Diversity in immunogenomics: the value and the challenge

Immunogenomics studies have been largely limited to individuals of European ancestry, restricting the ability to identify variation in human adaptive immune responses across populations. Inclusion of a greater diversity of individuals in immunogenomics studies will substantially enhance our understanding of human immunity.

Kerui Peng, Yana Safonova, Mikhail Shugay, Alice B. Popejoy, Oscar L. Rodriguez, Felix Breden, Petter Brodin, Amanda M. Burkhardt, Carlos Bustamante, Van-Mai Cao-Lormeau, Martin M. Corcoran, Darragh Duffy, Macarena Fuentes-Guajardo, Ricardo Fujita, Victor Greiff, Vanessa D. Jönsson, Xiao Liu, Lluis Quintana-Murci, Maura Rossetti, Jianning Xie, Gur Yaari, Wei Zhang, Malak S. Abedalthagafi, Khalid O. Adekoya, Rahaman A. Ahmed, Wei-Chiao Chang, Clive Gray, Yusuke Nakamura, William D. Lees, Purvesh Khatri, Houda Alachkar, Cathrine Scheepers, Corey T. Watson, Gunilla B. Karlsson Hedestam and Serghei Mangul

Acknowledgements

K.R.S. gratefully acknowledges grant funding from the NIH (R01NS080752, R01CA164371, U54CA143970, U54CA193489, U01CA220378, U54CA210180, U01CA250481), the James S. McDonnell Foundation, the Ben & Catherine Ivy Foundation, the Zicariello Foundation, the Arizona Biomedical Research Commission and the Mayo Clinic. M.M.M. gratefully acknowledges grant support from NIH R01DA039862 and R01MH091424. J.B.R. gratefully acknowledge grant funding from NCI (R01CA174737, P01CA245705).

Author contributions

K.M.B. drafted the manuscript and M.M.M., J.B.R. and K.R.S. oversaw the completion of the work.

Competing interests

K.R.S. is a co-founder of Precision Oncology Insights, Inc.

Additional information

Peer review information Nature Methods thanks the anonymous reviewers for their contribution to the peer review of this work.
Since then, studies and colleagues applied AIRR-seq to human immune receptor repertoire sequencing diversity of the HLA and KIR systems,13,14, extensive efforts to explore the genetic immunoglobulin repertoires, somatic (CDR) diversity; and, in the case of complementarity-determining region profiles; germline V, D and J gene usage; provides information on expression immune repertoires in different settings.8. Studies of TCR and immunoglobulin could be integrated into the AIRR-seq and the knowledge gained from these efforts ends (5′ rapid amplification of cDNA ends (5′ RACE) PCR to amplify TCR cDNA and to characterize TCR repertoires.6 Weinstein and colleagues sequenced an antibody repertoire in zebrafish6 in 2009, creating the foundation of adaptive immune receptor repertoire sequencing (AIRR-seq) technologies. In 2010, Boyd and colleagues applied AIRR-seq to human immunoglobulins.7 Since then, studies including AIRR-seq have seen exponential growth, and findings from these studies have shaped our understanding of human immune repertoires in different settings.4.

A critical step in AIRR-seq studies is the identification of germline TCR and immunoglobulin genes from AIRR-seq data. Computational methods to infer germline TCR and immunoglobulin genes from AIRR-seq data are expected to accelerate these efforts4–24 (Table 1). Comparisons are also needed between results obtained from methods for inferring germline gene variants from AIRR-seq repertoires and from direct sequencing of genomic DNA5, such as the sequencing and assembly of large-insert clones (for example, bacterial artificial chromosome (BAC) and fosmid clones)25 and, more recently, whole-genome sequencing and targeted long-read sequencing26.

The most widely used reference database for immunogenomics data, the international ImMunoGeneTics information system (IMGT)27, has been a valuable resource. However, it lacks a comprehensive set of human TCR and immunoglobulin alleles representing diverse populations worldwide. Further uncertainty stems from descriptions of sample populations in databases being based on geography or self-identified race and/or ethnicity of study subjects, rather than genetic ancestry. As a result, we have a limited understanding of population-level TCR and immunoglobulin germline gene variation. However, progress is being made.

The AIRR Community (AIRR-C; http://www.airr-community.org) is an international community of bioinformaticians and immunogeneticists that has been formed to develop standards and protocols to promote sharing and common analysis approaches for AIRR-seq data, including the AIRR Data Commons28. As a means to enrich available germline gene sets, the AIRR-C established the Inferred Allele Review Committee (IARC; https://www.antibodysoociety.org/the-airr-community/airr-subcommittees/inferred-allele-review-committee-iarc) to review and curate new immunoglobulin or TCR germline genes inferred from AIRR-seq data. Its work is underpinned by the Open Germline Receptor Database, which provides submission and review workflows. IARC-affirmed sequences are published in this database, together with supporting evidence. VDJbase was also recently launched as a public database that allows users to access population-level immunoglobulin and TCR germline data, including reports and summary statistics on germline genes, alleles, single nucleotide and structural variants, and haplotypes of interest derived from AIRR-seq and genomic sequencing data. It currently contains AIRR-seq data from 421 human

### Table 1 | Tools for inference of germline TCR and immunoglobulin genes from AIRR-seq data

| Tool | Type of receptors | Type of inferring genes | Needs gene database for inference | Comment |
|------|------------------|-------------------------|----------------------------------|---------|
| TlgGER6 | Ig | V | Yes | TlgGER and Partis assign AIRR-seq reads to V genes from the database and report a list of V gene alleles (both known alleles and alleles with modifications) |
| Partis6 | Ig | V | Yes | TlgGER and Partis assign AIRR-seq reads to V genes from the database and report a list of V gene alleles (both known alleles and alleles with modifications) |
| IgDiscover21 | Ig, TCR | V, J | Yes | IgDiscover uses the database for annotation of AIRR-seq reads, clusters reads with similar annotations, and reports both known and previously unobserved V genes |
| IMPre22 | Ig, TCR | V, J | No | IMPre infers V and J genes from clusters of similar AIRR-seq reads and uses a germline database (if available) for annotation of the inferred genes |
| IgScout23 | Ig | D, J | No | Both IgScout and MINING-D infer D genes as abundant substrings of CDR3s of AIRR-seq reads and use a germline database (if available) for annotation of the inferred genes |
| MINING-D24 | Ig, TCR | D | No | Both IgScout and MINING-D infer D genes as abundant substrings of CDR3s of AIRR-seq reads and use a germline database (if available) for annotation of the inferred genes |

Ig, immunoglobulin; TCR, T cell receptor.

High-throughput sequencing techniques in the late 2000s. Freeman and colleagues employed 5′ rapid amplification of cDNA ends (5′ RACE) PCR to amplify TCR cDNA and to characterize TCR repertoires.6 Weinstein and colleagues sequenced an antibody repertoire in zebrafish6 in 2009, creating the foundation of adaptive immune receptor repertoire sequencing (AIRR-seq) technologies. In 2010, Boyd and colleagues applied AIRR-seq to human immunoglobulins.7 Since then, studies including AIRR-seq have seen exponential growth, and findings from these studies have shaped our understanding of human immune repertoires in different settings.4.

A critical step in AIRR-seq studies is the identification of germline TCR and immunoglobulin genes from AIRR-seq data. Computational methods to infer germline TCR and immunoglobulin genes from AIRR-seq data are expected to accelerate these efforts4–24 (Table 1). Comparisons are also needed between results obtained from methods for inferring germline gene variants from AIRR-seq repertoires and from direct sequencing of genomic DNA5, such as the sequencing and assembly of large-insert clones (for example, bacterial artificial chromosome (BAC) and fosmid clones)25 and, more recently, whole-genome sequencing and targeted long-read sequencing26.

The most widely used reference database for immunogenomics data, the international ImMunoGeneTics information system (IMGT)27, has been a valuable resource. However, it lacks a comprehensive set of human TCR and immunoglobulin alleles representing diverse populations worldwide. Further uncertainty stems from descriptions of sample populations in databases being based on geography or self-identified race and/or ethnicity of study subjects, rather than genetic ancestry. As a result, we have a limited understanding of population-level TCR and immunoglobulin germline gene variation. However, progress is being made.

The AIRR Community (AIRR-C; http://www.airr-community.org) is an international community of bioinformaticians and immunogeneticists that has been formed to develop standards and protocols to promote sharing and common analysis approaches for AIRR-seq data, including the AIRR Data Commons28. As a means to enrich available germline gene sets, the AIRR-C established the Inferred Allele Review Committee (IARC; https://www.antibodysoociety.org/the-airr-community/airr-subcommittees/inferred-allele-review-committee-iarc) to review and curate new immunoglobulin or TCR germline genes inferred from AIRR-seq data. Its work is underpinned by the Open Germline Receptor Database, which provides submission and review workflows. IARC-affirmed sequences are published in this database, together with supporting evidence. VDJbase was also recently launched as a public database that allows users to access population-level immunoglobulin and TCR germline data, including reports and summary statistics on germline genes, alleles, single nucleotide and structural variants, and haplotypes of interest derived from AIRR-seq and genomic sequencing data. It currently contains AIRR-seq data from 421 human
donors, representing 724 immunoglobulin heavy chain gene alleles. The integration of TCR datasets is in progress. Together these initiatives will help pave the way for the development of approaches that extend germline curation efforts to include more data types and ultimately ensure that population-level metadata can be more effectively captured and leveraged.

**Recommendations for the immunology community**

The immunology community should make targeted efforts to include non-European populations in AIRR-seq and other immunogenomics studies. Already, AIRR-seq studies in more diverse populations have uncovered evidence for extensive genetic heterogeneity. For example, in a study of South Africans with HIV, Scheepers and colleagues discovered many immunoglobulin heavy chain variable (IGHV) alleles that were not represented in IMGT\(^{16}\), information of relevance to HIV vaccine design aimed at germline-targeting immunogens\(^{16}\). In a study in the Papua New Guinea population, I new IGHV gene and 16 IGHV allelic variants were identified from AIRR-seq data\(^{16}\). These discoveries of alleles indicate the need for further population-based AIRR-seq datasets and the identification and validation of the presence of new alleles so that they can be added to public databases. It will be critical to conduct studies in various human populations if we are to fully understand how AIRR-seq can be leveraged to make improvements in a wide range of applications, including vaccine design.

Further, we suggest that extant open AIRR-seq datasets could be used to augment immunoglobulin and TCR germline databases and inform AIRR-seq and other immunogenomics studies across diverse populations. It may be possible in the future to use AIRR-seq data to infer genetic ancestry, but such bioinformatics methods are yet to be developed and thus the utility of genetic ancestry in this field has yet to be demonstrated. Conclusions about new germline variants discovered through non-targeted sequencing data, including RNA-seq based on short read sequences, should be drawn with caution owing to the complexity of the adaptive immune receptor loci\(^{16}\), as described above. New methodologies and computational approaches should be developed to facilitate the inclusion of diverse population datasets into existing databases, with the aim of enhancing our knowledge base to reflect global genomic immunological diversity in populations around the globe. Such enriched databases would provide researchers with baseline resources to design and implement the next generation of personalized and precision immunodiagnostics and therapeutics\(^{16}\).

At the current stage of the global COVID-19 pandemic, many vaccine trials and programs are underway worldwide, offering opportunities to investigate the role of genetic factors in vaccine-mediated immune responses. Such investigations will require careful study designs to effectively address potential confounding factors such as environmental, economic and social determinants of health that systematically differ between populations defined by self-identified measures of diversity and that are correlated with continental-level ancestry\(^{16}\). Incomplete representation of diverse populations limits our capacity to address the impact of genetics on clinical phenotypes, and ideally this should be investigated alongside non-genetic risk factors for disease. Different genetic variants in an etiologic pathway modify the clinical presentation of disease, and these effects can differ by genomic background\(^{16}\). Specific immunoglobulin germline genes, and in some cases alleles, have been found to be preferentially used in the response to pathogens, suggesting a degree of convergence in the antibody response, as observed for influenza\(^{2}\), HIV-1\(^{16}\), Zika virus\(^{16}\) and SARS CoV-2\(^{16}\). Therefore, in addition to environmental factors, genetic variability in immune genes is likely to drive differential effects in vaccine effectiveness and infection outcomes\(^{16}\).

Our interdisciplinary group consists of leading researchers from 17 regions, including the United States, Canada, Norway, France, Sweden, the United Kingdom, Russia, Saudi Arabia, Israel, South Africa, Nigeria, Chile, Peru, China, Japan, Taiwan and French Polynesia, who share concerns about the lack of diversity in immunogenomics and embrace a need to tackle these challenges. As an interdisciplinary group with expertise in biomedical and translational research, population and public health genetics, health disparities, computational biology and immunogenomics, we wish to raise awareness about the value of including diverse populations in AIRR-seq and immunogenomics research.

---

**Authors**

- Kerui Peng\(^{1}\), Yana Safanova\(^{2,3}\), Mikhail Shugay\(^{4,5}\), Alice B. Popejoy\(^{6}\), Oscar L. Rodriguez\(^{6}\), Felix Breiden\(^{6}\), Petter Brodin\(^{6,7}\), Amanda M. Burkhartd\(^{6}\), Carlos Bustamante\(^{6}\), Van-MaiCao-Lormeau\(^{8}\), Martin M. Corcoran\(^{9}\), Darragh Duffy\(^{10}\), Macarena Fuentes-Guajardo\(^{11}\), Ricardo Fujita\(^{12}\), Victor Greiff\(^{12}\), Vanessa D. Jönsson\(^{15}\), Xiao Liu\(^{14}\), Lluís Quintana-Murci\(^{18,39}\), Maura Rossetti\(^{20}\), Jiaming Xie\(^{21}\), Gur Yaari\(^{22}\), Wei Zhang\(^{23}\), Malak S. Abdalathagafi\(^{24}\), Khalid O. Adekoya\(^{25}\), Rahaman A. Ahmed\(^{26}\), Wei-Chiao Chang\(^{26,27}\), Clive Gray\(^{28,29}\), Yusuke Nakamura\(^{30}\), William D. Lees\(^{31}\), Purvesh Khatri\(^{32,33}\), Houda Alachkar\(^{36}\), Cathrine Scheepers\(^{34,35,36}\), Corey T. Watson\(^{37}\), Gunilla B. Karlsson Hedestam\(^{38,39}\) and Serghei Mangul\(^{38,39}\)

1. Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, USA. 2. Computer Science and Engineering Department, University of California San Diego, San Diego, CA, USA. 3. Department of Biochemistry and Molecular Genetics, University of Louisville School of Medicine, Louisville, KY, USA. 4. Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia. 5. Progov Russian National Research Medical University, Moscow, Russia. 6. Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. 7. Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada. Science for Life Laboratory, Department of Women’s and Children Health, Karolinska Institutet, Stockholm, Sweden. 8. Pediatric Rheumatology, Karolinska University Hospital, Stockholm, Sweden. 9. Laboratory of Research on Infectious Vector-borne Diseases, Institut Louis Malardé, Papeete, French Polynesia. 10. Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden. 11. Translational Immunology Laboratory, Institut Pasteur, Paris, France. 12. Departamento de Tecnología Médica, Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, Chile. 13. Centro de Genética y Biología Molecular, Universidad de San Martín de Porres, La Molina, Lima, Perú. 14. Department of Immunology, University of Oslo, Oslo, Norway. 15. Departments of Computational and Quantitative Medicine and Hematology, Beckman Research Institute, City of Hope, Duarte, CA, USA. 16. Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. 17. Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, CNRS, Paris, France. 18. Department of Human Genomics and Evolution, Collège de France, Paris, France. 19. UCLA Immunogenetics Center, Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA. 20. Department of Pharmacology & Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, CA, USA. 21. Faculty of Engineering, Bar Ilan Institute of Nanotechnologies and Advanced Materials, Bar Ilan University, Ramat Gan, Israel. 22. Department of Computer Science, City University of Hong Kong, Hong Kong, China. 23. Genomics Research Department, Saudi Human Genome Project, King Fahad Medical City and King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia. 24. Department of Cell Biology and Genetics, University of Lagos, Lagos, Nigeria. 25. Department of Clinical
comment

Pharmacy, School of Pharmacy, Taipei Medical University, New Taipei City, Taiwan. 21Division of Nephrology, Department of Internal Medicine, Taipei Medical University-Shuang Ho Hospital, New Taipei City, Taiwan. 22Division of Immunology, Institute of Infectious Disease and Molecular Medicine and Department of Pathology, University of Cape Town, Cape Town, South Africa. 23Laboratory of Tissue Immunology, National Health Laboratory Services and Groote Schuur Hospital, Cape Town, South Africa. 24Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan. 25Institute of Structural and Molecular Biology, Birkbeck College, London, UK. 26Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, CA, USA. 27Center for Biomedical Research, Department of Medicine, School of Medicine, Stanford University, Stanford, CA, USA. 28Centre for HIV and STIs, National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa. 29South African Medical Research Council Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa. 30These authors contributed equally: Houda Alachkar, Cathrine Scheepers, Corey T. Watson, Gunilla B. Karlsson Hedestam, Serghei Mangul.

e-mail: serghei.mangul@gmail.com

Published online: 17 May 2021
https://doi.org/10.1038/s41592-021-01169-5

References
1. Sirugo, G., Williams, S. M. & Tubiolo, S. A. Cell 177, 1080 (2019).
2. Ben-Egham, C. et al. Nature 585, 184–186 (2020).
3. Ghoudhary, A. et al. Nature 586, 741–748 (2020).
4. Retzballe, G. et al. Am. J. Hum. Genet. 102, 733–743 (2018).
5. Freeman, J. D., Warren, R. L., Webb, J. R., Nelson, B. H. & Holt, R. A. Genome Res. 19, 1817–1824 (2009).
6. Weinstein, J. A., Jiang, N., White, R. A. III, Fisher, D. S. & Quake, S. R. Science 324, 807–808 (2009).
7. Boyd, S. D. et al. J. Immunol. 184, 6986–6992 (2010).
8. Briney, R., Inderbitzin, A., Joyce, C. & Burton, D. R. Nature 566, 393–397 (2019).
9. Avrin, Y. et al. Sci. Rep. 6, 20842 (2016).
10. Sharrow, G. V., Serebrovskaya, E. G., Yuzhakova, D. V., Britanova, O. V. & Chadakov, D. M. Nat. Rev. Immunol. 20, 294–307 (2020).
11. Bashford-Rogers, R. J. M. et al. Nature 574, 122–126 (2019).
12. Gage, D. et al. Nature 577, 399–404 (2020).
13. Singh, K. M. et al. Immunogenetics 64, 97–109 (2012).
14. Dhillon, A., Cox, C., Ijpb, Z., Nelson, M. R. & McVean, G. Nat. Genet. 47, 682–688 (2015).
15. Schepers, C. et al. J. Immunol. 194, 4371–4378 (2015).
16. Watson, C. T. et al. Am. J. Hum. Genet. 92, 536–546 (2013).
17. Watson, C. T., Glansville, J. & Marasco, W. A. Trends Immunol. 38, 459–470 (2017).
18. Quintana-Marcu, L. Cell 177, 184–199 (2019).
19. Gudala-Maria, D. et al. Front. Immunol. 10, 129 (2019).
20. Ralph, D. K. & Matisen, F. A. IV. PLOS Comput. Biol. 15, e1007133 (2019).
21. Corcoran, M. M. et al. Nat. Commun. 7, 13642 (2016).
22. Zang, W. et al. Front. Immunol. 7, 657 (2016).
23. Safonova, Y. & Pezner, P. A. Front. Immunol. 10, 987 (2019).
24. Bhardwaj, V., Francescetti, M., Rao, R., Pezner, P. A. & Safonova, Y. PLOS Comput. Biol. 16, e1007837 (2020).
25. Ohlin, M. et al. Front. Immunol. 10, 435 (2019).
26. Rodriguez, O. L. et al. Front. Immunol. 11, 2136 (2020).
27. Blandino, G. et al. Nature 526, 513–523 (2015).
28. Christley, S. et al. Front. Big Data 3, 22 (2020).
29. Burton, D. R. Nat. Rev. Immunol. 19, 77–78 (2019).
30. Wang, Y. et al. Immunogenetics 63, 259–265 (2011).
31. Greiff, Y., Yuari, G. & Cowell, L. G. Curr. Opin. Syst. Biol. 24, 109–110 (2020).
32. Ioannidou, J. P. A., Pouto, N. R. & Yanc, C. J. Am. Med. Assoc. 325, 623–624 (2021).
33. Severe Covid-19 GWAS Group et al. N. Engl. J. Med. 383, 1522–1534 (2020).
34. Huang, J. et al. Immunity 45, 1108–1121 (2016).
35. Robbiano, P. E. et al. Cell 169, 597–609 e11 (2017).
36. Yuan, M. et al. Science 369, 1119–1123 (2020).

Acknowledgements
We thank Nicky Mulder for comments that greatly improved the manuscript. Y.S. was supported by the National Science Foundation EAGER award (no. 2032783). M.S. is supported by Ministry of Science and Higher Education of the Russian Federation grant no. 075-15-2020-807. A.B.P. and C.D.B. are supported by an award from the National Human Genome Research Institute of the National Institutes of Health (U41HG009649). C.T.W. and O.L.R. are supported in part by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under award numbers R24AI138963 and R21AI124590. V.G. is supported by a UO World-Leading Research Community grant, the UO:LifeScience Convergence Environment Immunologiso, EU Horizon 2020 (no. 825821) and a Research Council of Norway FRIPRO project (#300740). V.D. was supported by an award from the National Cancer Institute of the National Institutes of Health (K12CA001727). The laboratory of L.Q.-M. is supported by the Institut Pasteur, the Collège de France, the CNRS, the Fondation Allianz-Institut de France and the French Government’s Investissement d’Avenir program, Laboratoires d’Excellence ‘Integrative Biology of Emerging Infectious Diseases’ (ANR-10-LABX-62-IBEID) and ‘Mélié Intérieur’ (ANR-10-LABX-69-01). R.A.A. is supported by the Fogarty International Center of the National Institute of Health under award number D43TW010934. P.K. is supported by the Bill and Melinda Gates Foundation (OPP1113682), National Institute of Allergy and Infectious Diseases (NIAID) (1U19AI136662), and the National Institutes of Health under award number U10AI057229, 5R01AI125197). Department of Defense (W911WH1910235) and Cancer and Transformational Awards from the Dr. Ralph & Marian Falk Medical Research Trust. C.S. is supported by NIAID of the National Institutes of Health under award number U01AI136677. G.K.H is supported by a grant from the Swedish Research Council (award number 532 2017-00968). S.M. is partially supported by National Science Foundation grants 2041984.

Competing interests
V.G. declares advisory board positions in anNET GmbH and Enpicom B.V. P.K. is a co-founder of Inflammatix, Inc. He is also a consultant to Inflammatix, Inc., Vir Biotechnology, Cepheid, and Genentech. G.K.H and M.C. are founders of ImmuneDiscover Sweden AB.