

ePlatypus: an ecosystem for computational analysis of immunogenomics data

Victor Kreiner^{1,*}, Andreas Agrafiotis^{1,2,*}, Tudor-Stefan Cotet¹, Raphael Kuhn¹, Danielle Shlesinger¹, Marcos Manero-Carranza¹, Keywan Khodaverdi¹, Solène Massery¹, Lorenzo Guerci¹, Kai-Lin Hong¹, Jiami Han¹, Kostas Stiklioraitis¹, Vittoria Martinolli D'Arcy¹, Raphael Dizerens¹, Samuel Kilchenmann¹, Lucas Stalder¹, Leon Nissen¹, Basil Vogelsanger¹, Stine Anzböck¹, Daria Laslo¹, Melinda Kondorosy¹, Marco Venerito¹, Alejandro Sanz García¹, Isabelle Feller¹, Annette Oxenius², Sai T. Reddy¹, Alexander Yermanos^{1,2,3,4,#}

¹Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland

²Institute of Microbiology, ETH Zurich, Zurich, Switzerland

³Department of Pathology and Immunology, University of Geneva, Geneva, Switzerland

⁴Center for Translational Immunology, University Medical Center Utrecht, Utrecht, Netherlands

*equal contribution

#Correspondence: ayermanos@gmail.com

Accessibility: <https://alexyermanos.github.io/Platypus/index.html>

Abstract

The maturation of systems immunology methodologies requires novel and transparent computational frameworks capable of integrating diverse data modalities in a reproducible manner. 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 a web-based platform and provides programming tutorials and an integrative database that elucidates selection patterns of adaptive immunity. 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.

Main

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 analysis such as differential gene expression, cell clustering and transcriptional phenotyping (Efremova et al. 2020; Satija et al. 2015). 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 (Borcherding et al., 2020; Yaari & Kleinstein, 2015; Yermanos et al., 2021).

Here, we present ePlatypus, a computational immunology ecosystem that expands upon Platypus, a previously developed immunogenomics software. The ePlatypus ecosystem (Figure 1) consists of an R package hosted on CRAN and possesses hundreds of functions, including those most relevant for single-cell immunogenomics (transcriptome and immune repertoires) as well as many other aspects of computational immunology. These include the following: i) profiling the relationship between clonal expansion and transcriptional phenotypes to identify distinct activation phenotypes of expanded lymphocytes (Figure S1), ii) pseudobulk differential expression pipelines to robustly characterize transcriptional clusters leveraging methods originally designed for bulk RNA-sequencing (Figure S2), iii) immune repertoire diversity metrics to characterize clonal distributions and to ensure sufficient sampling depths have been recovered (Figure S3), iv) phylogenetics to identify evolutionary trajectories and intraclonal network properties of B cells during infection (Figure S4), v) B and T cell sequence similarity networks to identify fundamental principles of lymphocyte repertoire architecture in the course of an immune response (Figure S4), vi) machine-learning guided classification to predict BCR and TCR specificity and further uncover feature importance of antigen-specific sequences (Figure S5), vii) predicting ligand-receptor interactions under homeostatic and disease conditions using the CellphoneDB repository (Efremova et al. 2020) (Figure S6), viii) spatial transcriptomics to spatially interrogate gene expression patterns and further integrate clonal selection and clonal evolution of adaptive immune responses (Figure S7),

ix) and structural modeling of immune receptor sequences and repertoires using multiple external tools including AlphaFold, IgFold, and DeepAb (Ruffolo et al. 2022; Jumper et al. 2021) (Figure S8). Importantly, ePlatypus currently hosts an online portal with 20 tutorials and walk-throughs (Figure S9), each of which contain code, comments, and explanatory text (Figure S10) for various computational immunology frameworks (Table S1, Figure S1).

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 (Figure 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.

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. 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; Kuhn et al., 2022; Merkenschlager et al., 2021; Shlesinger et al., 2022), homeostatic aging (Yermanos, et al., 2021) and experimental autoimmune encephalomyelitis (Shlesinger et al., 2022) (Table S2). 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) (Figure S11A, S11B). 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* (Figures S11B, S12). 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 (Figures S11C, S11D, S11E, S13). To highlight the potential to link repertoire features with transcriptional heterogeneity, we visualized the most expanded T cell clones on the transcriptional landscape (Figure S11F). This demonstrated diverse levels of clonal expansion within the database, with those expanded clones corresponding to a relative upregulation of activation markers compared to those with lower expansion (Figure S12).

Next, we used ePlatypus to investigate whether similar transcriptional heterogeneity could be detected for B cells present in the PlatypusDB. Multiple datasets derived from murine models of infection, immunization and autoimmune disease (Agrafiotis et al., 2021; Mathew et al., 2021b; Neumeier et al., 2021; Shlesinger et al., 2022; Yewdell et al., 2021) were integrated as previously described (Table S2). Transcriptional clustering suggested that common B cell clusters were present across multiple datasets (Figure S14A), which exhibited varying expression levels of markers relating to antibody secretion and B cell differentiation (e.g. *Cd138*, *Xbp1*, *Slamf7*) (Figure S15) and diverse isotype usage (Figure S14B). The analyses presented here highlights 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.

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 a core R package with hundreds of functions, programming tutorials, and a comprehensive database, helps extract deeper insight in immunogenomics while promoting open science.

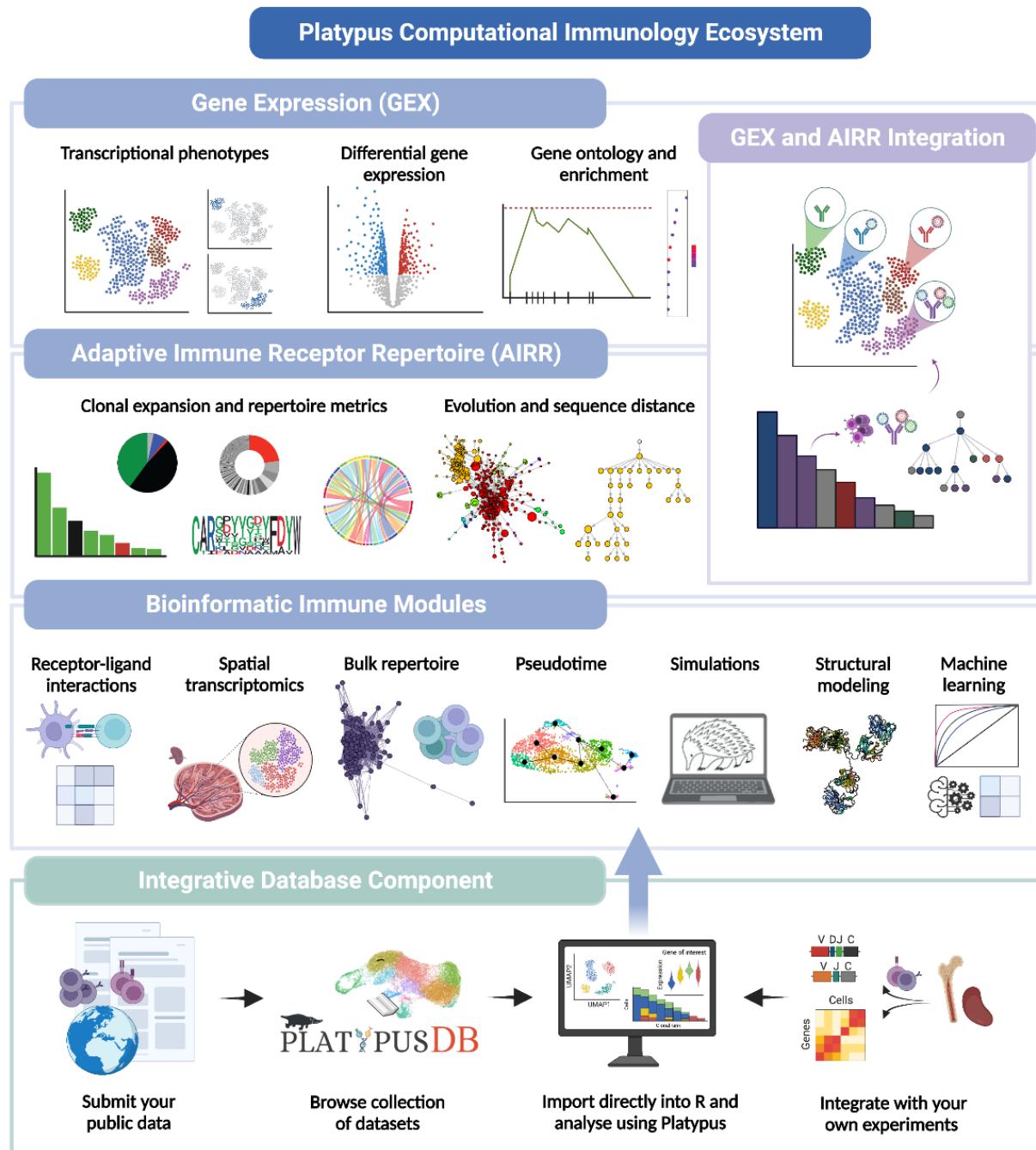


Figure 1. Breadth of the ePlatypus Computational Immunology Ecosystem. The ecosystem currently is composed of a core R package that has pipelines pertaining to immune repertoires, gene expression, receptor-ligand interactions, spatial transcriptomics, pseudotime, simulations, structural modeling and machine learning. Similarly, the ecosystem contains an integrated database and a website currently containing 20 tutorials with accompanying code.

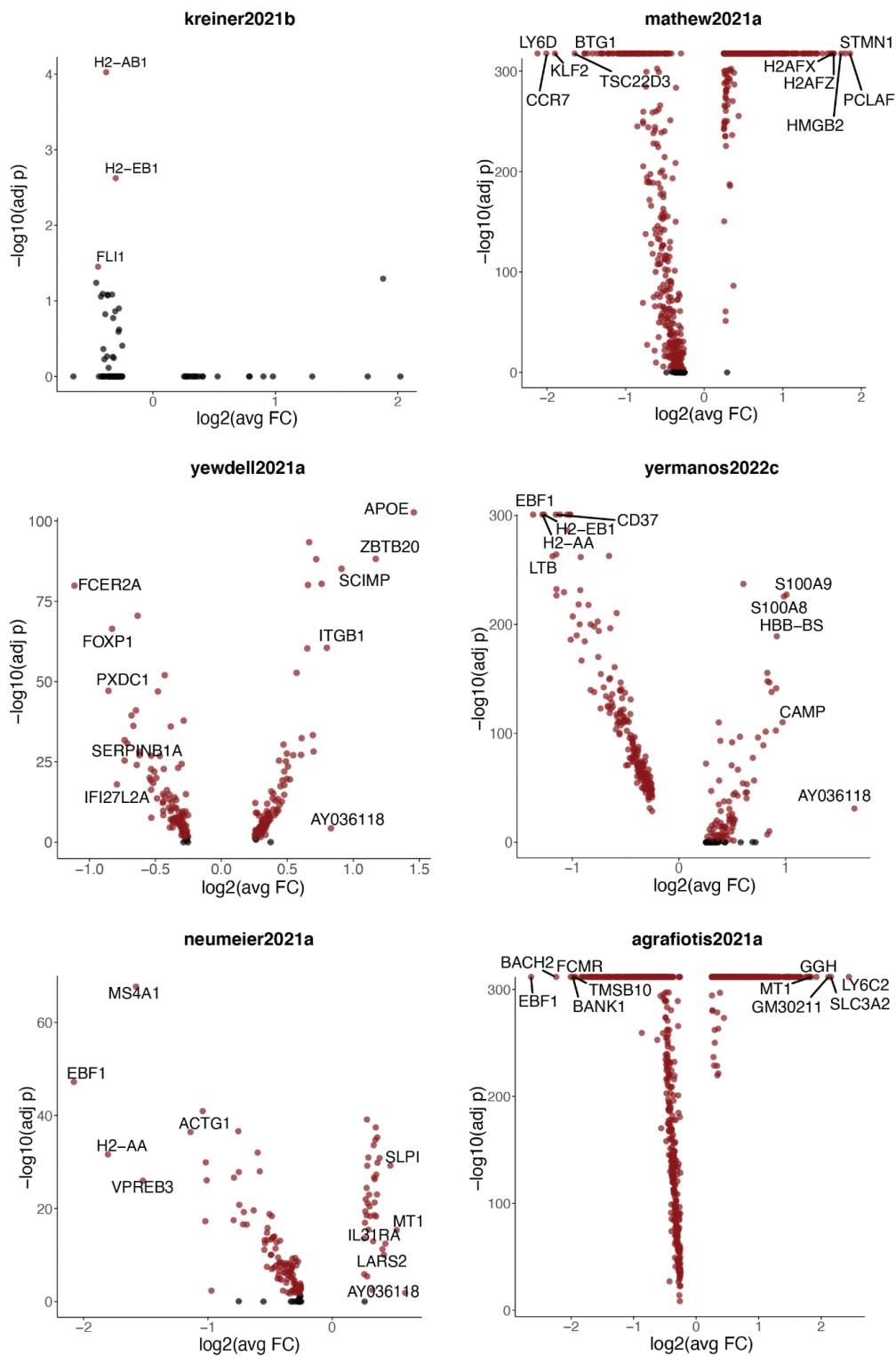


Figure S1. Differentially expressed genes between lowly ($n=1$ cell per clone) and highly ($n>1$ cell per clone) expanded clones for different B cell datasets. Positive $\log_2(\text{avgFC})$ indicates upregulated genes in the expanded clones for all datasets. Headers indicate the origin of the dataset as can be found in PlatypusDB.

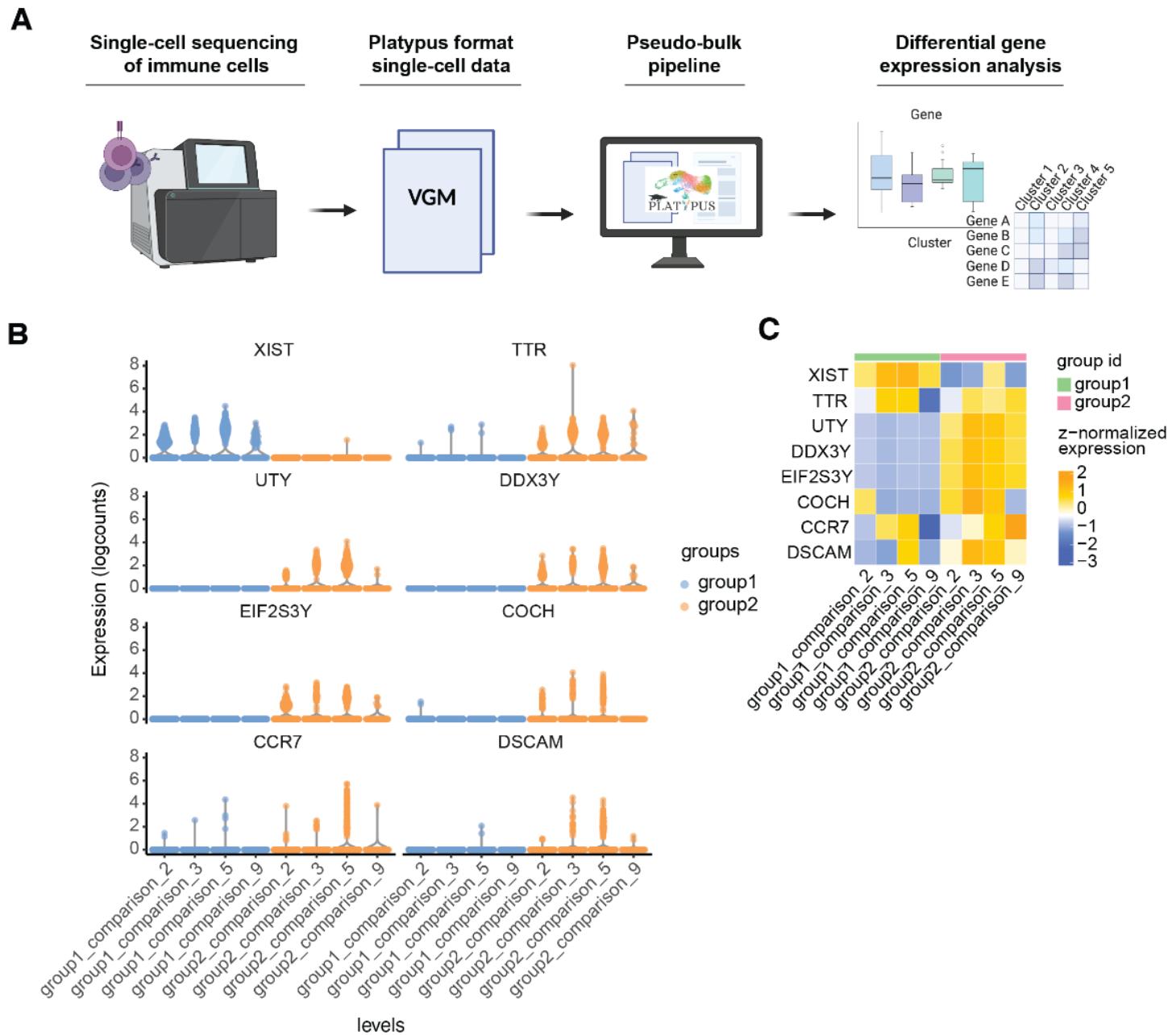


Figure S2. Pseudobulk pipeline within the ePlatypus ecosystem. A. Graphical overview of the Pseudobulk workflow. B. Violin plots displaying the expression levels of differentially expressed genes between two user-defined categories (group 1 and group 2). Single-cell transcriptional clusters 2, 3, 5 and 9 were grouped in comparison levels (or clusters) according to user-defined categories. Cells within the same comparison level belonging to different groups were tested for differential gene expression using the Platypus `pseudo_bulk_DE` function. C. Heatmap highlighting normalized expression levels of differentially expressed genes between the different comparison levels.

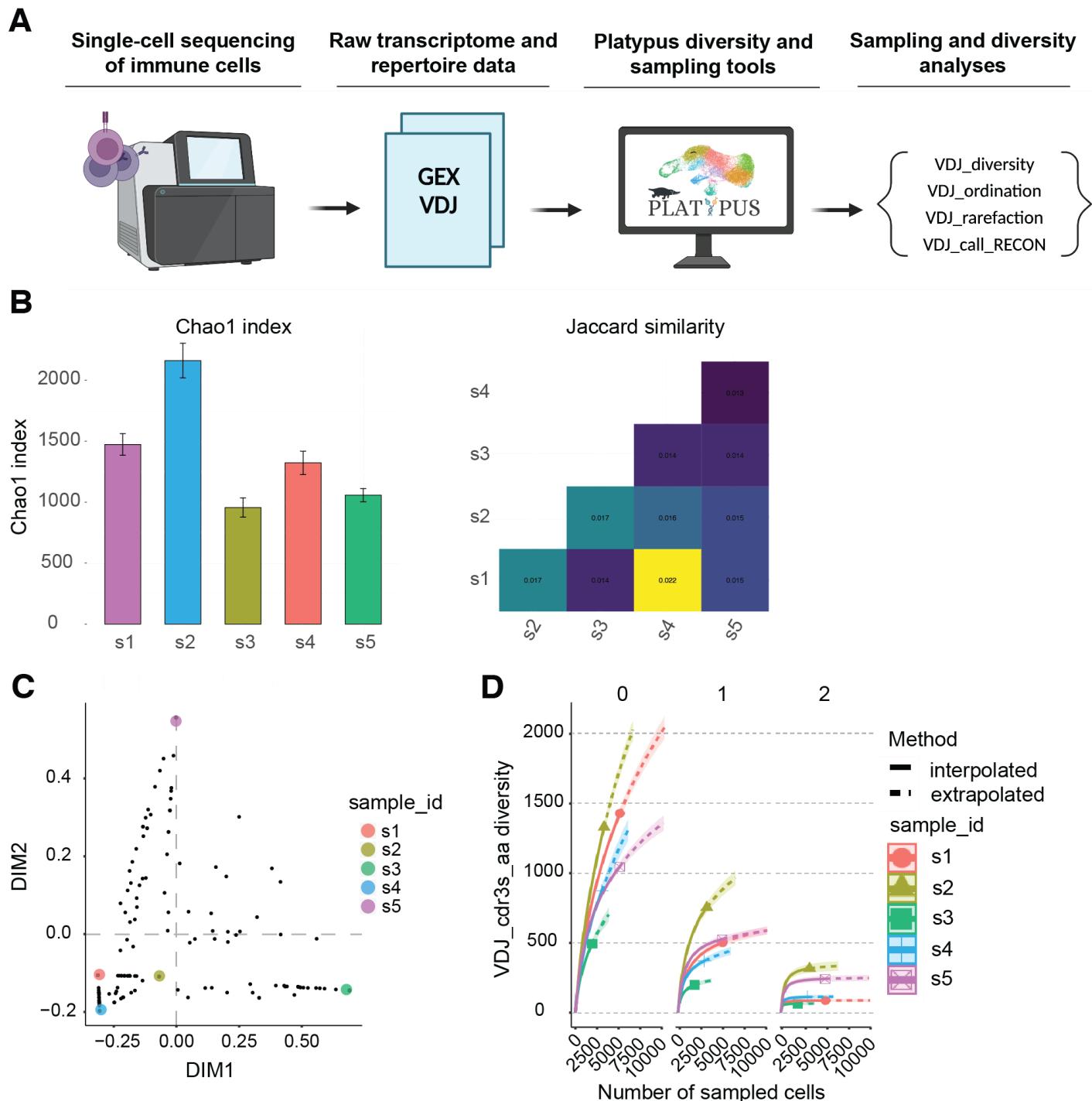


Figure S3. Sampling and diversity analysis pipelines. A. Graphical overview of the sampling and diversity workflow. Single-cell immune repertoire sequencing samples are formatted into a single VDJ_GEX_matrix object, which can be supplied to downstream calculations of diversity, rarefaction, and ordination analyses. B. VDJ_diversity includes a wide range of α and β -diversity calculators. α -diversity is displayed as per-sample bar plots for the Chao1 index, whereas β -diversity is shown as a heatmap. C. Dimensionality reduction on the abundance data can be performed using VDJ_ordination – several dimensionality reduction algorithms are available inside the function

(UMAP, t-SNE, PCA, PCoA). D. Rarefaction curves can be displayed using the VDJ_rarefaction function: in this example, the curve indicative of each Hill number is grouped by sample.

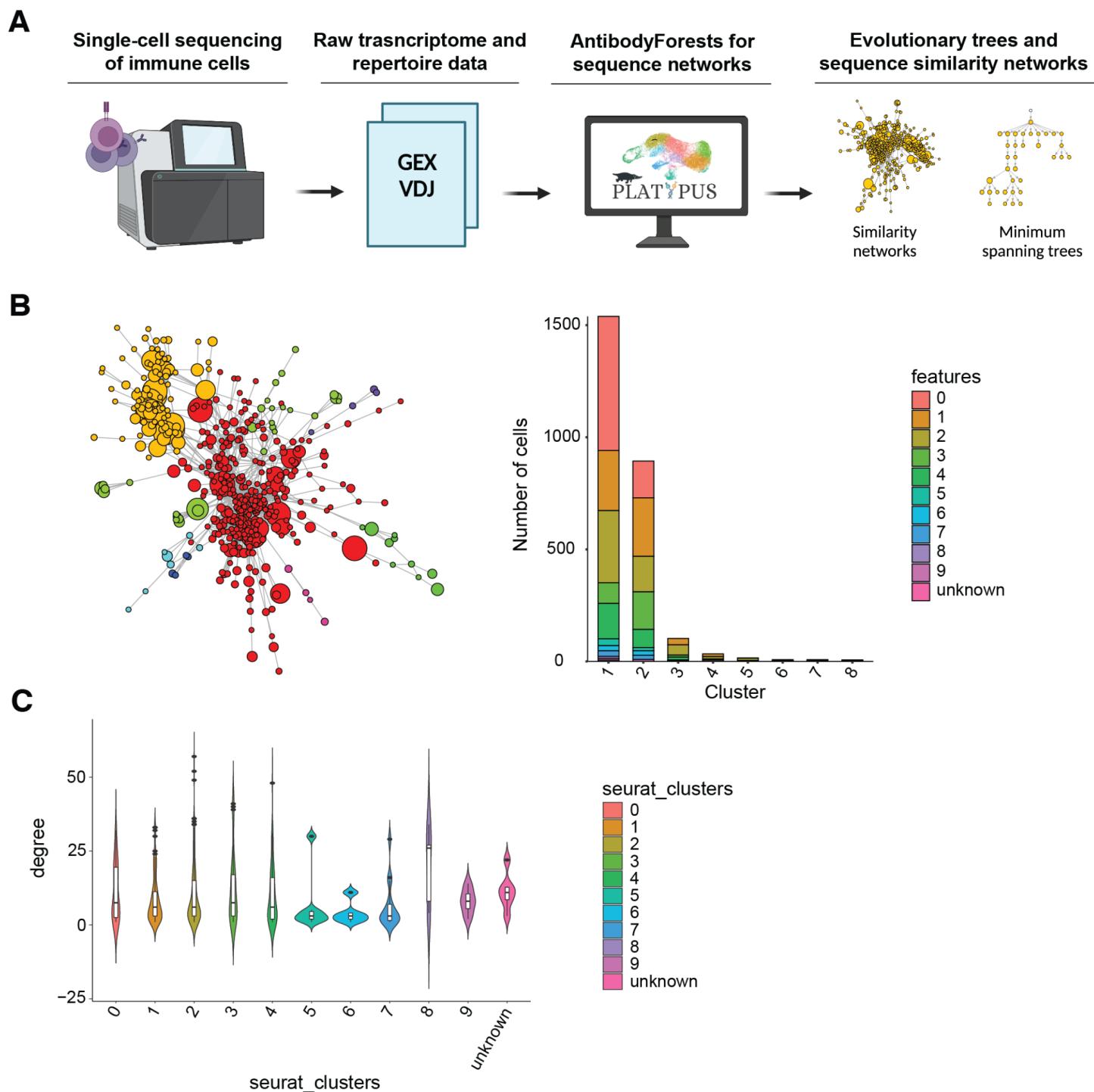


Figure S4. Repertoire similarity networks and quantification. A. The AntibodyForests pipeline designed for network inference, direct visualization, and network analysis of immune receptor minimum spanning trees or sequence similarity graphs. B. Clusters can be determined in a sequence similarity network using the `AntibodyForests_communities` function and then colored using the `AntibodyForests_plot` function. Moreover, the tool can produce per-cluster bar plots of various sequence or cell features (e.g., the Seurat transcriptomic cluster per sequence-similarity clusters). C. The `AntibodyForests_metrics` function calculates a wide range of node and edge metrics for the resulting AntibodyForests networks: in this case, a violin plot depicts the node degree distribution

across each Seurat transcriptomic cluster for the similarity graph showcased in B (each node was assigned the most frequent Seurat cluster across all cells with that particular receptor sequence).

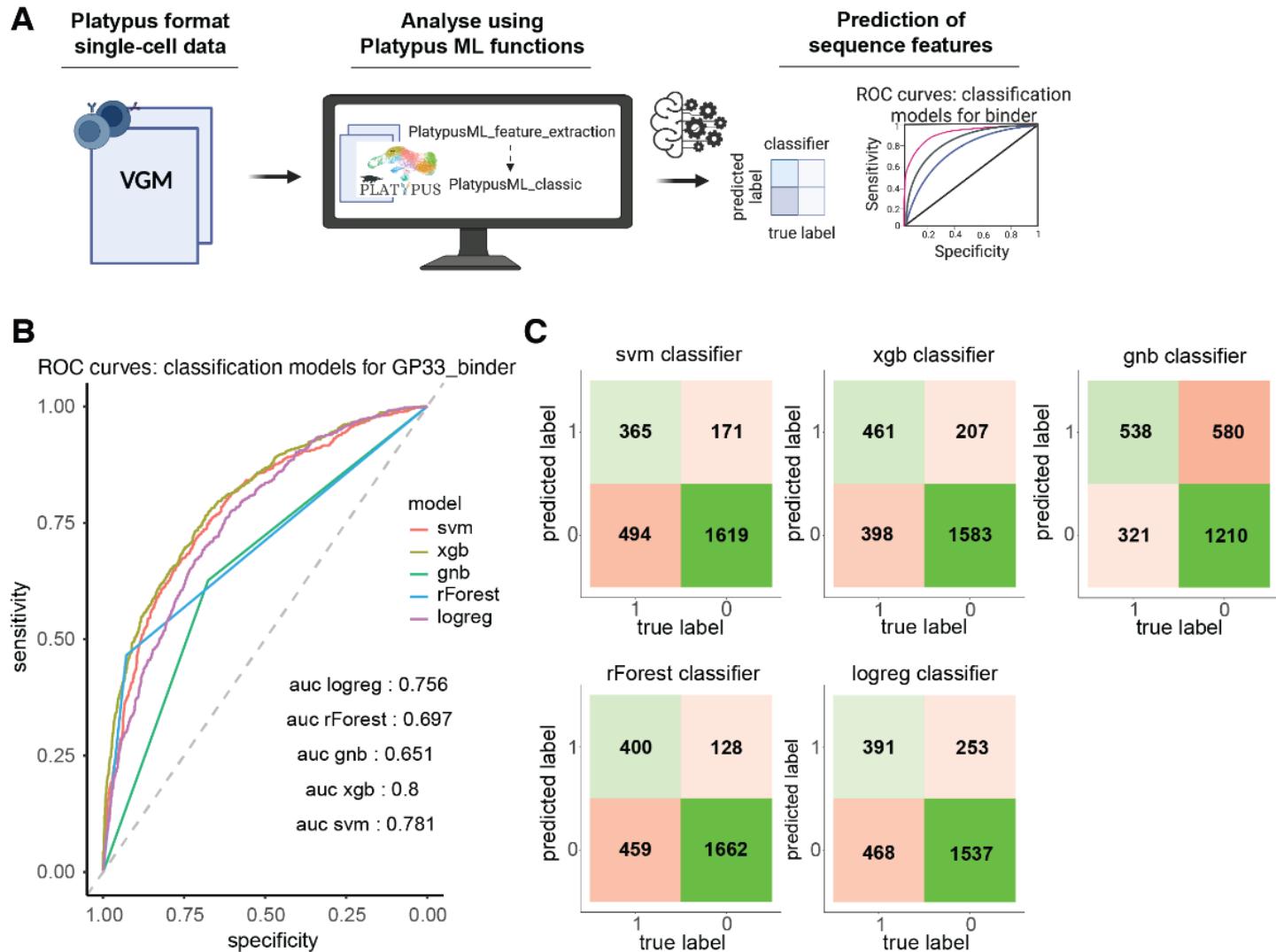


Figure S5. Machine learning pipelines for adaptive immune receptor classification. A. Workflow for machine learning (ML) classification of virus-specificity within the ePlatypus Ecosystem. B. Receiver operating characteristic (ROC) curves and area under the curve (AUC) scores for different classification models (logreg: logistic regression, rForest: random forest, gnb: gaussian naive bayes, xgb: XG Boost, svm: Support vector machine) for binary binding vs non-binding classification of virus-specific T cell receptors following viral infection. C. Confusion matrices for each of the models shown in (B). Green intensity is proportional to the number of true positives and true negatives, while red intensity is proportional to the number of false negatives and false positives.

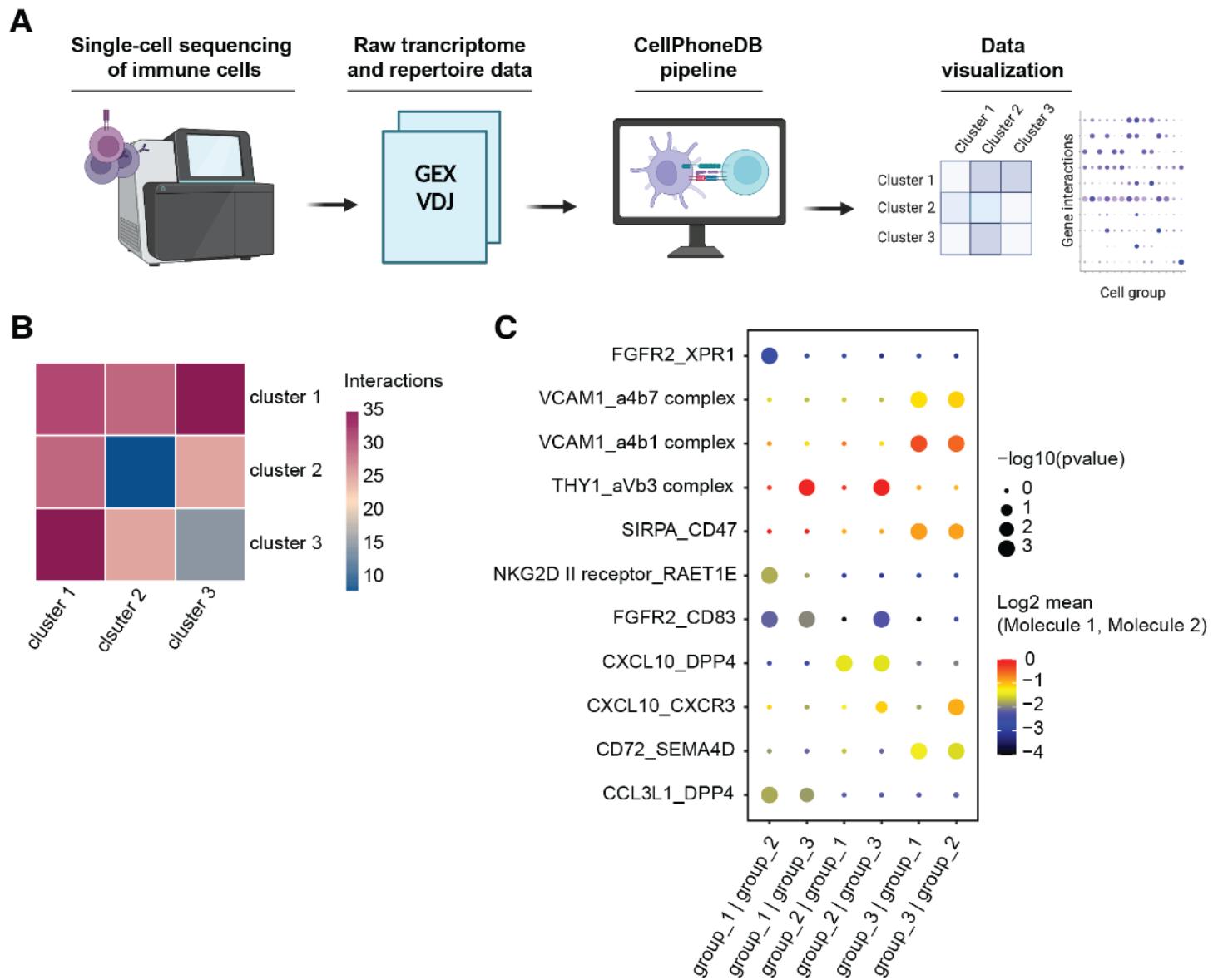
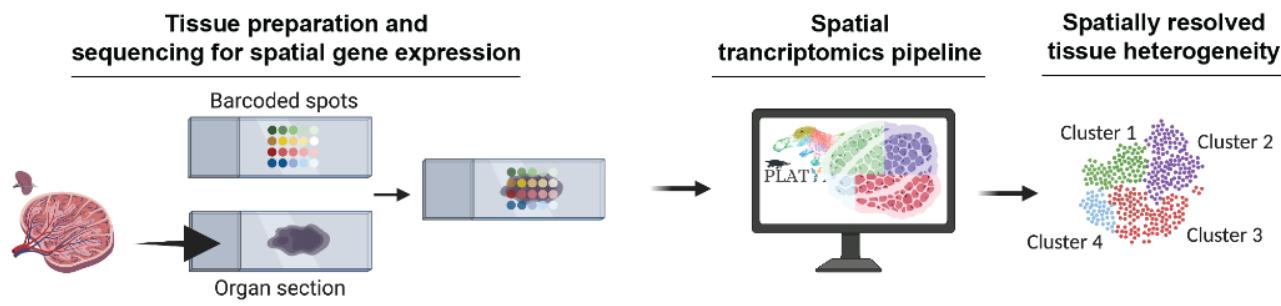


Figure S6. Predicting receptor-ligand interaction using CellphoneDB. A. Graphical overview for processing, generating, and analyzing ligand-receptor interactions. B. Heatmap displaying the level of interaction between cells belonging to different clusters. The interaction level is calculated on the single-cell gene expression data using the CellPhoneDB software. The heatmap is part of the output from the Platypus CellPhoneDB_analyse function. C. Dot plot depicting the interaction levels of the gene pairs reported (y-axis) in the cluster pairs reported (x-axis). Color of the dot indicates the logarithm of the ligand-receptor interaction mean value across cells of the cluster pair. Size of the dot indicates the negative logarithm of the p-value, highlighting which interactions are to be considered significant. The dot plot is part of the output of CellPhoneDB_analyse. A customized version of the plot can be generated using the dot_plot function by selecting the gene and cluster pairs to be displayed, in addition to a mean or p-value threshold.

A



B

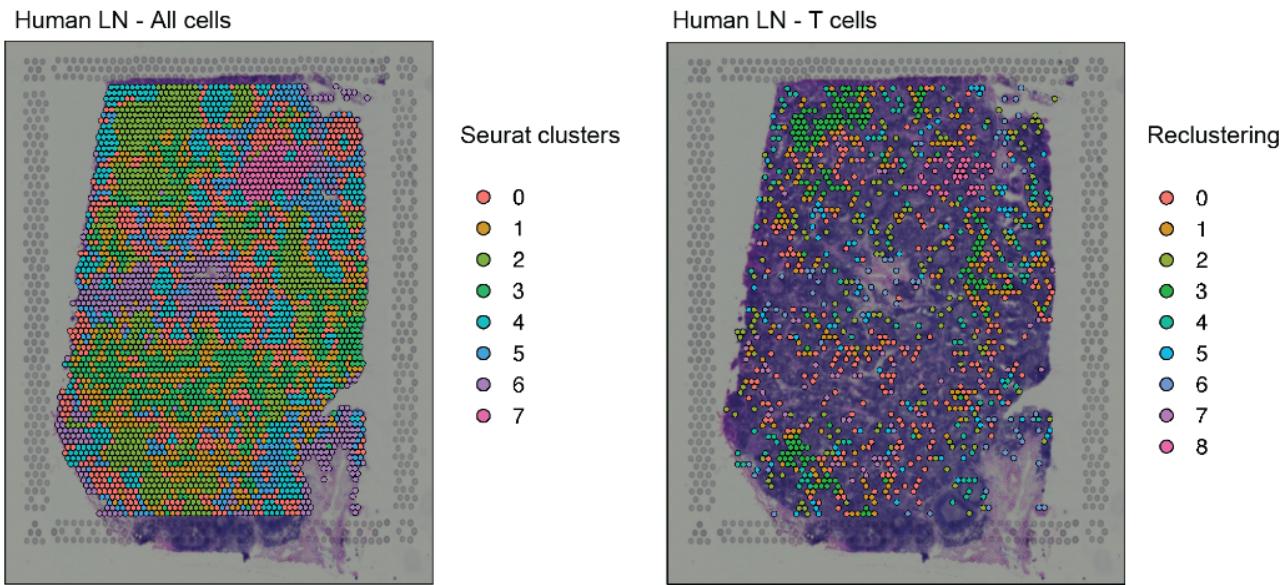


Figure S7. Spatial transcriptomics analysis pipeline. A. Graphical overview for the generation and analysis of spatial transcriptomics experiments. B. Unsupervised transcriptional clustering of cells from a human lymph node (LN) sample. The colors indicate the cluster to which the cell belongs for either all cells (right) or only T cells (right).

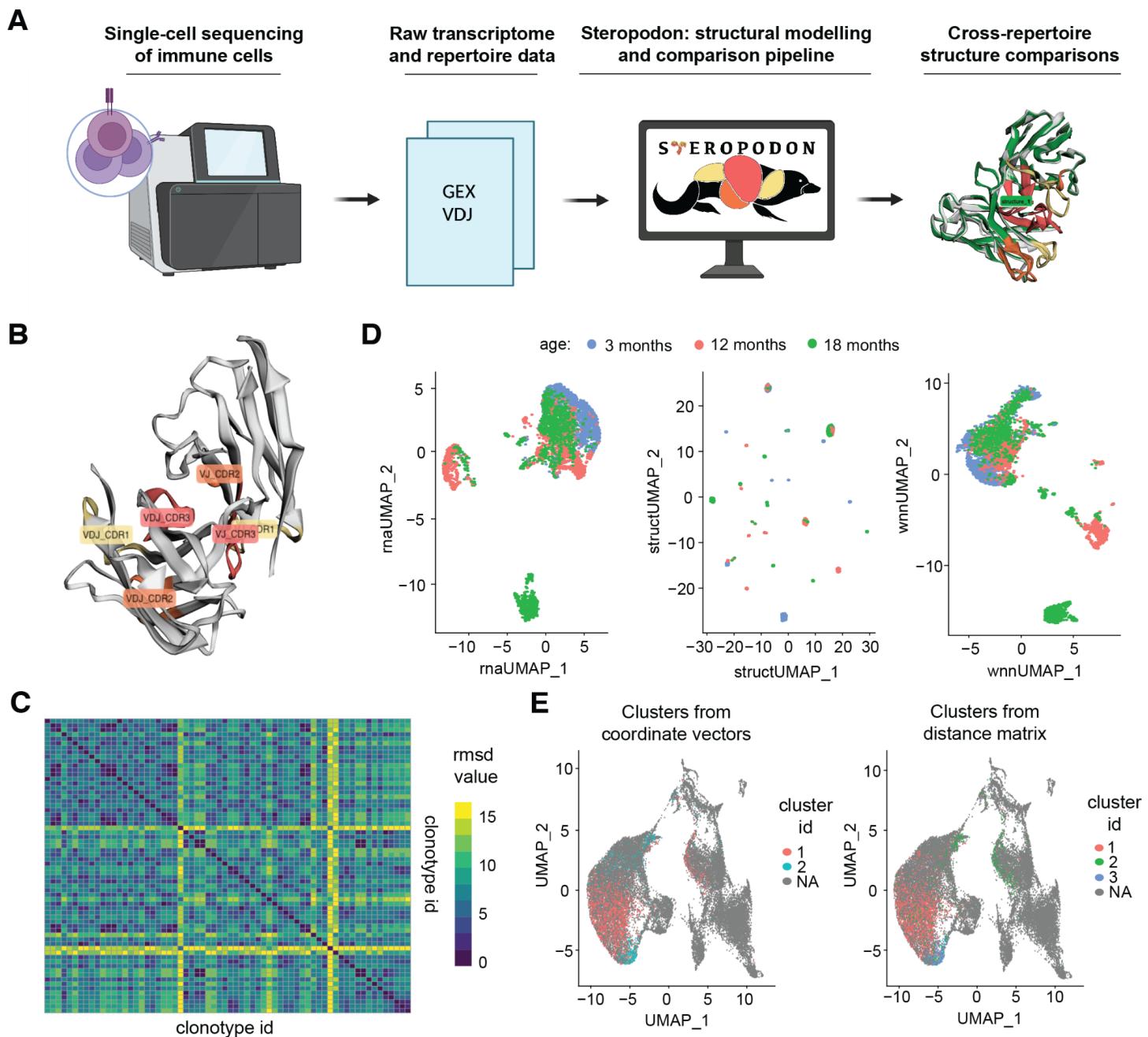
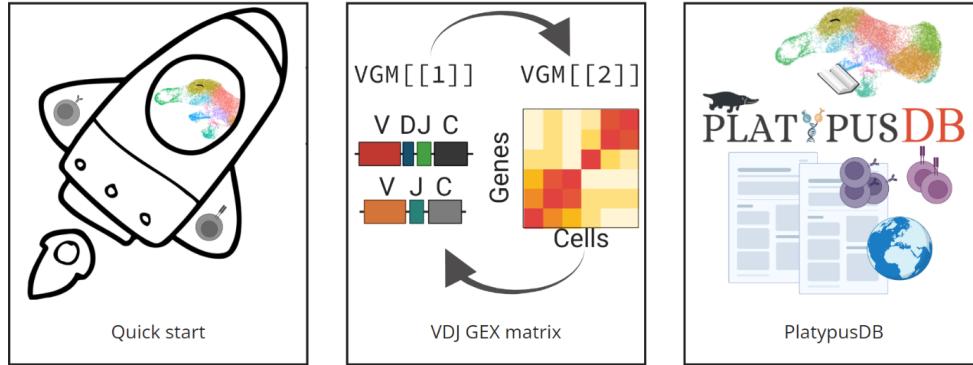


Figure S8. The Steropodon pipeline for immune receptor structural modeling and integrated, cross-repertoire analysis. A. The Steropodon workflow for obtaining receptor structures, starting from the Platypus VGM object. B. A sample structure from the TNFR2 dataset obtained via Steropodon_model and visualized using Steropodon_visualize, with the CDR regions labelled. C. A heatmap of the root mean square deviation (RMSD) between structures from the 10 most frequent clonotypes in 6 samples from the TNFR2 dataset, showcasing the overall structural similarities at the intra and cross-repertoire levels. D. UMAP projection of the gene expression vectors for each cell corresponding to the structures modeled from the 10 most expanded clonotypes from the TNFR2

dataset (left), UMAP projection of the coordinate vectors of the modeled structures after structural alignment (middle), and UMAP projection of the multimodal embeddings for the 2 modalities from before, integrated using Seurat's weighted nearest neighbors algorithm. E. UMAP projection of the GEX values of all cells in the 6 samples selected from the TNFR2 dataset, with cells highlighted by the structural clusters obtained on the coordinate feature vectors from D and the distance matrix from C , respectively, following hierarchical clustering.

Platypus and PlatypusDB vignettes

Here we provide a comprehensive overview of the functions of the Platypus package in form of concise vignettes by topic. If you are new to Platypus we suggest to go through in order. As Platypus included many wrapper functions employing Seurat for GEX analysis, please also refer to the corresponding [documentation](#).



Links

- [View on CRAN](#)
- [Browse source code](#)
- [Report a bug at](#)
- [Cite Platypus](#)
- [Cite Platypus DB](#)
- [License](#)
- [GPL-2](#)
- [Citation](#)
- [Citing Platypus](#)
- [Developers](#)
- Alexander Yermanos
Author, maintainer
- and the [Platypus Team](#)
- [More about authors...](#)

Figure S9. Example of three vignettes available within the ePlatypus Computational Immunology Ecosystem. All other available vignettes presented on the website are detailed in Table S1.

3. Creating phylogenetic trees via VDJ_phylogenetic_trees

1. Choosing the sequence type

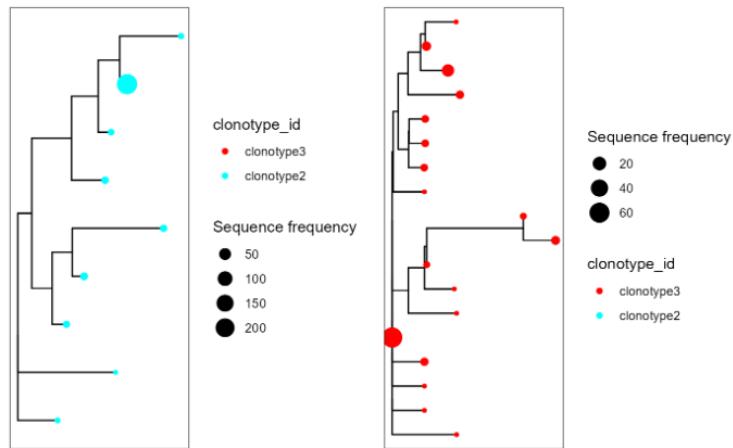
The sequence type can be controlled via the `sequence.type`, `as.nucleotide`, and `trimmed` parameters in `VDJ_phylogenetic_trees`. For `sequence.type`, the following options are available: 'VDJ' for the VDJ full sequence, 'VJ' for the full VJ sequence, 'VDJ.VJ' for the combined VDJ/VJ sequences, 'cdr3' for the CDR3 region, 'cdrh3' for the CDRH3 region. To infer trees for protein sequences, use `as.nucleotide = F`. If `trimmed` is set to `T`, `VDJ_phylogenetic_trees` will look for the trimmed full sequences - ensure `trim.and.align` was set to `T` when obtaining your VGM using the `VDJ_GEX_matrix` function; else, trees will be inferred for the raw sequences.

We will next infer trees for the CDR3 sequences, with `as.nucleotide` set to `T`. We will only create 3 trees for a single sample (`maximum.lineages = 3` which will subset the VDJ for the 3 most abundant clonotypes before creating the trees) for easier visualization. However, the usual output of `VDJ_phylogenetic_trees` is a nested list of `tidytree` objects (per sample, per clonotype), whereas the output of `VDJ_phylogenetic_trees_plot` is a list of intraclonal trees for each clonotype/lineage obtained from `VDJ_phylogenetic_trees`.

We will also set the `maximum.sequences` parameter to 20 to only obtain trees with a maximum of 20 sequences (the top 20 most frequent sequences per clonotype) and `minimum.sequences` to 1 (avoiding singletons). Moreover, we will set `include.germline` to `F` (otherwise `VDJ_phylogenetic_trees` will look for germlines/10x references in the VDJ and `VJ_trimmed_ref` columns).

```
s1_subset <- VDJ[VDJ$sample_id == 's1',]
p <- s1_subset %>% VDJ_phylogenetic_trees(sequence.type = 'cdr3',
                                             as.nucleotide = T,
                                             trimmed = F,
                                             minimum.sequences = 3,
                                             maximum.sequences = 20,
                                             maximum.lineages = 3,
                                             include.germline = F) %>%
  VDJ_phylogenetic_trees_plot()

cowplot::plot_grid(plotlist = p)
```



We can see how the first clonotype was filtered out, as it had fewer sequences than the minimum in `minimum.sequences`.

Figure S10. Example of one walk-through present in the ePlatypus Computational Immunology Ecosystem.

Contents

- 1 Introduction
- 2. Loading the VGM
- 3. Creating phylogenetic trees via VDJ_phylogenetic_trees
 - 1. Choosing the sequence type
 - 2. Phylogenetic tree algorithms
 - 3. Per-repertoire/global trees
- 4. Custom tree plots via VDJ_phylogenetic_trees_plot
- 5. Version information

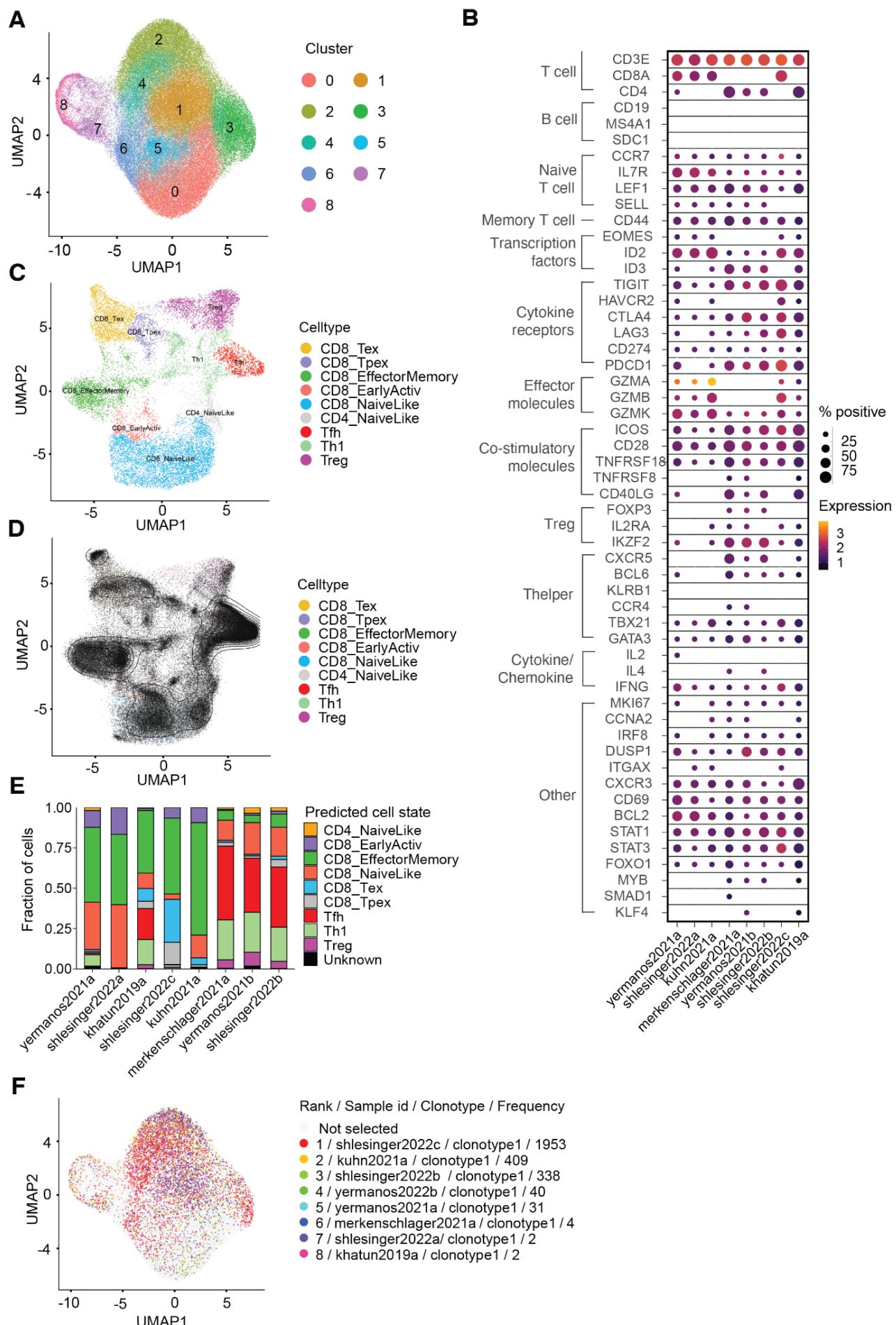


Figure S11. Repertoire and phenotypic breadth of PlatypusDB. A. Uniform manifold approximation projection of murine T cells from various experimental conditions. Each point represents an individual cell and color corresponds to the Seurat assigned cluster. B. Dotplot showing T cell subset assignment across all datasets based on expression of specific gene lists. C. Projection of reference dataset from projectTILS coloured by T cell phenotype. D. Projection of the complete T cell repertoire (black) over reference dataset from projectTILS (Figure S8C). E. Distribution of T-cell types across experiments from projectTILS. F. Uniform manifold approximation projection of murine T cells from various experimental conditions where each point represents an individual cell. Cells belonging to the most expanded clones overall are highlighted. Clones were defined as containing identical CDRb3+CDRa3 sequences.

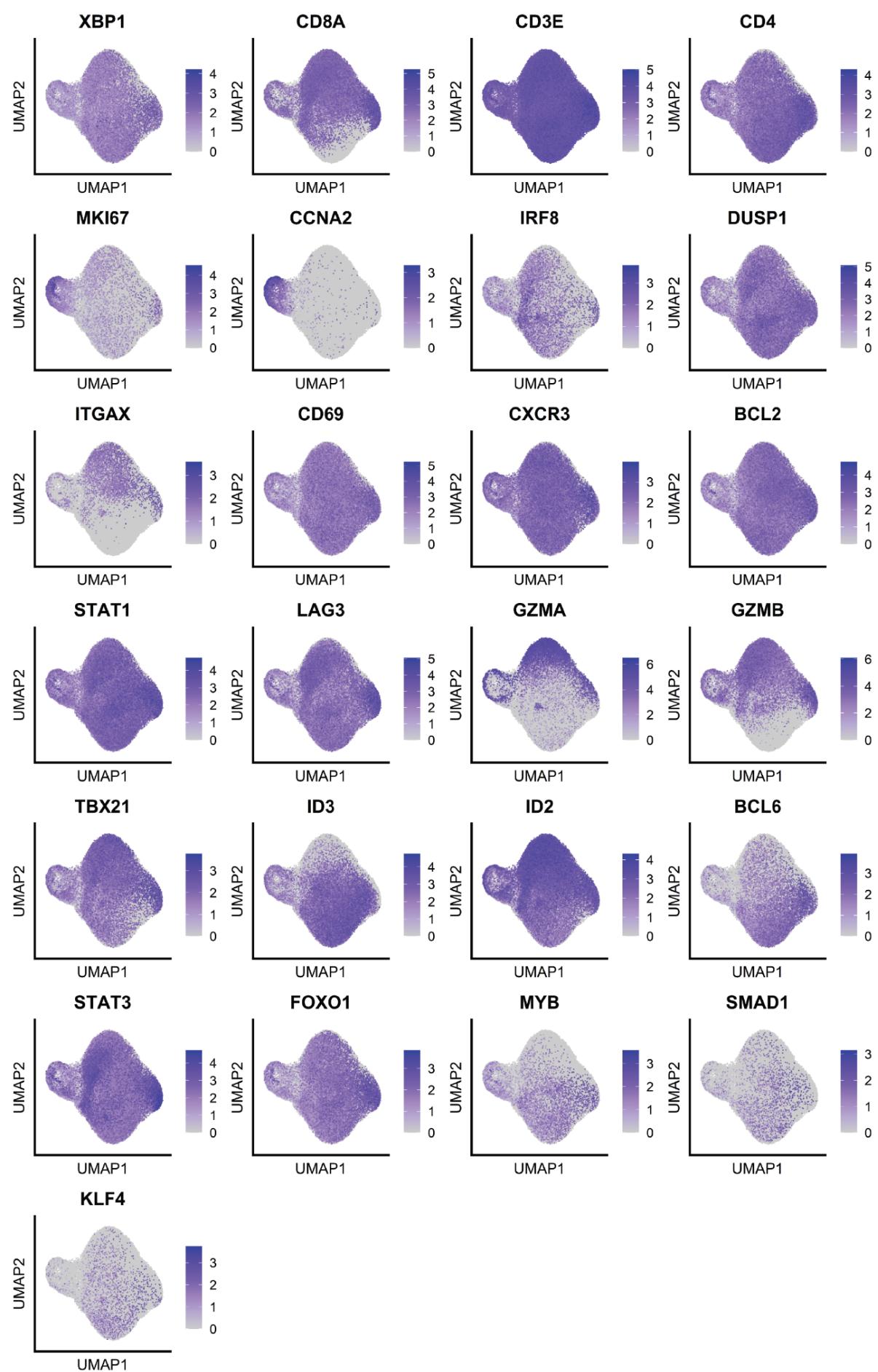


Figure S12. Uniform manifold approximation projection (UMAP) plots showing gene expression for selected T cell associated genes.

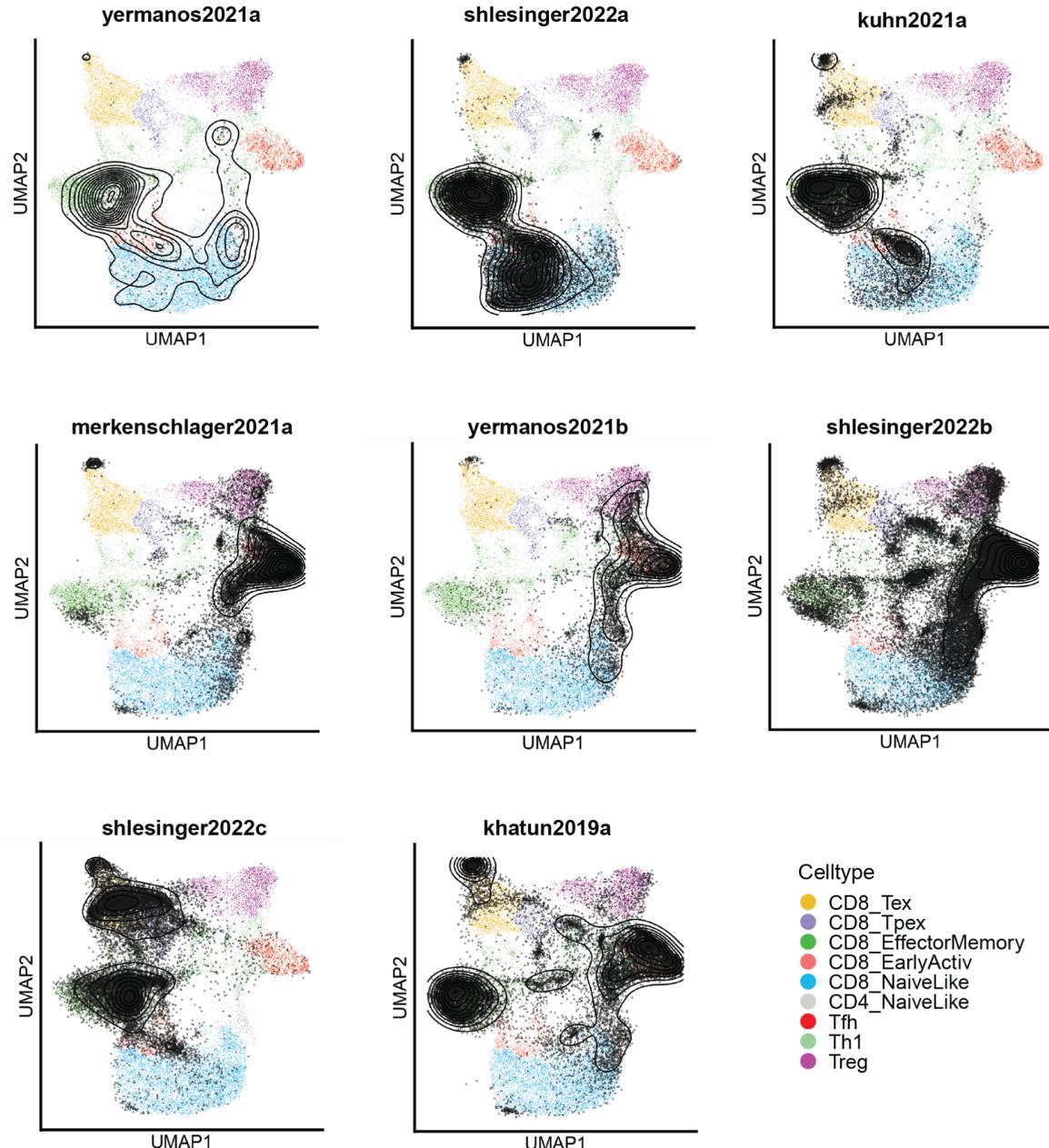


Figure S13. Uniform manifold approximation projection of each individual experiment over the projectTILS reference dataset.

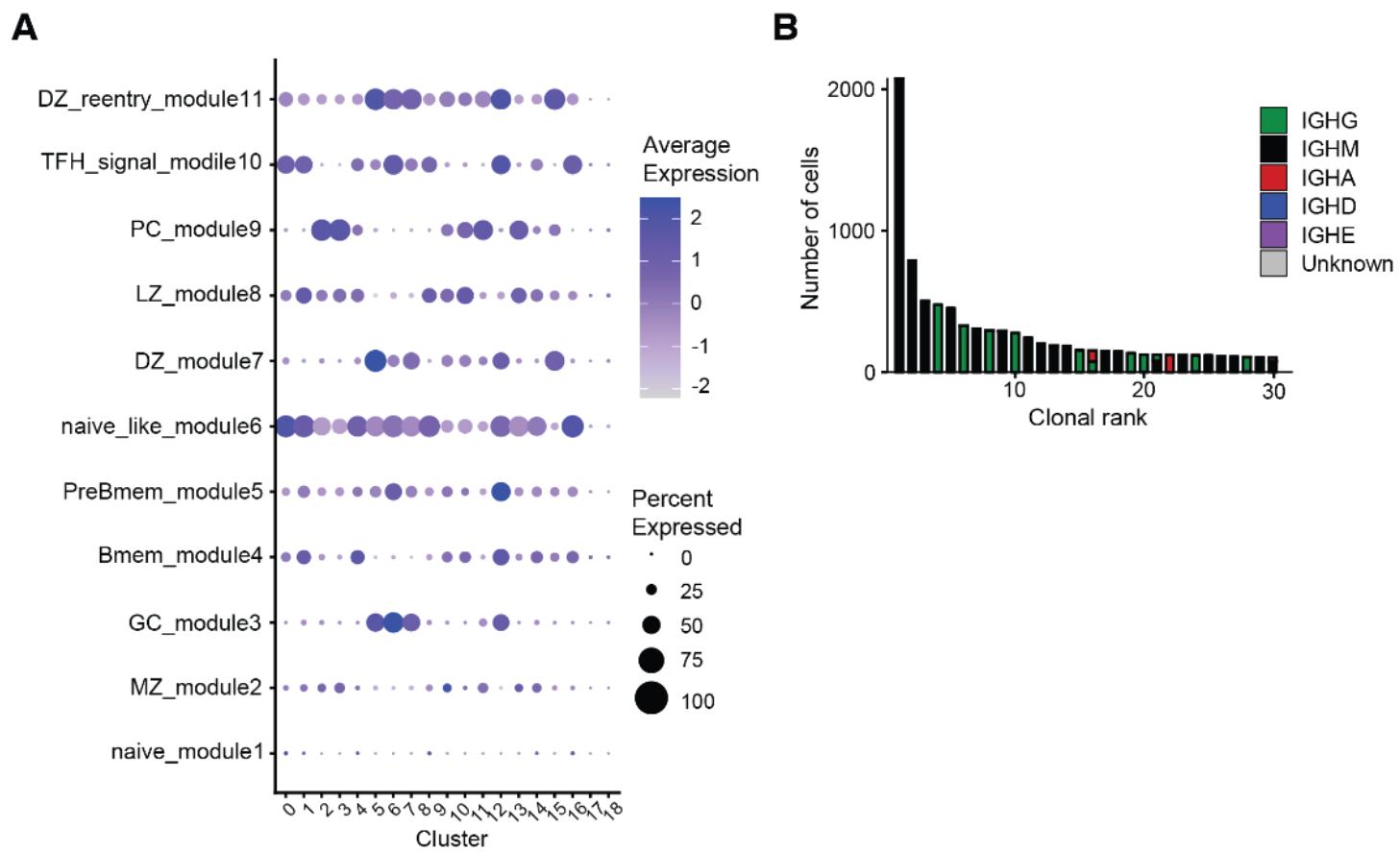


Figure S14. PlatypusDB contains a diverse collection of B cell phenotypes. A. Dotplot showing B cell subset assignment across all clusters based on expression of specific gene lists (Mathew et al. 2021). B. Analysis of the top 30 most highly expanded clones across all experimental conditions. Clonotyping was performed based on those B cells containing identical HCDR3 + LCDR3 amino acid sequences. The clones were colored by isotype expression.



Figure S15. Uniform manifold approximation projection (UMAP) plots showing gene expression for selected B cell genes.

Vignette	References of implemented tools	Content overview
Quickstart		Overview of Platypus functions starting from import modes, central processing via VDJ_GEX_matrix and basic GEX, VDJ and integrated VDJ-GEX analysis
VDJ-GEX matrix		Deep dive into the standard Platypus VDJ GEX matrix format (VGM) and advanced parameters during preprocessing
PlatypusDB		Tutorial for downloading, processing and merging datasets from PlatypusDB as well as local datasets.
Clonotyping		Deep dive into clonotyping strategies available in Platypus and options for dealing cells with less or more than 1 VDJ 1 VJ chains
VDJ-GEX integration		Overview of functions to co-analyze VDJ and GEX information and methods to transfer custom annotations
AIRR compatibility	TRUST4 (Song et al., 2021) AIRR community standards (Vander Heiden et al., 2018)	Single-cell immune repertoire and transcriptome of FACS sorted Tfh cells in acute and chronic LCMV infection
Echidna	Echidna (Han et al., 2022)	Repertoire and GEX data simulations
Antibody Forests		Full guide on network inference and analysis of immune receptor repertoires
Phylogenetic trees		Flexible pipeline for the generation and visualisation of phylogenetic trees of BCR and antibody lineages
Bulk repertoires		Guide to analysing MiXCR- or MAF-generated bulk repertoire datasets within the Platypus framework
Pseudo-bulk		Tutorial for Platypus functions that allow GEX pseudo-bulking and flexible differential gene analysis.
TCR lookup and specificity	VDJdb (Shugay et al., 2018) McPAS-TCR (Tickotsky et al., 2017) PIRD (Zhang et al., 2020)	Framework to annotate TCRs using public specificity databases (VDJdb, McPAS-TCR and PIRD TBAdb) for epitope specificity
Diversity and sampling		Functions for repertoire diversity and comparative metrics
Sequence Kmers		Guide to kmer and motif analysis across repertoires
VDJ dynamics		Functions to analyse and visualise longitudinal repertoire dynamics and tracking of clones
CellphoneDB	CellphoneDB (Efremova et al., 2020)	Tutorial to run the CellphoneDB receptor-ligand interaction analysis pipeline within the Platypus framework
ProjecTILs and pseudotime	ProjecTILs (Andreatta et al., 2021) Monocle3 (Trapnell et al., 2014) Velocyto (La Manno et al., 2018)	Guide to using and visualising projection and pseudotime algorithms and pipelines within the Platypus framework
Machine Learning		Tutorial for extraction of features, encoding and predicting repertoire features within the VDJ GEX matrix object. Features may include receptor specificity and others.
Structural modelling	AlphaFold (Jumper et al. 2021)	Pipeline for structural modelling of immune receptors via AlphaFold on computational clusters.
Spatial transcriptomics		Guide to analysis and visualisation of 10x Genomics-generated spatial transcriptomics data within the Platypus framework

Table S1. Overview of current tutorials present within the ePlatypus Computational Immunology Ecosystem.

PlatypusDB ID	Citation	DOI	Dataset or publication title
khatun2019a	Khatun et al., 2021	10.1084/jem.20200650	Single-cell lineage mapping of a diverse virus-specific naive CD4 T cell repertoire
kuhn2021a	Kuhn et al., 2022	10.3389/fimmu.2022.782441	Clonally expanded virus specific CD8 T cells acquire diverse transcriptional phenotypes during acute, chronic and latent infections
Merkenschlager 2021a	Merkenschlager et al., 2021	10.1038/s41586-021-03187-x	Dynamic regulation of Tfh selection during germinal centre reaction
mathew2021a	Mathew et al. 2021	10.1016/j.celrep.2021.109286	Temporal dynamics of persistent germinal centers and memory B cell differentiation following respiratory virus infection
shlesinger2022a	Shlesinger et al., 2022	10.1101/2022.02.07.479381	Single-cell immune repertoire and transcriptome of GP33+ Tetramer sorted CD8 T cells from single mouse infected with acute LCMV or chronic MCMV infection
shlesinger2022b	Shlesinger et al., 2022	10.1101/2022.02.07.479381	Single-cell immune repertoire and transcriptome of FACS sorted Tfh cells in acute and chronic LCMV infection
shlesinger2022c	Shlesinger et al., 2022	10.1101/2022.02.07.479381	Single-cell immune repertoire and transcriptome of NP396-tetramer sorted CD8 T cells in a rechallenge déjà vu model
yermanos2021a	Yermanos et al., 2021	10.1098/rspb.2020.2793	Single-cell immune repertoire and transcriptome sequencing reveals that clonally expanded and transcriptionally distinct lymphocytes populate the aged central nervous system in mice
yermanos2022b	NA	NA	Single-cell immune repertoire and transcriptome sequencing of Tfh cells in Influenza infection
kreiner2021a	Shlesinger et al., 2022	10.1101/2022.02.07.479381	Characterization of immune repertoires and phenotypes of T cells in experimental autoimmune encephalomyelitis by single cell sequencing
kreiner2021b	Shlesinger et al., 2022	10.1101/2022.02.07.479381	Characterization of immune repertoires and phenotypes of B cells in experimental autoimmune encephalomyelitis by single cell sequencing
yewdell2021a	Yewdell et al., 2021	10.1016/j.celrep.2021.109961	Temporal dynamics of persistent germinal centers and memory B cell differentiation following respiratory virus infection
neumeier2021a	Neumeier et al., 2021	10.1002/eji.202149331	Single-cell sequencing reveals clonally expanded plasma cells during chronic viral infection produce virus-specific and cross-reactive antibodies
neumeier2021b	Neumeier et al., 2021	10.1002/eji.202149331	Single-cell sequencing reveals clonally expanded plasma cells during chronic viral infection produce virus-specific and cross-reactive antibodies
agrafiotis2021a	Agrafiotis et al., 2021	10.1101/2021.11.09.467876	B cell clonal expansion is correlated with antigen-specificity in young but not old mice

Table S2. List of datasets used for the examples and analysis in this manuscript. The PlatypusDB ID is used in code to download datasets. In some cases, more than one dataset was published within the same publication. In that case the title corresponds to the dataset, while the citation and doi correspond to the overarching publication.

Online Methods

PlatypusDB architecture: The computational infrastructure for PlatypusDB was developed based on the analysis package Platypus and the Google Cloud storage API. The process of uploading a dataset includes the following steps: First, raw sequencing files are sourced locally, downloaded from public repositories such as GEO, or acquired directly from another research group. Raw reads were then aligned with Cellranger 6.0.1 to the following 10x Genomics reference genomes: refdata-gex-mm10-2020-A, refdata-cellranger-vdj-GRCm38-alts-ensembl-5.0.0, refdata-gex-GRCh38-2020-A, and refdata-cellranger-vdj-GRCh38-alts-ensembl-5.0.0. Raw output files were then uploaded as compressed directories to the PlatypusDB Google Cloud storage database. Raw outputs were then loaded and processed in R resulting in two formats. Firstly a per-sample list object containing main Cellranger output tables and secondly a VDJ-GEX-matrix object from Platypus v3.2.2. This object was generated using the VDJ_GEX_matrix function with default settings, if not otherwise noted. All output objects were uploaded to PlatypusDB using the package googleCloudStorageR. To allow for easy access to the database, download of R objects as well as compressed directories is available directly via URL without the need to install Google Cloud storage compatibility packages for R. The URL for a database lookup table is delivered with Platypus and allows for a single access point to the database, which remains constant as more datasets will be added in the future.

Data analysis: The filtered feature matrix directory was supplied as input to the VDJ_GEX_matrix function in the R package Platypus (v3.3) (Yermanos, Agrafiotis, et al. 2021), which uses the transcriptome analysis workflow of the R package Seurat (Satija et al. 2015). Only those cells containing less than 20% of mitochondrial reads were retained in the analysis. Genes involved in the adaptive immune receptor (e.g., TRB, TRBV1-1), were removed from the count matrix to prevent clonal relationships from influencing transcriptional phenotypes. Gene expression was normalized using the “harmony” argument in the VDJ_GEX_matrix function. 2000 variable features were selected using the “vst” selection method and used as input to principal component analysis (PCA) using the first 10 dimensions. Graph-based clustering using the Louvain modularity optimization and hierarchical clustering was performed using the functions FindNeighbors and FindClusters in Seurat using the first ten dimensions and a cluster resolution of 0.5. UMAP was similarly inferred using the

first ten dimensions. The FindMarkers function from Seurat was used when calculating differentially expressed genes (both across groups or across clusters) with logfc.threshold set to 0 and minimum number of cells expressing each gene set to 0.25 and subsequently supplied to the GEX_volcano function from Platypus. Mitochondrial and ribosomal genes were removed when visualizing DE genes. Feature plots were produced by supplying genes of interest to the function FeaturePlot in Seurat. Module scores for public gene sets (Mathew et al. 2021) were calculated using the AddModuleScore from Seurat. Cells containing no or more than one α /heavy and β /light chain were filtered out for TCR/BCR repertoire analysis. Clones were defined by identical CDR3 α /CDRH3 and CDR3 β /CDRL3 sequence (nucleotide or amino acid sequence) across all repertoires. Clones represented by more than one cell were considered highly-expanded clones, while single-celled clones were defined as lowly-expanded. The projection of cells onto reference UMAPs and cell state predictions were done using the R package ProjecTILs (Andreatta et al. 2021) under default conditions. Experiments were either individually or all together projected onto the ProjecTILs atlas. For Figures S1 to S8, single-cell immune repertoire sequencing experiments present in PlatypusDB were formatted into a single VDJ_GEX_matrix object that was then supplied to downstream analyses pipelines. Specifically, the pseudobulk analysis was performed using the pseudo_bulk_DE function from Platypus. Sampling and diversity analyses were performed using the VDJ_diversity and VDJ_rarefaction. The sequence similarity network of clusters was generated using the AntibodyForests_communities function and then colored using the AntibodyForests_plot. Node and edge metrics were calculated using the AntibodyForests_metrics function from Platypus. The PlatypusML_classification function from Platypus, which takes input the encoded features obtained from the PlatypusML_extract_features function, was used to run cross validation on a specified number of folds for different classification models (XGBoost, SVM, Random Forest, Logistic Regression & Gaussian Naive Bayes), outputting the AUC scores, ROC curve and confusion matrix for each classification model. Prediction of receptor-ligand interaction was calculated on the single-cell gene expression data using the CellPhoneDB software. The heatmap and dot plot were generated as an output of the CellPhoneDB_analyse function from Platypus. Receptor structures in the Steropodon workflow were obtained using the Steropodon_model function and visualized using the Steropodon_visualize function in Platypus.

Data visualization: Figure 1 and the supplementary graphical overviews were created with Biorender.com. Feature plots were produced using “FeaturePlot” (Seurat 4.0). Volcano plots were produced using “GEX_volcano” (Platypus v3.3). Dottile plots were produced using DotPlot (Seurat 4.0). All other figures were produced using Prism v9 (Graphpad).

Data availability: The accession numbers and publications for the sequencing data used in this manuscript are located in table S1. Platypus code used in this manuscript can be found at github.com/alexyermanos/Platypus.

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.

Funding: This work was supported by the ETH Zurich Research Grants (to STR and AO), an ETH Seed Grant (AY) and support from the "la Caixa" Foundation (ID 100010434, fellowship code: LCF/BQ/EU20/11810041) to MMC.

Competing Interests: There are no competing interests.

References

Agrafiotis, A., Neumeier, D., Hong, K.-L., Chowdhury, T., Ehling, R., Kuhn, R., Sandu, I., Kreiner, V., Starkie, D., Lightwood, D. J., Oxenius, A., Reddy, S. T., & Yermanos, A. (2021). B cell clonal expansion is correlated with antigen-specificity in young but not old mice. *BioRxiv*, 2021.11.09.467876. <https://doi.org/10.1101/2021.11.09.467876>

Andreatta, M., Corria-Osorio, J., Müller, S., Cubas, R., Coukos, G., & Carmona, S. J. (2021). Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nature Communications*, 12(1), 2965. <https://doi.org/10.1038/s41467-021-23324-4>

Borcherding, N., Bormann, N. L., & Kraus, G. (2020). scRepertoire: An R-based toolkit for single-cell immune receptor analysis. *F1000Research*, 9. <https://doi.org/10.12688/F1000RESEARCH.22139.2>

Efremova, Mirjana, Miquel Vento-Tormo, Sarah A. Teichmann, and Roser Vento-Tormo. 2020. “CellPhoneDB: Inferring Cell-Cell Communication from Combined Expression of Multi-Subunit Ligand-Receptor Complexes.” *Nature Protocols* 15 (4): 1484–1506.

Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89.

Horns, F., Dekker, C. L., & Quake, S. R. (2020). Memory B Cell Activation, Broad Anti-influenza Antibodies, and Bystander Activation Revealed by Single-Cell Transcriptomics. *Cell Reports*, 30(3), 905–913.e6. <https://doi.org/10.1016/J.CELREP.2019.12.063>

Khatun, A., Kasmani, M. Y., Zander, R., Schauder, D. M., Snook, J. P., Shen, J., Wu, X., Burns, R., Chen, Y. G., Lin, C. W., Williams, M. A., & Cui, W. (2021). Single-cell lineage mapping of a diverse virus-specific naive CD4 T cell repertoire. *The Journal of Experimental Medicine*, 218(3). <https://doi.org/10.1084/JEM.20200650>

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. ru, & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12), 1289–1296. <https://doi.org/10.1038/S41592-019-0619-0>

Kreutzfeldt, M., Bergthaler, A., Fernandez, M., Brück, W., Steinbach, K., Vorm, M., Coras, R., Blümcke, I., Bonilla, W. v., Fleige, A., Forman, R., Müller, W., Becher, B., Misgeld, T., Kerschensteiner, M., Pinschewer, D. D., & Merkler, D. (2013). Neuroprotective intervention by interferon- γ blockade prevents CD8+ T cell–mediated dendrite and synapse loss. *The Journal of Experimental Medicine*, 210(10), 2087. <https://doi.org/10.1084/JEM.20122143>

Kuhn, R., Sandu, I., Agrafiotis, A., Hong, K. L., Shlesinger, D., Neimeier, D., Merkler, D., Oxenius, A., Reddy, S. T., & Yermanos, A. (2022). Clonally Expanded Virus-Specific CD8 T Cells Acquire Diverse Transcriptional Phenotypes During Acute, Chronic, and Latent Infections. *Frontiers in Immunology*, 13, 210. <https://doi.org/10.3389/FIMMU.2022.782441>

Mathew, N. R., Jayanthan, J. K., Smirnov, I. v., Robinson, J. L., Axelsson, H., Nakka, S. S., Emmanouilidi, A., Czarnewski, P., Yewdell, W. T., Schön, K., Lebrero-Fernández, C., Bernasconi, V., Rodin, W., Harandi, A. M., Lycke, N., Borcherding, N., Yewdell, J. W., Greiff, V., Bemark, M., & Angeletti, D. (2021a). Single-cell BCR and transcriptome analysis after influenza infection reveals spatiotemporal dynamics of antigen-specific B cells. *Cell Reports*, 35(12), 109286. <https://doi.org/10.1016/J.CELREP.2021.109286>

Mathew, N. R., Jayanthan, J. K., Smirnov, I. v., Robinson, J. L., Axelsson, H., Nakka, S. S., Emmanouilidi, A., Czarnewski, P., Yewdell, W. T., Schön, K., Lebrero-Fernández, C., Bernasconi, V., Rodin, W., Harandi, A. M., Lycke, N., Borcherding, N., Yewdell, J. W., Greiff, V., Bemark, M., & Angeletti, D. (2021b). Single-cell BCR and transcriptome analysis after influenza infection reveals spatiotemporal dynamics of antigen-specific B cells. *Cell Reports*, 35(12). <https://doi.org/10.1016/J.CELREP.2021.109286>

Merkenschlager, J., Finkin, S., Ramos, V., Kraft, J., Cipolla, M., Nowosad, C. R., Hartweger, H., Zhang, W., Olinares, P. D. B., Gazumyan, A., Oliveira, T. Y., Chait, B. T., & Nussenzweig, M. C. (2021). Dynamic regulation of TFH selection during the germinal centre reaction. *Nature* 2021 591:7850, 591(7850), 458–463. <https://doi.org/10.1038/s41586-021-03187-x>

Miho, E., Yermanos, A., Weber, C. R., Berger, C. T., Reddy, S. T., & Greiff, V. (2018). Computational strategies for dissecting the high-dimensional complexity of adaptive immune repertoires. *Frontiers in Immunology*, 9(FEB), 224. <https://doi.org/10.3389/FIMMU.2018.00224>

Neumeier, D., Pedrioli, A., Genovese, A., Sandu, I., Ehling, R., Hong, K. L., Papadopoulou, C., Agrafiotis, A., Kuhn, R., Shlesinger, D., Robbiani, D., Han, J., Hauri, L., Csepregi, L., Greiff, V., Merkler, D., Reddy, S. T., Oxenius, A., & Yermanos, A. (2022). Profiling the specificity of clonally expanded plasma cells during chronic viral infection by single-cell analysis. *European Journal of Immunology*, 52(2), 297–311. <https://doi.org/10.1002/EJI.202149331>

Neumeier, D., Yermanos, A., Agrafiotis, A., Csepregi, L., Chowdhury, T., Ehling, R. A., Kuhn, R., Brisset-Di Roberto, R., di Tacchio, M., Antoniali, R., Starkie, D., Lightwood, D. J., Oxenius, A., & Reddy, S. T. (n.d.). *Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells*. <https://doi.org/10.1101/2021.07.16.452687>

Neumeier, D., Yermanos, A., Agrafiotis, A., Csepregi, L., Chowdhury, T., Ehling, R. A., Kuhn, R., Roberto, R. B.-D., Tacchio, M. di, Antoniali, R., Starkie, D., Lightwood, D. J., Oxenius, A., & Reddy, S. T. (2021). Phenotypic determinism and stochasticity in antibody repertoires of clonally expanded plasma cells. *BioRxiv*, 2021.07.16.452687. <https://doi.org/10.1101/2021.07.16.452687>

Ramesh, A., Schubert, R. D., Greenfield, A. L., Dandekar, R., Loudermilk, R., Sabatino, J. J., Koelzer, M. T., Tran, E. B., Koshal, K., Kim, K., Pröbstel, A. K., Banerji, D., Guo, C. Y., Green, A. J., Bove, R. M., DeRisi, J. L., Gelfand, J. M., Cree, B. A. C., Zamvil, S. S., ... Wilson, M. R. (2020). A pathogenic and clonally expanded B cell transcriptome in active multiple sclerosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117(37), 22932–22943. <https://doi.org/10.1073/pnas.2008523117>

Ruffolo, Jeffrey A., Lee-Shin Chu, Sai Pooja Mahajan, and Jeffrey J. Gray. 2022. “Fast, Accurate Antibody Structure Prediction from Deep Learning on Massive Set of Natural Antibodies.” *bioRxiv*. <https://doi.org/10.1101/2022.04.20.488972>.

Rubelt, F., Busse, C. E., Bukhari, S. A. C., Bürckert, J. P., Mariotti-Ferrandiz, E., Cowell, L. G., Watson, C. T., Marthandan, N., Faison, W. J., Hershberg, U., Laserson, U., Corrie, B. D., Davis, M. M., Peters, B., Lefranc, M. P., Scott, J. K., Breden, F., Luning Prak, E. T., & Kleinstein, S. H. (2017). Adaptive Immune Receptor Repertoire Community recommendations for sharing

immune-repertoire sequencing data. *Nature Immunology* 2017 18:12, 18(12), 1274–1278. <https://doi.org/10.1038/ni.3873>

Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. “Spatial Reconstruction of Single-Cell Gene Expression Data.” *Nature Biotechnology* 33 (5): 495–502. <https://doi.org/10.1038/nbt.3192>

Shlesinger, D., Hong, K.-L., Shammas, G., Page, N., Sandu, I., Agrafiotis, A., Kreiner, V., Fonta, N., Vincenti, I., Wagner, I., Piccinno, M., Mariotte, A., Klimek, B., Dizerens, R., Manero-Carranza, M., Kuhn, R., Ehling, R., Frei, L., Khodaverdi, K., ... Yermanos, A. (2022). Single-cell immune repertoire sequencing of B and T cells in murine models of infection and autoimmunity. *BioRxiv*, 2022.02.07.479381. <https://doi.org/10.1101/2022.02.07.479381>

Wherry, E. J., & Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nature Reviews. Immunology*, 15(8), 486. <https://doi.org/10.1038/NRI3862>

Yaari, G., & Kleinstein, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine* 2015 7:1, 7(1), 1–14. <https://doi.org/10.1186/S13073-015-0243-2>

Yermanos, A. (2022). Single-cell immune repertoire and transcriptome sequencing of Tfh cells after Influenza infection. *Unpublished Dataset*.

Yermanos, A., Agrafiotis, A., Kuhn, R., Robbiani, D., Yates, J., Papadopoulou, C., Han, J., Sandu, I., Weber, C., Bieberich, F., Vazquez-Lombardi, R., Dounas, A., Neumeier, D., Oxenius, A., & Reddy, S. T. (2021). Platypus: an open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. *NAR Genomics and Bioinformatics*, 3(2). <https://doi.org/10.1093/NARGAB/LQAB023>

Yermanos, A., Neumeier, D., Sandu, I., Borsa, M., Waindok, A. C., Merkler, D., Oxenius, A., & Reddy, S. T. (2021). 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 B*, 288(1945). <https://doi.org/10.1098/RSPB.2020.2793>

Yewdell, W. T., Smolkin, R. M., Belcheva, K. T., Mendoza, A., Michaels, A. J., Cols, M., Angeletti, D., Yewdell, J. W., & Chaudhuri, J. (2021). Temporal dynamics of persistent germinal centers and memory B cell differentiation following respiratory virus infection. *Cell Reports*, 37(6). <https://doi.org/10.1016/J.CELREP.2021.109961>

Yost, K. E., Satpathy, A. T., Wells, D. K., Qi, Y., Wang, C., Kageyama, R., McNamara, K. L., Granja, J. M., Sarin, K. Y., Brown, R. A., Gupta, R. K., Curtis, C., Bucktrout, S. L., Davis, M. M., Chang, A. L. S., & Chang, H. Y. (2019). Clonal replacement of tumor-specific T cells following PD-1

blockade. *Nature Medicine* 2019 25:8, 25(8), 1251–1259.

<https://doi.org/10.1038/s41591-019-0522-3>

Zhao, Y., Kilian, C., Turner, J. E., Bosurgi, L., Roedl, K., Bartsch, P., Gnirck, A. C., Cortesi, F., Schultheiß, C., Hellmig, M., Enk, L. U. B., Hausmann, F., Borchers, A., Wong, M. N., Paust, H. J., Siracusa, F., Scheibel, N., Herrmann, M., Rosati, E., ... Krebs, C. F. (2021). Clonal expansion and activation of tissue-resident memory-like Th17 cells expressing GM-CSF in the lungs of severe COVID-19 patients. *Science Immunology*, 6(16), 6692.

<https://doi.org/10.1126/SCIMMUNOL.ABF6692>