Data availability of open T-cell receptor repertoire data, a systematic assessment

Authors:
Yu-Ning Huang
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA
yuninghu@usc.edu
ORCID: https://orcid.org/0000-0003-1697-4267

Naresh Amrat Patel
Department of Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA
naresham@usc.edu

Jay Himanshu Mehta
Department of Pharmaceutical Sciences, School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA
jayhiman@usc.edu

Srishti Ginjala
School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi, Himachal Pradesh, India
srishti.ginjala@gmail.com
ORCID: https://orcid.org/0000-0001-7269-2382

Petter Brodin, PhD
Department of Immunology and Inflammation, Imperial College London, United Kingdom.
Department of Women's and Children's Health, Karolinska Institutet, Sweden
petter.brodin@ki.se
ORCID: https://orcid.org/0000-0002-8103-0046

Clive M Gray, MSc, PhD
Division of Molecular Biology and Human Genetics, Biomedical Research Institute, Stellenbosch University, Cape Town, South Africa
cgray@sun.ac.za
clive.gray@uct.ac.za
ORCID: https://orcid.org/0000-0002-9293-901X

Yesha M Patel
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, 1540 Alcazar Street, Los Angeles, CA 90033, USA
ypatel@usc.edu
ORCID: https://orcid.org/0000-0002-4983-1222

Lindsay G. Cowell, PhD
Division of Biomedical Informatics, Department of Population and Data Sciences, Department of Immunology, University of Texas Southwestern Medical Center at Dallas, USA
Lindsay.Cowell@utsouthwestern.edu
ORCID: https://orcid.org/0000-0003-1617-8244

Amanda M. Burkhardt
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, CA, USA
aburkhar@usc.edu
ORCID: https://orcid.org/0000-0002-7326-474X

Serghei Mangul
Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, CA, USA
serghei.mangul@gmail.com
ORCID: https://orcid.org/0000-0003-4770-3443
Twitter: @smangul1

Abstract

The improvement of next-generation sequencing technologies has promoted the field of immunogenetics and produced numerous immunogenomics data. Modern data-driven research has the power to promote novel biomedical discoveries through secondary analysis of such data. Therefore, it is important to ensure data-driven research with great reproducibility and robustness for promoting a precise and accurate secondary analysis of the immunogenomics data. In scientific research, rigorous conduct in designing and conducting experiments is needed, specifically in scientific and articulate writing, reporting and interpreting results. It is also crucial to make raw data available, discoverable, and well described or annotated in order to promote future re-analysis of the data. In order to assess the data availability of published T cell receptor (TCR) repertoire data, we examined 11,918 TCR-Seq samples corresponding to 134 TCR-Seq studies ranging from 2006 to 2022. Among the 134 studies, only 38.1% had publicly available raw TCR-Seq data shared in public repositories. We also found a statistically significant association between the presence of data availability statements and the increase in raw data availability (p=0.014). Yet, 46.8% of studies with data availability statements failed to share the raw TCR-Seq data. There is a pressing need for the biomedical community to increase awareness of the importance of promoting raw data availability in scientific research and take immediate action to improve its raw data availability enabling cost-effective secondary analysis of existing immunogenomics data by the larger scientific community.

Introduction

Advanced high throughput sequencing technologies have reshaped the landscape of contemporary immunology research providing researchers with a rich set of tools and methods to study fundamental aspects of immune responses across a variety of disciplines. Raw TCR-Seq data is produced accompanying the development and progress of immunogenomics research. Availability of raw TCR-Seq data allows effective re-analysis of the data[1] (also known as secondary analysis) which in turn may accelerate novel biomedical discoveries[2]. Using public repositories to share raw data substantially simplifies discovering and retrieving datasets of interest, as one has scalable, programmatic access to a large number of studies. Over the last decade, there has been tremendous progress to improve sharing of raw immunogenomics data, which allows researchers to easily obtain the various types of data stored in public repositories[3] (e.g. SRA[4]). But to unlock the full potential of publicly available raw TCR-Seq data for the secondary analyses, it is crucial for a study to be conducted accurately with reproducible and reliable laboratory practices and accurate results[5]. Moreover, the research should report publicly available raw data accompanying metadata for secondary analysis[6,7]. One of the merits of science is to be able to leverage published data for subsequent novel scientific discovery[8]. To ensure the results are reproducible and have sufficient quality for accurate and precise secondary analysis, the data should be shared under the FAIR principle–Findability, Accessibility, Interoperability, and Reusability–to ensure the quality of the raw data to be reused accurately and efficiently in research[8,9]. Scientific communities, peer-reviewed journals, funding agencies, and government agencies have emphasized the importance of making raw omics data publicly available with free access to the public[7]. As such, establishing protocols, guidelines, and authors' checklists for reporting the raw immunogenomics data in research, and peer-review journals mandating authors to upload the data to public repositories are potential means to increase the raw data availability[3,7,10].

It is also crucial to ensure studies share the raw data used for publication and analyses. In the fields of immunogenetics studies since the development of next-generation sequencing tools promotes research on genetic sequencing and production of raw immunongenomics data due to the reduction in price for genetic sequencing[11]. T cell receptor sequencing (TCR-Seq) studies allow researchers to profile human immune systems providing updated insights into T cell receptors. TCR-Seq data allows researchers to examine a unique individual's immune status, immune responses, and compare the T cell populations between individuals in healthy or disease states, such as autoimmune diseases, infectious diseases, and cancer[12–14]. Additionally, TCR-Seq studies allow the development of novel therapeutics and biomarkers, including diagnostics[15] for autoimmune disease[16] and cancer[17–19], CAR-T cell therapy[20], vaccines[21], and monoclonal and therapeutic antibodies[22]. Due to the production of vast TCR-Seq data in the field of TCR repertoire studies, there is a pressuring need to ensure that raw sequencing data from TCR-Seq studies is provided with full access to the public to facilitate further computational and bioinformatics research in the field. In this study, we examined TCR-Seq studies for the availability of the raw sequencing data across 134 TCR-Seq studies over 11,918 samples and examined the public genomic repositories where the researchers stored the TCR-Seq data from their analyses. We found, only 38.1% of TCR-Seq studies have made the overall raw data available, and that the majority of studies (35 studies) with available raw data stored the raw data in Sequence Read Archive[4]. Further, 61.9% of the TCR-Seq studies did not share the raw data in the original publications or that the raw data will only be available upon request. We discussed the potential barriers researchers are facing that deter them from sharing the raw data and the potential ways to improve the availability of the raw data in the field of TCR-Seq studies.

The availability of raw sequencing data in TCR-Seq studies is limited

We investigated 134 published TCR-Seq studies across 11,918 samples in PubMed for the raw sequencing data availability ranging from 2006 to 2022. The studies were considered as having available raw sequencing data if we could acquire the samples of the studies' raw FASTQ or FASTA files. According to our results, only 38.1% (51 out of 134 studies) of the TCR-Seq studies shared raw TCR-Seq data in the original publications at public genomic repositories (Figure 1a). Conversely, 61.9% (83 out of 134 studies) of the TCR-Seq studies have unavailable raw RNA-Seq data or have raw data available upon request (Figure 1a). We observed a similar trend of raw data availability among the 11,918 samples of the 134 TCR-Seq studies, in which 25% of the samples among the TCR-Seq studies had available raw data (Figure 1a). We also observed that the raw TCR-Seq data availability has increased over the past decade (Figure 1b). Among the 134 studies, 89.6% (120 studies) of the raw data were generated in medical research institutes, including medical schools, hospitals, medical centers, private health/disease research institutes, and government health research institutes, such as the National Institutes of Health (Figure 1c). The remaining raw data were generated in engineering schools and other research institutes (Figure 1c). Additionally, we examined the types of TCR-Seq chains in the available raw sequencing data. Among the samples with available raw sequencing data, 65.2% of the TCR-Seq data were TCR beta (TCRβ) chain, 30.7% of the TCR-Seq data were TCR alpha (TCRα) chain, and 4.1% of the TCR-Seq data were TCR gamma (TCRγ) or TCR delta (TCRδ) chain (Figure 1d).

We further investigated the specific reasons that the raw data was unavailable among the 83 studies without available raw TCR-Seq data. Among the 83 studies with unavailable raw data, rather than sharing raw sequencing data, 44 studies only shared summary data, such as summary data on ImmuneACCESS® (Adaptive Biotechnologies), VDJdb, and supplementary files. Other 34 studies did not include statements about the raw data or provide raw sequencing data in the original publications and five studies indicated that the raw sequencing data will be available upon making direct requests to the authors (Figure 2a).

The platforms used to store the raw TCR-Seq data

We further examined the platforms researchers used to share raw TCR-Seq data. Among the TCR-Seq studies that shared the raw sequencing data, 35 of the 51 studies (68.6%) of the raw data are shared in Sequence Read Archive (SRA)[4] and ten of the 51 studies (19.6%) shared the raw data in other repositories, including Gene Expression Omnibus[23] (GEO). The remaining studies shared the raw sequencing data in various online repositories, including European Genome-phenome Archive[24] (EGA), and National Genomics Data Center, China[25] (NGDC) (Figure 2a).

The presence of the “Data availability statement” increases the availability of raw TCR-Seq data

As part of promoting more transparent and reproducible research, many journals do require data availability statements for studies to be published in the journals[26,27]. We examined the impact of the presence of the “Data availability statement” in research articles on the availability of raw TCR-Seq data. According to the results, the presence of data availability statements improves the raw data availability by 23.3% (26 out of 87 studies without data availability statements) to 53.2% (25 out of 47 studies with data availability statements) (Supplementary table 1).

There are 47 studies containing “Data availability statements” in the corresponding publications. Among the 47 studies with “Data availability statements”, 53.2% (25 out of the 47 studies) studies shared the raw TCR-Seq data in the publications while the rest of the studies did not share raw TCR-Seq data in the original publications or would be available upon request. Conversely, there are 87 studies that did not have “Data availability statements” in the original publications. For the 87 publications that do not have the “Data availability statement”, only 29.9% (26 out of the 87 studies) of the studies shared the raw TCR-Seq data in the publications. There are three studies with data availability statements but provide unavailable accession numbers, so we cannot access the raw data of the three studies. Therefore, we categorized them into the studies with available data availability statements and raw data availability is not mentioned. We conducted a Pearson's chi-squared test (χ^2) to examine the associations between the presence of data availability statements and the availability of raw data and found a statistically significant relationship between the two($p=0.014$) - the data availability statements are improving access to raw data availability .

We also examined parts of the publications where the raw data availability is mentioned. Among the studies that shared the raw TCR-Seq data, authors included the data availability statements in various locations of the research articles. 49.0% (25 out of 51 studies) had the data availability statements in the “Data availability statement” portion of the research articles. Eight out of 51 studies (15.7%) had the data availability statements in the footnote of the studies, whereas the remaining 18 out of 51 studies (35.3%) mentioned their raw data availability directly in the main text of the studies.

Discussion

Our study is the first to assess the raw data availability of TCR-Seq studies. According to the results, the raw data availability is alarmingly low, with only 38.1% of the TCR-Seq studies sharing the raw data (Figure 2b). We investigated the association between the presence of data availability statements in the articles and the raw data availability. We discovered that the data availability statements presented have a statistically significant association with the raw data availability (p=0.014). The presence of data availability statements improves the raw data availability by 23.3%. While data availability statements help to improve data availability, we noticed that it is still possible to have such a statement and not share raw data or only share summary data or only will be available upon request (46.8%) (Supplementary table 1). Some studies only shared the summary data in the data availability statement section of the publications. Additionally, three of the studies shared an erroneous SRA accession number in the data availability statement sections of the publications so we were not able to access the raw data of the articles.

According to our analysis, the primary reason for unavailable raw data is that the studies did not indicate the statements of raw data in the research articles nor shared the raw data in the article (34 studies), which did not share the raw sequencing data of the analyses. Secondly, most authors only shared summary data in the original publications (44 studies) reckoning that they shared the data in their research. For example, 23 studies shared the summary data in the corporate-owned repository, ImmuneACCESS® by Adaptive Biotechnologies. Unfortunately, the ImmuneACCESS® repository only offers access to the summary data of the studies, thus the raw data files are still inaccessible to the public, making the researchers unable to re-analyze the raw data generated by Adaptive Biotechnologies for novel biomedical discoveries. Thirdly, several studies mentioned that the data will only be available by making direct requests of the data from the authors (5 studies). However, it has been previously shown that the statements mentioning that the raw data will be available upon request did not guarantee data availability and instead were not a sustainable and practical way to improve research reproducibility and raw data availability[26,27]. The specific reasons for not sharing raw TCR-Seq data are beyond the scope of this manuscript and need to be investigated in future studies. The perceptual and technical barriers researchers are facing when sharing the data are yet to be determined. Previous work has suggested that the reasons for authors to not share the raw data might be because they are unaware of the importance of sharing the raw data, there may be culturally related or technical barriers that deter or prevent authors from sharing the raw data[27].

Individual researchers, research institutes, and journals should all take part in ensuring raw data availability[27]. Individual researchers and research institutes more specifically should hold the burden of making the raw data publicly available. Journals also have a critical role in promoting raw data availability, and many journals are already taking actions to promote data availability via data availability statement requests within articles upon publication. However, more stringent measures need to be imposed by the journals to ensure raw data availability[28]. Since existing policies for making the data availability statement mandatory in the article are improving raw data availability, our results suggest that journal policies might have the largest potential impact on improving raw data availability. It is known that many journals have already mandated the authors to share the raw data of the studies[29]. Journal's policies in mandating authors to share the raw data might be a feasible way to improve the raw data availability[30]. In conclusion, the pressing need to increase awareness of enhancing raw data availability in scientific research can enable cost-effective secondary analysis of existing immunogenomics data for novel biomedical discoveries. Therefore, we recommend that all members of the biomedical community, including individual researchers, research institutions, and journals, should contribute to increasing the awareness of raw data sharing and improve the raw data availability in future studies.

Reference

1. Brito, J. J. *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *GigaScience* **9**, giaa056 (2020).
2. Johnston, M. P. Secondary Data Analysis: A Method of which the Time Has Come. *Qual. Quant. Methods Libr.* **3**, 619–626 (2017).
3. Breden, F. *et al.* Reproducibility and Reuse of Adaptive Immune Receptor Repertoire Data. *Front. Immunol.* **8**, 1418 (2017).
4. Kodama, Y., Shumway, M., Leinonen, R., & on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).

5. Miyakawa, T. No raw data, no science: another possible source of the reproducibility crisis. *Mol. Brain* **13**, 24 (2020).
6. Rajesh, A. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biol.* **22**, 106 (2021).
7. F, R. *et al.* Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* **18**, 1274–1278 (2017).
8. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
9. Wass, M. N., Ray, L. & Michaelis, M. Understanding of researcher behavior is required to improve data reliability. *GigaScience* **8**, giz017 (2019).
10. Bishop, M. Building the foundation for future research through Open data, code and protocols. *The Official PLOS Blog* https://theplosblog.plos.org/2019/12/building-the-foundation-for-future-research-through-open-data-code-and-protocols/ (2019).
11. van Dijk, E. L., Auger, H., Jaszczyszyn, Y. & Thermes, C. Ten years of next-generation sequencing technology. *Trends Genet. TIG* **30**, 418–426 (2014).
12. Dziubianau, M. *et al.* TCR repertoire analysis by next generation sequencing allows complex differential diagnosis of T cell-related pathology. *Am. J. Transplant. Off. J. Am. Soc. Transplant. Am. Soc. Transpl. Surg.* **13**, 2842–2854 (2013).
13. Hou, D., Chen, C., Seely, E. J., Chen, S. & Song, Y. High-Throughput Sequencing-Based Immune Repertoire Study during Infectious Disease. *Front. Immunol.* **7**, 336 (2016).
14. Benichou, J., Ben-Hamo, R., Louzoun, Y. & Efroni, S. Rep-Seq: uncovering the immunological repertoire through next-generation sequencing. *Immunology* **135**, 183–191 (2012).
15. Arnaout, R. A., Prak, E. T. L., Schwab, N., Rubelt, F., & Adaptive Immune Receptor Repertoire Community. The Future of Blood Testing Is the Immunome. *Front. Immunol.* **12**, 626793 (2021).
16. Ostmeyer, J. *et al.* Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. *BMC Bioinformatics* **18**, 401 (2017).
17. Linette, G. P. *et al.* Immunological ignorance is an enabling feature of the oligo-clonal T cell response to melanoma neoantigens. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 23662–23670 (2019).

18. Cowell, L. G. The Diagnostic, Prognostic, and Therapeutic Potential of Adaptive Immune Receptor Repertoire Profiling in Cancer. *Cancer Res.* **80**, 643–654 (2020).
19. Ostmeyer, J. *et al.* Biophysicochemical motifs in T cell receptor sequences as a potential biomarker for high-grade serous ovarian carcinoma. *PloS One* **15**, e0229569 (2020).
20. Sheih, A. *et al.* Clonal kinetics and single-cell transcriptional profiling of CAR-T cells in patients undergoing CD19 CAR-T immunotherapy. *Nat. Commun.* **11**, 219 (2020).
21. Lee, J. *et al.* Molecular-level analysis of the serum antibody repertoire in young adults before and after seasonal influenza vaccination. *Nat. Med.* **22**, 1456–1464 (2016).
22. Richardson, E. *et al.* A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-pertussis toxoid antibodies. *mAbs* **13**, 1869406 (2021).
23. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991-995 (2013).
24. Freeberg, M. A. *et al.* The European Genome-phenome Archive in 2021. *Nucleic Acids Res.* **50**, D980–D987 (2022).
25. CNCB-NGDC Members and Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic Acids Res.* **49**, D18–D28 (2021).
26. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci.* **115**, 2584–2589 (2018).
27. Tedersoo, L. *et al.* Data sharing practices and data availability upon request differ across scientific disciplines. *Sci. Data* **8**, 192 (2021).
28. Grant, R. & Hrynaszkiewicz, I. The Impact on Authors and Editors of Introducing Data Availability Statements at Nature Journals. *Int. J. Digit. Curation* **13**, 195–203 (2018).
29. Kim, J., Kim, S., Cho, H.-M., Chang, J. H. & Kim, S. Y. Data sharing policies of journals in life, health, and physical sciences indexed in Journal Citation Reports. *PeerJ* **8**, e9924 (2020).
30. Deshpande, D. *et al.* A comprehensive analysis of code and data availability in biomedical research. (2021) doi:10.31219/osf.io/uz7m5.

Figure

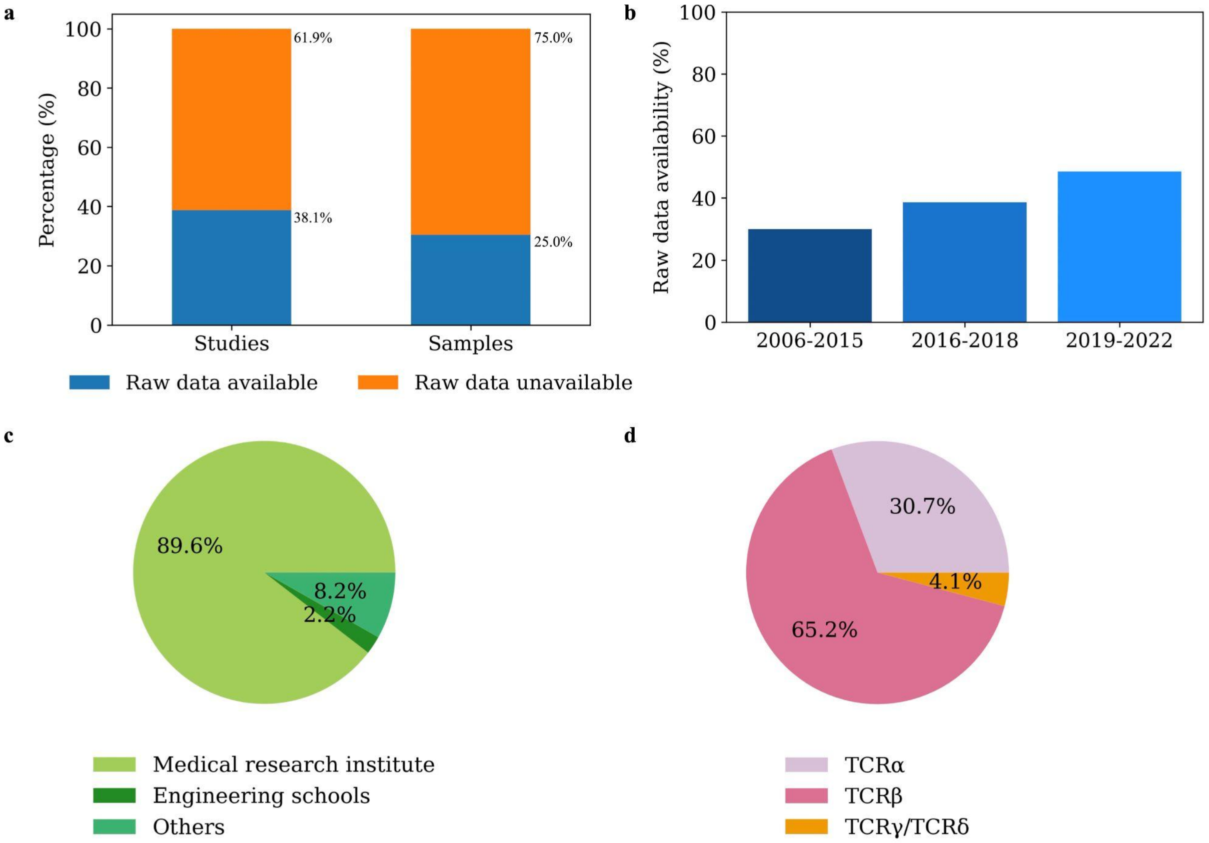


Figure 1. The raw data availability across the 134 TCR-Seq studies. a, The proportion of the raw sequencing data available in the TCR-Seq studies. Left: The raw sequencing data's availability of the 134 TCR-Seq studies; Right: The raw sequencing data's availability of the 11,918 samples in the 134 TCR-Seq studies. b, Bar plot of the changes in the raw TCR-Seq data availability from 2006 to 2022. c, The pie chart depicting the types of institutes where the data was generated. d, The pie chart depicting the types of the chain types of TCR-Seq data.

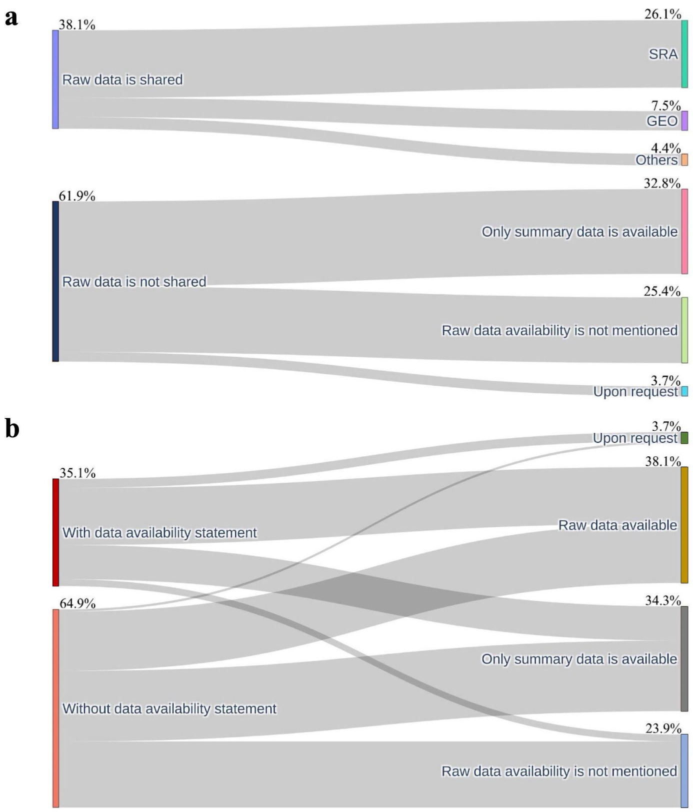
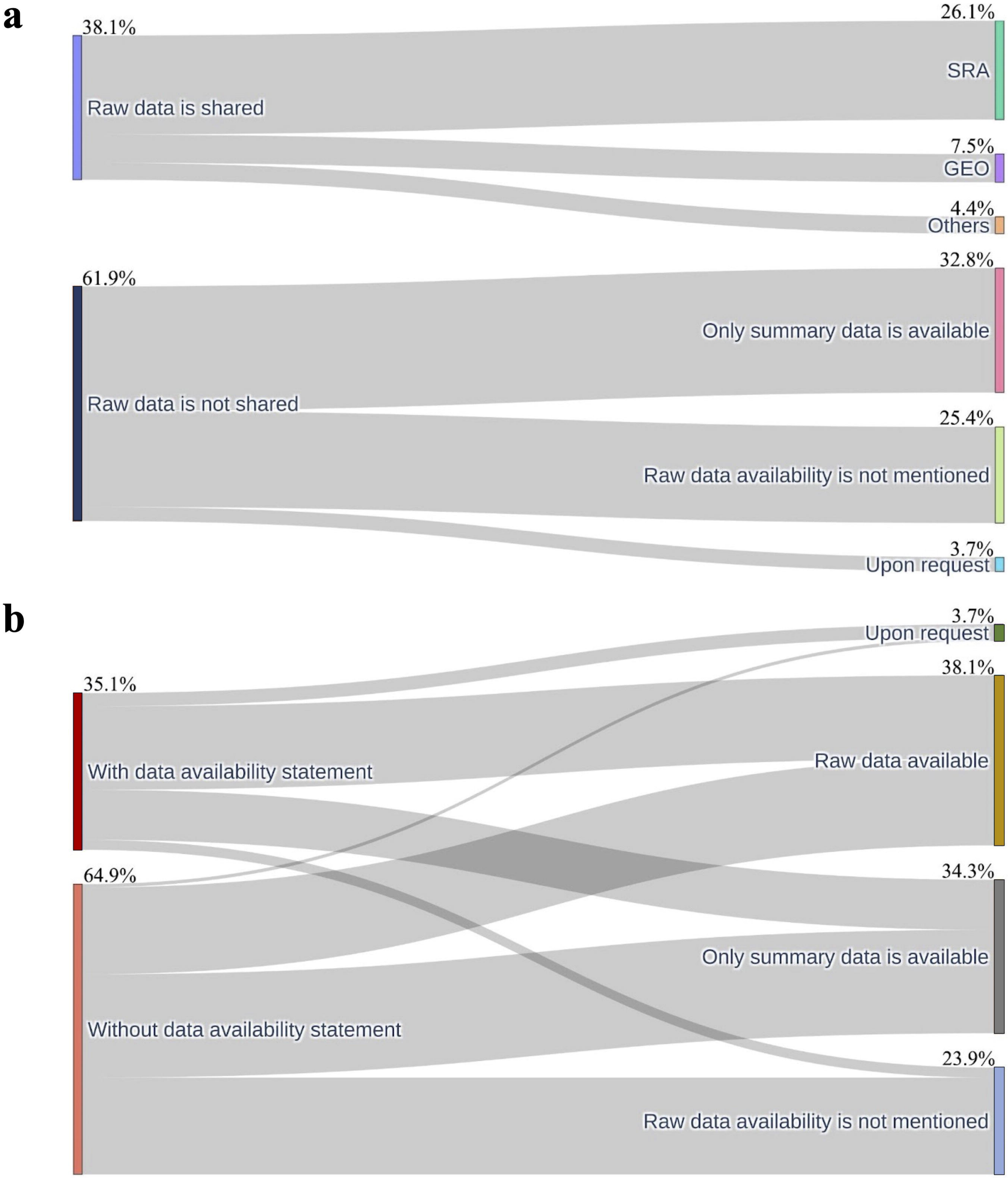


Figure 2. The Sankey plot of the raw data availability across the 134 TCR-Seq studies. a, The Sankey plot depicting the proportion of raw data availability (left) and platforms where available raw data is stored and the specific reasons why the raw data is unavailable (right). b, The Sankey plot shows the proportion of TCR-Seq studies with available "Data availability statements" in the text of the article (left) and the proportion of the studies with available raw data, studies reporting only summary data, and studies mentioning that data is available upon request, and studies not mentioning data availability information.

a
Percentage (%)
0
20
40
60
80
100
61.9%
38.1%
75.0%
25.0%
Studies
Samples
Raw data available
Raw data unavailable
b
Raw data availability (%)
0
20
40
60
80
100
2006-2015
2016-2018
2019-2022
c
89.6%
8.2%
2.2%
Medical research institute
Engineering schools
Others
d
30.7%
4.1%
65.2%
TCRα
TCRβ
TCRγ/TCRδ


a
38.1%
Raw data is shared
61.9%
Raw data is not shared
26.1%
SRA
7.5%
GEO
4.4%
Others
32.8%
Only summary data is available
25.4%
Raw data availability is not mentioned
3.7%
Upon request
b
35.1%
With data availability statement
64.9%
Without data availability statement
3.7%
Upon request
38.1%
Raw data available
34.3%
Only summary data is available
23.9%
Raw data availability is not mentioned