Immunity

CellPress

Letter

Ancestral diversity is limited in published T cell receptor sequencing studies

Yu-Ning Huang,[1,14] Kerui Peng,[1,14] Alice B. Popejoy,[2] Jieting Hu,[3] Theodore Scott Nowicki,[4] Stefan M. Gold,[5,6,7] Lluis Quintana-Murci,[8,9] Macarena Fuentes-Guajardo,[10] Mikhail Shugay,[11,12] Victor Greiff,[13] Amanda M. Burkhardt,[1] Houda Alachkar,[1] and Serghei Mangul[1,*]

[1]Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA 90089, USA
[2]Department of Biomedical Data Science, Stanford University, Stanford, CA 94305, USA
[3]Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
[4]Department of Pediatrics, Division of Pediatric Hematology/Oncology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[5]Institute for Neuroimmunology and Multiple Sclerosis, Universitätsklinikum Hamburg-Eppendorf, Hamburg, D-20246, Germany
[6]Department of Psychiatry, Campus Benjamin Franklin, Charité – Universitätsmedizin Berlin, Berlin, D-12203, Germany
[7]Medical Department, Campus Benjamin Franklin, Charité – Universitätsmedizin Berlin, Berlin, D-12203, Germany
[8]Human Evolutionary Genetics Unit, CNRS UMR2000, Institut Pasteur, Paris, 75015, France
[9]Department of Human Genomics and Evolution, Collège de France, Paris, 75231, France
[10]Departamento de Tecnología Médica, Facultad de Ciencias de la Salud, Universidad de Tarapacá, Arica, 1000009, Chile
[11]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, 117997, Russia
[12]Pirogov Russian National Research Medical University, Moscow, 117997, Russia
[13]Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, 0372, Norway
[14]These authors contributed equally to this work
*Correspondence: serghei.mangul@gmail.com
https://doi.org/10.1016/j.immuni.2021.09.015

The emergence of high-throughput sequencing techniques and the development of bioinformatics tools provide efficient ways to profile the human adaptive immune receptor repertoire (AIRR) including repertoires of T cell receptors (TCRs) and immunoglobulins (Bradley and Thomas, 2019). Despite these advancements, the representation of populations in the AIRR sequencing (AIRR-seq) studies remains unknown. AIRR data from populations with diverse genetic ancestry could potentially elucidate genomic and phenotypic similarities or differences in the human immune system and immune-related diseases across populations (Peng et al., 2021). It was previously demonstrated that the majority of the participants in genome-wide association studies (GWASs) were of European ancestry (Peterson et al., 2019). However, the level of ancestral diversity in AIRR-seq stand is not clear and has not been studied in detail (Peng et al., 2021). To address the emerging issue of the lack of diversity in immunogenomics studies described in our previous work (Peng et al., 2021) and to further develop a strategic roadmap to engage diverse populations in AIRR-seq studies and highlight this issue for the community, we investigated the ancestral diversity (AD) in TCR sequencing (TCR-seq) studies currently reported in published studies and databases. We also discuss the challenges and opportunities of including underrepresented populations in TCR-seq research and how this information may guide future efforts to increase diversity in the field of immunogenomics.

We surveyed AD information from 114 TCR-seq studies that included 3,261 study participants. We examined the current state of AD information availability across TCR-seq studies and showed the proportion of population groups from various aspects based on the available information. Information was extracted from the text of TCR-seq publications or directly acquired from the corresponding authors. Details on how studies were selected and methods used to obtain information about study participants are described in the supplemental information.

We first examined the availability of study participants' AD information in TCR-seq studies with publicly available sequencing data from publications in online public repositories. Fewer than 20% of the surveyed studies included such information in the text of the paper (Figure S1A). We then obtained study participants' AD information from an additional 21 studies by sending inquiries to corresponding authors. In total, we collected the study participants' AD information from 42 TCR-seq studies across 3,261 individuals for the subsequent analysis (Figure S1A). We further explored the availability of population information through PubMed Medical Subject Headings (MeSH) terms search. Fewer than 4% of 2,480 immunogenetics studies included both "immunogenetics" and "population" MeSH terms. Only 0.3% of 37,824 TCR-seq studies (publications tagged with "Receptors, Antigen, T-Cell" MeSH term) included a "population" MeSH term. Observations were similar between the manual text check on publications and the MeSH term search: most of the TCR-seq studies lack study population information.

Next, we investigated the reasons for the unavailability of study participants' AD information. Among studies without available study participants' AD information, 6% of the authors did not share data due to privacy concerns and limitations in study designs. First, the authors had concerns about violating the US Health Insurance Portability and Accountability Act (HIPAA) rules. Second, the authors were not able to share participants' information due to the limitations of the study protocols that were already approved by the Institutional Review Board (IRB). For example, AD information was not included during the recruitment phase, or the study participants' information was not approved for sharing with researchers that were from other institutions. Last, some studies utilized de-identified samples, which prevented authors from obtaining the study participants' information.

Furthermore, we analyzed the AD among the examined TCR-seq studies. Nearly 60% of the studies included European or European descent participants, while only 10% of the studies included African or African descent participants (Figure S1B). From the perspective of the study participants, more than 80% of study participants were reported to be European AD groups, followed by 9% of participants from Asian AD groups and 4% of participants of African AD groups (Figure S1B).

We also examined the proportion of TCR-seq studies that were conducted in a single AD group versus multiple AD groups. In total 33 out of 42 studies were conducted in a single self-reported ancestry group (Figure S1C). While European-based studies were dominant and followed by Asian-based studies, notably, no studies were conducted in African populations alone. Nine of the examined studies included study participants from multiple AD groups, all of which predominantly consisted of European populations.

We then focused on the ethnic information among 12 US-based TCR studies. Only 4% of study participants were self-reported as Hispanic, which indicated that Hispanics were highly underrepresented in TCR-seq studies, relative to their proportion of the US population (18.5%) based on the 2019 United States Census Bureau population estimate (Figure S1D). Furthermore, we investigated the temporal dynamics of AD diversity in TCR-seq studies. From 2009 to 2021, there was a 10% increase in the proportion non-European individuals (Figure S1E). Despite the increase, the distribution of study participants based on their ancestry is still highly disproportionate in TCR-seq studies.

We examined the relationship between sample sizes and the population component in TCR-seq studies. The results showed that the average number of study participants of European ancestry was much greater than the average number of study participants of Asian and African ancestries (Figure S1F). Lastly, in light of the COVID-19 pandemic, the pattern of skewed diversity in TCR-seq studies on COVID patients has also been observed. Substantial bias toward participants from European or of European ancestry was observed across all six COVID TCR-seq studies that we evaluated (Figure S1G) and five studies mainly included European populations (Figure S1H).

This survey provides insight into AD information availability and AD in TCR-seq studies. Similar to GWASs, our findings reveal that individuals of non-European ancestry were severely underrepresented in TCR-seq studies. The disproportionate distribution of study participants' self-reported ancestry information in TCR-seq studies may restrict our understanding of disease pathology in diverse populations, confine the discovery of immunogenomics variants across populations (Peterson et al., 2019), and hinder the advancement in precision medicine (Greiff et al., 2020). Furthermore, there was a severe lack of available AD information in most of the TCR-seq studies. This could potentially limit the reuse of TCR-seq data for secondary analysis to infer novel population-specific TCR alleles for improving the representation of the diverse populations in current reference databases.

According to previously published work, over 50% of 448 researchers and clinical genetics professionals surveyed considered AD important in clinical settings (Popejoy et al., 2020). The human leukocyte antigen (HLA) system elicits the adaptive immune response mediated by TCR. While substantial research has been done to examine the HLA diversity across diverse populations (Nemat-Gorgani et al., 2019), the current understanding of the AD in TCR is more limited. With the expanded inclusivity in AIRR-seq studies, we will have a more comprehensive understanding of AIRR in diverse populations (Peng et al., 2021). This knowledge will accelerate its translational and clinical applications, and eventually promote health equity. For example, the Moderna and Pfizer-BioNTech COVID-19 vaccines showed different efficacy in study participants of different ancestry (Pilishvili et al., 2021). Although the reasons for the variation in efficacy across different populations remain unknown, AIRR might potentially play an important role. The investigation of the relationship between AIRR and vaccine-mediated immune response may advance the development of future vaccines or therapeutics.

There are a few limitations of our study. First, the unavailability of genetic ancestry information may prevent us from accurately categorizing the study participants. Ultimately, it is preferable to examine ancestry by genetics-based methodologies. The genetic ancestry depicts the single-nucleotide variants across geographic origin groups and depicts the extent of single-nucleotide variants among individuals of different ancestries (Jorde and Bamshad, 2020). However, none of the studies we examined performed genotyping or other computational methods to infer the study participants' genetic ancestry information. The use of self-reported ancestry in TCR-seq studies may be considered reasonable at the current stage (Oni-Orisan et al., 2021). Second, the un-unified protocol of reporting study participants' AD information may introduce bias in our analysis. The lack of standardized terminologies around the AD information might cause inconsistency among researchers and make it challenging to conduct secondary analyses. Developing and adopting standardized experimental protocols and computational methods to report or infer genetic ancestry in the field of immunogenomics are urgently needed (Morales et al., 2018).

To promote sharing of the study participants' AD information, we recommend careful consideration of the data-sharing options when submitting IRB applications. For example, we recommend that researchers specify that it is acceptable to share data with external investigators. Furthermore, we advocate for an emphasis on data sharing from scientific journals and/or funding agencies. To address the discrepancies in the AD information reporting, we recommend that the scientific community establish standardized protocols or guidelines in reporting study participants' AD information. For example, the sources of the AD information including self-reported or genetically verified are worth noting. Additionally, Human Ancestry Ontology (HANCESTRO) provides a systematic description of ancestry.

We hope that this letter highlights for the community which population groups are underrepresented in TCR-seq studies and raises awareness of this issue. An enrichment in diversity of TCR-seq studies is needed and we advocate for the broadened knowledge in this field by studying diverse populations.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.immuni.2021.09.015.

ACKNOWLEDGMENTS

S.M. is partially supported by National Science Foundation grant 2041984. M.S. is supported by the Ministry of Science and Higher Education of the Russian Federation grant no. 075-15-2020-807. S.M.G. acknowledges support by the Deutsche Forschungsgemeinschaft (KFO296 Fetomaternal Immune Cross-Talk, grants GO1357/8-2 and FR1720/8-2). H.A. is supported by University of Southern California School of Pharmacy Seed Fund, The Norris Cancer Center IRG-ACS pilot fund, STOP Cancer pilot fund, The Ming Hsieh Institute foundation grants, NIH-NCI 1R01CA248381-01A1, and in part by NIH grant 5P30CA014089-45. We thank all the authors of the surveyed studies (Dr. Anastasia A. Minervina, Dr. Mikhail V. Pogorelyy, Dr. Grigory A. Efimov, Dr. Zheming Lu, Dr. Yang Ke, Dr. Xiao Liu, Dr. Mascha Binder, Dr. Dmitriy M. Chudakov, Dr. Nathalie Bedard, Dr. Jun S. Liu, Dr. Michelle Miron, Dr. Maura Rossetti, Dr. Satu Mustjoki, Dr. Atsunari Kawashima, Dr. Matthew A. Brown, Dr. Jorge Correale, and Dr. Tae Jin Kim) who shared the information of the study participants with us.

AUTHOR CONTRIBUTIONS

Y.H. collected and analyzed the data. A.B.P. shared expertise in the use of race, ethnicity, and ancestry in genetics. Y.H. and J.H. contacted the authors of surveyed studies. M.S. provided insights in TCR-seq studies on COVID-19 patients. K.P. and Y.H. wrote the manuscript with input from all other authors. S.M. conceived and supervised the study.

DECLARATION OF INTERESTS

V.G. declares advisory board positions in aiNET GmbH and Enpicom B.V. V.G. is a consultant for Roche/Genentech. S.M.G. declares honoraria from Mylan GmbH, Almirall S.A., and Celgene and research grants from Biogen outside the submitted work. He receives research funding from the Deutsche Forschungsgemeinschaft, Bundesministerium für Bildung und Forschung, Bundesministerium für Gesundheit, the National MS Society, and the European Commission.

REFERENCES

Bradley, P., and Thomas, P.G. (2019). Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition. Annu. Rev. Immunol. *37*, 547–570.

Greiff, V., Yaari, G., and Cowell, L.G. (2020). Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. Curr. Opin. Syst. Biol. *24*, 109–119.

Jorde, L.B., and Bamshad, M.J. (2020). Genetic Ancestry Testing: What Is It and Why Is It Important? JAMA *323*, 1089–1090.

Morales, J., Welter, D., Bowler, E.H., Cerezo, M., Harris, L.W., McMahon, A.C., Hall, P., Junkins, H.A., Milano, A., Hastings, E., et al. (2018). A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. Genome Biol. *19*, 21.

Nemat-Gorgani, N., Guethlein, L.A., Henn, B.M., Norberg, S.J., Chiaroni, J., Sikora, M., Quintana-Murci, L., Mountain, J.L., Norman, P.J., and Parham, P. (2019). Diversity of KIR, HLA Class I, and Their Interactions in Seven Populations of Sub-Saharan Africans. J. Immunol. *202*, 2636–2647.

Oni-Orisan, A., Mavura, Y., Banda, Y., Thornton, T.A., and Sebro, R. (2021). Embracing Genetic Diversity to Improve Black Health. N. Engl. J. Med. *384*, 1163–1167.

Peng, K., Safonova, Y., Shugay, M., Popejoy, A.B., Rodriguez, O.L., Breden, F., Brodin, P., Burkhardt, A.M., Bustamante, C., Cao-Lormeau, V.M., et al. (2021). Diversity in immunogenomics: the value and the challenge. Nat. Methods *18*, 588–591.

Peterson, R.E., Kuchenbaecker, K., Walters, R.K., Chen, C.Y., Popejoy, A.B., Periyasamy, S., Lam, M., Iyegbe, C., Strawbridge, R.J., Brick, L., et al. (2019). Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell *179*, 589–603.

Pilishvili, T., Fleming-Dutra, K.E., Farrar, J.L., Gierke, R., Mohr, N.M., Talan, D.A., Krishnadasan, A., Harland, K.K., Smithline, H.A., Hou, P.C., et al.; Vaccine Effectiveness Among Healthcare Personnel Study Team (2021). Interim Estimates of Vaccine Effectiveness of Pfizer-BioNTech and Moderna COVID-19 Vaccines Among Health Care Personnel - 33 U.S. Sites, January-March 2021. MMWR Morb. Mortal. Wkly. Rep. *70*, 753–758.

Popejoy, A.B., Crooks, K.R., Fullerton, S.M., Hindorff, L.A., Hooker, G.W., Koenig, B.A., Pino, N., Ramos, E.M., Ritter, D.I., Wand, H., et al.; Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (2020). Clinical Genetics Lacks Standard Definitions and Protocols for the Collection and Use of Diversity Measures. Am. J. Hum. Genet. *107*, 72–82.