



# ePlatypus: an ecosystem for computational analysis of immunogenomics data

## Journal Article

### Author(s):

Cotet, Tudor-Stefan; Agrafiotis, Andreas; Kreiner, Victor; Kuhn, Raphael; Shlesinger, Danielle; [Manero Carranza, Marcos](#) ; Khodaverdi, Keywan; Kladis, Evgenios; Desideri Perea, Aurora; Maassen-Veeters, Dylan; Glänzer, Wiona; Massery, Solene; Guerci, Lorenzo; Hong, Kai-Lin; Han, Jiami; Stikloraitis, Kostas; D'Arcy, Vittoria Martinolli; Dizerens, Raphael; Kilchenmann, Samuel; Stalder, Lucas; Nissen, Leon; Vogelsanger, Basil; Anzböck, Stine; Laslo, Daria; Bakker, Sophie; Kondorosy, Melinda; Venerito, Marco; Sanz García, Alejandro; Feller, Isabelle; [Oxenius, Annette](#) ; Reddy, Sai T.; Yermanos, Alexander

### Publication date:

2023-09

### Permanent link:

<https://doi.org/10.3929/ethz-b-000635030>

### Rights / license:

[Creative Commons Attribution 4.0 International](#)

### Originally published in:

Bioinformatics 39(9), <https://doi.org/10.1093/bioinformatics/btad553>

## Systems biology

# ePlatypus: an ecosystem for computational analysis of immunogenomics data

Tudor-Stefan Cotet<sup>1,†</sup>, Andreas Agrafiotis<sup>1,2,†</sup>, Victor Kreiner<sup>1,†</sup>, Raphael Kuhn<sup>1</sup>, Danielle Shlesinger<sup>1</sup>, Marcos Manero-Carranza<sup>1</sup>, Keywan Khodaverdi<sup>1</sup>, Evgenios Kladis<sup>1</sup>, Aurora Desideri Perea<sup>3</sup>, Dylan Maassen-Veeters<sup>3</sup>, Wiona Glänzer<sup>1</sup>, Solène Massery<sup>1</sup>, Lorenzo Guerci<sup>1</sup>, Kai-Lin Hong<sup>1</sup>, Jiami Han<sup>1</sup>, Kostas Stikloraitis<sup>1</sup>, Vittoria Martinolli D'Arcy<sup>1</sup>, Raphael Dizerens<sup>1</sup>, Samuel Kilchenmann<sup>1</sup>, Lucas Stalder<sup>1</sup>, Leon Nissen<sup>1</sup>, Basil Vogelsanger<sup>1</sup>, Stine Anzböck<sup>1</sup>, Daria Laslo<sup>1</sup>, Sophie Bakker<sup>3</sup>, Melinda Kondorosy<sup>1</sup>, Marco Venerito<sup>1</sup>, Alejandro Sanz García<sup>1</sup>, Isabelle Feller<sup>1</sup>, Annette Oxenius<sup>2</sup>, Sai T. Reddy<sup>1</sup>, Alexander Yermanos<sup>1,2,3,4,\*</sup>

<sup>1</sup>Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, Basel 4058, Switzerland

<sup>2</sup>Institute of Microbiology, ETH Zurich, Vladimir-Prelog-Weg 4, Zurich 8093, Switzerland

<sup>3</sup>Center for Translational Immunology, University Medical Center Utrecht, Lundlaan 6, Utrecht 3584 EA, The Netherlands

<sup>4</sup>Department of Pathology and Immunology, University of Geneva, 24 rue du Général-Dufour, Geneva 1211, Switzerland

\*Corresponding author. Department of Biosystems Science and Engineering, Mattenstrasse 26, Basel 4058, Switzerland. E-mail: ayermanos@gmail.com (A.Y.)

<sup>†</sup>Equal contribution.

Associate Editor: Janet Kelso

## Abstract

**Motivation:** The maturation of systems immunology methodologies requires novel and transparent computational frameworks capable of integrating diverse data modalities in a reproducible manner.

**Results:** Here, we present the ePlatypus computational immunology ecosystem for immunogenomics data analysis, with a focus on adaptive immune repertoires and single-cell sequencing. ePlatypus is an open-source web-based platform and provides programming tutorials and an integrative database that helps elucidate signatures of B and T cell clonal selection. Furthermore, the ecosystem links novel and established bioinformatics pipelines relevant for single-cell immune repertoires and other aspects of computational immunology such as predicting ligand–receptor interactions, structural modeling, simulations, machine learning, graph theory, pseudotime, spatial transcriptomics, and phylogenetics. The ePlatypus ecosystem helps extract deeper insight in computational immunology and immunogenomics and promote open science.

**Availability and implementation:** Platypus code used in this manuscript can be found at [github.com/alexermanos/Platypus](https://github.com/alexermanos/Platypus).

## 1 Introduction

The fields of systems and computational immunology have advanced substantially in recent years, most notably through progress in genomics and single-cell sequencing, which are transforming the measurement of adaptive immune responses from qualitative to quantitative science. In recent years, a number of bioinformatic software tools have been developed that provide rapid and facile exploration of single-cell RNA sequencing (scSeq) data and perform analyses such as differential gene expression, cell clustering and transcriptional phenotyping (Satija *et al.* 2015, Efremova *et al.* 2020). However, in the context of immunogenomics, lymphocytes (B and T cells) and their transcriptomes and immune receptor repertoires (B cell receptor, BCR and T cell receptor, TCR), there is a lack of software enabling the simultaneous interrogation and integration of multiple approaches capable of deconstructing high-dimensional immune responses, such as phylogenetics, machine learning, graph theory, and structural modeling. Moreover, although deep sequencing of immune repertoires has become a common method in modern

immunology, locating, downloading, and integrating data across experiments and research groups remains challenging. Finally, most immunogenomics software tools require computational expertise involved in analyzing such feature-rich datasets (Yaari and Kleinstein 2015, Yermanos *et al.* 2017, Borcherding *et al.* 2020).

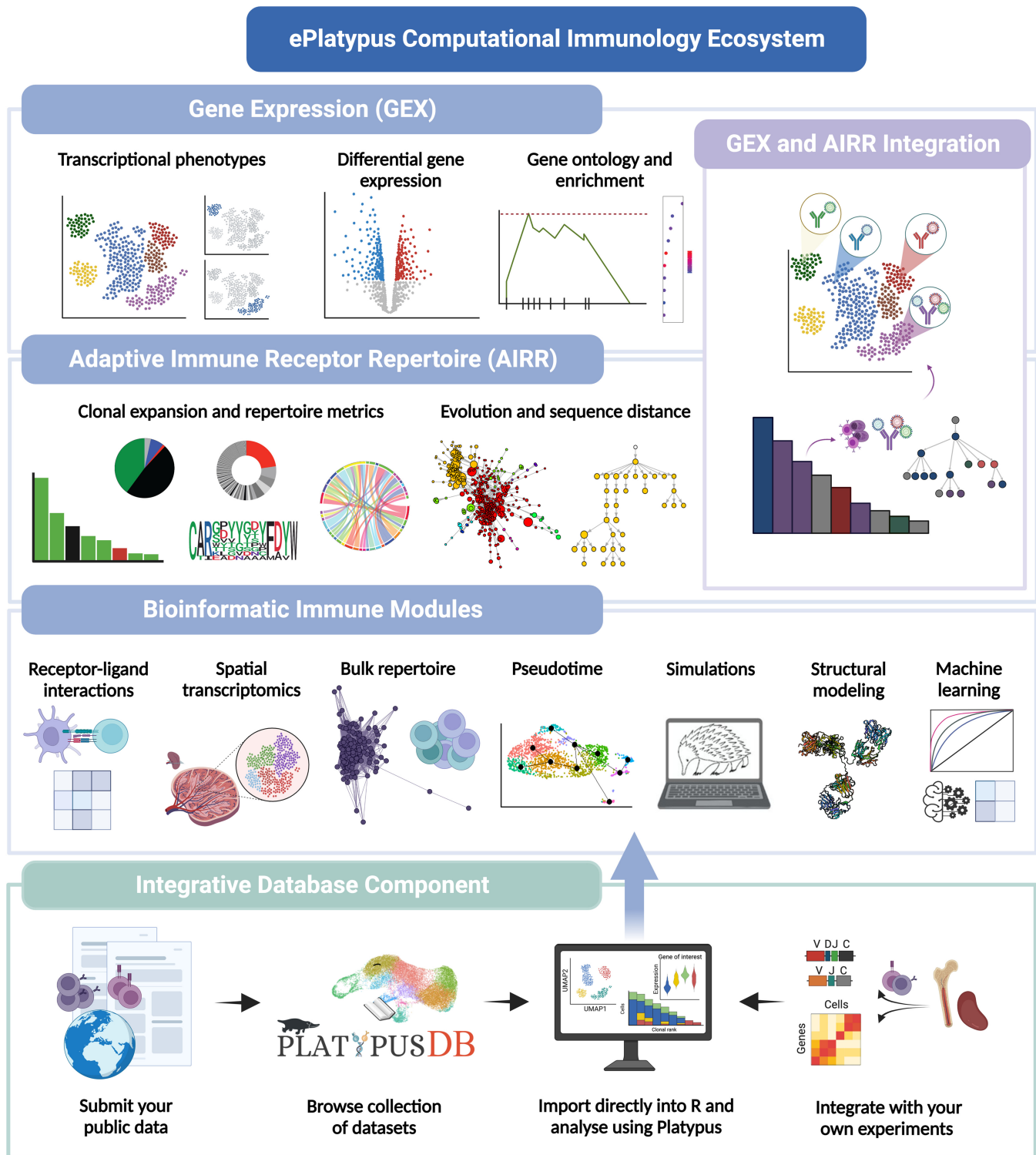
## 2 Ecosystem overview

Here, we present ePlatypus, a computational immunology ecosystem that expands upon Platypus (Yermanos *et al.* 2021a), a previously developed immunogenomics software. The ePlatypus ecosystem (Fig. 1) consists of hundreds of R and python functions, including those most relevant for single-cell immunogenomics (transcriptomes and immune repertoires) as well as many other aspects of computational immunology. More specifically, this novel ecosystem represents a complete rework from the original Platypus R package (Yermanos *et al.* 2021a), and has been rebuilt around a central data object that is now compatible with R and python

Received: 11 November 2022; Revised: 8 August 2023; Editorial Decision: 24 August 2023; Accepted: 6 September 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



**Figure 1.** Breadth of the ePlatypus computational immunology ecosystem. The ecosystem currently is composed of a core R package that has pipelines pertaining to immune repertoires, gene expression, receptor–ligand interactions, spatial transcriptomics, pseudotime, simulations, structural modeling, and machine learning. Similarly, the ecosystem contains an integrated database and a website currently containing 21 tutorials with accompanying code.

and can directly store and integrate features such as gene expression, immune receptors, spatial coordinates, and structural information (Supplementary Table S1). This central object can be directly supplied as input for novel downstream applications and modules spanning a wide-range of immunogenomics and bioinformatics applications (Supplementary Tables S1 and S2). Additionally, the ePlatypus ecosystem

contains a database component, PlatypusDB, that directly integrates into the R programming language, thereby allowing the rapid analysis and integration of B and T cells containing both adaptive immune receptor information (VDJ) and single-cell transcriptomes (GEX). PlatypusDB both stores raw output files from the commonly used aligner tool Cellranger (10x Genomics) and also holds the immune-relevant data in the

form of an R object that can be loaded directly into the R environment without explicitly requiring file download. Importantly, the data is stored as both the processed aligned output and as a preprocessed R object that contains transcriptome, immune repertoire, and metadata information. Within the programming interface, the user has the ability to perform the following actions: (i) download entire public sequencing datasets, (ii) download individual samples from publications, and (iii) download and integrate public repertoires with samples stored locally (Fig. 1). While the ePlatypus development team will continuously update the ecosystem with newly published datasets, external users can also submit their preprocessed immune receptor repertoires directly for manual curation and addition to the database.

### 3 Usage and application

To demonstrate several use cases of the ePlatypus computational ecosystem, we integrated and analyzed multiple single-cell transcriptomes and immune receptor repertoires across different disease conditions, viral infections, and vaccination studies (Supplementary Fig. S1 and Supplementary Table S3). These datasets were used to highlight various modules including: (i) pseudobulking differential expression pipelines to robustly characterize transcriptional clusters leveraging methods originally designed for bulk RNA-sequencing (Supplementary Fig. S2), (ii) immune repertoire diversity metrics to characterize clonal distributions and to ensure sufficient sampling depths have been recovered (Supplementary Fig. S3), (iii) phylogenetics to identify evolutionary trajectories and intraclonal network properties of B cells during infection (Supplementary Fig. S4), (iv) B and T cell sequence similarity networks to identify fundamental principles of lymphocyte repertoire architecture in the course of an immune response (Supplementary Fig. S4), (v) machine-learning guided classification to predict BCR and TCR specificity and further uncover feature importance of antigen-specific sequences (Supplementary Fig. S5), (vi) predicting ligand–receptor interactions under homeostatic and disease conditions using the CellphoneDB repository (Efremova *et al.* 2020) (Supplementary Fig. S6), (vii) spatial transcriptomics to spatially interrogate gene expression patterns and further integrate clonal selection and clonal evolution of adaptive immune responses (Supplementary Fig. S7), and (viii) structural modeling of immune receptor sequences and repertoires with the Steropodon pipeline using multiple external tools including AlphaFold, IgFold, and DeepAb (Jumper *et al.* 2021) (Supplementary Fig. S8). Furthermore, ePlatypus now supports python functionality for the implementation of repertoire analyses such as investigating clonal expansion and isotype distribution. This python pipeline can also be supplied to more advanced machine learning and artificial intelligence workflows such as the use of protein language models, including both foundational and receptor-specific language models such as ProtBERT, Sapiens (Prihoda *et al.* 2022), AntiBERTy (Ruffolo *et al.* 2021), ESM-1B (Lin *et al.* 2023), Ablang (Olsen *et al.* 2022), and TCR-BERT (Wu *et al.* 2021) for repertoire feature visualization and classification (Supplementary Fig. S9). Importantly, ePlatypus currently hosts an online portal with 21 educational tutorials and walk-throughs (Supplementary Fig. S10), each of which contain code, comments, and explanatory text (Supplementary Fig. S11) for various computational immunology frameworks (Supplementary Table S2).

To further demonstrate several use cases of the ePlatypus computational ecosystem and accompanying database, we integrated and analyzed multiple single-cell transcriptomes and immune receptor repertoires across different disease conditions, viral infections, and vaccination studies (Supplementary Fig. S1 and Supplementary Table S3). We directly downloaded murine T cell repertoires from previously published datasets containing both CD4 and CD8 T cells from conditions such as acute and chronic viral infections (Khatun *et al.* 2021, Merckenschlager *et al.* 2021, Kuhn *et al.* 2022, Shlesinger *et al.* 2022), homeostatic aging (Yermanos *et al.* 2021b), and experimental autoimmune encephalomyelitis (Shlesinger *et al.* 2022) (Supplementary Fig. S1 and Supplementary Table S3). Following transcriptional integration with Harmony (Korsunsky *et al.* 2019), which aims to reduce batch effects across different datasets, we visualized all cells using uniform manifold approximation projection (UMAP) (Supplementary Fig. S12A). This demonstrated two major transcriptional regions, dominated by either Cd4 or Cd8 gene expression, which could be simultaneously interrogated with other known gene markers of activation or exhaustion such as Cd44, Ifng, Pdcd1, Lag3, and Il7r (Supplementary Fig. S12B). Supplementing this focused analysis with ProjectTILS, a recently developed reference atlas which helps resolve murine T cell heterogeneity of tumor-infiltrating T cells (Andreatta *et al.* 2021), demonstrated that T cells from PlatypusDB almost entirely cover the ProjectTILS main reference dataset (Supplementary Figs S12C–E and S13).

Next, we explored whether transcriptional heterogeneity could similarly be detected for B cells present in PlatypusDB. Multiple datasets derived from murine models of infection, immunization, and autoimmune disease (Merckenschlager *et al.* 2021, Yewdell *et al.* 2021, Neumeier *et al.* 2022, Shlesinger *et al.* 2022, Agrafiotis *et al.* 2023) were integrated as previously performed with T cells (Supplementary Fig. S1 and Supplementary Table S3). Transcriptional analysis using both canonical B cell markers and previously reported B cell gene signatures highlighted the presence of diverse B cell subtypes present in PlatypusDB across multiple datasets (Supplementary Fig. S14A and B). For example, our database contains a large number of ASCs, identified based on expression of Sdc1 (Cd138), Xbp1, and Slamf7, which exhibited varying expression levels of markers relating to chemokine receptors (Cxcr3 and Cxcr4) and B cell proliferation (Mki67) (Supplementary Fig. S14C).

### 4 Concluding remarks

The analyses presented here highlight the breadth of B and T cell phenotypes and selection patterns already available within ePlatypus, which will only continue to grow as more user-supplied public datasets are added. Lastly, we computed the runtime of several pipelines within the ePlatypus ecosystem on datasets of varying size and cell numbers, highlighting the scalability and speed of our software (Supplementary Table S4).

The maturation of systems immunology methodologies requires novel and transparent computational frameworks capable of integrating diverse data modalities in a reproducible manner. The ePlatypus ecosystem, composed of hundreds of R and python functions, programming tutorials, and a comprehensive database, helps extract deeper insight in immunogenomics while promoting open science.

## Acknowledgements

We acknowledge and thank Dr Christian Beisel, Elodie Burcklen, Ina Nissen, and Mirjam Feldkamp at the ETH Zurich D-BSSE Genomics Facility Basel for excellent support and assistance. We also thank Nathalie Oetiker and Franziska Wagen for excellent experimental support.

## Supplementary data

Supplementary data are available at *Bioinformatics* online.

## Conflict of interest

None declared.

## Funding

This work was supported by the ETH Zurich Research Grants to S.T.R. and A.O.; an ETH Seed Grant to A.Y.; the “la Caixa” Foundation [ID 100010434, fellowship code: LCF/BQ/EU20/11810041] to M.M.-C.

## Data availability

The accession numbers and publications for the sequencing data used in this manuscript are located in [Supplementary Table S1](#). Platypus code used in this manuscript can be found at [github.com/alexeyermanos/Platypus](https://github.com/alexeyermanos/Platypus).

## References

- Agrafiotis A, Neumeier D, Hong K-L *et al.* Generation of a single-cell B cell atlas of antibody repertoires and transcriptomes to identify signatures associated with antigen specificity. *iScience* 2023;**26**:106055.
- Andreatta M, Corria-Osorio J, Müller S *et al.* Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nat Commun* 2021;**12**:2965.
- Borcherding N, Bormann NL, Kraus G. scRepertoire: an R-based toolkit for single-cell immune receptor analysis. *F1000Res* 2020;**9**:47.
- Brandes N, Ofer D, Peleg Y *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10.
- Efremova M, Vento-Tormo M, Teichmann SA *et al.* CellPhoneDB: inferring cell-cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat Protoc* 2020;**15**:1484–506.
- Jumper J, Evans R, Pritzel A *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–9.
- Khatun A, Kasmani MY, Zander R *et al.* Single-cell lineage mapping of a diverse virus-specific naive CD4 T cell repertoire. *J Exp Med* 2021;**218**:e20200650.
- Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with harmony. *Nat Methods* 2019;**16**:1289–96.
- Kuhn R, Sandu I, Agrafiotis A *et al.* Clonally expanded virus-specific CD8 T cells acquire diverse transcriptional phenotypes during acute, chronic, and latent infections. *Front Immunol* 2022;**13**:782441.
- Lin Z, Akin H, Rao R *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30.
- Merkenschlager J, Finkin S, Ramos V *et al.* Dynamic regulation of T selection during the germinal Centre reaction. *Nature* 2021;**591**:458–63.
- Neumeier D, Pedrioli A, Genovese A *et al.* Profiling the specificity of clonally expanded plasma cells during chronic viral infection by single-cell analysis. *Eur J Immunol* 2022;**52**:297–311.
- Olsen TH, Moal IH, Deane CM. AbLang: an antibody language model for completing antibody sequences. *Bioinform Adv* 2022;**2**:vbac046.
- Prihoda D, Maamary J, Waight A *et al.* BioPhi: a platform for antibody design, humanization, and humanness evaluation based on natural antibody repertoires and deep learning. *MAbs* 2022;**14**:2020203.
- Ruffolo JA, Gray JJ, Sulam J. Deciphering antibody affinity maturation with language models and weakly supervised learning. arXiv [q-bio:BM], 2021, preprint: not peer reviewed.
- Satija R, Farrell JA, Gennert D *et al.* Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* 2015;**33**:495–502.
- Shlesinger D, Hong K-L, Shammass G *et al.* Single-cell immune repertoire sequencing of B and T cells in murine models of infection and autoimmunity. *Genes Immun* 2022;**23**:183–195.
- Vander Heiden JA, Marquez S, Marthandan N *et al.*; AIRR Community. AIRR community standardized representations for annotated immune repertoires. *Front Immunol* 2018;**9**:2206.
- Wu K, Yost KE, Daniel B *et al.* TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-binding analyses. bioRxiv, <https://doi.org/10.1101/2021.11.18.469186>, 2021.
- Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med* 2015;**7**:121.
- Yermanos A, Agrafiotis A, Kuhn R *et al.* Platypus: an open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. *NAR Genom Bioinform* 2021a;**3**:lqab023.
- Yermanos A, Greiff V, Krautler NJ *et al.* Comparison of methods for phylogenetic B-cell lineage inference using time-resolved antibody repertoire simulations (AbSim). *Bioinformatics* 2017;**33**:3938–46.
- Yermanos A, Neumeier D, Sandu I *et al.* Single-cell immune repertoire and transcriptome sequencing reveals that clonally expanded and transcriptionally distinct lymphocytes populate the aged Central nervous system in mice. *Proceedings of the Royal Society. Proc Biol Sci* 2021b;**288**:20202793.
- Yewdell WT, Smolkin RM, Belcheva KT *et al.* Temporal dynamics of persistent germinal centers and memory B cell differentiation following respiratory virus infection. *Cell Rep* 2021;**37**:109961.