Title

Diversity in immunogenomics: the value and the challenge.

Journal

Nature Methods, 18(6)

Authors

Peng, Kerui

Safonova, Yana

Shugay, Mikhail

et al.

Publication Date

2021-06-01

DOI

10.1038/s41592-021-01169-5

Peer reviewed

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

Nat Methods. 2021 June ; 18(6): 588–591. doi:10.1038/s41592-021-01169-5.

Diversity in immunogenomics: the value and the challenge

Kerui Peng[1], **Yana Safonova**[2,3], **Mikhail Shugay**[4,5], **Alice B. Popejoy**[6], **Oscar L. Rodriguez**[3], **Felix Breden**[7], **Petter Brodin**[8,9], **Amanda M. Burkhardt**[1], **Carlos Bustamante**[6], **Van-Mai Cao-Lormeau**[10], **Martin M. Corcoran**[11], **Darragh Duffy**[12], **Macarena Fuentes-Guajardo**[13], **Ricardo Fujita**[14], **Victor Greiff**[15], **Vanessa D. Jönsson**[16], **Xiao Liu**[17], **Lluis Quintana-Murci**[18,19], **Maura Rossetti**[20], **Jianming Xie**[21], **Gur Yaari**[22], **Wei Zhang**[23], **Malak S. Abedalthagafi**[24], **Khalid O. Adekoya**[25], **Rahaman A. Ahmed**[25], **Wei-Chiao Chang**[26,27], **Clive Gray**[28,29], **Yusuke Nakamura**[30], **William D. Lees**[31], **Purvesh Khatri**[32,33], **Houda Alachkar**[1,36], **Cathrine Scheepers**[34,35,36], **Corey T. Watson**[3,36], **Gunilla B. Karlsson Hedestam**[11,36], **Serghei Mangul**[1,36,✉]

[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]Pirogov 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

[8]Science for Life Laboratory, Department of Women's and Children Health, Karolinska Institutet Stockholm, Sweden

[9]Pediatric Rheumatology, Karolinska University Hospital, Stockholm, Sweden

[10]Laboratory of Research on Infectious Vector-borne Diseases, Institut Louis Malardé, Papeete, French Polynesia

[11]Department of Microbiology, Tumor and Cell Biology, Karolinska Institutet, Stockholm, Sweden

[12]Translational Immunology Laboratory, Institut Pasteur, Paris, France

[13]Departamento de Tecnología Médica, Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, Chile

✉ serghei.mangul@gmail.com .

Competing interests

V.G. declares advisory board positions in aiNET 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.

[14]Centro de Genética y Biología Molecular, Universidad de San Martín de Porres, La Molina, Lima, Perú

[15]Department of Immunology, University of Oslo, Oslo, Norway

[16]Departments of Computational and Quantitative Medicine and Hematology, Beckman Research Institute, City of Hope, Duarte, CA, USA

[17]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

[18]Human Evolutionary Genetics Unit, Institut Pasteur, UMR 2000, CNRS, Paris, France

[19]Department of Human Genomics and Evolution, Collège de France, Paris, France

[20]UCLA Immunogenetics Center, Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California Los Angeles, Los Angeles, CA, USA

[21]Department of Pharmacology & Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, CA, USA

[22]Faculty of Engineering, Bar Ilan Institute of Nanotechnologies and Advanced Materials, Bar Ilan University, Ramat Gan, Israel

[23]Department of Computer Science, City University of Hong Kong, Hong Kong, China

[24]Genomics Research Department, Saudi Human Genome Project, King Fahad Medical City and King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia

[25]Department of Cell Biology and Genetics, University of Lagos, Lagos, Nigeria

[26]Department of Clinical Pharmacy, School of Pharmacy, Taipei Medical University, New Taipei City, Taiwan

[27]Division of Nephrology, Department of Internal Medicine, Taipei Medical University-Shuang Ho Hospital, New Taipei City, Taiwan

[28]Division of Immunology, Institute of Infectious Disease and Molecular Medicine and Department of Pathology, University of Cape Town, Cape Town, South Africa

[29]Laboratory of Tissue Immunology, National Health Laboratory Services and Groote Schuur Hospital, Cape Town, South Africa

[30]Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan

[31]Institute of Structural and Molecular Biology, Birkbeck College, London, UK

[32]Institute for Immunity, Transplantation and Infection, School of Medicine, Stanford University, CA, USA

[33]Center for Biomedical Research, Department of Medicine, School of Medicine, Stanford University, Stanford, CA, USA

[34]Centre for HIV and STIs, National Institute for Communicable Diseases of the National Health Laboratory Service, Johannesburg, South Africa

[35]South African Medical Research Council Antibody Immunity Research Unit, School of Pathology, University of the Witwatersrand, Johannesburg, South Africa

[36]These authors contributed equally: Houda Alachkar, Cathrine Scheepers, Corey T. Watson, Gunilla B. Karlsson Hedestam, Serghei Mangul

Abstract

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 immunology.

Current state of diversity in genomics studies

Genomic studies have mainly used samples from individuals of European ancestry, at the expense of learning from the largest and most genetically diverse populations. For example, 78% of individuals included in genome-wide association studies (GWAS) reported in the GWAS Catalog (https://www.ebi.ac.uk/gwas/home) through January 2019 are of European descent[1], while Asian populations account for 59.5% of the world population based on the Population Reference Bureau's World Population Data Sheet (https://www.prb.org/datasheets/). Though this is partially due to inadequate sampling of non-European populations, researchers tend to exclude data from minority groups when conducting statistical analyses[2] even when diverse datasets are available. The limited inclusion of samples from diverse populations hinders the equitable advancement of genomic medicine as a result of persistent uncertainty with respect to the genetic etiology of disease across populations, as well as differential rates of adverse drug events, treatment outcomes and other health disparities.

In recent years there has been an increased awareness of the limited generalizability of findings across populations and the benefits for the discovery and interpretation of gene–trait associations brought about by the inclusion of diverse populations in genomic studies. This has motivated the inclusion of diverse, multiethnic populations in large-scale genomic studies. For example, whole-genome sequencing in individuals of African descent[3] and whole-exome sequencing in a southern African population[4] have improved understanding of genetic variation in under-represented populations. Additional efforts have been made to establish reference genome datasets for research in diverse populations; these include the GenomeAsia 100K Project, Human Heredity and Health in Africa (H3Africa) initiative, Taiwan Biobank, Population Architecture Using Genomics and Epidemiology (PAGE) Consortium, Trans-Omics for Precision Medicine (TOPMed) program, Clinical Sequencing Evidence-Generating Research (CSER) consortium, Human Genome Reference Program (HGRP) and All of Us Research Program. However, the field of immunogenomics, especially that related to adaptive immune receptors, has yet to benefit from a similar growth in diversity.

The need for diversity in immunogenomics

Central to immunity are the repertoires of T cell receptors (TCRs), immunoglobulins, human leukocyte antigens (HLAs) and killer cell immunoglobulin-like receptors (KIRs). Thus, analyses of the loci that encode these molecules are critical to immunogenomics studies.

T cells and B cells recognize antigens through their TCRs and immunoglobulins, which are formed through the process of V(D)J (variable, diversity and joining region) recombination. Capturing the vast diversity of recombined, expressed TCR and immunoglobulin repertoires was not possible until the development of 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[5]. Weinstein and colleagues sequenced an antibody repertoire in zebrafish[6] 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[8].

AIRR-seq analysis and other immunogenomics studies offer new opportunities to deepen our understanding of the immune system in the context of a variety of human diseases, including infectious diseases[9], cancer[10], autoimmune conditions[11] and neurodegenerative diseases[12]. Furthermore, AIRR-seq data provides information on expression profiles; germline V, D and J gene usage; complementarity-determining region (CDR) diversity; and, in the case of immunoglobulin repertoires, somatic hypermutation levels. There have been extensive efforts to explore the genetic diversity of the HLA and KIR systems[13,14], and the knowledge gained from these efforts could be integrated into the AIRR-seq studies of TCR and immunoglobulin germline gene diversity.

As in the field of genomics, greater diversity in immunogenomics research has the potential to enable the discovery of novel genetic traits associated with immune system phenotypes that are common or different across populations. While evidence for extensive diversity in germline TCR and immunoglobulin genes have been reported in the human population[15,16], most AIRR-seq studies that use sequencing to study T and B cell receptor repertoires have been conducted in individuals of European descent, leaving other populations under-represented[17]. Exclusion of non-European populations in genomics research limits our understanding of how pathogens have exerted selective pressures on immune-related genes in populations living in different environments, and thus on infectious disease manifestation[18].

Germline gene diversity and databases

A critical step in AIRR-seq studies is germline gene assignments, which requires reliable and comprehensive databases of germline V(D)J alleles representing different populations. So far, such databases are lacking because the genetic regions encoding these genes have been exceptionally challenging to characterize at the genomic level. Not only do these loci contain a mixture of functional genes and pseudogenes with high similarity, but they

are also characterized by considerable structural variation, with deletions and duplications occurring at high frequency in different populations. Given the complexity of the TCR and immunoglobulin genomic loci and deficits in existing germline databases, the determination of immune receptor germline gene usage from bulk RNA-seq or whole-genome sequencing is often inaccurate. Efforts to improve germline databases are therefore critical for improved coverage of diversity in immune repertoire analysis. Computational methods to infer germline TCR and immunoglobulin genes from AIRR-seq data are expected to accelerate these efforts[19–24] (Table 1). Comparisons are also needed between results obtained from methods for inferring germline gene variants from AIRR-seq repertoires[25] and from direct sequencing of genomic DNA[15], such as the sequencing and assembly of large-insert clones (for example, bacterial artificial chromosome (BAC) and fosmid clones)[16] and, more recently, whole-genome sequencing and targeted long-read sequencing[26].

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 Commons[28]. As a means to enrich available germline gene sets, the AIRR-C established the Inferred Allele Review Committee (IARC; https://www.antibodysociety.org/the-airr-community/airr-subcomittees/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[15], information of relevance to HIV vaccine design aimed at germline-targeting immunogens[29]. In a study in the Papua New Guinea population, 1 new IGHV gene and 16 IGHV allelic variants were identified from AIRR-seq data[30]. 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[26], 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[31].

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[32]. 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[33]. 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[9], HIV-1[34], Zika virus[35] and SARS CoV-2[36]. Therefore, in addition to environmental factors, genetic variability in immune genes is likely to drive differential effects in vaccine effectiveness and infection outcomes[17].

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.

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 R21AI142590. V.G. is supported by a UiO World-Leading Research Community grant, the UiO:LifeScience Convergence Environment Immunolingo, EU Horizon 2020 iReceptorplus (no. 825821) and a Research Council of Norway FRIPRO project (#300740). V.D.J. 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 College 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 'Milieu 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) (1U19AI109662, U19AI057229, 5R01AI125197), Department of Defense (W81XWH1910235) and Catalyst 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.

References

1. Sirugo G, Williams SM & Tishkoff SA Cell 177, 1080 (2019). [PubMed: 31051100]
2. Ben-Eghan C et al. Nature 585, 184–186 (2020). [PubMed: 32901124]
3. Choudhury A et al. Nature 586, 741–748 (2020). [PubMed: 33116287]
4. Retshabile G et al. Am. J. Hum. Genet 102, 731–743 (2018). [PubMed: 29706352]
5. Freeman JD, Warren RL, Webb JR, Nelson BH & Holt RA Genome Res. 19, 1817–1824 (2009). [PubMed: 19541912]
6. Weinstein JA, Jiang N, White RA III, Fisher DS & Quake SR Science 324, 807–810 (2009). [PubMed: 19423829]
7. Boyd SD et al. J. Immunol 184, 6986–6992 (2010). [PubMed: 20495067]
8. Briney B, Inderbitzin A, Joyce C & Burton DR Nature 566, 393–397 (2019). [PubMed: 30664748]
9. Avnir Y et al. Sci. Rep 6, 20842 (2016). [PubMed: 26880249]
10. Sharoncw GV, Serebrowskaya EO, Yuzhakova DV,Britanova OV & Chudakov DM Nat. Rev. Immunol 20, 294–307 (2020). [PubMed: 31988391]
11. Bashford-Rogers RJM et al. Nature 574, 122–126 (2019). [PubMed: 31554970]
12. Gate D et al. Nature 577, 399–404 (2020). [PubMed: 31915375]
13. Singh KM et al. Immunogenetics 64, 97–109 (2012). [PubMed: 21898189]
14. Dilthey A, Cox C, Iqbal Z, Nelson MR & McVean G Nat. Genet 47, 682–688 (2015). [PubMed: 25915597]
15. Scheepers C et al. J. Immunol 194, 4371–4378 (2015). [PubMed: 25825450]
16. Watson CT et al. Am. J. Hum. Genet 92, 530–546 (2013). [PubMed: 23541343]
17. Watson CT, Glanville J & Marasco WA Trends Immunol. 38, 459–470 (2017). [PubMed: 28539189]
18. Quintana-Murci L Cell 177, 184–199 (2019). [PubMed: 30901539]
19. Gadala-Maria D et al. Front. Immunol 10, 129 (2019). [PubMed: 30814994]

20. Ralph DK & Matsen FA IV PLOS Comput. Biol 15, e1007133 (2019). [PubMed: 31329576]
21. Corcoran MM et al. Nat. Commun 7, 13642 (2016). [PubMed: 27995928]
22. Zhang W et al. Front. Immunol 7, 457 (2016). [PubMed: 27867380]
23. Safonova Y & Pevzner PA Front. Immunol 10, 987 (2019). [PubMed: 31134072]
24. Bhardwaj V, Franceschetti M, Rao R, Pevzner PA & Safonova Y PLOS Comput. Biol 16, e1007837 (2020). [PubMed: 32339161]
25. Ohlin M et al. Front. Immunol 10, 435 (2019). [PubMed: 30936866]
26. Rodriguez OL et al. Front. Immunol 11, 2136 (2020). [PubMed: 33072076]
27. Lefranc M-P et al. Nucleic Acids Res. 43, D413–D422 (2015). [PubMed: 25378316]
28. Christley S et al. Front. Big Data 3, 22 (2020). [PubMed: 33693395]
29. Burton DR Nat. Rev. Immunol 19, 77–78 (2019). [PubMed: 30560910]
30. Wang Y et al. Immunogenetics 63, 259–265 (2011). [PubMed: 21249354]
31. Greiff V, Yaari G & Cowell LG Curr. Opin. Syst. Biol 24, 109–119 (2020).
32. Ioannidis JPA, Powe NR & Yancy 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). [PubMed: 32558485]
34. Huang J et al. Immunity 45, 1108–1121 (2016). [PubMed: 27851912]
35. Robbiani DF et al. Cell 169, 597–609.e11 (2017). [PubMed: 28475892]
36. Yuan M et al. Science 369, 1119–1123 (2020). [PubMed: 32661058]

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
TIgGER[19]	Ig	V	Yes	TIgGER 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)
Partis[20]	Ig	V	Yes	
IgDiscover[21]	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
IMPre[22]	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
IgScout[23]	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-D[24]	Ig, TCR	D	No	

Ig, immunoglobulin; TCR, T cell receptor.