, see Fig E8 in this article's Online Repository at www.jacionline.org). We defined atopy as tIgE level greater than 80 KU/L. Previous work has shown that the genetic structure of the Barbados population is similar to that of African Americans, with subjects from Barbados having on average higher proportions of African ancestry compared with African Americans.[19,31] Asthma cases were defined by reported and documented histories of current or past physician-diagnosed asthma, whereas controls reported a negative history of asthma. All participants provided written informed consent. This study is an initiative of CAAPA, which was funded by the National Institutes of Health (grant no. R01 HL104608). National Institutes of Health guidelines for conducting human genetic research were followed. The institutional review boards of Johns Hopkins University (for the Genomic Research on Asthma in the African Diaspora [GRAAD], BAGS, and The Asthma Biorepository For Integrative Genomic Exploration [BRIDGE] cohorts), Howard University (for the GRAAD and HUFS cohorts), Wake Forest University (for the Severe Asthma Research Program [SARP] cohort), the University of Chicago (for the Chicago Asthma Genetics [CAG] cohort), University of the West Indies, Cave Hill Campus, Barbados (for the BAGS cohort), and University of Mississippi Medical Center (for the Jackson Heart Study [JHS] and Atherosclerosis Risk in Communities Study [ARIC] cohorts) all reviewed and approved this study.

Genotyping and SNP imputation

Details of genotyping and SNP imputation can be found in the GWAS article.[31] Briefly, each CAAPA study was separately genotyped on various GWAS chips, as well as genotyped on an African-ancestry specific chip (African Diaspora Power Chip).[33] All these genomic data sets were imputed using the CAAPA whole-genome sequence reference panel on the Michigan imputation server (https://imputationserver.sph.umich.edu).[34]

HLA allele SNP-based imputation and haplotyping

The HIBAG R package is designed to impute *HLA* alleles from SNP genotypes using a reference panel (or bagging set) built with an attribute bagging machine-learning technique.[35,36] Appropriate reference panels matching the population of interest are crucial to obtain accurate imputation. Several reference panels are available publicly (http://www.biostat.washington.edu/~bsweir/HIBAG/) but are mostly suitable for European ancestry imputation.

Because of the complexity and diversity of African-ancestry genomes, large reference panels are necessary and are yet to be collected, with the currently available panel being derived from only approximately 150 African-ancestry individuals. To maximize imputation quality, we created our own African-ancestry reference panel from a subset of 917 CAAPA individuals from whom we had high-resolution genotyped *HLA* alleles (second field or 4-digit) and *MHC* SNP genotypes (accessed from dbGAP, phs001123.v1.p1). The *HLA* alleles were called with the Omixon software (Budapest, Hungary) from whole-genome sequencing data, and we selected 29,970 *MHC* SNPs overlapping between the whole-genome sequencing data and the GWAS data. From this large HLA + SNP data set, we used HIBAG to train statistical models called bags[35] that will serve as a reference for *HLA* allele imputation from SNP data. With this strategy, we obtained 5 bagging sets corresponding to the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1*, and *HLA-DRB1* genes (available on request). The bagging step is computationally very demanding because it required 30,000 central processing unit hours on 700 central processing units to build a full reference panel on the 5 HLA genes (see Table E2 in this article's Online Repository at www.jacionline.org). We then imputed *HLA* alleles in all available CAAPA samples (N = 4993; Table E1), and selected a postprobability threshold greater than or equal to 0.5 as the recommended compromise between good accuracy imputation and call rate from Zheng et al[35] (see Tables E2 and E3 in this article's Online Repository at www.jacionline.org). Choosing a higher postprobability threshold results in marginal accuracy improvements but in a large call rate drop.[35]

The imputed *HLA* alleles were then uploaded in our in-house Web application, Easy-HLA (http://hla.univ-nantes.fr/),[37] to infer the *HLA* 5-gene haplotypes (*HLA-A*∼*HLA-B*∼*HLA-C*∼*HLA-DQB1*∼*HLA-DRB1*) by extending the method of Gourraud et al.[38] We identified a total of 4077 unique 5-gene haplotypes, including 2698 haplotypes with only 1 occurrence (frequency = 1.1×10^{-4}) and up to 1 haplotype (30:01∼42:01∼17:01∼03:02∼04:02) with 149 occurrences (frequency = 0.017). To increase the power of detecting an association with *HLA* haplotypes, we chose to restrict the regression analyses to 27 haplotypes with at least 20 occurrences, representing a frequency of 0.2% in the pooled CAAPA individuals.

A complete workflow of the analyses is available as Fig E1 in this article's Online Repository at www.jacionline.org.

Ancestry estimation

Details of local ancestry inference can be found in the GWAS article.[31] Briefly, after performing SNP phasing using ShapeIT,[39] the 4993 genomes were merged with African and European genomes from the 1000 Genomes project.[40] RFMix[41] inferred 15,824 local ancestry segments across the CAAPA genomes, including 20 segments in the *MHC* region on chromosome 6. The mean proportion of African ancestry was calculated from the RFMix estimation both genome-wide and in the *MHC* region. The local ancestry estimate was further used to correct for population differences in statistical analyses.

Statistical analysis

We performed association analyses on *HLA* alleles with good imputation accuracy (postprobability threshold ≥ 0.5) that were exhibiting an overall (case + controls) frequency of greater than or equal to 2% (54 *HLA* alleles in total; see Table E4 in this article's Online Repository at www.jacionline.org). For each CAAPA study and each *HLA* allele, we performed a case-control (overall: 1607 cases vs 3365 controls; see Table E1 and Fig E1) analysis in R[42] using logistic regression models to test for association with asthma susceptibility. From RFMix inference (see Ancestry estimation above), we have the information of local ancestry within *MHC*. To account for ancestry, we applied 2 strategies: (1) stratification on local ancestry only from African origin and (2) local ancestry as covariate in the regressions. The cases with asthma were further divided between childhood-onset (N = 804) and adult-onset (N = 445) asthma and compared with controls.

Studentized residuals of $\log_{10}$-transformed tIgE (adjusted for age and sex as described previously[32]) was available from 4 of 8 CAAPA studies. We further transformed this adjusted tIgE level into a *z* score[43] to allow comparison between cohorts (see Fig E8). Each *HLA* allele was then tested for association with tIgE levels separately in those with asthma (N = 972, Table E1) and those without asthma (N = 816) using a linear regression model. We accounted for ancestry as described for case-control analyses. Atopy was defined as tIgE level greater than 80 KU/L and was explored only in cases with asthma (atopic individuals = 725, and nonatopic individuals = 247).

We then performed a meta-analysis using METAL[44] for both case-control and tIgE statistical tests. We used the default sample size–based method from the METAL tool. Z scores in Tables E5 and E6 in this article's Online Repository at www.jacionline.org are a weighted sum of *Z* scores across studies, with each *Z* score reflecting the direction of effect and *P* value.[44] We also computed a weighted effect size using the inverse variance–based method from the METAL tool for the most significant alleles. We established a stringent Bonferroni threshold of significance at $P = 9.3 \times 10^{-4}$ (0.05/54) accounting for the number of alleles tested.[45,46]

RESULTS

Description of the study population

In our study, we had access to a total of 4993 individuals of African ancestry (the United States and Barbados) from 8 CAAPA studies (Table E1). Overall, our patients were 39.3% males and were recruited at age 38.7 years on average (for a description of the demographic characteristics in each cohort, see Table E1). Our analyses only focused on unrelated individuals, with a total of 1607 subjects with asthma and 3365 controls (BAGS and HUFS cohorts included some family members, and these individuals were excluded from our data set). Finally, tIgE data were available for 972 subjects with asthma and 816 controls. In this subset group, median (interquartile range) tIgE level was 214.0 (69.0-590.7) IU/mL for cases and 66.4 (25.0-212.2) IU/mL for controls (for a description in each cohort, see Fig E8 and Table E1). The workflow of the analysis is described in Fig E1.

Admixture in the *MHC* region

Because of the recent history of admixture between African and European populations, individuals living in the Americas with African ancestry show a high proportion of European ancestry (on average 20% of European ancestry across the genome for African Americans and African Caribbean).[47] The specific admixture structure within the *MHC* genomic region has not been studied in detail before. To understand the local ancestry structure in the *MHC* and the sensitivity of association results to differences in local ancestry, we inferred the local ancestry for the unrelated CAAPA individuals (full data set, N = 4993) both across the whole genome and specifically within the *MHC* region, and compared their admixture structure (Fig 1, *A*).

The whole-genome admixture observed in CAAPA was typical of African-ancestry individuals with a history of European admixture (Fig 1, *A*, *top panel*), and represents 81.9% on average with a median of 83.9% (interquartile range, 77.3-89.0) African ancestry across the genome for CAAPA individuals (Fig 1, *B*), as previously described in Mathias et al.[47] As expected, the *MHC* region showed a nonnormally distributed admixture among CAAPA individuals with a particular trimodal distribution (Fig 1, *A*, *bottom panel*) that reflects the diploid origin of a small portion of the genome (1/1000). Indeed, we obtained similar trimodal distribution patterns with similarly sized chromosome 6 segments (5Mb, Fig 1, *C*). Within the *MHC*, we observed European/

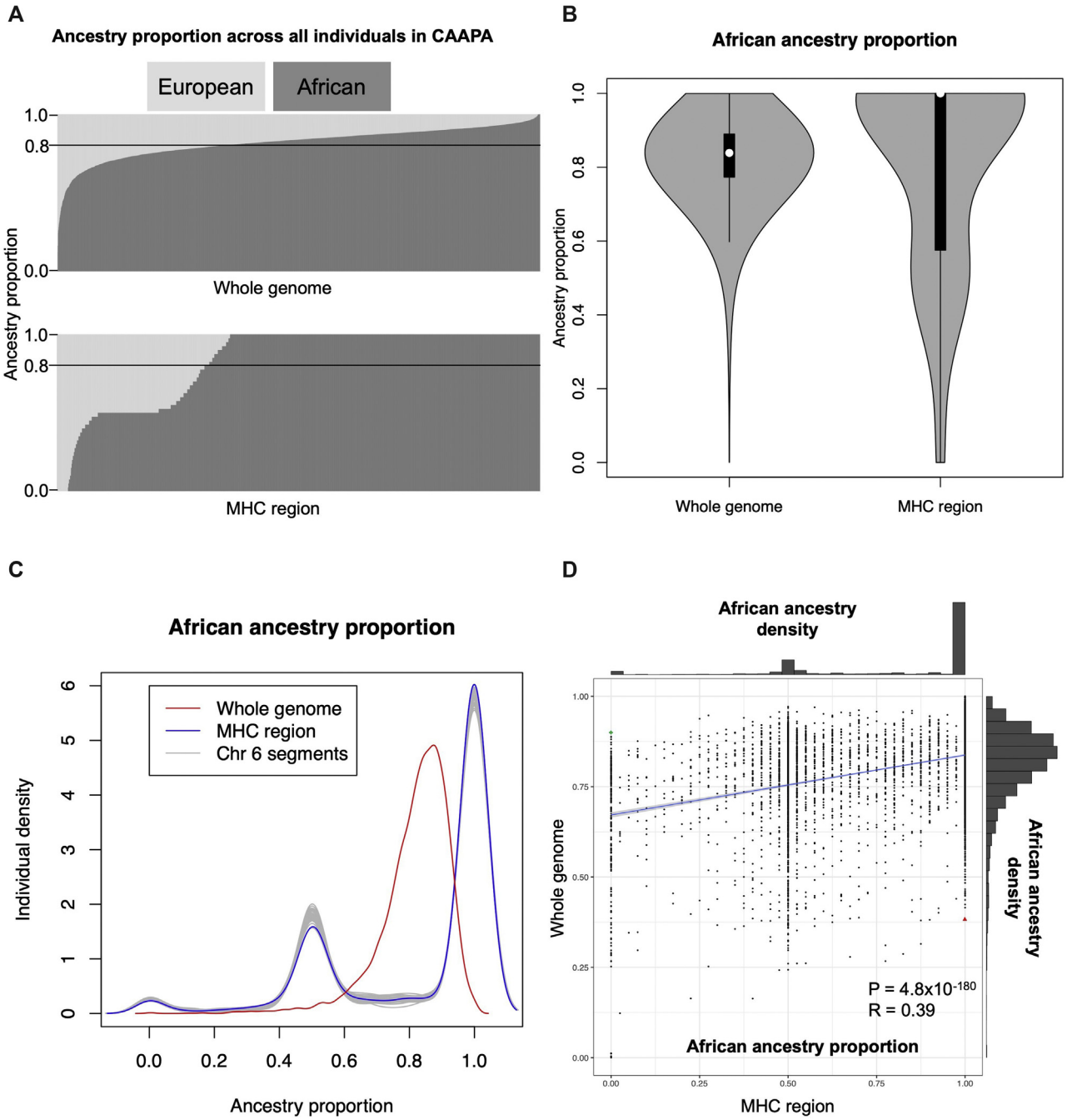

FIG 1. Ancestry proportions are equivalent between whole-genome and *MHC* genomic region. **A,** CAAPA individuals' ancestry proportion (full data set, N = 4993). X axis: each individual. Y axis: ancestry proportion. Black horizontal line: 80% African ancestry. The *MHC* only represents a thousandth of the genome (5Mb on chromosome 6). Mean local ancestry: light gray for European, dark gray for African. **B,** Violin plot comparing African- ancestry proportion. Median: white dots. Interquartile range: black boxes. Black lines: 95% CI. **C,** Density plot comparing African-ancestry proportion between whole-genome (red), MHC region (blue), and similarly sized chromosome 6 segments (gray, N = 42 segments). **D,** The graph combines (1) a scatter plot comparing African-ancestry proportion across all CAAPA individuals and (2) a density plot showing the proportion of individuals with a specific level of African ancestry in whole-genome (*right*) and *MHC* genomic region (*top*). The blue line was drawn with the correlation coefficient. Pearson correlation: $P = 4.8 \times 10^{-180}$, $R = 0.39$ (95% CI, 0.37-0.41). Two extreme individuals are represented: red triangle (100% African ancestry MHC, 41.5% whole genome), green diamond (0% African ancestry MHC, 89.9% whole genome).

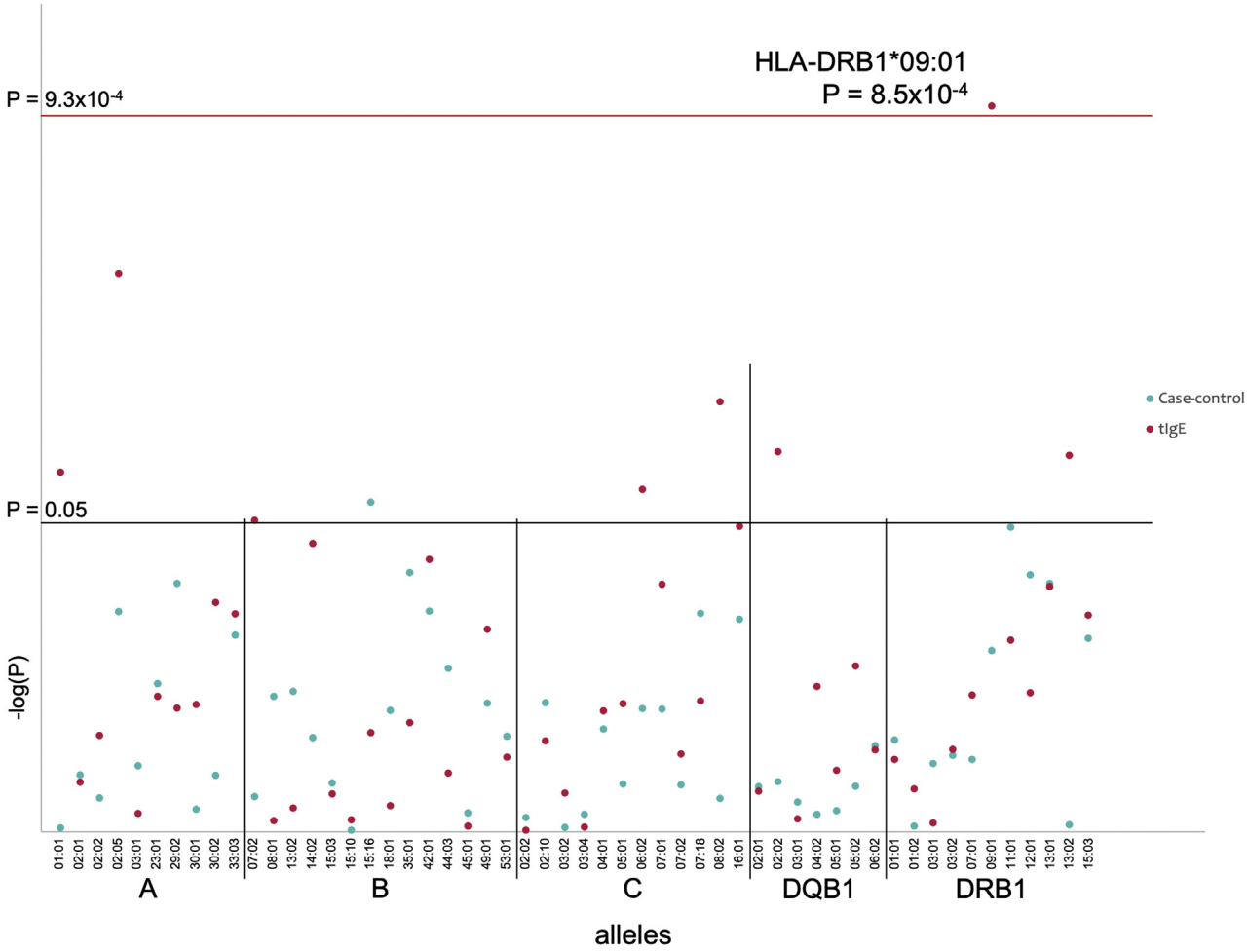

FIG 2. *HLA-DRB1*09:01* allele associates with asthma atopy. Manhattan plot of each *HLA* allele association for case-control analysis (blue) and tIgE analysis (red). *P* values are meta-analysis from regression analyses stratified for African-ancestry chromosomes and represented as −log10 of *P* values. Red horizontal line represents Bonferroni threshold of significance ($P = 9.3 \times 10^{-4}$). Black horizontal line represents nominal threshold of significance ($P = .05$). Each *HLA* allele tested is represented.

European ancestry (100% light gray, 2.3% of the population), African/African ancestry (100% dark gray, 64.1% of the population), European/African ancestry (50/50, 12.6% of the population), and admixed individuals carrying a mosaic of European and African *MHC* fragments emphasizing the recombination events in that locus (21.0% of the population, Fig 1, *A*, *bottom panel*). Importantly, similar distributions were observed across the individual CAAPA studies (see Fig E3 in this article's Online Repository at www.jacionline.org), indicating their homogeneity regarding admixture structure.

Although the distribution patterns looked different, the whole-genome African ancestry proportion correlated significantly with the *MHC* region African-ancestry proportion (Fig 1, *D*; Pearson correlation: $R = 0.39$; 95% CI, 0.37-0.41; $P = 4.8 \times 10^{-180}$). This result indicates that the *MHC* admixture distribution correlates to the whole-genome distribution: individuals above whole-genome African-ancestry proportion median (83.9%) are carrying a homozygous *MHC* region segment of African origin for 77.6% of them compared with 50.8% for individuals below the median (fold increase of 1.5). However, it is important to emphasize that this very significant correlation does not preclude specific extreme cases such as the following 2 cases: (1) the red triangle on Fig 1, *D*, represents an individual with 100% African ancestry within the MHC region but 41.5% across the whole genome; (2) at the opposite, the green diamond on Fig 1, *B*, represents an individual with 0% African ancestry within the MHC region but 89.9% across the whole genome.

HLA alleles' imputation

The imputation of *HLA* alleles is a powerful computational technique to infer *HLA* alleles only from genotyped SNPs of the *MHC* using a reference panel.[35] The accuracy of the imputation depends on the reference panel size and matching with the population of interest. Because the publicly available African ancestry reference panel was small (~150) and did not yield accurate *HLA* allele imputation, and because the allelic diversity is larger in African ancestry individuals compared with Europeans, we developed our own reference panel from a subset of 917 CAAPA individuals for whom we had genotyped *HLA* alleles in addition to the *MHC* SNP genotypes. Our large matching reference panel provided a much-improved quality of imputation (Table E3), with an overall call rate of 81% (compared with

TABLE I. *P* values of previously found asthma-associated *HLA* alleles tested in our study

HLA	Case vs control meta-analysis *P* value	tIgE level meta-analysis *P* value	Population	PubMed ID
DQB1*05:01	.82	.55	Hispanic, European	10051703,[9] 17686102[14]
DRB1*01:02	.95	.66	Hispanic, European	12878360,[15] 22397267,[13] 17686102[14]
DRB1*13:01	.09	.09	European	17686102[14]

P values are from the meta-analysis on case-control and tIgE levels in CAAPA individuals with asthma using an allelic model and stratifying for 2 African-ancestry chromosomes.

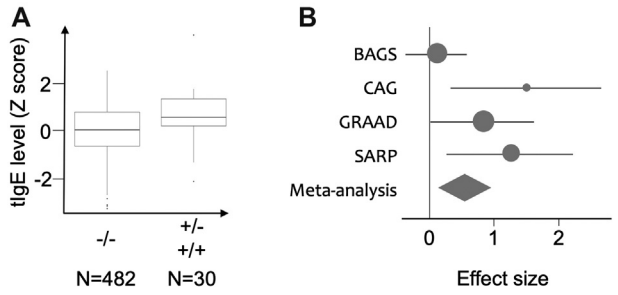

FIG 3. *HLA-DRB1*09:01* allele associates with higher tIgE. **A,** Box plot comparing tIgE level between those with asthma with 1 or 2 *HLA-DRB1*09:01* allele (+/− and +/+) and those with asthma without HLA-DRB1*09:01 allele (−/−). tIgE level is expressed in *z* score. The N representing the number of individuals in each group is depicted on the graph; these numbers exclude *HLA-DRB1* alleles with a postprobability of less than 0.5 and individuals with 2 chromosomes from African ancestry (see Fig E2 in this article's Online Repository at www.jacionline.org). **B,** Forest plot representing the association between HLA-DRB1*09:01 and tIgE in the different CAAPA cohorts where tIgE is available. weighted effect size 0.51 (0.15 to 0.87). BAGS effect size 0.07 (−0.41 to 0.56). CAG effect size 1.48 (0.31 to 2.64). GRAAD effect size 0.80 (0.00 to 1.60). SARP effect size 1.23 (0.24 to 2.21). *GRAAD,* Genomic Research on Asthma in the African Diaspora.

42% with the public embedded reference panel). This result emphasizes the need to develop and share large and diverse reference panels to improve *HLA* imputation accuracy (our panel is available on request). This innovative approach therefore generated accurate *HLA* alleles for all CAAPA individuals, including 54 common *HLA* alleles with a frequency greater than or equal to 2% that were further tested for association with asthma-related phenotypes.

Asthma susceptibility

For each *HLA* allele, we conducted a logistic regression analysis comparing those with asthma and those without asthma while stratifying by 2 African ancestry chromosomes (resulting in 1178 cases and 2421 controls).

First, we assessed the association of *HLA* alleles under an allelic model with asthma susceptibility in each CAAPA cohort using logistic regression. The results obtained in each study (Table E1) were subsequently combined by meta-analysis (Fig 2). The Quantile Quantile plot showed no inflation of *P* values, demonstrating the fitness of our statistical model (lambda 0.93, META in Fig E4 in this article's Online Repository at www.jacionline.org). No *HLA* allele passed the Bonferroni threshold of significance ($P \leq 9.3 \times 10^{-4}$), and the best association was identified for *HLA-B*15:16* ($P = .04$; weighted effect size, 0.43 [0.04-0.83]; Fig 2). In this setting, we expect a statistical power of greater than 90% to detect HLA associations with an allele frequency of 5% and an effect size of 1.5. Running the analysis by correcting for local ancestry (see Fig E6 in this article's Online Repository at www.jacionline.org) or under a dominant model (see Figs E5 and E7 in this article's Online Repository at www.jacionline.org) did not reveal any significant signals (full results presented in Table E5).

We then inferred *HLA* 5-gene haplotypes (*HLA-A∼HLA-B∼HLA-C∼HLA-DQB1∼HLA-DRB1*) for each CAAPA individual using the local version of an HLA-oriented Web application (http://hla.univ-nantes.fr/). No statistically significant associations were observed.

Next, we investigated age-at-onset by running a stratified case-control association analysis (N = 804 childhood onset and N = 445 adult onset) but no significant associations were observed.

Finally, we focused specifically on the *HLA* alleles that were previously associated with asthma-related phenotypes in European and Asian populations[1,9,13-15] (Table I), but did not identify any significant associations.

Total IgE levels in individuals with asthma

We next explored *HLA* associations with asthma atopy by focusing on the tIgE levels (available in 4 of 8 CAAPA studies) among subjects with asthma. tIgE levels are higher in those with asthma compared with those without asthma (Fig E8), and high tIgE levels associate with more severe asthma and lower baseline lung function.[48] For each *HLA* allele, we conducted a linear regression analysis in those with asthma using tIgE as a continuous variable while stratifying for 2 African ancestry chromosomes (resulting in 724 individuals). In the meta-analysis, *HLA-DRB1*09:01* was significantly associated with elevated tIgE levels under an allelic model ($P = 8.5 \times 10^{-4}$; weighted effect size, 0.51 [0.15-0.87]; moderate heterogeneity across studies $P = .04$; Figs 2 and 3; see Fig E9 in this article's Online Repository at www.jacionline.org), indicating that this allele is associated with degree of atopy in individuals with asthma. Association effects were in the same direction across all cohorts and *HLA-DRB1*09:01* reached nominal significance in 2 cohorts ($P = .02$ in CAG and $P = .02$ in SARP). We obtained similar results by correcting for local ancestry (see Fig E11 in this article's Online Repository at www.jacionline.org) or using a dominant model (see Figs E10 and E12 in this article's Online Repository at www.jacionline.org; see Table E6 for full results). This association is specific to the population with asthma, because no association was found between *HLA-DRB1*09:01* (or any other *HLA* allele) and tIgE in the controls (N = 816). *HLA-DRB1*09:01* had a frequency of 3.8% in the overall CAAPA study (Table E4), 3.0% in the tIgE CAAPA subgroup (Table E6), 3.1% in the *HLA* genotyped CAAPA subset (N = 917), and 3.0% in a reference African American data set (http://www.allelefrequencies.net, USA National Marrow Donor Program African American),[49] confirming the quality of our imputation and the confidence in this

association signal. To test for possible biases between cohorts with and without tIgE availability, we performed an asthma case/control analysis focusing on the 4 cohorts with tIgE-level data. The results did not significantly differ from the whole data analysis and there was still no significant signal identified (see Table E7 in this article's Online Repository at www.jacionline.org). In a sensitivity analysis, we increased the postprobability threshold to test whether our signal would hold. The direction and strength of the effect is conserved up to a greater than 0.7 and greater than 0.6 threshold, respectively; higher thresholds result in a large call rate drop (71% of HLA-DRB1 alleles reach a postprobability > 0.5 and 30% a postprobability > 0.8; see Table E8 in this article's Online Repository at www.jacionline.org).

Finally, the *HLA* alleles that were previously associated with asthma-related phenotypes in European and Asian populations[1,9,13-15] were not associated with tIgE levels in our population (Table I). Using tIgE levels, we also stratified our cohort into atopic (N = 725) and nonatopic (N = 247) individuals. Performing atopic versus nonatopic analysis did not reveal any significant results, likely due to a lack of statistical power.

DISCUSSION

Our study focused for the first time on *HLA* allele associations with asthma susceptibility and atopy in a large African ancestry cohort of individuals with asthma (CAAPA, N = 4993). We identified the *HLA-DRB1*09:01* allele associated with elevated tIgE levels in individuals with asthma.

Thus far, most *HLA* studies in asthma either had small sample sizes or assessed associations only at the SNP level, therefore not fully capturing the complexity and biological relevance of the *HLA* genes in the *MHC* region. Moreover, because previous studies focused on populations of European and Asian descent, studying a large population of African ancestry is of great interest because it can shed new light on the relationship between HLA and asthma. In the recently published CAAPA GWAS (of the same data),[31] Daya et al[31] did not identify any significant associations for SNPs inside the *HLA* region (best SNP: rs9272346, P = .03). This reinforced our strategy to study *HLA* alleles. To investigate the specific role of HLA in asthma, we used a powerful machine-learning–based computational technique to infer *HLA* alleles. We developed a new reference panel matching individuals of African ancestry to increase imputation accuracy in this (genetically diverse) population. We inferred *HLA* alleles with at least 86% accuracy for the *HLA-A*, *HLA-B*, *HLA-C*, *HLA-DQB1*, and *HLA-DRB1* genes in all CAAPA studies (Table E2).

Because the African American and Barbados African Caribbean genomes are a mosaic of European and African ancestral genomes, we first explored the specific admixture structure within the *MHC* genomic region. Because *HLA* genes are under strong selection pressure,[50] one could have expected differences between the *MHC* region and the rest of the genome. Our analyses revealed a similar proportion of admixture as in the rest of the genome (around 20% of European ancestry), and an admixture pattern that is comparable to similarly sized segments on chromosome 6 (Fig 1, *C*). This indicates that the *MHC* does not seem to be under ancestry-specific selective pressure compared with the rest of the genome and that we can analyze this genomic region as any other part of the genome. Yet, we must acknowledge the fairly recent history of admixture in African Americans and African Caribbeans, having occurred only in the last few centuries, which does not exclude the possibility of detecting ancestry-specific selection in the future, after many more generations. In our analyses to account for admixture, we applied 2 strategies: (1) stratification on local ancestry only from African origin and (2) local ancestry as covariate in the regressions (see the Methods section).

Our case-control analyses (1607 vs 3365) did not reveal any significant *HLA* associations with asthma susceptibility. Focusing on tIgE levels to study atopy in individuals with asthma, we restricted our sample size to only 972 patients with asthma from 4 studies (724 after African local ancestry stratification) and found *HLA-DRB1*09:01* associated with an increased level of tIgE in blood ($P = 8.5 \times 10^{-4}$). This P value reaches our Bonferroni threshold (9.3×10^{-4}) of significance calculated by dividing 0.05 by the number of tested alleles (N = 54). This association was not found in control individuals, indicating the specificity of the signal for atopy in individuals with asthma. These results illustrate how an extreme phenotype (high tIgE levels associated with increased asthma severity[48]) such as atopy can maximize the statistical power of detection by limiting the noise due to the diversity of asthmatic phenotypes.

We have to acknowledge the difficulty of finding additional replication cohorts, with CAAPA standing as an exceptionally large data set. The *HLA-DRB1*09:01* allele is relatively rare in African and European populations (1%-3%) but frequent in East Asian populations (15%-25%).[28] In a validating cohort of 544 Japanese subjects including 103 patients with asthma with tIgE,[51] we did not identify any significant association (tIgE association with *HLA-DRB1*09:01*, P = .49, frequency = 11.8%). This is probably due to lack of power, but also could be explained by genetic background difference between African ancestry and Asian ancestry individuals or dissimilar environmental effects between the Japanese islands and the Americas.

Previous reports showed that *HLA-DRB1*09:01* associated with several autoimmune diseases (rheumatoid arthritis, systemic lupus erythematosus,[52] type 1 diabetes—especially late-onset,[53] and dermatomyositis[54]), especially in Asian populations where the allele is more frequent. This suggests that HLA-DRB1*09:01 carriers are at increased risk for self-peptide presentation and/or exacerbated chronic inflammation. This is in line with the pleiotropic effects observed in autoimmune diseases where some factors, especially genetics, share pathophysiological mechanisms across a span of autoimmune diseases.[55]

Interestingly, *HLA-DRB1*09:01* was identified as a susceptibility allele to food allergy (peach allergy).[56] This correlates with our findings of increased tIgE levels for *HLA-DRB1*09:01* carriers. Finally, in the immune epitope database (iedb.org),[57] $CD4^+$ T-cell epitopes against common environmental allergens were identified using Tetramer Guided Epitope Mapping. Among 86 epitopes, 4 Derp2 overlapping epitopes showed evidence of specific binding with HLA-DRB1*09:01 tetramers (direct submission to the immune epitope database, reference ID: 1024966).[57,58] Derp2 is a well-known allergen from house-dust mite and a strong inducer of allergic asthma in susceptible individuals, which can lead to the immune system activation and elevation of tIgE level.[59,60] In the context of HLA-DRB1*09:01, this represents another lead in the possible mechanism of action of this allele in asthma atopy. The allele could be responsible for an increased activation of $CD4^+$ T cells

compared with other alleles, by presentation of specific dust mite peptides, which would trigger an amplified inflammation environment and result in an increased production of IgE by B cells. Further experiments are needed to confirm the association and define the functional relationship between the HLA-DRB1*09:01 allele and tIgE.

We have to acknowledge some limitations in our study. Even if this is the largest ever African ancestry–oriented study of association of HLA alleles with asthma, this study remains restricted to a relatively small sample size, especially for samples with tIgE levels available (N = 972 individuals with asthma). However, this study conveys the ambitious will of gathering data from several research groups across multiple cities and countries, which appears to be the future of genetics studies. The large diversity of cohorts is also a limitation both in terms of clinical definitions (different clinicians, definitions evolve through time) and in terms of genetics and environmental backgrounds. These effects were limited here by using a meta-analysis technique and corrections from genetically computed ancestry. The main difficulty remains to find an appropriate replication cohort when the goal of CAAPA was already to assemble as many cohorts as possible.

In conclusion, we identified for the first time *HLA-DRB1*09:01* allele association with tIgE levels in African ancestry individuals with asthma, potentially through a mechanism involving specific peptide presentation and/or increasing inflammation.

We thank Labex IGO (ANR-11-LABX-0016-01) and IHU-CESTI for their support. We gratefully acknowledge the contributions of Pissamai and Trevor Maul, Paul Levett, Anselm Hennis, P. Michele Lashley, Raana Naidu, Malcolm Howitt, and Timothy Roach (BAGS), Audrey Grant, Eduardo Viera Ponte, Alvaro A. Cruz, and Edgar Carvalho (BIAS), Susan Balcer-Whaley, Maria Stockton-Porter, and Mao Yang (GRAAD), Delmy-Aracely Mejía-Mejía, Mario Meraz, Jaime Nuñez, and Eileen Fabiani Herrera Mejía (HONDAS), Trevor Ferguson and Deanna Ashley (JAAS), Silvia Jimenez, Nathalie Acevedo, and Dilia Mercado (PGCA), Ann Jedlicka (REACH), Addison K. May, Caroline Gilmore, and Patricia Minton (Vanderbilt University), Qun Niu (University of Chicago), and Adeyinka Falusi and Abayomi Odetunde (University of Ibadan, Nigeria). We also acknowledge the support of John Jay Shannon (Cook County Health Systems) and Kevin Weiss (Northwestern University), Regina Miranda and the Indians Zenues guards (San Basilio de Palenque, Bolivar, Colombia), Ulysse Ateba Ngoa (Leiden University), and Charles Rotimi, Adeyemo Adebowale, Floyd J. Malveaux, and Elena Reece (Howard University). We thank the numerous health care providers and community clinics and coinvestigators who assisted in the phenotyping and collection of DNA samples, and the families and patients for generously donating DNA samples to BAGS, BIAS, BREATHE, CAG, GRAAD, HONDAS, Jackson Heart Study, REACH, VALID, SARP, COPDGene, JAAS, PGCA, AEGS, and the asthma studies in Gabon and Palenque, Colombia. Special thanks to community leaders, teachers, doctors, and personnel from health centers at the Garifuna communities for organizing the medical brigades and to the medical students at Universidad Católica de Honduras, Campus San Pedro y San Pablo for their participation in the fieldwork related to HONDAS; study coordinator Sandra Salazar; and health liaisons and public health officers of the main Conde office, Adaliudes Conceição, Luciana Quintela, Ivanice Santos, Analú Lima, Benivaldo Valber Oliveira Silva, and Iraci Santos Araujo, and students from the Federal University of Bahia who assisted in data collection in BIAS: Rafael Santana, Roberta Barbosa, Ana Paula Santana, Charlton Barros, Marcele Brandão, Ludmila Almeida, Thiago Cardoso, and Daniela Costa. We are grateful for the support from the international state governments and universities from Honduras, Colombia, Brazil, Gabon, Nigeria, the Netherlands, Jamaica, Barbados, and the United States who made this work possible.

Key messages

- **Using a machine learning–based method and a matching reference population of African ancestry, we were able to impute HLA alleles from SNP data with good accuracy in the CAAPA data set. The reference population was instrumental to reach high prediction accuracy.**
- **The *HLA-DRB1*09:01* allele significantly associated with an increase in tIgE levels in patients of African ancestry with asthma.**

REFERENCES

1. Kontakioti E, Domvri K, Papakosta D, Daniilidis M. HLA and asthma phenotypes/endotypes: a review. Hum Immunol 2014;75:930-9.
2. Wenzel SE. Asthma phenotypes: the evolution from clinical to molecular approaches. Nat Med 2012;18:716-25.
3. Hansen TH, Bouvier M. MHC class I antigen presentation: learning from viral evasion strategies. Nat Rev Immunol 2009;9:503-13.
4. Jones EY, Fugger L, Strominger JL, Siebold C. MHC class II proteins and disease: a structural perspective. Nat Rev Immunol 2006;6:271-82.
5. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42:D1001-6.
6. Trowsdale J, Knight JC. Major histocompatibility complex genomics and human disease. Annu Rev Genomics Hum Genet 2013;14:301-23.
7. Brostoff J, Mowbray JF, Kapoor A, Hollowell SJ, Rudolf M, Saunders KB. 80% of patients with intrinsic asthma are homozygous for HLA W6. Is intrinsic asthma a recessive disease? Lancet Lond Engl 1976;2:872-3.
8. Bruce CA, Bias WB, Norman PS, Lichtenstein LM, Marsh DG. Studies of HLA antigen frequencies, IgE levels, and specific allergic sensitivities in patients having ragweed hayfever, with and without asthma. Clin Exp Immunol 1976;25:67-72.
9. Lara-Marquez ML, Yunis JJ, Layrisse Z, Ortega F, Carvallo-Gil E, Montagnani S, et al. Immunogenetics of atopic asthma: association of DRB1*1101 DQA1*0501 DQB1*0301 haplotype with Dermatophagoides spp.-sensitive asthma in a sample of the Venezuelan population. Clin Exp Allergy 1999;29:60-71.
10. Robinson J, Halliwell JA, Hayhurst JD, Flicek P, Parham P, Marsh SGE. The IPD and IMGT/HLA database: allele variant databases. Nucleic Acids Res 2015;43:D423-31.
11. Hirota T, Takahashi A, Kubo M, Tsunoda T, Tomita K, Doi S, et al. Genome-wide association study identifies three new susceptibility loci for adult asthma in the Japanese population. Nat Genet 2011;43:893-6.
12. Moffatt MF, Gut IG, Demenais F, Strachan DP, Bouzigon E, Heath S, et al. A large-scale, consortium-based genomewide association study of asthma. N Engl J Med 2010;363:1211-21.
13. Ivković-Jureković I, Zunec R, Balog V, Grubić Z. The distribution of HLA alleles among children with atopic asthma in Croatia. Coll Antropol 2011;35:1243-9.
14. Munthe-Kaas MC, Carlsen KL, Carlsen KH, Egeland T, Håland G, Devulapalli CS, et al. HLA Dr-Dq haplotypes and the TNFA-308 polymorphism: associations with asthma and allergy. Allergy 2007;62:991-8.
15. Torío A, Sánchez-Guerrero I, Muro M, Villar LM, Minguela A, Marín L, et al. HLA class II genotypic frequencies in atopic asthma: association of DRB1*01-DQB1*0501 genotype with *Artemisia vulgaris* allergic asthma. Hum Immunol 2003;64:811-5.
16. Demenais F, Margaritte-Jeannin P, Barnes KC, Cookson WOC, Altmüller J, Ang W, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat Genet 2018;50:42-53.
17. Galanter JM, Gignoux CR, Torgerson DG, Roth LA, Eng C, Oh SS, et al. Genome-wide association study and admixture mapping identify different asthma-associated loci in Latinos: the Genes-environments & Admixture in Latino Americans study. J Allergy Clin Immunol 2014;134:295-305.
18. Li X, Ampleford EJ, Howard TD, Moore WC, Torgerson DG, Li H, et al. Genome-wide association studies of asthma indicate opposite immunopathogenesis direction from autoimmune diseases. J Allergy Clin Immunol 2012;130:861-8.e7.
19. Mathias RA, Grant AV, Rafaels N, Hand T, Gao L, Vergara C, et al. A genome-wide association study on African-ancestry populations for asthma. J Allergy Clin Immunol 2010;125:336-46.e4.

20. Noguchi E, Sakamoto H, Hirota T, Ochiai K, Imoto Y, Sakashita M, et al. Genome-wide association study identifies HLA-DP as a susceptibility gene for pediatric asthma in Asian populations. PLoS Genet 2011;7:e1002170.
21. Pino-Yanes M, Gignoux CR, Galanter JM, Levin AM, Campbell CD, Eng C, et al. Genome-wide association study and admixture mapping reveal new loci associated with total IgE levels in Latinos. J Allergy Clin Immunol 2015;135:1502-10.
22. Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, et al. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet 2011;43:887-92.
23. Lasky-Su J, Himes BE, Raby BA, Klanderman BJ, Sylvia JS, Lange C, et al. HLA-DQ strikes again: genome-wide association study further confirms *HLA-DQ* in the diagnosis of asthma among adults. Clin Exp Allergy 2012;42:1724-33.
24. Ferreira MA, Vonk JM, Baurecht H, Marenholz I, Tian C, Hoffman JD, et al. Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. Nat Genet 2017;49:1752-7.
25. Shrine N, Portelli MA, John C, Soler Artigas M, Bennett N, Hall R, et al. Moderate-to-severe asthma in individuals of European ancestry: a genome-wide association study. Lancet Respir Med 2019;7:20-34.
26. Kim KW, Ober C. Lessons learned from GWAS of asthma. Allergy Asthma Immunol Res 2019;11:170-87.
27. Nicodemus-Johnson J, Laxman B, Stern RK, Sudi J, Tierney CN, Norwick L, et al. Maternal asthma and microRNA regulation of soluble HLA-G in the airway. J Allergy Clin Immunol 2013;131:1496-503.e4.
28. Gourraud P-A, Khankhanian P, Cereb N, Yang SY, Feolo M, Maiers M, et al. HLA diversity in the 1000 genomes dataset. PloS One 2014;9:e97282.
29. Limou S, Coulonges C, Herbeck JT, van Manen D, An P, Le Clerc S, et al. Multiple-cohort genetic association study reveals CXCR6 as a new chemokine receptor involved in long-term nonprogression to AIDS. J Infect Dis 2010;202:908-15.
30. Limou S, Zagury J-F. Immunogenetics: genome-wide association of non-progressive HIV and viral load control: HLA genes and beyond. Front Immunol 2013;4:118.
31. Daya M, Rafaels N, Brunetti TM, Chavan S, Levin AM, Shetty A, et al. Association study in African-admixed populations across the Americas recapitulates asthma risk loci in non-African populations. Nat Commun 2019;10:880.
32. Levin AM, Mathias RA, Huang L, Roth LA, Daley D, Myers RA, et al. A meta-analysis of genome-wide association studies for serum total IgE in diverse study populations. J Allergy Clin Immunol 2013;131:1176-84.
33. Johnston HR, Hu Y-J, Gao J, O'Connor TD, Abecasis GR, Wojcik GL, et al. Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. Sci Rep 2017;7:46398.
34. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 2016;48:1279-83.
35. Zheng X, Shen J, Cox C, Wakefield JC, Ehm MG, Nelson MR, et al. HIBAG–HLA genotype imputation with attribute bagging. Pharmacogenomics J 2014;14:192-200.
36. Pappas DJ, Lizee A, Paunic V, Beutner KR, Motyer A, Vukcevic D, et al. Significant variation between SNP-based HLA imputations in diverse populations: the last mile is the hardest. Pharmacogenomics J 2018;18:367-76.
37. Geffard E, Limou S, Walencik A, Daya M, Watson H, Torgerson D, et al. EasyHLA, a validated web application suite to reveal the full details of HLA typing [published online ahead of print November 21, 2019]. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz875.
38. Gourraud P-A, Lamiraux P, El-Kadhi N, Raffoux C, Cambon-Thomsen A. Inferred HLA haplotype information for donors from hematopoietic stem cells donor registries. Hum Immunol 2005;66:563-70.
39. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 2013;10:5-6.
40. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature 2015;526:68-74.
41. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet 2013;93:278-88.
42. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.R-project.org/.
43. Kreyszig E. Advanced engineering mathematics. 4th ed. New York: Wiley; 1979:939.
44. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genome-wide association scans. Bioinformatics 2010;26:2190-1.
45. Vince N, Bashirova AA, Lied A, Gao X, Dorrell L, McLaren PJ, et al. HLA class I and KIR genes do not protect against HIV type 1 infection in highly exposed uninfected individuals with hemophilia A. J Infect Dis 2014;210:1047-51.
46. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961;56:52-64.
47. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat Commun 2016;7:12522.
48. Naqvi M, Choudhry S, Tsai H-J, Thyne S, Navarro D, Nazario S, et al. Association between IgE levels and asthma severity among African American, Mexican, and Puerto Rican patients with asthma. J Allergy Clin Immunol 2007;120:137-43.
49. Maiers M, Gragert L, Klitz W. High-resolution HLA alleles and haplotypes in the United States population. Hum Immunol 2007;68:779-88.
50. Meyer D, C Aguiar VR, Bitarello BD, C Brandt DY, Nunes K. A genomic perspective on HLA evolution. Immunogenetics 2018;70:5-27.
51. Morii W, Sakai A, Ninomiya T, Kidoguchi M, Sumazaki R, Fujieda S, et al. Association of Japanese cedar pollinosis and sensitization with HLA-DPB1 in the Japanese adolescent. Allergol Int 2018;67:61-6.
52. Shimane K, Kochi Y, Suzuki A, Okada Y, Ishii T, Horita T, et al. An association analysis of HLA-DRB1 with systemic lupus erythematosus and rheumatoid arthritis in a Japanese population: effects of *09:01 allele on disease phenotypes. Rheumatology 2013;52:1172-82.
53. Murao S, Makino H, Kaino Y, Konoue E, Ohashi J, Kida K, et al. Differences in the contribution of HLA-DR and -DQ haplotypes to susceptibility to adult- and childhood-onset type 1 diabetes in Japanese patients. Diabetes 2004;53:2684-90.
54. Lin JM, Zhang YB, Peng QL, Yang HB, Shi JL, Gu ML, et al. Genetic association of HLA-DRB1 multiple polymorphisms with dermatomyositis in Chinese population. HLA 2017;90:354-9.
55. Anaya J-M. Common mechanisms of autoimmune diseases (the autoimmune tautology). Autoimmun Rev 2012;11:781-4.
56. Khor S-S, Morino R, Nakazono K, Kamitsuji S, Akita M, Kawajiri M, et al. Genome-wide association study of self-reported food reactions in Japanese identifies shrimp and peach specific loci in the HLA-DR/DQ gene region. Sci Rep 2018;8:1069.
57. Vita R, Overton JA, Greenbaum JA, Ponomarenko J, Clark JD, Cantrell JR, et al. The immune epitope database (IEDB) 3.0. Nucleic Acids Res 2015;43:D405-12.
58. Wambre E, James EA, Kwok WW. Characterization of CD4+ T cell subsets in allergy. Curr Opin Immunol 2012;24:700-6.
59. Wang X, Yang Q, Wang P, Luo L, Chen Z, Liao B, et al. Derp2-mutant gene vaccine inhibits airway inflammation and up-regulates Toll-like receptor 9 in an allergic asthmatic mouse model. Asian Pac J Allergy Immunol 2010;28:287-93.
60. Genc S, Eroglu H, Kucuksezer UC, Aktas-Cetin E, Gelincik A, Ustyol-Aycan E, et al. The decreased CD4+CD25+ FoxP3+ T cells in nonstimulated allergic rhinitis patients sensitized to house dust mites. J Asthma 2012;49:569-74.

Bioinformatics, 36(7), 2020, 2157–2164
doi: 10.1093/bioinformatics/btz875
Advance Access Publication Date: 21 November 2019
Original Paper

Genetics and population analysis

Easy-HLA: a validated web application suite to reveal the full details of HLA typing

Estelle Geffard[1,*], Sophie Limou[1], Alexandre Walencik[1,2], Michelle Daya[3], Harold Watson[4], Dara Torgerson[5], Kathleen C. Barnes on behalf of CAAPA[3], Anne Cesbron Gautier[2], Pierre-Antoine Gourraud[1,*] and Nicolas Vince[1]

[1]Nantes Université, Centrale Nantes, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, UMR 1064, ITUN, Nantes F-44000, France, [2]Laboratoire d'Histocompatibilité et d'Immunogénétique, EFS Centre—Pays de la Loire, Nantes F-44000, France, [3]Department of Medicine, University of Colorado Denver, Aurora, CO 80045, USA, [4]Faculty of Medical Sciences Cave Hill Campus, The University of the West Indies, Bridgetown BB11000, Barbados and [5]McGill University and Genome Quebec Innovation Centre, Montreal, QC H3A 0G1, Canada

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on December 6, 2018; revised on September 19, 2019; editorial decision on November 14, 2019; accepted on November 20, 2019

Abstract

Motivation: The HLA system plays a pivotal role in both clinical applications and immunology research. Typing HLA genes in patient and donor is indeed required in hematopoietic stem cell and solid-organ transplantation, and the histocompatibility complex region exhibits countless genetic associations with immune-related pathologies. Since the discovery of HLA antigens, the HLA system nomenclature and typing methods have constantly evolved, which leads to difficulties in using data generated with older methodologies.

Results: Here, we present Easy-HLA, a web-based software suite designed to facilitate analysis and gain knowledge from HLA typing, regardless of nomenclature or typing method. Easy-HLA implements a computational and statistical method of HLA haplotypes inference based on published reference populations containing over 600 000 haplotypes to upgrade missing or partial HLA information: 'HLA-Upgrade' tool infers high-resolution HLA typing and 'HLA-2-Haplo' imputes haplotype pairs and provides additional functional annotations (e.g. amino acids and KIR ligands). We validated both tools using two independent cohorts (total $n = 2500$). For HLA-Upgrade, we reached a prediction accuracy of 92% from low- to high-resolution of European genotypes. We observed a 96% call rate and 76% accuracy with HLA-2-Haplo European haplotype pairs prediction. In conclusion, Easy-HLA tools facilitate large-scale immunogenetic analysis and promotes the multi-faceted HLA expertise beyond allelic associations by providing new functional immunogenomics parameters.

Availability and implementation: Easy-HLA is a web application freely available (free account) at: https://hla.univ-nantes.fr.

Contact: easyhla@gmail.com or pierre-antoine.gourraud@univ-nantes.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

HLA genes from the major histocompatibility complex (MHC) encode a specific group of cell surface molecules mediating recognition of non-self-antigens by the immune system. HLA plays key roles in transplantation management and success. HLA matching between a patient and potential donors is essential in hematopoietic stem cell transplantation (HSCT) (Copelan, 2006; Loiseau *et al.*, 2007) and solid-organ transplantations (Held *et al.*, 1994). Donor-recipient compatibility is defined by the number of alleles shared across HLA-A, -B, -C, -DRB1 and -DQB1 genes. The chance of graft success is optimal when donor and recipient are fully compatible and have the lowest number of HLA alleles mismatches (Lee *et al.*, 2007; Zachary and Leffell, 2016). The level of typing resolution is positively correlated with the probability of allele matching during donor search. Additionally, time restrictions in solid-organ transplantation from deceased donors often make HLA allele high-resolution typing impossible, and only intermediate resolution or even low-resolution genotyping may be available at the time of organ allocation. Beyond these major clinical impacts, HLA has

Table 1. Common nomenclature reporting HLA types

Name	Typing	Resolution	Nomenclature
Broad serology	Phenotyping (lymphocytotoxicity)	Low	B14
Split serology	Phenotyping (lymphocytotoxicity)	Low	B64 B65
First-field	Genotyping (PCR SSP)	Low	B*14
NMDP code	Genotyping (PCR SSO)	Intermediate	14:HUJ
Second-field	Genotyping (Sanger sequencing and/or NGS)	High	B*14:01 B*14:02

Note: *HLA* alleles nomenclature established by the World Health Organization Nomenclature Committee (http://hla.alleles.org/nomenclature/committee.html). Nomenclature is regularly updated. Here, we consider *HLA*-B*14:01:01 as an example. 'NMDP codes' allele codes narrow the list of alleles that must be considered at a given locus by eliminating some possibilities (e.g. B*14:HUJ means that the typing is either B*14:01 or B*14:02). 'NMDP codes' are implemented and updated by the NMDP (https://bioinformatics.bethematchclinical.org/hla-resources/allele-codes/allele-code-lists/allele-code-list-in-alphabetical-order/). PCR SSO: sequence specific oligonucleotide. PCR SSP: sequence specific primers (Howell *et al.*, 2010).

been frequently associated with numerous immune-related pathologies (MacArthur *et al.*, 2017; Tian et al., 2017; Vince *et al.*, 2014).

MHC genomic region on chromosome 6 (6p21.3) is the most complex and polymorphic locus of the human genome (Howell *et al.*, 2010). The MHC counts >220 genes (Horton *et al.*, 2004), including 21 polymorphic HLA genes from the classical HLA class I (e.g. HLA-A, HLA-B and HLA-C) and HLA class II (e.g. HLA-DRB1 and HLA-DQB1). The HLA system comprises >22 000 described alleles (Robinson *et al.*, 2015) (https://www.ebi.ac.uk/ipd/imgt/hla/stats.html). HLA alleles correspond to a specific sequence of HLA genes and can be considered as single nucleotide variants (including tetra-allelic ones as well as insertions and deletions) haplotypes. HLA haplotypes can be constructed from these HLA alleles; here, we consider the five main genes HLA-A∼B∼C∼DRB1∼DQB1 for haplotyping, and named the 5-gene haplotypes as the following example:
A*34:02∼B*14:01∼C*08:02∼DRB1*04:05∼DQB1*03:02. The complexity of this region is not only due to its diversity but also to its linkage disequilibrium (LD). LD is defined as the non-random association of neighboring polymorphisms, i.e. the difference between the observed frequency of allele combinations (haplotypes) and the expected frequency under random transmission. LD and haplotype frequencies are shaped by selective pressure, genetic drift, non-random mating, recombination events and shared genetic effect between alleles (Ahmad *et al.*, 2003; Goodin *et al.*, 2018).

HLA typing techniques have considerably evolved over the years with a wide array of methods providing increasing levels of resolution (Erlich, 2012, Table 1). Historically, phenotyping was performed by detecting HLA proteins on cell surface with specific antibodies. These serology-based methods have progressively been replaced with DNA-based typing. Today, full-length HLA genes sequenced through Next-Generation Sequencing (NGS) provides the highest standard and resolution. In parallel with HLA typing methods, nomenclature has greatly evolved (Table 1), which nowadays significantly hampers retrospective analyses and HSCT compatible donor search from archived HLA datasets recorded in low/mid-resolutions with possibly some missing genes (e.g. HLA-C or HLA-DQB1) (Hurley *et al.*, 2004). This major pitfall for clinics and biomedical research highlights the crucial need for high-resolution allele imputation from low or intermediate resolution in order to reduce allele ambiguity by simultaneously increasing genotype resolution and imputing unknown genes (Madbouly *et al.*, 2014).

Finally, many current typing technologies are not designed to deliver full-length HLA haplotypes. Knowledge of haplotype pairs can be particularly useful to determine if unrelated individuals have a chance to be haplo-identical in an HSCT clinical setting; and in research, haplotypes are necessary for functional annotations. Familial explorations can be performed to determine haplotypes from parental genotypes; however, this technique is expensive and challenging to implement as it requires access to relatives' DNA. Beyond this family-based approach, computational haplotype inference based on probabilistic models from genotypic data has been proposed (Salem *et al.*, 2005). Several methods for haplotype inference exist, from algorithms based on parsimony (Clark, 1990) or on likelihood [such as the Expectation-Maximization (EM) algorithms] (Excoffier and Slatkin, 1995) to Bayesian algorithms (Stephens and Donnelly, 2003). Overall, the most commonly used methods to compute HLA haplotypes are EM-based algorithms (Eberhard *et al.*, 2013), which can accommodate several loci with an arbitrary number of alleles for a large number of individuals with ambiguous haplotypes (Eberhard *et al.*, 2013; Salem *et al.*, 2005). However, they show limited performance with small sample size and do not support haplotype determination from a unique individual. Moreover, results are dependent on inherent dataset characteristics: individuals genetic heterogeneity, number of loci and genotype resolution (Eberhard *et al.*, 2013). These methods are not always straightforward or need powerful computation (Salem *et al.*, 2005). Previously, a maximum likelihood-based HLA haplotype imputation technique was validated on several datasets for unrelated HSCT donor search (Gourraud *et al.*, 2005). This method computes the most likely haplotype pair from HLA genotypes based on HLA genotype frequencies throughout donor transplant registries. Most of the reference haplotype frequencies come from the large reference population of the National Marrow Donor Program (NMDP). NMDP designates the US voluntary bone marrow donor registry. This registry has proposed several breakthroughs in the field of bone marrow transplantation by making available the large HLA haplotype database used in the current study, and also, by creating a specific nomenclature: 'NMDP codes'. These codes allow to describe HLA typing with some allele ambiguity represented by two to five letter codes (Table 1).

Following this strategy, we developed Easy-HLA, a user-friendly web application designed to deliver a complete suite of HLA annotations (freely available through a secure connection at https://hla.univ-nantes.fr). From HLA genotypes and regardless of resolution level, Easy-HLA can statistically resolve HLA genotype ambiguity, and increase HLA data resolution and functional annotations. Easy-HLA facilitates the use of HLA data collected from both classical and historic laboratory procedures. In this article, we present our application and its validation using independent cohorts delivering optimized information for immunogenetic investigations.

2 Implementation

Easy-HLA is a web-based application suite designed to predict haplotypes from HLA genotypes. The input HLA genotypes can be entered with low/mid-resolution and/or can contain ambiguities, in a single request (one individual genotype) or batch mode (several individuals genotypes). Regarding security and data storage, the loaded data files are deleted immediately after analysis completion, and the output data files are safely conserved on our server for 1 week.

We implemented our tools with web scripting languages using PHP (Hypertext Preprocessor) combined with the pgSQL procedural language. The pgSQL language is used to interrogate the haplotype database and find haplotype pairs corresponding to the input genotype. PHP functions were designed to query multiple databases (serological identity, NMDP nomenclature equivalence) to translate

the HLA nomenclature complexity. We used estimates from maximum likelihood-based statistical method to infer HLA haplotypes and subsequently predict unavailable HLA information.

2.1 Database

Easy-HLA main algorithm is based on HLA haplotype frequencies from a large reference population, these frequencies were obtained with a maximum likelihood-based HLA haplotype imputation technique previously validated (Gourraud *et al.*, 2005). We stored our data in a PostgreSQL database. The core reference haplotype frequencies come from the NMDP published in 2013 for uses in clinical transplant and immunological research (Gragert *et al.*, 2013). From the HLA genotypes of 6.59 million US subjects, the NMDP estimated high-resolution HLA haplotype frequencies in five ancestral populations using an EM algorithm (Schaid *et al.*, 2002). The large sample size allows an accurate estimation of rare alleles and haplotype frequencies. The NMDP haplotype database thereby reports frequencies of over 600 000 haplotypes divided into five ancestral populations (African-Americans: 198 216; Asian and Pacific Islanders: 158 307; Europeans: 304 697; Hispanics: 220 020 and Native-Americans: 36 417). We completed this large dataset with RFGM (Registre France Greffe de Moelle), a French population database containing >16 000 haplotypes (Gourraud *et al.*, 2015; Pappas *et al.*, 2015). The user has the possibility to choose the best matching reference population with his/her input individual(s) ancestry among these six reference datasets.

2.2 Algorithm

From each HLA genotype, our algorithm enumerates each possible haplotype pair and computes the corresponding likelihood. Considering a diploid genotype (G) for three HLA genes (A, B and C) and two alleles per gene (upper and lower cases), we obtain four distinct theoretical haplotype pairs [or diplotypes, $d1$–4, Equation (1)]. We can generalize the computation of N theoretical diplotypes from a diploid genotype (G) for x genes (with heterozygous alleles) with the equation $N = 2^{x-1}$.

Enumeration of diplotypes

$$G(\mathrm{Aa} \sim \mathrm{Bb} \sim \mathrm{Cc}) \begin{cases} \mathrm{d1}\ (\mathrm{A} \sim \mathrm{B} \sim \mathrm{C}, \mathrm{a} \sim \mathrm{b} \sim \mathrm{c}) \\ \mathrm{d2}\ (\mathrm{A} \sim \mathrm{b} \sim \mathrm{C}, \mathrm{a} \sim \mathrm{B} \sim \mathrm{c}) \\ \sout{\mathrm{d3}\ (\mathrm{A} \sim \mathrm{b} \sim \mathrm{c}, \mathrm{a} \sim \mathrm{B} \sim \mathrm{C})} \\ \mathrm{d4}\ (\mathrm{A} \sim \mathrm{B} \sim \mathrm{c}, \mathrm{a} \sim \mathrm{b} \sim \mathrm{C}) \end{cases} \quad (1)$$

Our algorithm is founded on a reference database of HLA haplotype frequencies (f) in different populations: haplotypes not reported in the reference dataset are removed from the haplotype list [$a \sim B \sim C$ strikethrough in $d3$ in Equation (1)], resulting in n previously observed pairs of haplotypes (here, $n = 3$) and therefore reducing the space of haplotypes to explore.

We calculated genotypic frequencies from haplotype frequencies by following Hardy Weinberg's genetic distribution law. When a diplotype is homozygous, the likelihood (L) is the squared value of the haplotype frequency ($f2$). When the diplotype is heterozygous the likelihood (L) is:

Likelihood of an enumerated heterozygous diplotype

$$L(d1(A \sim B \sim C, a \sim b \sim c)) = 2 * f(A \sim B \sim C) * f(a \sim b \sim c) \quad (2)$$

where $f(A \sim B \sim C)$ and $f(a \sim b \sim c)$ corresponds to the respective frequencies of each haplotype. The number of homozygous diplotypes is low (Gragert *et al.*, 2013); therefore, we consider only the likelihood for heterozygous diplotypes.

When HLA genotypes are specified with allelic ambiguities (low-resolution) and/or untyped loci (incomplete genotype), multiple alternative diplotypes can be inferred. For allele ambiguity, the Easy-HLA imputation algorithm produces all possible HLA genotypes associated with the ambiguous input. Correspondingly, when a locus is missing [in our example, the HLA-B gene was not typed and is recorded as XX—Equation (3)], Easy-HLA generates all possible alleles for this missing gene (B, b and β). Equation (3) displays only haplotypes pairs reported in our reference database with a frequency above the user-defined threshold. Indeed, we do not show every possible theoretical haplotype pairs as many are not observed in our population datasets, and would therefore have a null estimated frequency.

Enumeration of diplotypes from an incomplete genotype with a missing locus using all compatible haplotypes present in our database

$$G(A\ a \sim XX \sim Cc) \begin{cases} d1\ (A \sim B \sim C, a \sim b \sim c) \\ d2\ (A \sim b \sim C, a \sim B \sim c) \\ d3\ (A \sim \beta \sim c, a \sim b \sim C) \\ d4\ (A \sim \beta \sim c, a \sim \beta \sim C) \end{cases} \quad (3)$$

From an incomplete HLA genotype, Easy-HLA algorithm hence produces all possible diplotypes and then computes their corresponding likelihood. For each diplotype, a confidence measurement named post-probability (Post-P) is calculated as the ratio of likelihood of a particular diplotype ($L(dt)$) among the likelihood of all n possible diplotypes ($L(di)$):

Post-probability of each possible diplotype

$$\mathrm{Post} - P(dt) = \frac{L(dt)}{\sum_{i=1}^{n} L(di)} \quad (4)$$

where i is an index for enumerating the different diplotypes and n is the number of possible diplotypes. The post-probability of the most likely diplotype is then:

Post-probability of the most likely diplotype

$$\mathrm{Post} - P = \frac{\max(L(di))}{\sum_{i=1}^{n} L(di)} \quad (5)$$

The diplotype with the highest post-probability is by definition dependent of the haplotype frequencies in the reference dataset. When interpreting the output, one has to be cautious when top post-probabilities are close, as the real haplotype pair might then not always be the most likely.

From the likelihood of each predicted diplotype, Easy-HLA can then infer a high-resolution genotype for the incomplete or ambiguous input genotype. The likelihood (L) of the imputed high-resolution genotype is:

Likelihood of the imputed high-resolution genotype

$$L(G(A\ a \sim Bb \sim Cc)) = \sum_{i=1}^{n} L(di) \quad (6)$$

where i is an index for enumerating the different diplotypes di, n is the number of possible diplotypes and L is the likelihood of a diplotype obtained from haplotype frequencies f.

2.3 Software presentation

Easy-HLA input is an HLA genotype for each gene (HLA-A, -B, -C, -DRB1 and -DQB1), accepting various levels of HLA nomenclature [Table 1—serology resolution, generic HLA genotyping obtained by molecular biology or codes gathering different HLA alleles (NMDP/MAC UI codes)], as well as missing or incomplete genotypes (Fig. 1). After logging into a personal account, the user has to enter a genotype and select the reference population matching his/her data among six: African-Americans, Asian/Pacific Islanders, Europeans, Hispanics, Native-Americans or French. Alternatively, it is possible to run a search on all combined populations, in that case the output does not provide any frequencies, but indicates the population in which the haplotypes are the most frequent. All data are securely collected, processed and stored. In batch mode, the user uploads a file containing the set of genotypes (automatically deleted after the imputation). In addition, the post-probability threshold (confidence value) should be chosen carefully as it impacts the call rate (chance to have a result) and output accuracy (see below and Fig. 2). On the user interface, a field is available to specify a frequency threshold to further restrict the list of possible diplotypes. Indeed, genotypes with

a post-probability below the selected threshold are automatically excluded to prevent an over-representation of rare genotypes. Overall, Easy-HLA displays the different high-resolution genotypes with their likelihood and post-probability starting with the most probable one (Fig. 1A). Optionally, it is possible to select only the most probable high-resolution genotype (in batch mode). When our algorithm does not find a corresponding pair of haplotypes in the reference dataset for a given genotype, it cannot make a prediction and returns an information message: 'No matching donor found with your selected criteria'. However, in the HLA-2-Haplo module, if one reported haplotype corresponds to half the given genotype, the algorithm infers the missing second haplotype to propose a haplotype pair solving the genotype.

The HLA-Upgrade module can predict a full 5-loci HLA-A, -B, -C, -DRB1 and -DQB1 genotyping at high-resolution level by statistically resolving missing, low-resolution or ambiguous typings such as NMDP codes. By updating HLA genotypes, HLA-Upgrade empowers the analysis of old cohorts or cohorts with a long delay of recruitment (Fig. 1B), for which HLA-C and HLA-DQB1 genes are often missing (only recently added in routine genotyping).

The HLA-2-Haplo module predicts the most likely haplotype pair from a given genotype (Fig. 1B) and provides their frequencies in different populations. HLA-2-Haplo can be a particularly useful tool to determine if unrelated individuals have a chance to be haploidentical in an HSCT clinical setting and to provide Supplementary Material for research. To solve phasing ambiguity for a given HLA genotype, our algorithm compares the potential haplotype pairs with the previously reported haplotypes stored in our large reference database, and can impute the second haplotype if only one haplotype from the diplotype was previously observed. Three additional functions, only available in batch mode, are offered with this module: (i) HLA-expr delivers HLA-C predicted expression (based on allele specific mean HLA-C expression, see (Vince *et al.*, 2016); (ii) HLA-AA provides HLA alleles amino acids and (iii) HLA-KIRlig indicates the KIR (Killer-cell immunoglobulin-like receptor) binding group (C1/C2 groups, Bw4/Bw6 or KIR2DL2 ligands) for each individual HLA allele.

3 Performance

3.1 HLA-Upgrade validation

We validated the HLA-Upgrade module using two independent cohorts of unrelated individuals with high-resolution (second-field) HLA genotyping for HLA-A∼B∼C∼DRB1∼DQB1 loci: (i) 1579 Europeans from the Nantes blood center and (ii) 917 individuals of African ancestry from Consortium on Asthma among African-ancestry Populations in the Americas (Barnes *et al.*, 1996; Mathias *et al.*, 2016). To evaluate the database exhaustiveness on the presence of haplotypes from both cohorts in the database, we tested HLA genotypes at high-resolution from both cohorts in HLA-Upgrade. We found 96.5% of the European cohort genotypes and 70.1% of the African-American cohort genotypes represented in the database. From these full high-resolution datasets, we simulated low-resolution HLA genotypes for both cohorts by creating 12 different situations based on four resolution levels [first-field, second-field, serology and NMDP simulated with correspondence table (https://www.ebi.ac.uk/ipd/imgt/hla/)] and on three levels of input loci (HLA-A∼B∼DRB1, HLA-A∼B∼C∼DRB1 and HLA-A∼B∼C∼DRB1∼DQB1). We used HLA-Upgrade to predict full HLA-A∼B∼C∼DRB1∼DQB1 high-resolution genotypes for each of the 12 simulated datasets, and defined accuracy as the percentage of correct predictions compared to the original HLA typing. We defined call rate as the number of predictions compared to the total number of expected predictions.

The resolution level impacts prediction accuracy, prediction is almost twice as good for intermediate-resolution (NMDP) and high-resolution (second-field) genotypes compared to low-resolution genotype (serology and first-field) (Supplementary Fig. S1). For the HLA-A∼B∼C∼DRB1∼DQB1 genotype, the prediction accuracy was 58.1%, 58.6%, 97.6% and 100% for first-field, serology, NMDP and second-field resolution level, respectively. Interestingly, split serology was as accurate as first-field HLA genotyping, meaning that the erroneous predictions are not based on split antigens. Second, results obtained from NMDP codes and second-field HLA typing did not differ drastically, emphasizing the typing precision of the high-definition sequence specific oligonucleotide. Similarly, the level of input loci impacts the prediction accuracy (Supplementary Fig. S1 and Fig. 2A): inputs lacking alleles from one gene (HLA-C or HLA-DQB1) resulted in a drop of accuracy. On average, HLA-C prediction was 20 points better when HLA-C was provided in input (Fig. 2A). We showed similar results from inputs lacking HLA-DQB1 in the African cohort. The prediction accuracy per locus in the European cohort from first-field resolution HLA-A∼B∼C∼DRB1∼DQB1 genotype inputs was 97% for HLA-A, 93% for HLA-B, 96% for HLA-C, 86% for HLA-DRB1 and 93% for HLA-DQB1 (Fig. 2A, left panel). As a comparison, the prediction accuracy per locus in the African-American cohort was 87%, 90%, 89%, 78% and 92% for HLA-A, -B, -C, -DRB1 and -DQB1, respectively (Fig. 2A, right panel). The prediction accuracy by locus for the European cohort was on average 10 points higher than for the African-derived cohort, probably because of the smaller sample size and larger HLA haplotype diversity in African-ancestry populations.

For each input level (3, 4 or 5 genes), prediction accuracy (Fig. 2B) and call rate (Fig. 2C) of the full genotype change almost linearly with an increasing post-probability threshold (confidence measure). For the 5-loci input level, accuracy increased from 58% to 92% in Europeans and from 35% to 64% in African-Americans for a post-probability threshold going from 0% (accepts everything) to 90% (includes genotypes with post-probability > 90%), respectively. On the contrary, call rate decreased from 100% to 20% in Europeans and from 100% to 10% in African-Americans for a threshold increasing from 0% to 90%, respectively. Correspondingly for the 3-loci input level, accuracy increased from 38% to 94% in Europeans and from 16% to 67% in African-Americans and call rate decreased from 100% to 3% in Europeans and from 100% to 0% in African-Americans for a post-probability threshold of 0% and 90%, respectively. Results are consistent with the additional validation presented in Supplementary Fig. S4.

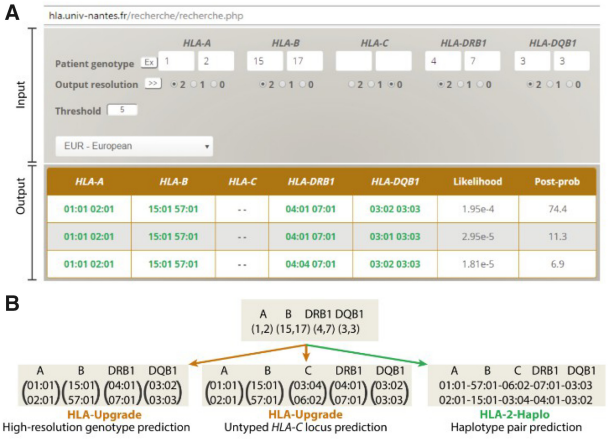

Fig. 1. Easy-HLA software presentation. (**A**) Example of the single query mode. The patient genotype is entered for each gene in first or second-field, serology, NMDP codes or left empty. The user must choose the output resolution (2: second-field, 1: first-field, 0: empty), the post-probability threshold and reference population. The output table contains full mid to high-resolution genotypes with their respective likelihood and post-probability. (**B**) Easy-HLA delivers updated *HLA* information from low-resolution *HLA* typing. In this example, we start with a classical *HLA* serological genotype (A∼B∼DRB1∼DQB1). HLA-Upgrade statistically predicts high-resolution genotypes (left panel), and can also predict an untyped locus, such as *HLA-C* (middle panel). Finally, HLA-2-Haplo imputes the most likely haplotype pair. These predictions are all done *in silico* and as such prevent from additional genotyping in the laboratory

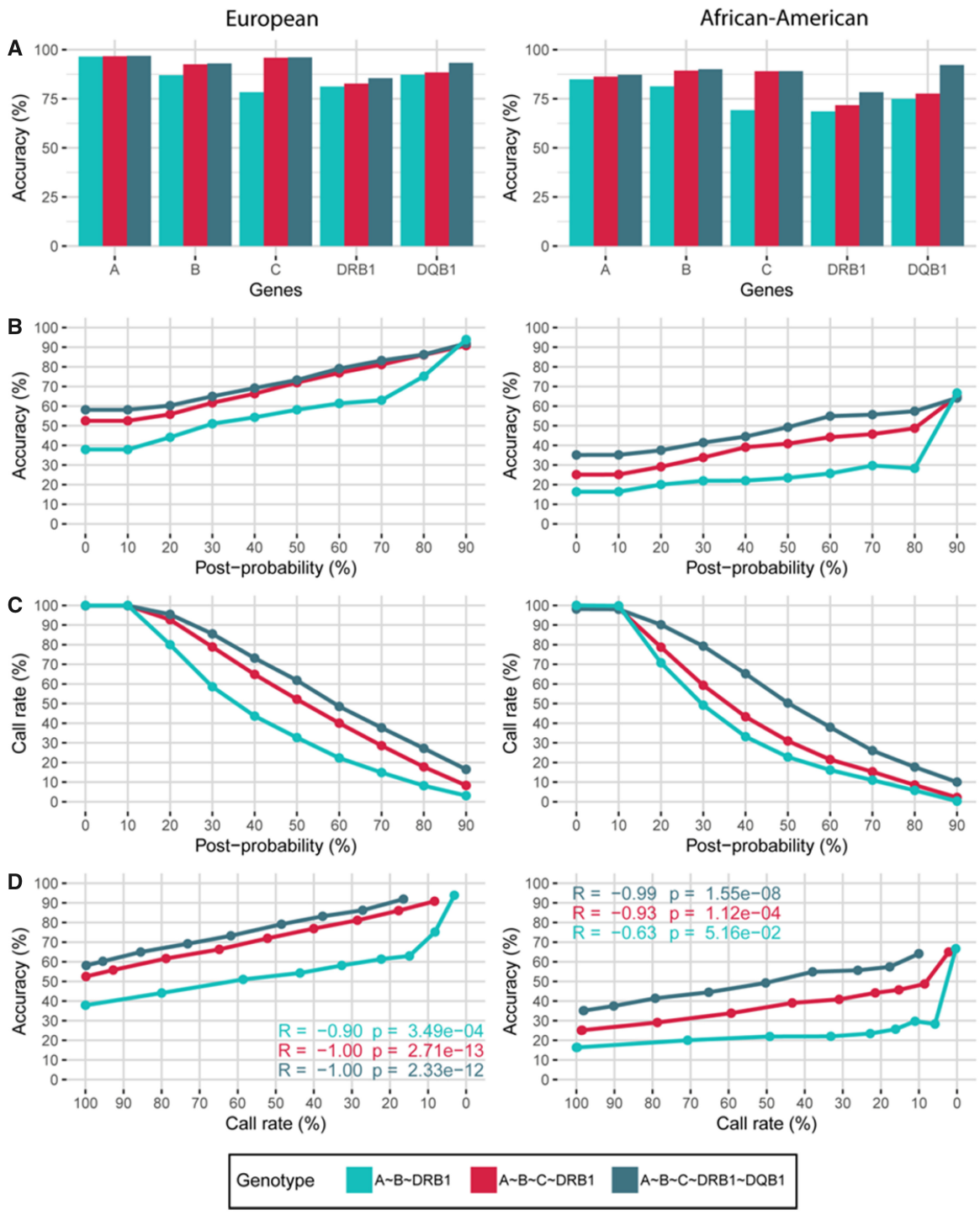

Fig. 2. Validation of the HLA-Upgrade module in the European (EUR, left) and African-American (AFA, right) populations (post-probability threshold set at 0%). HLA-A∼B∼C∼DRB1∼DQB1 high-resolution genotypes were predicted from different gene combinations of first-field genotypes: HLA-A∼B∼DRB1 (blue), HLA-A∼B∼C∼DRB1 (red), HLA-A∼B∼C∼DRB1∼DQB1 (dark blue). (**A**) Prediction accuracy per locus. (**B**) Prediction accuracy according to genotype post-probability. (**C**) Call rate according to genotype post-probability. (**D**) Prediction accuracy according to call rate. (Color version of this figure is available at *Bioinformatics* online.)

The prediction accuracy increases as the call rate declines (Fig. 2D), exemplifying a challenging risk/benefit balance that limits error risk (increased accuracy) at the cost of a lower number of output results (low call rate). The post-probability parameter is therefore crucial for HLA-Upgrade prediction performances. By default, we recommend a post-probability threshold set at 10% for exploratory research with HLA-Upgrade, advocating for more results with a compromise on allele accuracy. At this threshold, we have a 100% call rate, but we eliminate genotypes with very low frequencies for European- and African-ancestry populations.

3.2 HLA-2-Haplo validation

We validated HLA-2-Haplo module using two independent cohorts with high-resolution HLA-A∼B∼C∼DRB1∼DQB1: (i) 273 European-ancestry (genotyping) and (ii) 116 African-ancestry

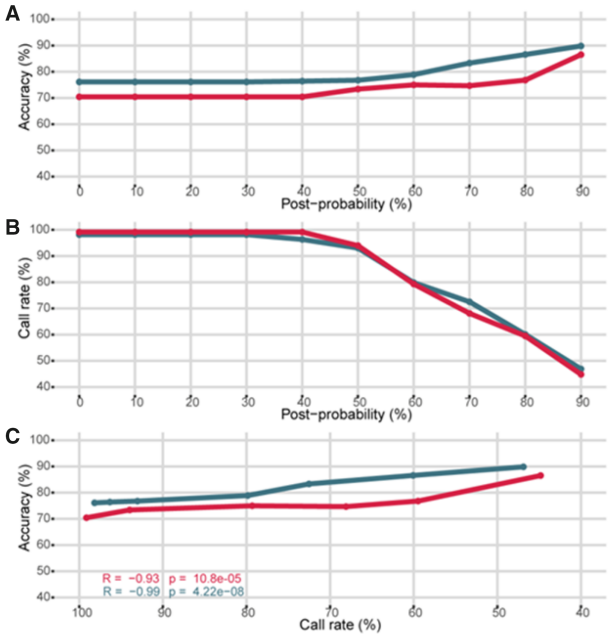

Fig. 3. Validation of the HLA-2-Haplo module in the European (EUR, dark blue) and African-American (AFA, red) populations. (**A**) Accuracy of haplotypes pairs prediction according to the calculated post-probability. (**B**) Call rate of haplotypes pairs prediction according to the calculated post-probability. (**C**) Accuracy of haplotypes pairs prediction according to the call rate. (Color version of this figure is available at *Bioinformatics* online.)

individuals [HLA imputation from single nucleotide polymorphism (SNP) genotypes] (Barnes *et al.*, 1996; Mathias *et al.*, 2016). African-ancestry individuals were previously genome-wide SNP genotyped (Johnston *et al.*, 2017) and we imputed HLA alleles for the 5-loci HLA-A, -B, -C, -DRB1 and -DQB1 with the HIBAG R package (Zheng *et al.*, 2014). For each individual, we deduced haplotype pairs from parental HLA typing (hereditary familial study with parents/child trios). The method used is segregation analysis. Families were selected with an ascending minimum of one first degree relative (parents/children, Supplementary Fig. S6).

For each validation cohort, we predicted haplotypes with HLA-2-Haplo from the 5-HLA loci in high-resolution (Fig. 3). Similar to HLA-Upgrade, post-probability measures the confidence of predicted haplotypes based on frequency. Prediction accuracy ranged from 76% to 90% in Europeans and from 70% to 86% in African-Americans with a post-probability from 0% to 90%, respectively (Fig. 3A). For both cohorts, we observed a drop of call rate after the post-probability threshold reached 40% and down to 45% for the 90% post-probability threshold (Fig. 3B). Overall, prediction accuracy is relatively stable across call rates (Fig. 3C). For HLA-2-Haplo, we recommend a default post-probability threshold of 30%, where the call rate reaches a maximum whereas ambiguities with rare genotypes are minimal: call rate is then 99% and 96%, and accuracy is 70% and 76% for the African- and European-ancestry populations, respectively.

Using the same validation cohorts, we compared performance of haplotype pair prediction between HLA-2-Haplo and the 'haplo.stats' R package (Schaid *et al.*, 2002), a statistical tool based on a maximum likelihood method for haplotyping when linkage phase is ambiguous in cohort studies ('haplo.em' function). Unlike the usual EM tools, their algorithm considers more than two alleles per locus, accept missing allele, and use a 'progressive insertion': the algorithm progressively inserts batches of loci into haplotypes of growing lengths before iterating over the EM steps.

We tested the impact of different sampling sizes for input genotypes (10, 50, 100 or all cohort) on the prediction accuracy (Supplementary Fig. S2 and Table S1). Compared to HLA-2-Haplo predictions for Europeans (76% accuracy) and African-Americans (70% accuracy), the 'haplo.stats' predictions were very dependent of sampling size and accuracy was systematically lower than HLA-2-Haplo. 'haplo.stats' accurately predicted 22.6% (21.7–23.5) with 10 genotypes, 36.8% (36.4–37.1) with 100 and up to 46.1% (46.0–46.2) with the Europeans whole cohort ($n = 273$) and 14.9% (14.1–15.7) with 10 genotypes, 40.5% (40.3–40.7) with 100 and up to 42.4% (42.2–42.6) with the African-Americans whole cohort ($n = 116$). Our results are therefore 'more reproducible', in the sense that a given genotype will always output the same results no matter what other individual observations are in the dataset. This methodological characteristic explains why sample size does not impact prediction accuracy (70% for AFA and 76% for EUR) with HLA-2-Haplo.

3.3 Execution

To test our tools' performance, we evaluated HLA-Upgrade runtime on under 48 conditions with a post-probability threshold of 0% and only the first result in output (Supplementary Fig. S3) including: two ancestral populations (European and African-American), four input file sizes (10 100, 1000 and 5000 genotypes), three resolutions (Split serological resolution, first-field and second-field) and two loci combinations (A∼B∼DRB1 and A∼B∼C∼DRB1∼DQB1).

HLA-Upgrade took ∼12 s to analyze files with 100-s field A∼B∼DRB1 genotypes of European ancestry, 8.6 min in first-field and 6.6 min in split serology. We observed a linear runtime progression with the different file sizes (Supplementary Fig. S3). First-field genotypes required a longer execution time than the other two resolutions, which can be explained by a larger number of possible haplotypes.

In addition to input file size and resolution, execution runtime was also impacted by the input level of missingness and the reference population database size. Indeed, a higher number of haplotypes to browse translates into increased runtime: runtime for A∼B∼DRB1 genotypes was 3-fold longer than for A∼B∼C∼DRB1∼DQB1 genotypes and Europeans (304 697 haplotypes in the reference database) took on average 3-fold longer than African-Americans (198 216 haplotypes in the reference database).

4 Discussion

Easy-HLA is a web application suite designed to facilitate large-scale immunogenetic analyses and gain knowledge from HLA typing, regardless of the variety of nomenclature or typing methods. In this report, we presented two tools, HLA-Upgrade and HLA-2-Haplo, based on a large HLA haplotype reference database. Our tools integrate published external data comprising a published set of haplotype frequencies estimated from bone marrow donor registries (>6.5 million individuals and 600 000 unique haplotypes) in order to facilitate an accurate interpretation of the input dataset.

HLA-Upgrade can successfully predict a full HLA-A, -B, -C, -DRB1 and -DQB1 high-resolution genotyping in different populations from low-resolution and/or partially known HLA typing. As expected, HLA-Upgrade performance positively correlates with HLA typing input resolution: when there is more uncertainty or missingness in input, prediction will be lower. Users must find a balance between highly confident results (high accuracy) and number of predicted genotypes (call rate). From our validation cohorts, we recommend a default post-probability threshold at 10%. At this threshold, from the first-field HLA-A∼B∼C∼DRB1∼DQB1 genotype, we predicted a high-resolution genotype with an accuracy of 78.5–92.3% and 85.5–96.9% per HLA locus and a call rate of 98.1% and 99.8% in African- and European-ancestry populations, respectively. We also validated HLA-Upgrade using the 1000 Genomes project HLA data and obtained consistent conclusions in Europeans, Africans, Hispanics and Asian-Pacific Islanders (Supplementary Fig. S4). We tested allele frequencies difference between imputed and non-imputed data: this shows a good correlation with a very limited loss of diversity toward frequent alleles (Supplementary Fig. S7). Currently, HLA-Upgrade outputs the post-probability (confidence measure) for the overall 5-loci prediction.

For future perspectives, we plan to implement an allele post-probability and a locus post-probability to underline the high allelic variability and emphasize the impact of rare alleles amongst the different genotypes. As a proof-of-concept, we carried out a preliminary study to weight the accuracy by genotype frequencies. Indeed, we can consider that rare alleles should not carry the same weight as common alleles in our computation as they will skew the accuracy results. With weighting, our prediction is 43% point better for the A∼B∼DRB1 genotype and 30% point for the A∼B∼C∼DRB1∼DQB1 genotype. Weighting the accuracy computation with HLA genotype frequency considerably improved accuracy for HLA-Upgrade in individuals of European ancestry (see Supplementary Fig. S5), so this strategy is very promising.

The HLA-2-Haplo module accurately predicts haplotype pairs from HLA genotypes. From high-resolution input and a 30% post-probability threshold, we obtained a 99% and 96% call rate and 70% and 76% prediction accuracy for African- and European-ancestry populations, respectively. Importantly, HLA-2-Haplo systematically outperforms 'haplo.stats', a pre-existing HLA haplotyping R package, independently of sample size.

Both Easy-HLA inference tools rely on a statistical algorithm based on HLA haplotype frequencies from a large reference database (>600 000 haplotypes from five different ancestral populations). One major strength of our strategy is the size and diversity of our reference registry including the biggest published haplotype frequency database from the NMDP (Gragert *et al.*, 2013) and the RFGM databases (Gourraud *et al.*, 2015). However, these databases also convey most of Easy-HLA's limitations: exhaustiveness (96.5% European and 70.1% African-American cohort genotypes are present in the database), population diversity coverage [mixed populations (REF)], typing errors, resolution level and HLA loci coverage. For example, the current haplotype frequency databases do not include HLA-DPB1. Our reference database samples HLA haplotypes from the USA and from France and therefore does not represent exhaustive HLA genetic diversity. However, we believe this bias is mostly compensated by the large size of the database, which allows both an accurate estimate of haplotype frequencies and the presence of many rare haplotypes, overall improving our predictions. Finally, Easy-HLA is flexible and our algorithms are fully compatible to evolve regularly with future database releases.

Here, we developed algorithms implemented in a user-friendly web application to facilitate the analysis and reveal the full details of HLA typing. Easy-HLA tools are of great interest both for biomedical research and clinical applications. First, HLA-Upgrade allows to update archived HLA cohorts recorded in low/mid-resolutions and/or with missing loci (such as HLA-C), which empowers the users to combine old and recent datasets to perform large immunogenetic association analyses with various immune-related pathologies. HLA-Upgrade could also be translated into clinics to assess HSCT compatibility between a patient and potential donors with low/mid-resolution. Second, HLA-2-Haplo predicts haplotypes that are the breeding-ground for further research investigations and for functional immunogenomic annotations: HLA-C expression (HLA-expr), amino acid equivalence (HLA-AA) and KIR ligand classification (HLA-KIRlig). Our tools are currently used in HSCT: first, for unrelated donors search using HLA-Upgrade to update an old typing in a donor database, and second, for haplo-identical transplantation using HLA-2-Haplo to predict haplotypes. Our tools have also been used in biomedical research: using HLA-Upgrade, we have updated two large solid-organ transplantation cohorts from low-resolution to high-resolution genotypes, hence empowering us to now carry out allele, haplotype and immunogenetic data (HLA-2-Haplo additional functionalities) associations with graft survival.

In conclusion, Easy-HLA (freely available online at: https://hla.univ-nantes.fr) facilitates large-scale analyses and promotes the multi-faceted HLA expertise beyond allelic associations. Our tool perfectly illustrates how computational and statistical modeling can relay and upgrade high-value experimental data to better enlighten clinical practice and sustain research.

Funding

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 number 846520. dbGaP Study Accession: phs001123.v1.p1.

Conflict of Interest: none declared.

References

Ahmad,T. *et al.* (2003) Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC. *Hum. Mol. Genet.*, **12**, 647–656.

Barnes,K.C. *et al.* (1996) Linkage of asthma and total serum IgE concentration to markers on chromosome 12q: evidence from Afro-Caribbean and Caucasian populations. *Genomics*, **37**, 41–50.

Clark,A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, **7**, 111–122.

Copelan,E.A. (2006) Hematopoietic stem-cell transplantation. *N. Engl. J. Med.*, **354**, 1813–1826.

Eberhard,H.-P. *et al.* (2013) Comparative validation of computer programs for haplotype frequency estimation from donor registry data. *Tissue Antigens*, **82**, 93–105.

Erlich,H. (2012) HLA DNA typing: past, present, and future. *Tissue Antigens*, **80**, 1–11.

Excoffier,L. and Slatkin,M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.*, **12**, 921–927.

Goodin,D.S. *et al.* (2018) Highly conserved extended haplotypes of the major histocompatibility complex and their relationship to multiple sclerosis susceptibility. *PLoS One*, **13**, e0190043.

Gourraud,P.-A. *et al.* (2005) Inferred HLA haplotype information for donors from hematopoietic stem cells donor registries. *Hum. Immunol.*, **66**, 563–570.

Gourraud,P.-A. *et al.* (2015) High-resolution HLA-A, HLA-B, and HLA-DRB1 haplotype frequencies from the French Bone Marrow Donor Registry. *Hum. Immunol.*, **76**, 381–384.

Gragert,L. *et al.* (2013) Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum. Immunol.*, **74**, 1313–1320.

Held,P.J. *et al.* (1994) The impact of HLA mismatches on the survival of first cadaveric kidney transplants. *N. Engl. J. Med.*, **331**, 765–770.

Horton,R. *et al.* (2004) Gene map of the extended human MHC. *Nat. Rev. Genet.*, **5**, 889–899.

Howell,W.M. *et al.* (2010) The HLA system: immunobiology, HLA typing, antibody screening and crossmatching techniques. *J. Clin. Pathol.*, **63**, 387–390.

Hurley,C.K. *et al.* (2004) Hematopoietic stem cell donor registry strategies for assigning search determinants and matching relationships. *Bone Marrow Transplant*, **33**, 443–450.

Johnston,H.R. *et al.* (2017) Identifying tagging SNPs for African specific genetic variation from the African Diaspora Genome. *Sci. Rep.*, **7**, 46398.

Lee,S.J. *et al.* (2007) High-resolution donor-recipient HLA matching contributes to the success of unrelated donor marrow transplantation. *Blood*, **110**, 4576–4583.

Loiseau,P. *et al.* (2007) HLA association with hematopoietic stem cell transplantation outcome: the number of mismatches at HLA-A, -B, -C, -DRB1, or -DQB1 is strongly associated with overall survival. *Biol. Blood Marrow Transplant*, **13**, 965–974.

MacArthur,J. *et al.* (2017) The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.*, **45**, D896–D901.

Madbouly,A. *et al.* (2014) Validation of statistical imputation of allele-level multilocus phased genotypes from ambiguous HLA assignments. *Tissue Antigens*, **84**, 285–292.

Mathias,R.A. *et al.* (2016) A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat. Commun.*, **7**, 12522.

Pappas,D.J. *et al.* (2015) Comparison of high-resolution human leukocyte antigen haplotype frequencies in different ethnic groups: consequences of sampling fluctuation and haplotype frequency distribution tail truncation. *Hum. Immunol.*, **76**, 374–380.

Robinson,J. *et al.* (2015) The IPD and IMGT/HLA database: allele variant databases. *Nucleic Acids Res.*, **43**, D423–D431.

Salem,R.M. *et al*. (2005) A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. *Hum. Genomics*, **2**, 39–66.

Schaid,D.J. *et al*. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet*., **70**, 425–434.

Stephens,M. and Donnelly,P. (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet*., **73**, 1162–1169.

Tian,C. *et al*. (2017) Genome-wide association and HLA region fine-mapping studies identify susceptibility loci for multiple common infections. *Nat. Commun*., **8**, 599.

Vince,N. *et al*. (2016) HLA-C Level is regulated by a polymorphic Oct1 binding site in the HLA-C promoter region. *Am. J. Hum. Genet*., **99**, 1353–1358.

Vince,N. *et al*. (2014) HLA class I and KIR genes do not protect against HIV type 1 infection in highly exposed uninfected individuals with hemophilia A. *J. Infect. Dis*., **210**, 1047–1051.

Zachary,A.A. and Leffell,M.S. (2016) HLA mismatching strategies for solid organ transplantation—a balancing act. *Front. Immunol*., **7**, 575.

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

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

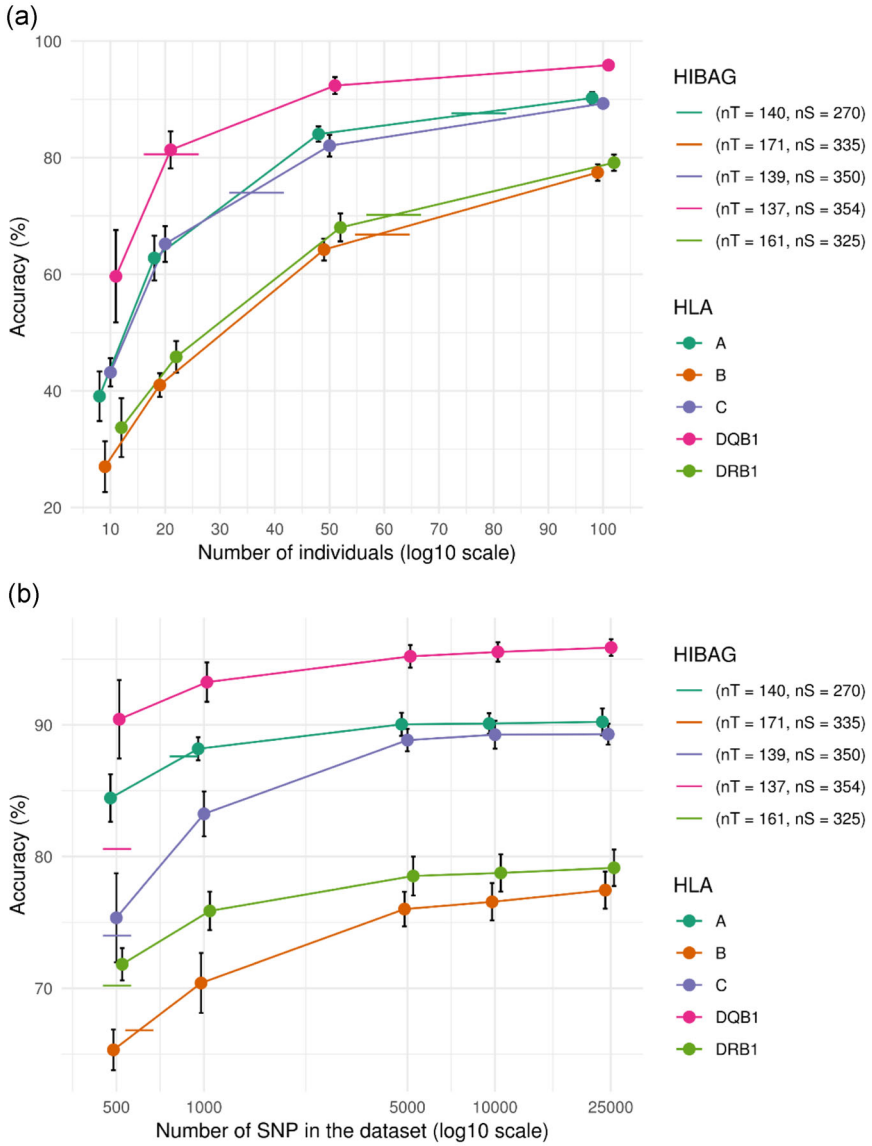

FIGURE 1 Influence of the number of individuals (a) and SNPs (b) in the HIBAG reference panel building on the accuracy of *HLA* alleles prediction. From the CAAPA data set ($N = 880$ and SNPs = 24,504), we produced a set of 10 training subsets ($n_{\text{training}} = 100$) and test ($n_{\text{test}} = 780$) 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

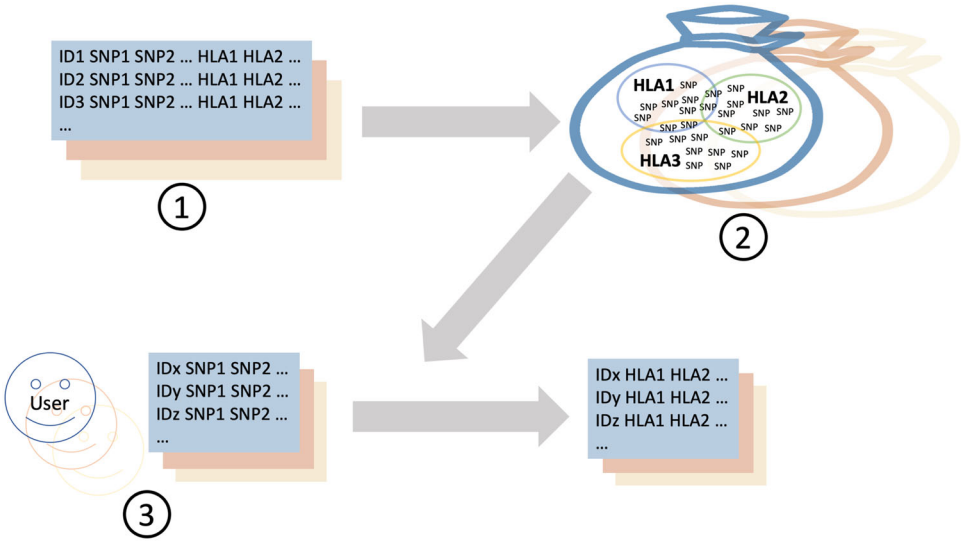

FIGURE 2 The SNP-HLA Reference Consortium (SHLARC) design. Aim 1: Increase the amount of SNP + HLA data available both in terms of quantity and diversity. Aim 2: Optimize SNP-HLA imputation methods. Aim 3: The SHLARC website will allow users from the scientific community to benefit from the data and 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;1–8. https://doi.org/10.1002/gepi.22334

frontiers in Genetics

REVIEW
published: 02 December 2021
doi: 10.3389/fgene.2021.774916

Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research

Venceslas Douillard[1], Erick C. Castelli[2], Steven J. Mack[3], Jill A. Hollenbach[4,5], Pierre-Antoine Gourraud[1], Nicolas Vince[1†] and Sophie Limou[1,6*†] on behalf of the Covid-19|HLA & Immunogenetics Consortium and the SNP-HLA Reference Consortium*

[1]Centre de Recherche en Transplantation et Immunologie, CHU Nantes, Inserm, Centre de Recherche en Transplantation et Immunologie, Université de Nantes, Nantes, France, [2]Unesp—Universidade Estadual Paulista, Botucatu, Brazil, [3]Division of Allergy, Immunology and Bone Marrow Transplantation, Department of Pediatrics, School of Medicine, University of California, San Francisco, San Francisco, CA, United States, [4]Department of Neurology, University of California, San Francisco, San Francisco, CA, United States, [5]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, United States, [6]Ecole Centrale de Nantes, Department of Computer Sciences and Mathematics in Biology, Nantes, France

OPEN ACCESS

Edited by:
Ramcés Falfán-Valencia, Instituto Nacional de Enfermedades Respiratorias-México (INER), Mexico

Reviewed by:
Gilberto Vargas Alarcón, Instituto Nacional de Cardiologia Ignacio Chavez, Mexico
Pengyu Hong, Brandeis University, United States
Martha Perez-rodriguez, Mexican Social Security Institute (IMSS), Mexico

***Correspondence:**
Nicolas Vince
nicolas.vince@univ-nantes.fr
Sophie Limou
sophie.limou@univ-nantes.fr

[†]These authors have contributed equally to this work and share last authorship.

Specialty section:
This article was submitted to Human and Medical Genomics, a section of the journal Frontiers in Genetics

Received: 13 September 2021
Accepted: 08 November 2021
Published: 02 December 2021

Citation:
Douillard V, Castelli EC, Mack SJ, Hollenbach JA, Gourraud P-A, Vince N and Limou S (2021) Approaching Genetics Through the MHC Lens: Tools and Methods for HLA Research. Front. Genet. 12:774916. doi: 10.3389/fgene.2021.774916

The current SARS-CoV-2 pandemic era launched an immediate and broad response of the research community with studies both about the virus and host genetics. Research in genetics investigated HLA association with COVID-19 based on *in silico*, population, and individual data. However, they were conducted with variable scale and success; convincing results were mostly obtained with broader whole-genome association studies. Here, we propose a technical review of HLA analysis, including basic HLA knowledge as well as available tools and advice. We notably describe recent algorithms to infer and call HLA genotypes from GWAS SNPs and NGS data, respectively, which opens the possibility to investigate HLA from large datasets without a specific initial focus on this region. We thus hope this overview will empower geneticists who were unfamiliar with HLA to run MHC-focused analyses following the footsteps of the Covid-19|HLA & Immunogenetics Consortium.

Keywords: Major Histocompatibility Complex (MHC), HLA, association analysis, imputation, immunogenetics

INTRODUCTION TO HUMAN LEUKOCYTE ANTIGENS: CREATING IMMUNITY FROM DIVERSITY

The classical HLA proteins are expressed on the surface of human cells. Although their primary role is to present exogenous and endogenous peptides, they were first described as "antigens" due to their interaction with T-cells in transplant rejection (Dausset, 1958). Along with other genes in the *MHC* region, the products of the *HLA* genes are essential in the adaptive immune response. By presenting peptides to both $CD8^+$ (HLA class I molecules) and $CD4^+$ T cells (HLA class II molecules), HLA proteins initiate an immune response against foreign (non-self) peptides which may be defective products of translation, neo-antigens generated by mutated genes from tumor cells, or pathogenic in origin. In addition, class I HLA proteins interact with the KIR ligands of NK cells, including KIR and LILRB, which are important in innate immunity (Carrington et al., 2008; Kulkarni et al., 2008; Trowsdale and Moffett, 2008). Thus, HLA molecules are key features of both innate and adaptive immune responses. *HLA* genes central role in immunity against

infectious diseases and their importance for transplantation have made them the subject of much study.

HLA proteins are coded by multiple genes on the short arm of chromosome 6 at the 6p21 locus; this region containing *HLA* genes is referred to as the Major Histocompatibility Complex (MHC) for its seminal role in transplantation (Dausset, 1981; Montgomery et al., 2018). Although there is a common confusion between the two terms *HLA* and *MHC*, *HLA* specifically refers to the genes involved in antigen processing and presentation whereas the *MHC* corresponds to a whole locus, with *HLA* and other immune-related genes such as the complement system. The *MHC* region is the most gene-dense region of the human genome, with 1% of the human coding genes (>200) found in 0.1% of the genome length (Shiina et al., 2009). The *MHC* region is commonly defined as a 4 Mb segment on chromosome 6 (*MOG* 29657002–33192499 *COL11A2*, GRCh38. p13 assembly) (Beck et al., 1999). However, due to extended patterns of linkage disequilibrium (LD), an extended MHC (xMHC) is often referred to in immunogenomics (25726063–33400556, GRCh38. p13 assembly) (Horton et al., 2004). The *MHC* region is divided into three regions based on gene sequence similarities and functions, class I, II, and III in which approximately 40% of the genes are immune-related. *HLA* genes are found in the class I and class II regions and are commonly divided in two categories: classical HLA proteins present peptides to T-cells, whereas non-classical HLA are mostly involved either in peptide presentation with other receptors, with immune modulation, or with various steps of classical HLA formation and loading.

The *MHC* class I region contains 12 *HLA* pseudogenes and 6 *HLA* genes (*HLA-A*, *-B*, *-C*, *-E*, *-F,* and *-G*), including three classical (*HLA-A*, *-B,* and *-C*) that are ubiquitously expressed as a heterodimer with beta-2 microglobulin at the cells' surface. Class I HLA molecules and their bound peptides are specifically recognized by $CD8^+$ T cells receptors. The non-classical HLA class I molecules (*HLA-G*, *-E*, and *-F*) present different expression patterns. *HLA-E* and *HLA-F* are usually ubiquitously expressed in low levels, and they interact with ligands in T and NK cells (such as HLA-E with NKG2A). *HLA-G* is predominantly expressed at the maternal-fetal interface and has primarily been associated with maternal–fetal tolerance by interacting CD8 from T cells and LILRB1, LILRB2, and KIR2DL4 from NK cells (Donadi et al., 2011).

The Class II region comprises four non-classical genes (*HLA-DMA*, *-DMB*, *-DOA*, *-DOB*), mostly related to peptide loading, and 17 classical *HLA* genes (e.g., *HLA-DRA*, *HLA-DRB1*, *HLA-DQA1*, *HLA-DQB1*, *HLA-DPA1*, and others) that are expressed in Antigen Presenting Cells (APC) such as B cells, monocytes, macrophages, dendritic cells as well as epithelial cells under inflammatory signals (Rock et al., 2016). Unlike class I HLA molecules, class II molecules are heterodimers, consisting of α and β chains, encoded by the corresponding *HLA* genes (e.g., *HLA-DPA1* and *HLA-DPB1* produce the HLA-DP molecule), which facilitates molecular diversity. The HLA-DR beta chain can be encoded by nine different genes (DRB1-9) with complex patterns of expression, and gene content adding additional layers of complexity (Faner et al., 2009). Finally, the class III region, located between the class I and II regions, is the most gene-dense region of the *MHC*; this region contains genes encoding elements of the complement system, chaperone genes, cytokines such as *TNF* and *LTA*, but no *HLA* genes.

Finally, there are other important non-*HLA* genes in the *MHC*, such as *TAP1* and *TAP2*, both related to peptide pumping from the cytoplasm to the endoplasmic reticulum (Praest et al., 2018), *MICA* and *MICB*, both induced in viral infections and tumors and activate NK-mediate killing (Ghadially et al., 2017), the tripartite motif (TRIM) family, related to cell cycle progression, autophagy, and viral replication restriction (van Tol et al., 2017), *PSORS1C1*, conferring susceptibility to psoriasis and systemic sclerosis (Allanore et al., 2011), and others.

In addition to their large number and potential for many combinations, the *HLA* genes display unparalleled genetic diversity, with more than 27,000 alleles and almost 17,000 unique proteins (June 03, 2021, https://www.ebi.ac.uk/ipd/imgt/hla/stats.html) identified for the five most polymorphic loci (*HLA-A*, *-C*, *-B*, *-DRB* and *-DQB1*). This diversity of HLA molecules is concentrated in the peptide-binding groove, which allows the presentation of peptides of various shapes and sizes, hence conferring broad protection against pathogens at the population level. At the same time, the polymorphic nature of *HLA* is also found on non-coding parts of the genes, such as the promoter and have an impact on expression (Kulkarni et al., 2011; Vince et al., 2016; Lima et al., 2019). Over evolutive time, together with founder effects, multiple pathogen-challenges have exerted selective pressures on *HLA* alleles in human populations across the globe (Meyer and Thomson, 2001; Spurgin and Richardson, 2010), shaping allele frequency differences and selecting very specific or even private *HLA* alleles in some populations (Brandt et al., 2018; Meyer et al., 2018). The progress of genomics, and immunogenomics over the last decade, had deepened our understanding of HLA role in human diseases though the use of genome-wide association studies (GWAS) (Kennedy et al., 2017; Dendrou et al., 2018).

The COVID-19 HLA and Immunogenetics Consortium (CHIC) has been created during the pandemic to coordinate efforts on *HLA* analysis. The CHIC provided a website with information on HLA data and current projects (The COVID-19 HLA and Immunogenetics Consortium, 2020a). It is supported by a database (The COVID-19 HLA and Immunogenetics Consortium, 2020b) and its role is the centralization of relevant HLA and clinical data for COVID-19 study. It contains HLA data of 2,892 individuals from nine projects. These data are freely available and new data can be easily uploaded upon account creation. In addition, the website allows HLA allele frequencies visualization, and use of HLA data management and analysis tools. An HLA Imputation Portal (HIP) is set up to allow geneticists to infer individuals HLA alleles using SNP genotyping data, relying on multi-ethnic models from Zheng et al. (Zheng et al., 2014). This tool may help leverage SNP data to gain power in HLA association studies. The CHIC also produced a broad review on immunogenetic parameters (e.g., HLA, KIR, complement, cytokines and chemokines receptors) and their role in COVID-19 (Aguiar et al., 2021). A more specific review of COVID-19 and HLA

associations (Douillard et al., 2021) highlights links between the pathology and HLA at different levels, from allele frequency correlation to HLA associations and haplotypes. The consortium will gradually improve its portal by providing access to additional and more diverse imputation reference panels, and by recruiting more individuals. Results from GWASs showed no association between HLA SNPs and COVID-19 infection (COVID-19 Host Genetics Initiative, 2021) but demonstrated an association with COVID-19 severity; dedicated HLA allele association studies identified potential signals of interest (Castelli et al., 2021). The spread of HLA tools, allowing HLA allele inference from whole-exome or whole-genome sequencing as well as from GWAS SNP data will significantly increase the sample size from available cohorts to maximize the statistical discovery power of HLA-centric studies. In this report, we pursue this effort to provide an overview of methods for generating HLA data along with several analytical strategies to capitalize on this genetic information. We will also cover additional immunogenomic parameters, as MHC-related associations still have much to reveal (Trowsdale and Knight, 2013). We hope this work will empower researchers to include HLA-focused investigations in their palette and will contribute to promote efforts for in-depth explorations of the relationship between HLA and immune-related outcomes in this pandemic era.

GENERATING AND WORKING WITH HLA DATA

Performing immunogenetic studies can be a challenge for those unfamiliar with the specifics of HLA nomenclature. An individual HLA genotype can be obtained through multiple molecular techniques, the complexity of its nomenclature allows the alleles in a genotype to be described in different styles, and these data can be stored in a variety of file standards. Overall, HLA information can take multiple forms, requiring a comprehensive understanding of the nomenclature in order to run proper statistical analyses and find relevant associations.

Generating HLA Data

Originally, immunologists conducted microlymphocytotoxicity assays, testing patients T/B cells (for HLA class I) or B cells (for HLA class II) against different anti-sera or monoclonal antibodies in the presence of complement. Sera or antibodies recognizing the HLA antigens on cells would activate the complement and lyse the cell; this serology staining would reveal the patient HLA serotype (Park and Terasaki, 2000). Serology was however limited by the underlying complexity of HLA and it resulted in poor performances in transplantation (Hurley, 2021). The need to improve this performance and technique evolution, with the advent of PCR, conducted HLA specialists to switch to molecular typing. Molecular techniques were adopted for *HLA* typing; these methods allowed systematic identification of *HLA* alleles, based on sequence polymorphisms, providing a 'higher resolution' result that distinguishes many more allele categories than serological methods. This molecular typing consistently improved in resolution throughout the years driving nomenclature evolution along the way. Sequence-specific (PCR-SSO) methods rely on the hybridization of hundreds of labeled SSO probes targeting unique sequences in polymorphic regions. Sequence-specific priming (PCR-SSP) methods directly amplify elements of the *HLA* genes with PCR primers containing sequence-specific 3′ end polymorphisms, resulting in less ambiguity (inability to distinguish alleles with similar nucleotide sequences), than SSO methods (Meral and Beksaç, 2007).

Sanger sequencing-based typing (PCR-SBT) methods initially provided sequences of the exons that encoded the peptide-binding groove, and later overlapping sets of sequences for entire genes. PCR-SBT was the gold standard for *HLA* genotyping until the development of next-generation sequencing (NGS) methods (Meral and Beksaç, 2007; De Santis et al., 2013). Except for the last one, PCR-SBT, previous methodologies were not suitable to detect new variants, and their goals were detecting known polymorphisms.

The application of NGS was explored in 2012, as part of the 16th International HLA and Immunogenetics Workshop (IHIW), but, given issues with mapping of short reads, allelic imbalance, phasing, and high costs, PCR-SBT remained the gold standard. More recently, the integration of NGS technologies with bioinformatic solutions for immunogenetics has improved the speed and accuracy of NGS *HLA* genotyping with lower error rates and fewer ambiguities than PCR-SBT (Baier et al., 2019; Jekarl et al., 2021), and the application of NGS was the focus of the 17th IHIW in 2017 (Vayntrub et al., 2020). Moreover, NGS is ideal to detect new *HLA* variants. Researchers now routinely identify novel *HLA* alleles (Nilsson et al., 2018; Ralazamahaleo et al., 2019; Loginova et al., 2020; Ananeva et al., 2021a; Ananeva et al., 2021b; Cheranev et al., 2021; Loginova et al., 2021) using NGS and confirm them using SBT with PCR-SBT error often responsible for non-concordance between the two. NGS-based sequencing of multiple exons and introns has led to increases in the growth of the IPD-IMGT-HLA Database collection (Robinson et al., 2015; Robinson et al., 2019). Unfortunately, the total number of new alleles may be underestimated as it is not uncommon for new alleles to be NGS-typed without Sanger validation.

So-called third-generation NGS generates unambiguous, phased *HLA* genotypes, using instruments like the PacBio SMRT (Mayor et al., 2015) or Oxford Nanopore Technology MinION (De Santis et al., 2020) to avoid multiple molecular techniques. This approach is faster than SBT and generates phased polymorphism with longer reads. Researchers using Oxford Nanopore Technology systems have successfully sequenced 11 *HLA* loci with low ambiguities in under 6 h (Mosbruger et al., 2020).

HLA Nomenclature

Soon after cell-surface antigens were identified as polymorphic between individuals, the WHO Nomenclature Committee for Factors of the HLA System was formed to develop a specific nomenclature for *HLA* genes, proteins and allelic variants (Allen et al., 1968). The original "HL-A" factor serologically

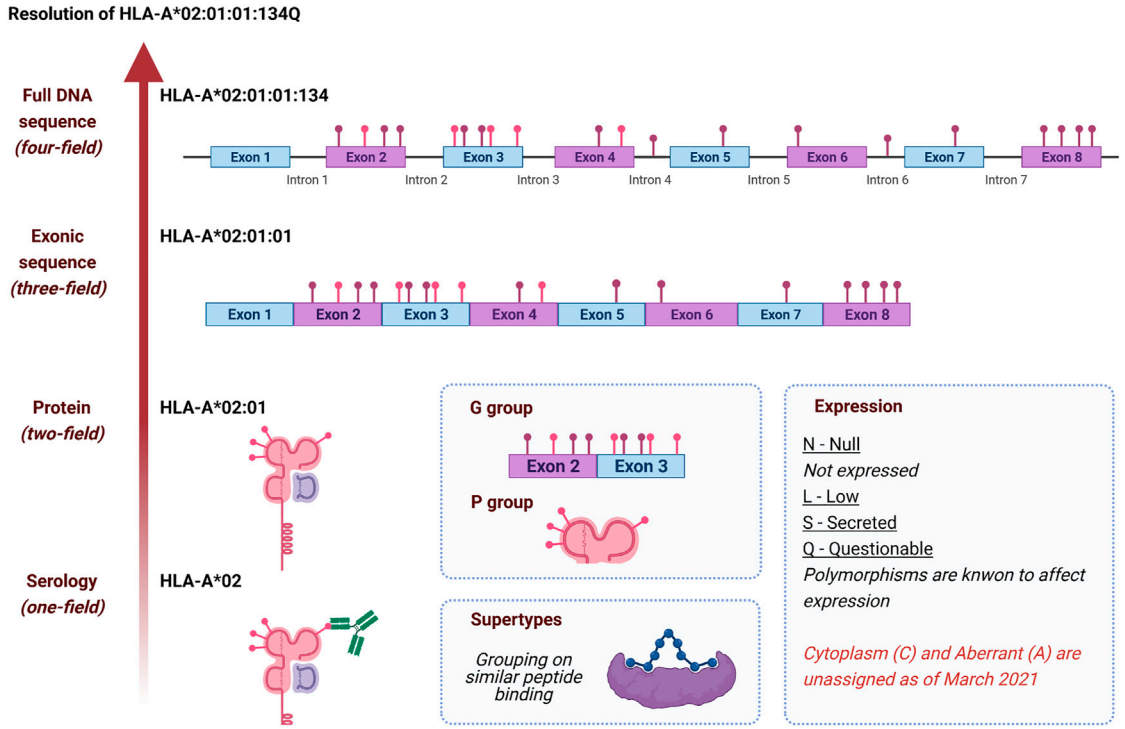

FIGURE 1 | History and development of HLA nomenclature as illustrated by *HLA-A*02:01:01:134Q*. Each level of resolution corresponds to a group of *HLA* alleles fitting the description, except for the full DNA sequence, a unique *HLA* allele. Colored pins represent non-synonymous polymorphism (pink) and synonymous or intronic polymorphisms (purple); the displayed polymorphism is only indicative and does not reflect *HLA-A*02:01:01:134Q* sequence. P and G groups are named with the lowest numbered two-field (HLA-A*02:01) and three-field (HLA-A*02:01:01) HLA allele name, respectively. Class II P and G groups are based on exon 2 only, while class I P and G groups are based on exons 2 and 3. Supertypes are not defined as part of the official nomenclature (Wang and Claesson, 2014; del Guercio et al., 1995; Sidney et al., 1995). Created with biorender.com.

typed with multiple antibodies with an individual type (e.g., HLA (1,2/7,8) identifying them as positive for factors 1,2,7,8, and confirmed two distinct haplotypes from parental typing. As dozens of *HLA* genes and thousands of alleles were identified, the nomenclature was expanded to accommodate new complexity while building on the historical serological vocabulary. In 1987, the nomenclature was updated to accommodate newly available protein and nucleotide sequences (Antigens, 1987). The modern locus names were adopted at this time, and four-digits names were assigned to alleles, which were only defines as protein variants. In 2010, the current field-delimited nomenclature was adopted to account for the growing number of silent and non-coding nucleotide variants (Marsh et al., 2010).

A modern HLA allele name consists of up to four "fields", each of which includes a two- or more digit number, each separated by a colon (**Figure 1**).

The first and second fields represent a historical serological group, and a unique protein sequence, respectively. All allele names have at least two fields. Alleles sharing the 1st, and 2nd fields with a different 3rd field encode the same protein but have unique silent-substitution in the exonic sequence, whereas sequence differences contained in the introns are written in the 4th field. The four fields of an allele name can also be suffixed with a single-letter "expression variant", identifying alleles that are either not expressed, expressed at a low or questionable level, or secreted. For example, *HLA-A*02:01:01* represents an exonic sequence shared by e.g., *HLA-A*02:01:01:01* and *HLA-A*02:01:01:134Q*. In the latter case, the expression of *HLA-A*02:01:01:134Q* is Questionable, due to a potential alternate splicing nucleotide variant in intron 2. Allele names can be truncated to fewer fields for different applications, with each truncation described as a level of "resolution" (e.g., HLA-A*02 is a one-field resolution allele).

In addition to this allele nomenclature, specific groups of alleles have been defined. P and G groups refer to multiple alleles sharing either the same peptide or nucleotide sequence for the peptide-binding groove, respectively. For instance, *HLA-A*02:01:01:134Q* and *HLA-A*02:252* both belong to the A*02:01P P group; the two proteins are globally different but share the same peptide-binding groove. *HLA-A*02:01:01:134Q* and *HLA-A*02:89:01* belong to the A*02:01:01G G group as they share identical peptide-binding groove encoding exon sequences.

HLA supertypes are groups of alleles sharing similar peptide-binding repertoires. Supertypes are defined by "structural similarities, shared peptide-binding motifs, and identification of cross-reacting peptides" (Wang and Claesson, 2014). Using this classification, HLA-A*02:01:01:134Q potentially belongs with HLA-A*02:02, A*02:05, A*69:01 in the A2 supertype. (del

TABLE 1 | Tools for HLA analyses.

HLA application name	Description	URL
Alphlard-nt (Hayashi et al., 2019)	Identification of somatic mutations in HLA molecules from whole-genome and exome data using Bayesian algorithms	—
BIGDAWG (Pappas et al., 2016)	Open-source R package for the case-control analysis of highly polymorphic data at the allele, haplotype and amino-acid level	https://CRAN.R-project.org/package=BIGDAWG
Easy-HLA (Geffard et al., 2020)	Website with HLA alleles haplotyping, upgrading and inference from HLA genotypes, prediction of HLA-C expression	http://hla.univ-nantes.fr/
HATK (Choi et al., 2021)	Open-source *Python* pipeline for HLA association studies, including tools for HLA data formatting	https://github.com/WansonChoi/HATK
HLA-check (Jeanmougin et al., 2017)	Perl tool evaluating the probability of accurate HLA genotype imputation by comparing it to SNP imputation in the exonic region of HLA.	https://github.com/mclegrand/HLA-check/
HLA-EMMA (Kramer et al., 2020)	Donor/recipient compatibility assessment based on solvent-accesible amino acids, based on intralocus comparisons	http://www.HLA-EMMA.com
HLA*fix*	Open-source R pipeline for HLA association studies. Performing SNP quality control steps, stratification, HLA imputation and representation of the results	https://univ-nantes.io/Nico_V/hlafix
HLAHapV (Osoegawa et al., 2016)	A Java-based HLA Haplotype Validator for quality assessments of HLA typing	https://github.com/nmdp-bioinformatics/ImmunogeneticDataTools
HLA-NET (Nunes et al., 2014)	Set of tools to manipulate HLA data, infer haplotypes, convert files format, and information about typing	https://hla-net.eu/
HLApers (Aguiar et al., 2020)	Genotyping and quantification of HLA expression from RNA-seq data	https://github.com/genevol-usp/HLApers
HLA-TAPAS (Luo et al., 2020)	Open-source *Python* pipeline for creation of reference panels and HLA association studies	https://github.com/immunogenomics/HLA-TAPAS
MergeReference (Cook and Han, 2017)	SNP2HLA compatible tool to concatenate multiple reference panels in order to gain accuracy during HLA imputation	http://software.buhmhan.com/MergeReference
pyHLA (Fan and Song, 2017)	Association analysis for HLA alleles in *Python* language	https://github.com/felixfan/PyHLA

Guercio et al., 1995). Some studies of the HLA molecules' evolution have interpreted HLA diversity differently. Kaufman et al. (Kaufman, 2018; Di et al., 2021) have proposed promiscuous and generalist HLA categories when Di et al. have challenged the concepts of supertypes and function peptide-binding groove groups.

HLA Data Formats

The modern and legacy nomenclature systems are still in use, which often makes data comparison and meta-analysis difficult. In addition, *HLA* alleles are stored in multiple formats which impact their use with bioinformatic tools. TSV or CSVs have been used to store HLA genotypes, usually organizing individuals in rows and *HLA* genes in columns (with two columns for each gene). Such files are often generated manually, but are used by multiple population genetic and disease-association applications (Lancaster et al., 2007; Excoffier and Lischer, 2010; Pappas et al., 2016). More strictly-defined bioinformatic-oriented formats include HLA PED (or HPED) (Choi et al., 2021), an HLA-focused extension of the PED format (Purcell et al., 2007); Variant Call Format (VCF), as used by BEAGLE (Browning et al., 2018), in which HLA allele names are recoded as multiple binary identifiers, and Histoimmunogenetic Markup Language (HML), an XML format developed specifically for exchanging HLA and Killer-cell Immunoglobulin-like Receptor (KIR) genotype data (Milius et al., 2015).

The IPD-IMGT/HLA Database releases new and updated reference sequences and allele names every 3 months. Individuals datasets may have been generated under any release version, which is why tools like the Allele Name Translation Tool (ANTT) have been developed to standardize datasets to a common release version. (Mack and Hollenbach, 2010). Development of a standardized means of storing and sharing data is still underway. In 2015, the MIRING reporting guideline (Mack et al., 2015) introduced standardized data elements and a controlled vocabulary for HLA genotype data and meta-data, which were implemented in HML (Milius et al., 2015). An HML message includes information on the IPD-IMGT/HLA Database version, the entity and how they generated the data, as well as references to external sources (e.g., reference sequences and aligned read). HML is used to transmit HLA genotyping data to the National Marrow Donor Program (and other similar registries and donor centers), but has yet to be adopted for genetic–analysis applications. Most of the existent HLA analysis applications require fewer data elements than are included in an HML message.

Given the number of different applications of HLA data, new informatics tools can influence the interpretation of this information. Multiple ancillary tools have been developed for HLA research. Whether they allow researchers to run rapid association analyses, extract new information from data, or link HLA genotypes to novel fields of translational research, all contribute to the advances in the HLA research (**Table 1**).

INFERRING AND IMPUTING *HLA* ALLELES: FROM COMPLEX READ-MAPPING TO THE STUDY OF LINKAGE DISEQUILIBRIUM

HLA inference is an umbrella term comprising multiple bioinformatic tools and statistical methods to obtain individuals' *HLA* genotypes. Inference implies using missing information to obtain *HLA* genotypes, this can generally refer

to using untargeted sequencing data, which have insufficient sequence read depth, to thoroughly recover the *HLA* alleles polymorphisms (Klasberg et al., 2019).

Inference From Whole-Genome Sequencing and Whole-Exome Sequencing

Unlike NGS typing techniques which targets *HLA* genes (as many commercial kits apply), untargeted sequencing does not focus on *HLA*. Whole-genome sequencing (WGS) methods aim to identify all genetic variations of an individual genome, while whole-exome sequencing (WES) is designed to target all exons. Initially, these methods did not support the calling of *HLA* alleles; low coverage and short read-lengths led to poor *HLA* typing accuracy (Bauer et al., 2016). Low coverage does allow identification of *HLA* alleles, due to their high levels of polymorphism and extensive conserved sequences among genes, and improvements were needed (Hosomichi et al., 2015). Moreover, general pipelines for analyzing NGS data from WGS do not work for *HLA* genes; because they present high sequence similarity, it is very common that a short read (a sequence generated in NGS procedures) from one gene aligns to another gene (cross-mapping), leading to genotyping errors (e.g., *HLA-A* and *HLA-H*, or *HLA-C* and *HLA-B*) (Castelli et al., 2018). The intense polymorphism observed in *HLA* genes may bias read alignment when using a single genome reference, especially when one individual presents too many modifications compared to the reference genome. This issue overestimates reference allele frequencies and causes genotyping errors (Brandt et al., 2015). Therefore, it is mandatory to use methods tailored for *HLA* genes to get reliable genotypes and haplotypes at the SNP level from NGS data.

Multiple algorithms have been developed and refined (Klasberg et al., 2019). These include: 1) classic read-mapping with HLA-specific quality control steps or different scores, hla-mapper (http://www.castelli-lab.net/apps/hla-mapper) (Castelli et al., 2018) which also works on KIR genes and provide genotyping and haplotyping at the SNP level, seq2HLA (Boegel et al., 2012) and HLAforest (Kim and Pourmand, 2013), among other tools; 2) population graph reference methods (e.g., HLA*PRG:LA), which identify probability edges between polymorphisms nodes and project read data onto these to evaluate the most likely alleles.

Recent reviews and tool comparisons on the optimal methods for non-*HLA* targeted sequencing data are already available (see (Klasberg et al., 2019; Chen et al., 2021)). In 2020, Chen et al. found that HLA-HD was the most accurate tool for producing HLA genotypes from WGS and WES. However, the study focused on the performance of five tools only. Notably, most of the tools they studied achieved much higher accuracies than previously reported by Bauer et al., 2016, which emphasizes a drastic improvement in read coverage and processing in the *MHC* region (Bauer et al., 2016). Finally, researchers successfully implemented these tools in association studies, promoting their importance for HLA-centric epidemiological studies (Juhos et al., 2015; Xie et al., 2017; Mimori et al., 2019; Vince et al., 2020a).

HLA Allele Imputation

HLA genotyping data can also be generated using *HLA* imputation tools, which generate genotypes for individuals on the basis of LD between GWAS-derived SNP data for the *MHC* region and specific *HLA* alleles. These methods ultimately rely on reference datasets of *HLA* and SNP genotypes for the same individuals, and have become increasingly accurate in their predictions as new algorithms are developed.

Following the opportunity brought by SNP to SNP imputation, SNP to *HLA* imputation algorithms offered a quick and easy way to obtain *HLA* genotypes from widely available GWAS SNP genotyping data (McCarthy et al., 2016). 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 (**Figure 2**).

The first published algorithms, SNP2HLA (Jia et al., 2013) and HLA*IMP (Dilthey et al., 2011), were based on different implementations of hidden Markov models; SNP2HLA used BEAGLE (Browning and Browning, 2009), a haplotyping and SNP genotype imputation tool. In 2014, Zheng et al. proposed HIBAG, an attribute bagging method tailored for *HLA* data (Zheng et al., 2014), which showed better performance than pre-existing tools, and at the time was the only method to provide population-specific reference panels for hundreds of individuals while enabling construction of personalized reference panels building. Initial independent reviews suggested that SNP2HLA performed better on 3,265 samples from BioVU, a de-identified electronic health record database coupled to a DNA biorepository (Karnes et al., 2017). However, later reviews (Kuniholm et al., 2016; Pappas et al., 2018) and studies (Ritari et al., 2020) have favored HIBAG for HLA imputation, notably on more complex HLA data.

In practice, both SNP2HLA and HIBAG are commonly used to conduct *HLA* imputation or creation of new reference panels. Overall accuracy differences are low for European panels that had been extensively assessed. An important point still under investigation is the impact of population diversity in reference panels. While some researchers advocate for the creation of exhaustive multi-ethnic reference panels (Degenhardt et al., 2019), others have shown that specific populations (e.g., insular or admixed require more restrained reference panels (Khor et al., 2015; Ritari et al., 2020).

The difficulty in determining if a reference panel is suitable for *HLA* imputation is related to how well it matches to target data, on the frequency of common alleles and the presence of rare *HLA* alleles, specific to some population (especially in underrepresented populations). This has led to the creation of reference panels with limited *HLA* diversity. While accuracy values are often reported as the ultimate answer to a model viability, these values can be misleading. For a rare *HLA* allele in a validation dataset, a 90% accuracy value can be achieved if that allele should be imputed 20 times out of 2,000 alleles (i.e., 1,000 individuals) but is never predicted. Therefore, other metrics (e.g., sensitivity, specificity, or F1 score (Cook et al., 2021)), must not be overlooked. Admixed populations are formed by individuals from different genetic backgrounds in

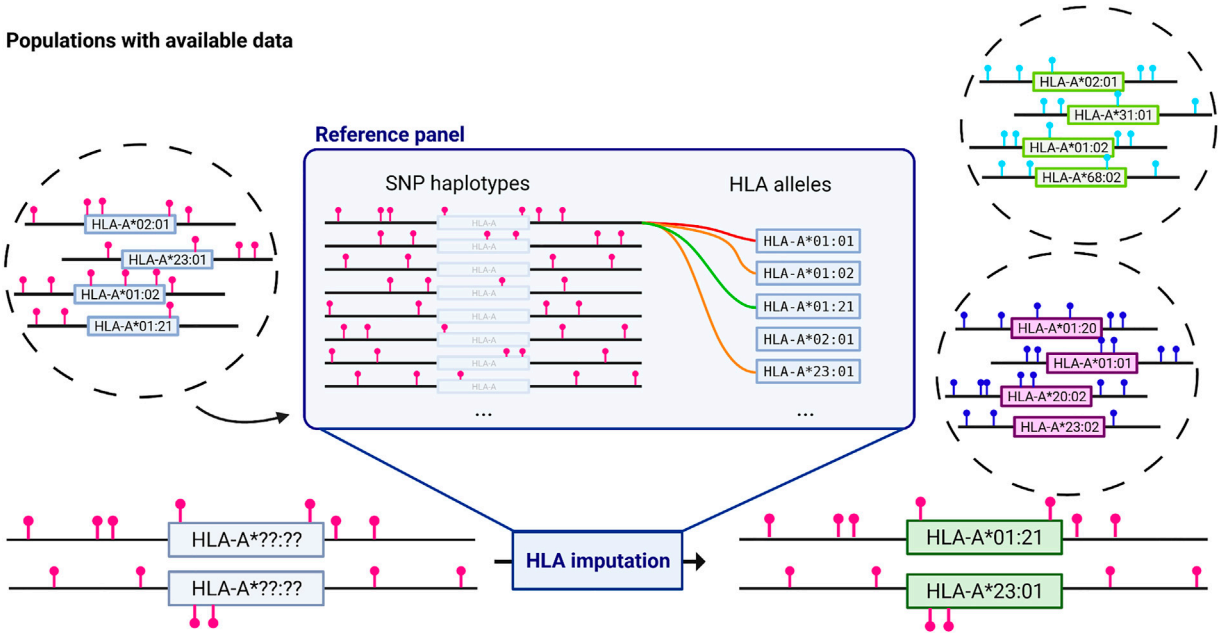

FIGURE 2 | *HLA* imputation from GWAS data. Reference panels are created from individuals with known SNP and *HLA* data. Depending on the method, an algorithm will deduce the probability of a specific *HLA* allele in the population given a SNP haplotype. These new found links are stored for that reference panel and applied to new SNP data to infer *HLA* genotypes. *HLA-A* is given as an example with a truncated list of alleles; other *MHC* genes are imputed using the same method. Different populations are represented in different circles and imply different allele frequencies. Pinpoints represent SNPs and are only indicative. *HLA* imputation results are highly dependent on the population chosen for the reference panel. Created with biorender.com.

variable proportions, and *HLA* imputation can be sub-optimal if the reference panel is only drawn from one of the ancestral populations. Conversely, a reference panel from an admixed population with a different overall genome proportion from the individuals being imputed may also provide inaccurate results.

To effect worldwide improvement in *HLA* imputation efforts, we led the creation of an international consortium, the SNP-HLA Reference Consortium (SHLARC), whose aim is to gather data to represent the extreme diversity of *HLA* alleles, fostering accurate imputation (Vince et al., 2020b). We further advocate for improvements to current *HLA* imputation tools and for the development of a platform promoting easy access to *HLA* imputation for immunogeneticists. Though *HLA* imputation is not yet suited for clinical settings, generalization of *HLA* association studies offers a new way to investigate immune pathologies (Meyer and Nunes, 2017).

New versions HLA*IMP (Motyer et al., 2016) and SNP2HLA have been released (e.g., MHC*IMP (Squire et al., 2020), CookHLA (Cook et al., 2021), and Deep-HLA (Naito et al., 2021)) that apply new algorithms. These highlight the community intense interest in *HLA* imputation. CookHLA is an updated version of SNP2HLA (based on the BEAGLE algorithm) that better accounts for LD in the *HLA* region and makes use of the genetic map option to better impute individuals who are not well represented in the reference panels. For its part, Deep-HLA seems especially promising as deep learning may lead to better imputation of rare alleles.

BIOINFORMATIC ANALYSES OF HLA INFORMATION

The pressing challenge of understanding the COVID-19 pandemic, given previous associations with infectious diseases, has led researchers to scrutinize *HLA* using any available resource. In addition to issues of nomenclature and on-going technological evolution of typing methods, the complexity of *HLA* analyses is also derived from the multiple forms these analyses can take. On the one hand, *HLA* allele frequencies and predicted binding affinity of pathogen peptides to *HLA* alleles allow for a first step in the HLA world, as they are easily available, but are limited to investigate its actual role. On the other hand, the in-depth implication of *HLA* is revealed when looking at SNP association in the MHC region, and specifically when looking at allele associations, but their realization is hindered by high costs and technical difficulties. The study of HLA is multi-layered, with a continuum of methods peaking with analysis of individual data and multi-locus haplotypes, all of which contributing to a comprehensive understanding of the role of HLA in a given analysis.

HLA Allele Frequencies

The diversity of *HLA* alleles across geographically separated populations is thought to be the result of balancing selection due to local pathogens (Meyer and Thomson, 2001). The allelefrequencies.net database has the most extensive collection of *HLA* allele frequencies in diverse populations (Middleton et al.,

2003). In addition, *HLA* typing conducted by bone marrow registries may constitute a local estimation of *HLA* allele distribution in a population (Sacchi et al., 2019; Schmidt et al., 2020). It is possible to statistically analyze the correlation (e.g., *via* linear regression or Pearson coefficient) between a quantitative value, such as the number of COVID-19 cases, and the *HLA* allele frequencies obtained from a different sample in every studied population (e.g., in a database or registry).

However, while these correlations are faster and easier to obtain than new *HLA* genotypes, they may result in spurious correlations because: 1) most of the *HLA* alleles (and observed haplotypes) have a low frequency. For example, according to allelefrequencies.net, in the 416,581 individuals from the African-American NMDP population in the United States, two-thirds of the 321 *HLA-B* alleles at two-field resolution have frequencies below 0.003% (24 or less occurrences). Assuming that reference population samples are representative is not always accurate. A possible solution is to focus on common *HLA* alleles; 2) statistical tests are often applied without multiple-testing correction, regardless of the number of tests; 3) the confounding variables, both genetic (e.g., ancestry) and environmental (e.g., comorbidities), are often overlooked.

In any case, *correlation* is not *causation.* Therefore, the high number of *HLA* alleles and biased frequencies are bound to create spurious links between their presence and any phenotype. Therefore, to thoroughly investigate the relationship between *HLA* and phenotype, it is of the utmost importance to conduct studies and control for other genetic factors such as population stratification, linkage disequilibrium, or comorbidities (some linked to *HLA* polymorphism itself such as diabetes). Statistical bias could also be reduced by working on a higher number of samples and correcting for multiple testing. It is also worth considering different resolution levels of information, from *in silico* studies to full haplotype information.

In silico Peptide Binding

HLA molecules present endogenous and exogenous peptides, however, affinities for these peptides vary greatly depending on the peptide conformation and the peptide cleft topology and chemistry. Whether an *HLA* allele presents several or few peptides derived from one specific pathogen is one mechanism potentially explaining the strong immune response or tolerance towards it. Researchers can use prediction tools, such as NetMHCpan (Nielsen and Andreatta, 2016; Jurtz et al., 2017), trained on binding affinity and elution assays, to evaluate the number of potentially bound peptides for any *HLA* class I allele. The "pan" methods, contrary to the "allele-specific" methods, use similarities in sequence data to predict the peptide binding capacity of HLA alleles for which no information is available. Other tools exist and have been reviewed by Mei et al., in 2020 (Mei et al., 2019). Such predictions, coupled with *HLA* genotype data of individuals, give a theoretical insight into the possible adaptive immune response of a person. In these tools, the peptidome of the studied pathogen is informatically divided into peptide sequences of limited size (8–12 residues to account for the size of peptides presented by class I molecules), and the number of alleles predicted to bind a large number of peptides is inferred to represent better presentation to T cells, and a protective role against the pathogen. However, the only way to definitely determine peptide binding affinity is through laboratory experiments.

Genome-wide Association Studies

Genotyping data obtained with SNP arrays has proven to be fast and inexpensive for investigating the genetic component of complex traits and diseases (Claussnitzer et al., 2020), compared to more thorough and exhaustive sequencing technologies. Without assumptions regarding the region potentially involved in the studied trait, GWAS helped discover protective and risk alleles, particularly in the *HLA* region (Kennedy et al., 2017). Contrary to the use of independent *HLA* allele frequencies for studying a pathology, association studies assess the difference between affected individuals and unaffected individuals or the distribution of a particular quantitative trait. Both genetic and phenotypic data are individual and not population-based, reducing biases. The statistically significant SNP (aka, top hits) are linked to genes by proximity, and investigation by pathway analysis can reveal additional biological information on their effect. More recently, transcriptome-wide association studies have allowed more accurate investigation of the impact of a SNP on the expression of genes (Wainberg et al., 2019). In addition, some SNPs can be highly correlated to an *HLA* allele (e.g., rs2395029 and *HLA-B*57:01* have been described multiple times as in complete linkage disequilibrium (de Bakker et al., 2006)), and therefore provide additional functional information for biological interpretation. Finally, statistical regression models can take into account potential confounding factors (e.g., genetic ancestry and population stratification, sex, age, comorbidities) to control for limiting biases.

Given the complex LD patterns across the *MHC* region, SNP association analyses are not usually precise enough to identify specific disease-associated *HLA* alleles. LD patterns may differ between populations. For example, the rs2395029 tags *HLA-B*57:01* in Europeans but displays reduced LD in African-Americans (Colombo et al., 2008). The complex LD patterns and the high number of genes in the *MHC* region, make it difficult to pinpoint an SNP to a specific *HLA* allele in most cases.

HLA Allele Association Studies

Association studies of *HLA* alleles offer a more relevant biological explanation, based on peptide presentation. *HLA* allele data can come from different sources, including various epochs of *HLA* typing and *HLA* imputation from SNPs (see above). These data can be analyzed as is, or low resolution HLA data can be "upgraded" using the HLA-Upgrade tool from the Easy-HLA website, which statistically impute the most probable two-field genotype based on a haplotype database (Geffard et al., 2020). Once HLA data from multiple sources have been standardized for allele content and resolution, a frequency cut-off value is usually applied to test only those alleles with sufficient occurrences in the dataset to guarantee statistical power in the analysis. *HLA* alleles being highly polymorphic, they often display lower frequencies, and a larger sample size is

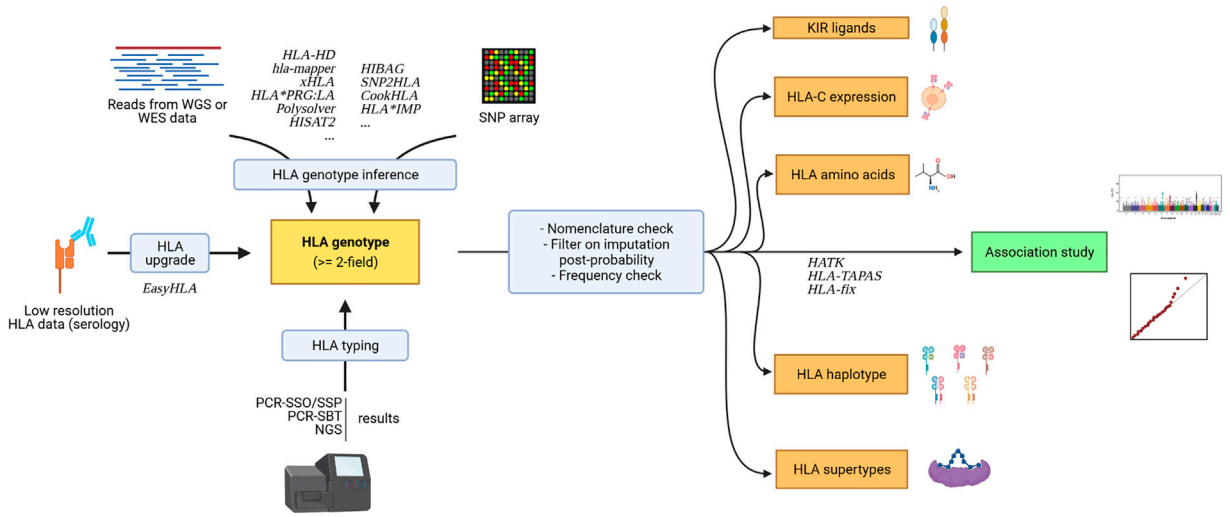

FIGURE 3 | Association study pipeline for HLA data and surrounding immunogenetic factors. Created with biorender.com.

usually required to obtain significant results compared to SNP analyses.

Regression models, which are commonly used for SNP association, are the most versatile and common statistical models implemented to test associations between individual *HLA* alleles and phenotypes of interest (linear models for continuous and logistic for discrete phenotypes, respectively). Regression models can work with multiple covariables, allowing the disentanglement of the HLA effect and confounding factors such as population stratification, sex, gender, and others. Similar to GWAS SNP analyses, *HLA* alleles are tested individually as biallelic markers for each *HLA* gene, as each individual can exhibit 0, 1, or 2 occurrences of a given allele. As HLA molecules are expressed co-dominantly (Hughes and Nei, 1988), the dominant genetic model is commonly preferred to allelic or recessive models to assess *HLA* allele associations. However, it should be mentioned that different alleles might present different expression levels due to promoter and 3′UTR variations and final protein stability. Indeed, this is another HLA world: the effect of variants in the expression levels, which sometimes are directly linked with disease susceptibility (Kulkarni et al., 2011).

As in GWAS analysis, the overall performance of a statistical model can be evaluated with a Quantile-Quantile (QQ) plot, representing the observed *p*-value distribution for each *HLA* allele compared to the expected distribution under the null hypothesis. Any deviation from this distribution is highlighted by a deviation from a straight line. (Murdoch et al., 2008). Different scenarios can be described: 1) observed *p*-values mostly follow the null hypothesis, indicating that the statistical model accurately fits the data; 2) observed *p*-values deviate below the null hypothesis line, indicating that the statistical model is probably underpowered; 3) observed *p*-values deviate above the null hypothesis line, indicating that the statistical model may not be well parameterized and some confounding factors are not enough considered. Once the robustness of the analysis is confirmed, it is important to obtain a comprehensive visualization of the results with Manhattan plots, for instance, displaying–log10 (*p*-value) along with the list of test *HLA* alleles ordered numerically (as seen in Vince et al. (Vince et al., 2020a)). Volcano plots can also display the significativity of alleles along with their effect size, allowing a global view of their impact. Finally, the significance threshold accounting for multiple testing can be determined with the Bonferroni correction (5% α threshold divided by the number of tests) or other corrections such as the FDR, or permutations.

Easy-HLA: Going Beyond *HLA* Alleles to *HLA* Genes Haplotypes, HLA Expression Levels, Specific HLA Amino Acids, KIR Ligand Groups

New tools have been developed to facilitate the analysis of additional immunogenetic parameters (e.g. KIR ligands, see **Figure 3**).

HLA genotypes can be used to infer additional immunogenetic parameters that can further be analyzed (see **Figure 3**.) to get a clearer understanding of the relationship between immunity and pathologies. While one *HLA* allele already represents a haplotype of SNPs within a gene, as it is a collection of polymorphisms in the gene of interest, researchers have demonstrated the importance of looking at multiple *HLA* alleles on the same chromosome, which is referred to as an *HLA* haplotype. Association studies can be done on haplotypes, but many haplotype frequencies can be even lower than constituent allele frequencies. In a clinical setting, the collection of haplotype information is also useful, notably in HSCT transplants, for identifying haploidentical individuals. These haplotypes can be inferred using the HLA-2-Haplo tool from Easy-HLA website (Geffard et al., 2020), for instance. A straightforward, reliable, but expansive strategy to get *HLA* gene haplotypes is the analysis of trios (mother, father, and offspring) or third-generation long-read sequencing such as PacBio SMRT.

Easy-HLA also infers HLA-C expression levels, HLA alleles amino acids, and KIR ligand groups. Recently, high HLA-C

expression levels were associated with better control of HIV (Apps et al., 2013; Vince et al., 2016). Class I *HLA* alleles have also been grouped according to their dependence on tapasin, a major actor in peptide loading, which proved to be an interesting subdivision for studying HIV-1 control (Bashirova et al., 2020). Moreover, testing HLA allele amino acids may indicate a specific function of a given residue across several alleles, as with this study by McLaren et al., again in HIV control (McLaren et al., 2012). Finally, studying KIR ligand groups along with KIR typing as previously described (Martin and Carrington, 2013; Vince et al., 2014) can reveal the binding patterns of specific HLA alleles. For example, HLA-A and HLA-B molecules bearing the Bw4+ motif bind specifically to KIR3DL1. Similarly, HLA-C group 1 (C1) allele-encoded molecules carry an asparagine at position 80 and specifically bind KIR2DL2/3, as opposed to group 2 (C2) allele-encoded molecules, which carry a lysine and specifically bind KIR2DL1 (Parham et al., 2012). Grouping *HLA* alleles according to different functional parameters can increase the power of detecting a true positive signal and represent an opportunity to come closer to the biological cause behind *HLA* genetic association with diseases.

CONCLUSION

However intricate it may be, the *MHC* region, and *HLA* in particular, is the perfect candidate to investigate infectious or auto-immune diseases, as its primary biological role is to present antigen to the immune system. HLA research was able to grow in different directions from *in silico* studies on peptide binding to association studies of *HLA* alleles, giving leads on HLA involvement in pathologies. That said, HLA-focused analysis requires special care because its immense diversity and low-frequency distribution may potentially result in spurious associations when tested incorrectly or in a small cohort. Fortunately, many tools have been and are still developed to obtain high-quality HLA information for a low cost with statistical inference, through *HLA* inference from NGS data or *HLA* imputation from SNP GWAS data or HLA resolution upgrading from HLA genotypes. Researchers considering to explore HLA should take advantage of existing resources and mobilize them when taking on new challenges, such as with the SARS-CoV-2 research.

AUTHOR CONTRIBUTIONS

VD contributed in writing the review and produced figures. EC, SM, JH, P-AG, NV, and SL contributed in writing and editing various sections of the review.

FUNDING

NV has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 846520. This work is supported by the ATIP-Avenir Inserm program, the Region Pays de Loire ConnectTalent. This work was also supported by United States, National Institutes of Health (NIH) National Institute of Allergy and Infectious Disease (NIAID) grants R01AI128775 (JH, SJM), and R01AI158861 (JH).

REFERENCES

Aguiar, V. R. C., Augusto, D. G., Castelli, E. C., Hollenbach, J. A., Meyer, D., Nunes, K., et al. (2021). An Immunogenetic View of COVID-19. *Genet. Mol. Biol.* 44 (1), 1–24. doi:10.1590/1678-4685-gmb-2021-0036

Aguiar, V. R. C., Masotti, C., Camargo, A. A., and Meyer, D. (2020). "HLApers: HLA Typing and Quantification of Expression with Personalized Index," in *Methods in Molecular Biology*. Editor S. Boegel (New York, NY: Springer US), Vol. 2120, 101–112. doi:10.1007/978-1-0716-0327-7_7

Allanore, Y., Saad, M., Dieudé, P., Avouac, J., Distler, J. H. W., Amouyel, P., et al. (2011). Genome-Wide Scan Identifies TNIP1, PSORS1C1, and RHOB as Novel Risk Loci for Systemic Sclerosis. *Plos Genet.* 7 (7), e1002091. doi:10.1371/journal.pgen.1002091

Allen, F. H., Amos, D. B., Batchelor, J. R., Bodmer, W. F., Ceppellini, R., Dausset, J., et al. (1968). Nomenclature for Factors of the HL-A System. *Bull. World Health Organ.* 39 (3), 483–486. Available at: http://www.ncbi.nlm.nih.gov/pubmed/5303912.

Ananeva, A., Sergeeva, I., Gusev, O., and Shagimardanova, E. (2021). Three Novel HLA-C Alleles Identified in Russian Individuals: C*04:01:124 , C*12:02:38, and C*12:03:64. *HLA* 97 (3), 237. doi:10.1111/tan.14178

Ananeva, A., Leksina, Y., Andryushkina, A., and Shagimardanova, E. (2021). The Novel HLA-A*02:941 Allele Was Identified during High-resolution HLA Typing. *Hla* 97 (2), 136–138. doi:10.1111/tan.14088

Antigens, T. (1987). Nomenclature for Factors of the HLA System. *Tissue Antigens* 32 (4), 177–187. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3217934.

Apps, R., Qi, Y., Carlson, J. M., Chen, H., Gao, X., Thomas, R., et al. (2013). Influence of HLA-C Expression Level on HIV Control. *Science (80-)* 340 (6128), 87–91. doi:10.1126/science.1232685

Baier, D. M., Hofmann, J. A., Fischer, H., Rall, G., Stolze, J., Ruhner, K., et al. (2019). Very Low Error Rates of NGS-Based HLA Typing at Stem Cell Donor Recruitment Question the Need for a Standard Confirmatory Typing Step before Donor Work-Up. *Bone Marrow Transpl.* 54 (6), 928–930. doi:10.1038/s41409-018-0411-2

Bashirova, A. A., Viard, M., Naranbhai, V., Grifoni, A., Garcia-Beltran, W., Akdag, M., et al. (2020). HLA Tapasin independence: Broader Peptide Repertoire and HIV Control. *Proc. Natl. Acad. Sci. USA* 117 (45), 28232–28238. doi:10.1073/pnas.2013554117

Bauer, D. C., Zadoorian, A., Wilson, L. O. W., and Thorne, N. P. (2016). Evaluation of Computational Programs to Predict HLA Genotypes from Genomic Sequencing Data. *Brief Bioinform* 19 (2), bbw097. doi:10.1093/bib/bbw097

Beck, S., Geraghty, D., Inoko, H., Rowen, L., Aguado, B., Bahram, S., et al. (1999). Complete Sequence and Gene Map of a Human Major Histocompatibility Complex. The MHC Sequencing Consortium. *Nature* 401 (6756), 921. doi:10.1038/44853

Boegel, S., Löwer, M., Schäfer, M., Bukur, T., de Graaf, J., Boisguérin, V., et al. (2012). HLA Typing from RNA-Seq Sequence Reads. *Genome Med.* 4 (12), 102. doi:10.1186/gm403

Brandt, D. Y. C., Aguiar, V. R. C., Bitarello, B. D., Nunes, K., Goudet, J., and Meyer, D. (2015). Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3 (Bethesda)* 5 (5), 931–941. doi:10.1534/g3.114.015784

Brandt, D. Y. C., César, J., Goudet, J., and Meyer, D. (2018). The Effect of Balancing Selection on Population Differentiation: A Study with HLA Genes. *G3 (Bethesda)* 8 (8), 2805–2815. doi:10.1534/g3.118.200367

Browning, B. L., and Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-phase Inference for Large Data Sets of Trios and Unrelated Individuals. *Am. J. Hum. Genet.* 84 (2), 210–223. doi:10.1016/j.ajhg.2009.01.005

Browning, B. L., Zhou, Y., and Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am. J. Hum. Genet.* 103 (3), 338–348. doi:10.1016/j.ajhg.2018.07.015

Carrington, M., Martin, M. P., and van Bergen, J. (2008). KIR-HLA intercourse in HIV Disease. *Trends Microbiol.* 16 (12), 620–627. doi:10.1016/j.tim.2008.09.002

Castelli, E. C., de Castro, M. V., Naslavsky, M. S., Scliar, M. O., Silva, N. S. B., Andrade, H. S., et al. (2021). MHC Variants Associated with Symptomatic versus Asymptomatic SARS-CoV-2 Infection in Highly Exposed Individuals. *Front. Immunol.* 12 (September), 1–11. doi:10.3389/fimmu.2021.742881

Castelli, E. C., Paz, M. A., Souza, A. S., Ramalho, J., and Mendes-Junior, C. T. (2018). Hla-mapper: An Application to Optimize the Mapping of HLA Sequences Produced by Massively Parallel Sequencing Procedures. *Hum. Immunol.* 79 (9), 678–684. doi:10.1016/j.humimm.2018.06.010

Chen, J., Madireddi, S., Nagarkar, D., Migdal, M., Vander Heiden, J., Chang, D., et al. (2021). In Silico tools for Accurate HLA and KIR Inference from Clinical Sequencing Data Empower Immunogenetics on Individual-Patient and Population Scales. *Brief Bioinform* 22 (3), 1–11. doi:10.1093/bib/bbaa223

Cheranev, V., Loginova, M., Jankevic, T., Rebrikov, D., and Korostin, D. (2021). HLA-A *11: 382N , a Novel HLA-A Null Allele Identified by Next-generation Sequencing. *Hla* 97 (5), 448–449. doi:10.1111/tan.14185

Choi, W., Luo, Y., Raychaudhuri, S., and Han, B. (2021). HATK: HLA Analysis Toolkit. *Bioinformatics* 37 (3), 416–418. doi:10.1093/bioinformatics/btaa684

Claussnitzer, M., Cho, J. H., Collins, R., Cox, N. J., Dermitzakis, E. T., Hurles, M. E., et al. (2020). A Brief History of Human Disease Genetics. *Nature* 577 (7789), 179–189. doi:10.1038/s41586-019-1879-7

Colombo, S., Rauch, A., Rotger, M., Fellay, J., Martinez, R., Fux, C., et al. (2008). TheHCP5Single-Nucleotide Polymorphism: A Simple Screening Tool for Prediction of Hypersensitivity Reaction to Abacavir. *J. Infect. Dis.* 198 (6), 864–867. doi:10.1086/591184

Cook, S., Choi, W., Lim, H., Luo, Y., Kim, K., Jia, X., et al. (2021). Accurate Imputation of Human Leukocyte Antigens with CookHLA. *Nat. Commun.* 12, 1–11. doi:10.1038/s41467-021-21541-5

Cook, S., and Han, B. (2017). MergeReference: A Tool for Merging Reference Panels for HLA Imputation. *Genomics Inform.* 15 (3), 108–111. doi:10.5808/gi.2017.15.3.108

COVID-19 Host Genetics Initiative (2021). Mapping the Human Genetic Architecture of COVID-19. *Nature*. Available at: http://www.nature.com/articles/s41586-021-03767-x.

Dausset, J. (1958). Iso-leuco-anticorps. *Acta Haematol.* 20 (1–4), 156–166. doi:10.1159/000205478

Dausset, J. (1981). The Major Histocompatibility Complex in Man: Past, Present and Futur Concepts. *Science (80-)* 213 (September), 55–97. Available at: http://linkinghub.elsevier.com/retrieve/pii/B9780124169746000065.

de Bakker, P. I. W., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., et al. (2006). A High-Resolution HLA and SNP Haplotype Map for Disease Association Studies in the Extended Human MHC. *Nat. Genet.* 38 (10), 1166–1172. doi:10.1038/ng1885

De Santis, D., Dinauer, D., Duke, J., Erlich, H. A., Holcomb, C. L., Lind, C., et al. (2013). 16 Th IHIW : Review of HLA Typing by NGS. *Int. J. Immunogenet.* 40 (1), 72–76. doi:10.1111/iji.12024

De Santis, D., Truong, L., Martinez, P., and D'Orsogna, L. (2020). Rapid High-resolution HLA Genotyping by MinION Oxford Nanopore Sequencing for Deceased Donor Organ Allocation. *Hla* 96 (2), 141–162. doi:10.1111/tan.13901

Degenhardt, F., Wendorff, M., Wittig, M., Ellinghaus, E., Datta, L. W., Schembri, J., et al. (2019). Construction and Benchmarking of a Multi-Ethnic Reference Panel for the Imputation of HLA Class I and II Alleles. *Hum. Mol. Genet.* 28 (12), 2078–2092. doi:10.1093/hmg/ddy443

del Guercio, M. F., Sidney, J., Hermanson, G., Perez, C., Grey, H. M., Kubo, R. T., et al. (1995). Binding of a Peptide Antigen to Multiple HLA Alleles Allows Definition of an A2-like Supertype. *J. Immunol.* 154 (2), 685–693. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7529283.

Dendrou, C. A., Petersen, J., Rossjohn, J., and Fugger, L. (2018). HLA Variation and Disease. *Nat. Rev. Immunol.* 18 (5), 325–339. doi:10.1038/nri.2017.143

Di, D., Nunes, J. M., Jiang, W., and Sanchez-Mazas, A. (2021). Like Wings of a Bird: Functional Divergence and Complementarity between HLA-A and HLA-B Molecules. *Mol. Biol. Evol.* 38 (4), 1580–1594. doi:10.1093/molbev/msaa325

Dilthey, A. T., Moutsianas, L., Leslie, S., and McVean, G. (2011). HLA*IMP-an Integrated Framework for Imputing Classical HLA Alleles from SNP Genotypes. *Bioinformatics* 27 (7), 968–972. doi:10.1093/bioinformatics/btr061

Donadi, E. A., Castelli, E. C., Arnaiz-Villena, A., Roger, M., Rey, D., and Moreau, P. (2011). Implications of the Polymorphism of HLA-G on its Function, Regulation, Evolution and Disease Association. *Cell. Mol. Life Sci.* 68 (3), 369–395. doi:10.1007/s00018-010-0580-7

Douillard, V., Castelli, E., Mack, S. J., Hollenbach, J., Gourraud, P-A., Vince, N., et al. (2021). Approaching Genetics through the MHC Lens: Current HLA Investigations on SARS-CoV-2 and Perspectives. *Front. Genet.* In review.

Excoffier, L., and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: a New Series of Programs to Perform Population Genetics Analyses under Linux and Windows. *Mol. Ecol. Resour.* 10 (3), 564–567. doi:10.1111/j.1755-0998.2010.02847.x

Fan, Y., and Song, Y.-Q. (2017). PyHLA: Tests for the Association between HLA Alleles and Diseases. *BMC Bioinformatics* 18 (1), 90. doi:10.1186/s12859-017-1496-0

Faner, R., James, E., Huston, L., Pujol-Borrel, R., Kwok, W. W., and Juan, M. (2009). Reassessing the Role of HLA-DRB3 T-Cell Responses: Evidence for Significant Expression and Complementary Antigen Presentation. *Eur. J. Immunol.* 40 (1), 91–102. doi:10.1002/eji.200939225

Geffard, E., Limou, S., Walencik, A., Daya, M., Watson, H., Torgerson, D., et al. (2020). Easy-HLA: a Validated Web Application Suite to Reveal the Full Details of HLA Typing. *Bioinformatics* 36 (7), 2157–2164. doi:10.1093/bioinformatics/btz875

Ghadially, H., Brown, L., Lloyd, C., Lewis, L., Lewis, A., Dillon, J., et al. (2017). MHC Class I Chain-Related Protein A and B (MICA and MICB) Are Predominantly Expressed Intracellularly in Tumour and normal Tissue. *Br. J. Cancer* 116 (9), 1208–1217. doi:10.1038/bjc.2017.79

Hayashi, S., Moriyama, T., Yamaguchi, R., Mizuno, S., Komura, M., Miyano, S., et al. (2019). ALPHLARD-NT: Bayesian Method for Human Leukocyte Antigen Genotyping and Mutation Calling through Simultaneous Analysis of Normal and Tumor Whole-Genome Sequence Data. *J. Comput. Biol.* 26 (9), 923–937. doi:10.1089/cmb.2018.0224

Horton, R., Wilming, L., Rand, V., Lovering, R. C., Bruford, E. A., Khodiyar, V. K., et al. (2004). Gene Map of the Extended Human MHC. *Nat. Rev. Genet.* 5 (12), 889–899. doi:10.1038/nrg1489

Hosomichi, K., Shiina, T., Tajima, A., and Inoue, I. (2015). The Impact of Next-Generation Sequencing Technologies on HLA Research. *J. Hum. Genet.* 60 (11), 665–673. doi:10.1038/jhg.2015.102

Hughes, A. L., and Nei, M. (1988). Pattern of Nucleotide Substitution at Major Histocompatibility Complex Class I Loci Reveals Overdominant Selection. *Nature* 335 (6186), 167–170. doi:10.1038/335167a0

Hurley, C. K. (2021). Naming HLA Diversity: A Review of HLA Nomenclature. *Hum. Immunol.* 82 (7), 457–465. doi:10.1016/j.humimm.2020.03.005

Jeanmougin, M., Noirel, J., Coulonges, C., and Zagury, J.-F. (2017). HLA-check: Evaluating HLA Data from SNP Information. *BMC Bioinformatics* 18 (1), 334. doi:10.1186/s12859-017-1746-1

Jekarl, D. W., Lee, G. D., Yoo, J. Bin., Kim, J. R., Yu, H., Yoo, J., et al. (2021). HLA-A, -B, -C, -DRB1 Allele and Haplotype Frequencies of the Korean Population and Performance Characteristics of HLA Typing by Next-generation Sequencing. *HLA* 97 (3), 188–197. doi:10.1111/tan.14167

Jia, X., Han, B., Onengut-Gumuscu, S., Chen, W.-M., Concannon, P. J., Rich, S. S., et al. (2013). Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* 8 (6), e64683. doi:10.1371/journal.pone.0064683

Juhos, S., Vágó, T., Ferriola, D., Duke, J., Vörös, S., Brown, B. O., et al. (2015). Deriving HLA Genotyping from Whole Genome Sequencing Data Using Omixon HLA Twin(tm) in G3's Global Clinical Study. *Hum. Immunol.* 76, 131. doi:10.1016/j.humimm.2015.07.183

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J. Immunol.* 199 (9), 3360–3368. doi:10.4049/jimmunol.1700893

Karnes, J. H., Shaffer, C. M., Bastarache, L., Gaudieri, S., Glazer, A. M., Steiner, H. E., et al. (2017). Comparison of HLA Allelic Imputation Programs. *PLoS ONE* 12 (2), e0172444. doi:10.1371/journal.pone.0172444

Kaufman, J. (2018). Generalists and Specialists: A New View of How MHC Class I Molecules Fight Infectious Pathogens. *Trends Immunol.* 39 (5), 367–379. doi:10.1016/j.it.2018.01.001

Kennedy, A. E., Ozbek, U., and Dorak, M. T. (2017). What Has GWAS Done for HLA and Disease Associations? *Int. J. Immunogenet.* 44 (5), 195–211. doi:10.1111/iji.12332

Khor, S.-S., Yang, W., Kawashima, M., Kamitsuji, S., Zheng, X., Nishida, N., et al. (2015). High-Accuracy Imputation for HLA Class I and II Genes Based on High-Resolution SNP Data of Population-specific References. *Pharmacogenomics J.* 15 (6), 530–537. doi:10.1038/tpj.2015.4

Kim, H. J., and Pourmand, N. (2013). HLA Haplotyping from RNA-Seq Data Using Hierarchical Read Weighting. *PLoS ONE* 8 (6), e67885. doi:10.1371/journal.pone.0067885

Klasberg, S., Surendranath, V., Lange, V., and Schöfl, G. (2019). Bioinformatics Strategies, Challenges, and Opportunities for Next Generation Sequencing-Based HLA Genotyping. *Transfus. Med. Hemother* 46 (5), 312–325. doi:10.1159/000502487

Kramer, C. S. M., Koster, J., Haasnoot, G. W., Roelen, D. L., Claas, F. H. J., and Heidt, S. (2020). HLA-EMMA : A User-friendly Tool to Analyse HLA Class I and Class II Compatibility on the Amino Acid Level. *Hla* 96 (1), 43–51. doi:10.1111/tan.13883

Kulkarni, S., Martin, M. P., and Carrington, M. (2008). The Yin and Yang of HLA and KIR in Human Disease. *Semin. Immunol.* 20 (6), 343–352. doi:10.1016/j.smim.2008.06.003

Kulkarni, S., Savan, R., Qi, Y., Gao, X., Yuki, Y., Bass, S. E., et al. (2011). Differential microRNA Regulation of HLA-C Expression and its Association with HIV Control. *Nature* 472 (7344), 495–498. doi:10.1038/nature09914

Kuniholm, M. H., Xie, X., Anastos, K., Xue, X., Reimers, L., French, A. L., et al. (2016). Human Leucocyte Antigen Class I and II Imputation in a Multiracial Population. *Int. J. Immunogenet.* 43 (6), 369–375. doi:10.1111/iji.12292

Lancaster, A. K., Single, R. M., Solberg, O. D., Nelson, M. P., and Thomson, G. (2007). PyPop Update - a Software Pipeline for Large-Scale Multilocus Population Genomics. *Tissue Antigens* 69 (3), 192–197. doi:10.1111/j.1399-0039.2006.00769.x

Lima, T. H. A., Souza, A. S., Porto, I. O. P., Paz, M. A., Veiga-Castelli, L. C., Oliveira, M. L. G., et al. (2019). HLA-A Promoter, Coding, and 3'UTR Sequences in a Brazilian Cohort, and Their Evolutionary Aspects. *HLA* 93 (2–3), 65–79. doi:10.1111/tan.13474

Loginova, M., Smirnova, D., Kutyavina, S., Paramonov, I., and Zarubin, M. (2021). The Novel HLA-A Allele, HLA-A*01:354 , Identified in a Buryat Individual. *Hla* 97 (5), 435–436. doi:10.1111/tan.14170

Loginova, M., Smirnova, D., Paramonov, I., and Kozhemyako, O. (2020). The Novel HLA-DRB1*14:221 Allele Was Identified during High-resolution HLA Typing. *Hla* 96 (2), 231–232. doi:10.1111/tan.13868

Luo, Y., Kanai, M., Choi, W., Li, X., Yamamoto, K., Ogawa, K., et al. (2020). A High-Resolution HLA Reference Panel Capturing Global Population Diversity Enables Multi-Ethnic fine-mapping in HIV Host Response. *medRxiv*, 1–46. doi:10.1101/2020.07.16.20155606

Mack, S. J., and Hollenbach, J. A. (2010). Allele Name Translation Tool and Update NomenCLature: Software Tools for the Automated Translation of HLA Allele Names between Successive Nomenclatures. *Tissue Antigens* 75 (5), 457–461. doi:10.1111/j.1399-0039.2010.01477.x

Mack, S. J., Milius, R. P., Gifford, B. D., Sauter, J., Hofmann, J., Osoegawa, K., et al. (2015). Minimum Information for Reporting Next Generation Sequence Genotyping (MIRING): Guidelines for Reporting HLA and KIR Genotyping via Next Generation Sequencing. *Hum. Immunol.* 76 (12), 954–962. doi:10.1016/j.humimm.2015.09.011

Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Bontrop, R. E., Dupont, B., Erlich, H. A., et al. (2010). Nomenclature for Factors of the HLA System, 2010. *Tissue Antigens* 75 (4), 291–455. doi:10.1111/j.1399-0039.2010.01466.x

Martin, M. P., and Carrington, M. (2013). Immunogenetics of HIV Disease. *Immunol. Rev.* 254 (1), 245–264. doi:10.1111/imr.12071

Mayor, N. P., Robinson, J., McWhinnie, A. J. M., Ranade, S., Eng, K., Midwinter, W., et al. (2015). HLA Typing for the Next Generation. *PLoS One* 10 (5), e0127153. doi:10.1371/journal.pone.0127153

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., et al. (2016). A Reference Panel of 64,976 Haplotypes for Genotype Imputation. *Nat. Genet.* 48 (10), 1279–1283. doi:10.1038/ng.3643

McLaren, P. J., Ripke, S., Pelak, K., Weintrob, A. C., Patsopoulos, N. A., Jia, X., et al. (2012). Fine-mapping Classical HLA Variation Associated with Durable Host Control of HIV-1 Infection in African Americans. *Hum. Mol. Genet.* 21 (19), 4334–4347. doi:10.1093/hmg/dds226

Mei, S., Li, F., Leier, A., Marquez-Lago, T. T., Giam, K., Croft, N. P., et al. (2019). A Comprehensive Review and Performance Evaluation of Bioinformatics Tools for HLA Class I Peptide-Binding Prediction. *Brief. Bioinform.* 21, 1119–1135. Available at: https://academic.oup.com/bib/article/21/4/1119/5511798.

Meral, B. (2007). "Bone Marrow and Stem Cell Transplantation," in *Methods in Molecular Biology*. Editor M. Beksaç (New Jersey: Humana Press), Vol. 134, 313. doi:10.1007/978-1-4614-9437-9

Meyer, D., C. Aguiar, V. R., Bitarello, B. D., C. Brandt, D. Y., and Nunes, K. (2018). A Genomic Perspective on HLA Evolution. *Immunogenetics* 70 (1), 5–27. doi:10.1007/s00251-017-1017-3

Meyer, D., and Nunes, K. (2017). HLA Imputation, what Is it Good for? *Hum. Immunol.* 78 (3), 239–241. doi:10.1016/j.humimm.2017.02.007

Meyer, D., and Thomson, G. (2001). How Selection Shapes Variation of the Human Major Histocompatibility Complex: a Review. *Ann. Hum. Genet* 65 (1), 1–26. doi:10.1046/j.1469-1809.2001.6510001.x

Middleton, D., Menchaca, L., Rood, H., and Komerofsky, R. (2003). New Allele Frequency Database: http://www.allelefrequencies.Net. *Tissue Antigens* 61 (5), 403–407. doi:10.1034/j.1399-0039.2003.00062.x

Milius, R. P., Heuer, M., Valiga, D., Doroschak, K. J., Kennedy, C. J., Bolon, Y.-T., et al. (2015). Histoimmunogenetics Markup Language 1.0: Reporting Next Generation Sequencing-Based HLA and KIR Genotyping. *Hum. Immunol.* 76 (12), 963–974. doi:10.1016/j.humimm.2015.08.001

Mimori, T., Yasuda, J., Kuroki, Y., Shibata, T. F., Katsuoka, F., Saito, S., et al. (2019). Construction of Full-Length Japanese Reference Panel of Class I HLA Genes with Single-Molecule, Real-Time Sequencing. *Pharmacogenomics J.* 19 (2), 136–146. doi:10.1038/s41397-017-0010-4

Montgomery, R. A., Tatapudi, V. S., Leffell, M. S., and Zachary, A. A. (2018). HLA in Transplantation. *Nat. Rev. Nephrol.* 14, 558–570. doi:10.1038/s41581-018-0039-x

Mosbruger, T. L., Dinou, A., Duke, J. L., Ferriola, D., Mehler, H., Pagkrati, I., et al. (2020). Utilizing Nanopore Sequencing Technology for the Rapid and Comprehensive Characterization of Eleven HLA Loci; Addressing the Need for Deceased Donor Expedited HLA Typing. *Hum. Immunol.* 81 (8), 413–422. doi:10.1016/j.humimm.2020.06.004

Motyer, A., Vukcevic, D., Dilthey, A., Donnelly, P., McVean, G., and Leslie, S. (2016). Practical Use of Methods for Imputation of HLA Alleles from SNP Genotype Data. *bioRxiv*, 091009.

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2008). P-values Are Random Variables. *The Am. Statistician* 62 (3), 242–245. doi:10.1198/000313008x332421

Naito, T., Suzuki, K., Hirata, J., Kamatani, Y., Matsuda, K., Toda, T., et al. (2021). A Deep Learning Method for HLA Imputation and Trans-ethnic MHC fine-mapping of Type 1 Diabetes. *Nat. Commun.* 12 (1), 1–14. doi:10.1038/s41467-021-21975-x

Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; Improved Prediction of Binding to MHC Class I Molecules Integrating Information from Multiple Receptor and Peptide Length Datasets. *Genome Med.* 8 (1), 1–9. doi:10.1186/s13073-016-0288-x

Nilsson, L. L., Funck, T., Kjersgaard, N. D., and Hviid, T. V. F. (2018). Next-generation Sequencing ofHLA-Gbased on Long-Range Polymerase Chain Reaction. *Hla* 92 (3), 144–153. doi:10.1111/tan.13342

Nunes, J. M., Buhler, S., Roessli, D., and Sanchez-Mazas, A. (2014). TheHLA-net GENE[RATE]pipeline for Effective HLA Data Analysis and its Application to 145 Population Samples from Europe and Neighbouring Areas. *Tissue Antigens* 83 (5), 307–323. doi:10.1111/tan.12356

Osoegawa, K., Mack, S. J., Udell, J., Noonan, D. A., Ozanne, S., Trachtenberg, E., et al. (2016). HLA Haplotype Validator for Quality Assessments of HLA Typing. *Hum. Immunol.* 77 (3), 273–282. doi:10.1016/j.humimm.2015.10.018

Pappas, D. J., Lizee, A., Paunic, V., Beutner, K. R., Motyer, A., Vukcevic, D., et al. (2018). Significant Variation between SNP-Based HLA Imputations in Diverse Populations: the Last Mile Is the Hardest. *Pharmacogenomics J.* 18 (3), 367–376. doi:10.1038/tpj.2017.7

Pappas, D. J., Marin, W., Hollenbach, J. A., and Mack, S. J. (2016). Bridging ImmunoGenomic Data Analysis Workflow Gaps (BIGDAWG): An Integrated Case-Control Analysis Pipeline. *Hum. Immunol.* 77 (3), 283–287. doi:10.1016/j.humimm.2015.12.006

Parham, P., Norman, P. J., Abi-Rached, L., and Guethlein, L. A. (2012). Human-specific Evolution of Killer Cell Immunoglobulin-like Receptor Recognition of Major Histocompatibility Complex Class I Molecules. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367 (1590), 800. doi:10.1098/rstb.2011.0266

Park, I., and Terasaki, P. (2000). Origins of the First HLA Specificities. *Hum. Immunol.* 61 (3), 185–189. doi:10.1016/s0198-8859(99)00154-8

Praest, P., Luteijn, R. D., Brak-Boer, I. G. J., Lanfermeijer, J., Hoelen, H., Ijgosse, L., et al. (2018). The Influence of TAP1 and TAP2 Gene Polymorphisms on TAP Function and its Inhibition by Viral Immune Evasion Proteins. *Mol. Immunol.* 101 (May), 55–64. doi:10.1016/j.molimm.2018.05.025

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* 81 (3), 559–575. doi:10.1086/519795

Ralazamahaleo, M., Andreani, M., Giustiniani, P., Guidicelli, G., and Visentin, J. (2019). Characterization of the Novel HLA-DQA1*01:01:05 Allele by Sequencing-based Typing. *Hla* 94 (2), 172–173. doi:10.1111/tan.13569

Ritari, J., Hyvärinen, K., Clancy, J., Partanen, J., and Koskela, S. (2020). Increasing Accuracy of HLA Imputation by a Population-specific Reference Panel in a FinnGen Biobank Cohort. *NAR Genomics Bioinforma* 2 (2), 1–9. doi:10.1093/nargab/lqaa030

Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., and Marsh, S. G. E. (2019). IPD-IMGT/HLA Database. *Nucleic Acids Res.* 48 (D1), D948–D955. doi:10.1093/nar/gkz950

Robinson, J., Halliwell, J. A., Hayhurst, J. D., Flicek, P., Parham, P., and Marsh, S. G. (2015). The IPD and IMGT/HLA Database: Allele Variant Databases. *Nucleic Acids Res.* 43, D423–D431. doi:10.1093/nar/gku1161

Rock, K. L., Reits, E., and Neefjes, J. (2016). Present Yourself! by MHC Class I and MHC Class II Molecules. *Trends Immunol.* 37, 724–737. doi:10.1016/j.it.2016.08.010

Sacchi, N., Castagnetta, M., Miotti, V., Garbarino, L., and Gallina, A. (2019). High-resolution Analysis of the HLA-A, -B, -C and -DRB1 Alleles and National and Regional Haplotype Frequencies Based on 120 926 Volunteers from the Italian Bone Marrow Donor Registry. *Hla* 94 (3), 285–295. doi:10.1111/tan.13613

Schmidt, A. H., Sauter, J., Baier, D. M., Daiss, J., Keller, A., Klussmeier, A., et al. (2020). Immunogenetics in Stem Cell Donor Registry Work: The DKMS Example (Part 1). *Int. J. Immunogenet.* 47 (1), 13–23. doi:10.1111/iji.12471

Shiina, T., Hosomichi, K., Inoko, H., and Kulski, J. K. (2009). The HLA Genomic Loci Map: Expression, Interaction, Diversity and Disease. *J. Hum. Genet.* 54 (1), 15–39. doi:10.1038/jhg.2008.5

Sidney, J., del Guercio, M. F., Southwood, S., Engelhard, V. H., Appella, E., Rammensee, H. G., et al. (1995). Several HLA Alleles Share Overlapping Peptide Specificities. *J. Immunol.* 154 (1), 247–259. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7527812.

Spurgin, L. G., and Richardson, D. S. (2010). How Pathogens Drive Genetic Diversity: MHC, Mechanisms and Misunderstandings. *Proc. Biol. Sci.* 277, 979–988. doi:10.1098/rspb.2009.2084

Squire, D. M., Motyer, A., Ahn, R., Nititham, J., Huang, Z.-M., Oksenberg, J. R., et al. (2020). MHC*IMP - Imputation of Alleles for Genes in the Major Histocompatibility Complex. *bioRxiv*. doi:10.1101/2020.01.24.919191

The COVID-19 HLA and Immunogenetics Consortium (2020a). HLA|COVID-19. Available at: http://www.hlacovid19.org/.

The COVID-19 HLA and Immunogenetics Consortium (2020b). HLA|COVID-19 Database. Available at: https://database-hlacovid19.org/.

Trowsdale, J., and Knight, J. C. (2013). Major Histocompatibility Complex Genomics and Human Disease. *Annu. Rev. Genom. Hum. Genet.* 14 (1), 301–323. doi:10.1146/annurev-genom-091212-153455

Trowsdale, J., and Moffett, A. (2008). NK Receptor Interactions with MHC Class I Molecules in Pregnancy. *Semin. Immunol.* 20 (6), 317–320. doi:10.1016/j.smim.2008.06.002

van Tol, S., Hage, A., Giraldo, M., Bharaj, P., and Rajsbaum, R. (2017). The TRIMendous Role of TRIMs in Virus-Host Interactions. *Vaccines* 5 (3), 23. doi:10.3390/vaccines5030023

Vayntrub, T. A., Mack, S. J., and Fernandez-Viña, M. A. (2020). Preface: 17th International HLA and Immunogenetics Workshop. *Hum. Immunol.* 81, 52–58. doi:10.1016/j.humimm.2020.01.008

Vince, N., Bashirova, A. A., Lied, A., Gao, X., Dorrell, L., McLaren, P. J., et al. (2014). HLA Class I and KIR Genes Do Not Protect against HIV Type 1 Infection in Highly Exposed Uninfected Individuals with Hemophilia A. *J. Infect. Dis.* 210 (7), 1047–1051. doi:10.1093/infdis/jiu214

Vince, N., Douillard, V., Geffard, E., Meyer, D., Castelli, E. C., Mack, S. J., et al. (2020). SNP-HLA Reference Consortium (SHLARC): HLA and SNP Data Sharing for Promoting MHC-centric Analyses in Genomics. *Genet. Epidemiol.* 44 (7), 733–740. doi:10.1002/gepi.22334

Vince, N., Li, H., Ramsuran, V., Naranbhai, V., Duh, F.-M., Fairfax, B. P., et al. (2016). HLA-C Level Is Regulated by a Polymorphic Oct1 Binding Site in the HLA-C Promoter Region. *Am. J. Hum. Genet.* 99 (6), 1353–1358. doi:10.1016/j.ajhg.2016.09.023

Vince, N., Limou, S., Daya, M., Morii, W., Rafaels, N., Geffard, E., et al. (2020). Association of HLA-Drb1∗09:01 with tIgE Levels Among African-Ancestry Individuals with Asthma. *J. Allergy Clin. Immunol.* 146 (1), 147–155. doi:10.1016/j.jaci.2020.01.011

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., et al. (2019). Opportunities and Challenges for Transcriptome-wide Association Studies. *Nat. Genet.* 51 (4), 592–599. doi:10.1038/s41588-019-0385-z

Wang, M., and Claesson, M. H. (2014). "Classification of Human Leukocyte Antigen (HLA) Supertypes," in *Methods in Molecular Biology (Clifton, NJ)*, 309–317. doi:10.1007/978-1-4939-1115-8_17

Xie, C., Yeo, Z. X., Wong, M., Piper, J., Long, T., Kirkness, E. F., et al. (2017). Fast and Accurate HLA Typing from Short-Read Next-Generation Sequence Data with xHLA. *Proc. Natl. Acad. Sci. USA* 114 (30), 8059–8064. doi:10.1073/pnas.1707945114

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., et al. (2014). HIBAG-HLA Genotype Imputation with Attribute Bagging. *Pharmacogenomics J.* 14 (2), 192–200. doi:10.1038/tpj.2013.18

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.