nature methods

Review article

https://doi.org/10.1038/s41592-024-02243-4

Single-cell immune repertoire analysis

Received: 5 December 2023

Accepted: 12 March 2024

Published online: 18 April 2024

Check for updates

Sergio E. Irac[1,6], Megan Sioe Fei Soon[2,6], Nicholas Borcherding[3,4] & Zewen Kelvin Tuong[2,5]

Single-cell T cell and B cell antigen receptor-sequencing data analysis can potentially perform in-depth assessments of adaptive immune cells that inform on understanding immune cell development to tracking clonal expansion in disease and therapy. However, it has been extremely challenging to analyze and interpret T cells and B cells and their adaptive immune receptor repertoires at the single-cell level due to not only the complexity of the data but also the underlying biology. In this Review, we delve into the computational breakthroughs that have transformed the analysis of single-cell T cell and B cell antigen receptor-sequencing data.

The adaptive immune system relies on the immense diversity of the immune repertoire to recognize and respond to a wide range of pathogens and foreign substances. This is mediated by the vast array of surface-bound T cell antigen receptors (TCRs) and B cell antigen receptors (BCRs). The latter can also be subsequently secreted as soluble antibodies when a B cell differentiates into a plasma cell or plasmablast. The wide range of antigen specificities of these receptors also enable T cells and B cells to discern self from non-self-antigens, allowing the immune system to respond appropriately to threats while leaving the host unharmed. By studying the specificity of T cell and B cell responses, immunologists have been able to gain valuable insights into the dynamics of immune responses, immune cell diversity and even provide predictive and/or prognostic information, for both disease and treatment outcomes.

Each TCR or BCR is a dimer composed of two distinct chains. TCRs consist of either an α chain (TRA) paired with a β chain (TRB) in αβ T cells, or a γ chain (TRG) paired with a δ chain (TRD) in γδ T cells. BCRs consist of a heavy chain (IGH) and a light chain; the light chain is from either the κ (IGK) or λ (IGL) locus. These chains are the product of a sophisticated process involving genetic recombination of the variable (V), diversity (D) and joining (J) gene segments, orchestrated by the RAG1 and RAG2 proteins[1], which occurs in developing T cells and B cells in the thymus and bone marrow, respectively. Random non-templated (N) and/or palindromic (P) nucleotide insertions can also be introduced at the junctions of these segments, further adding to the complexity. These recombination events occur at the junction called the complementarity-determining region (CDR) 3, while CDR1 and CDR2 are found entirely within the V gene region. As CDRs are the regions that bind to cognate antigens, they, particularly the CDR3 region, have been the focus for most downstream analyses. Furthermore, after B cells are activated, BCRs can undergo (random) somatic hypermutations (SHMs) throughout the receptor, mediated by the activation-induced cytidine deaminase (AID). AID is also required for class switch recombination (CSR) of BCRs, a biological process that replaces the BCR constant gene encoding the isotype class with wide-ranging consequences on B cell maturation and overall humoral immunity. Collectively, these events ensure the diversity and uniqueness of the resulting TCRs and BCRs, which also allows us to use them as distinct molecular barcodes of T cell and B cell clones. This serves as a useful proxy for tracking antigen-specific responses over time and for correlating with cellular phenotypes and clinical outcome, for example, in identifying clonal patterns associated with improved anticancer treatment response[2,3], relating TCR repertoire diversity with HIV virus adaptation[4] and/or decoding the differences in BCR repertoire to coronavirus disease 2019 (COVID-19) infection compared to vaccination[5].

There are a variety of 'bulk' high-throughput technologies for immune repertoire sequencing, which involves the analysis of pooled BCRs or TCRs from a given tissue. 'Bulk' immune repertoire sequencing technology is largely restricted to analyzing a single chain (for example, TRB or IGH only) and does not capture the dimeric nature of the adaptive immune receptors. The library construction and sequencing strategies also prevent the recovery of bona fide paired-chain sequencing using the 'bulk' approach. However, paired-chain sequencing is now possible, most notably via single-cell technology, allowing the profiling of paired-chained adaptive immune receptors at scale. The variety of high-throughput technologies for sequencing TCRs and BCRs, as well

[1]Cancer Immunoregulation and Immunotherapy, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [2]Ian Frazer Centre for Children's Immunotherapy Research, Child Health Research Centre, Faculty of Medicine, The University of Queensland, Brisbane, Queensland, Australia. [3]Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO, USA. [4]Omniscope, Palo Alto, CA, USA. [5]Frazer Institute, Faculty of Medicine, The University of Queensland, Brisbane, QLD, Australia. [6]These authors contributed equally: Sergio E. Irac, Megan Sioe Fei Soon. e-mail: z.tuong@uq.edu.au

as the analytical tools for bulk repertoire approaches and progress in pairing or predicting antigen receptor recognition, have been reviewed elsewhere[6,7]. This Review will focus on the more recent advances in the bioinformatics analysis methods of single-cell TCR or BCR-sequencing (scTCR/BCR-seq) data.

Briefly, a typical workflow for processing scTCR/BCR-seq data is as follows (Fig. 1): after sequencing single cells, reads are aligned and reconstructed into TCR/BCR chains for each cell. There are also specialized tools that can reconstruct TCR/BCR contigs from full-length single-cell RNA-sequencing (scRNA-seq) data. The reconstructed contigs are then annotated for their respective V, D, J and constant genes using reference databases, for example, international ImMunoGeneTics information system (IMGT)[8]. The 'contig-level' data can then be paired to 'cell-level' data and various tools handle this differently, implementing further quality-control steps to retain analysis-ready contigs and/or cells for further analysis. A key quality-control step useful for most situations is the filtering of contigs so that each mature T cell/B cell only has one productive pair of TCRs/BCRs; there may be some situations where this may not be required (for example, in developing T cells/B cells). Other downstream analyses include differential V/D/J-gene usage analysis, clustering of identical/similar TCR/BCRs into clonotypes, generation of clonotype networks for diversity estimation, inferring lineage trees to trace clonotype phylogeny, integration with scRNA-seq data and more. The complexity of the data presents opportunities for innovative solutions for appropriate downstream analysis.

Initial scTCR/BCR-seq bioinformatics tools

The advent of scRNA-seq technology started in 2009 (ref. 9) and scTCR/BCR-seq quickly followed. The field initially focused on reconstructing the TCR/BCR chains from full-length scRNA-seq data.

TraCeR[10] (T cell receptor reconstruction and clonality inference from scRNA-seq) was one of the first tools introduced to reconstruct TCR sequences from full-length scRNA-seq data with high accuracy and sensitivity. TraCeR first aligns RNA-seq reads from each cell to a curated list including all possible V/J-gene combinations for TCRα/β chains. Contigs are then derived from de novo assembly of the aligned reads and annotated using IgBLAST[11] with IMGT[8] reference sequences. TRAPeS[12] is another scTCR reconstruction software, but their approach has some differences to that of TraCeR. Firstly, it aligns reads to V and J genes, identifying unaligned reads that were mapped to V/J or constant regions and reconstructs the unmapped reads into putative CDR3 sequences. These sequences were extended from both ends until they merge/overlap. TRAPeS showed a high success rate and even outperformed TraCeR[10] on identical library types[12]. However, TRAPeS has its own set of challenges. For example, the tool may struggle in reconstructing some CDR3 sequences due to high similarities in certain V and J segments[12]. There are several other tools such as TRUST4 (ref. 13) and MiXCR[14] that have since emerged that reconstruct TCR/BCR chains from 10x Genomics scRNA/TCR/BCR-seq data. TRUST4 also yielded more TCRs and BCRs and was able to reconstruct αβ-TCR, γδ-TCR and BCR contigs from the 10 × 5′ scRNA-seq gene expression data[13].

Similarly, several tools are available to reconstruct BCRs from full-length scRNA-seq data, including BASIC[15], BraCer[16] and VDJPuzzle[17]. A recent study[18] benchmarked six scBCR reconstruction tools using four experimental datasets and one simulated dataset with differing characteristics, for example, different B cell types, varying library lengths and sequencing depth/coverage. The methods were compared only against datasets generated using low-throughout plate-based scRNA-seq methods. In the simulated dataset, sequences were simulated with different levels of SHMs. The Immcantation framework[19] was used to quantify the number of introduced SHMs as a ground truth, and the tools were then compared on how the occurrence of SHMs may impact their performance. Every tool performed well in their reconstruction with some differences in overall sensitivity and accuracy. Based on their findings, the authors created a recommendation flowchart for tool selection according to the required library type, accuracy, sensitivity and SHM quantification. For example, BraCer[16] was benchmarked to be more suitable for long reads and the most accurate tool when assessing varying SHM frequencies[18]. Among the tools tested, only TRUST4 was capable of processing BCR data generated using high-throughput technologies, for example, 10x Genomics. While MiXCR[14] was not included in the test, it has the same capability.

The popularization of joint assessment of single-cell RNA expression with paired-chain TCR/BCR-sequencing data rose with the commercialization of the 10x Genomics 5′ immune profiling solution in 2018–2019. Understandably, analyses of the single-cell immune repertoire data generated using the 10x Genomics platform before 2020 were primarily based on the tabular outputs from cellranger. These outputs were also adapted for use in tools that were originally designed for 'bulk' TCR/BCR-seq analysis such as Immcantation suite[19] and vdjtools[20].

The Adaptive Immune Receptor Repertoire (AIRR) Community, which started in 2015, began the standardization of data formats and structures from high-throughput immune repertoire sequencing in 2017 (ref. 21). Inclusion of AIRR-standardized format for the 10x Genomics TCR/BCR data did not occur until later versions of cellranger. Several tools have emerged to provide different analysis options specifically designed to handle single-cell data, including Scirpy[22], Dandelion[23,24] and scRepertoire[25] (Table 1).

Single-cell immune repertoire data analysis overview

As the bulk of scTCR/BCR-seq data generation has been performed on the 10x Genomics platform, we will focus the following sections on how the data are generally processed and analyzed starting from that format. Having said that, other scTCR/BCR-seq data generation methods are emerging, and most outputs would conform to AIRR standards, allowing data to be analyzed in a similar fashion, for example, BD Rhapsody TCR/BCR assays provides an AIRR-standardized output file.

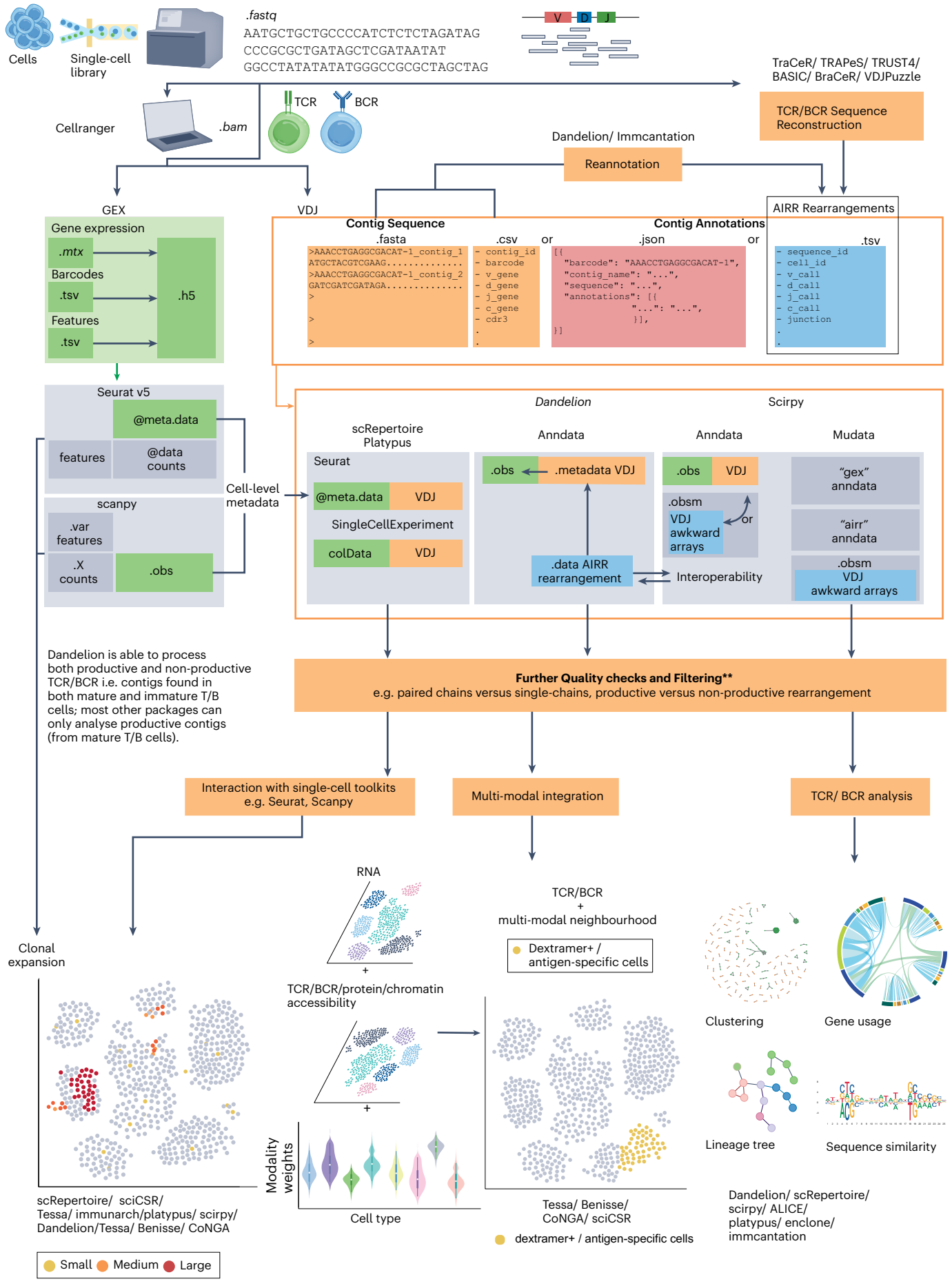
Data preparation

10x Genomics cellranger vdj provides a number of output files that are useful for scTCR/BCR-seq analysis. The information associated with each output file, together with some descriptions of the algorithms to achieve them, is listed on the 10x Genomics support website. Here, we provide a summary of relevant files, as well as their impact on analysis, in Table 2. Most of these files can be used for input into Dandelion, scRepertoire and Scirpy, with the latter also accepting other data formats such as TraCeR/BraCer's output.

In general, users would start with files that have the 'filtered_' prefix, which indicates that cellranger filtering has been applied to retain full-length and productive TCR/BCR sequences. This is especially relevant for studying mature T cells/B cells as their TCR/BCR have been successfully rearranged. Downstream analysis using the

Fig. 1 | Typical workflow of scTCR/BCR-seq. Cells are isolated as single cells followed by amplification of cDNA concentrated on the V(D)J locus to obtain TCR and BCR sequences. Raw sequencing reads are then aligned and reconstructed into TCR/BCR contigs, which are then annotated against reference V(D)J genes either through cellranger or other dedicated software. Gene expression data are typically read with single-cell analysis tool kits, while V(D)J data can be reannotated or directly read by scTCR/BCR-seq analysis tool kits. The data structure for each tool may differ but it has a common theme in linking the contig-level information to cell-level metadata. Further quality-control steps are then performed to filter for contigs that match user-defined parameters. Downstream analyses may include clonotype calling, differential V/D/J-gene usage, visualization of clonal expansion, integration with multimodal single-cell data and many more. Examples of dedicated tools for each step are appended to illustrate how they can be used together. Created with BioRender.com.

https://doi.org/10.1038/s41592-024-02243-4



https://doi.org/10.1038/s41592-024-02243-4

Table 1 | scTCR/BCR-seq analysis tools

Tool (source code)	Latest version	Platform/ language	TCR	BCR	Contig reconstruction	Reannotation	Filtering	Visualization	Clustering	Phylogenetic analysis	Multimodal integration	scRNA-seq tool kit interaction
TraCeR[10] (https://github.com/teichlab/tracer/)	v.0.6.0	Python	✔		✔	✔						
TRAPeS[12] (https://github.com/YosefLab/TRAPeS/)		Python, C++	✔		✔	✔						
TRUST4[13] (https://github.com/liulab-dfci/TRUST4/)	v1.0.12	Python	✔	✔	✔	✔						
MiXCR[14] (https://github.com/milaboratory/mixcr/)	v4.6.0	Java	✔	✔	✔	✔				✔		
BASIC[15] (https://github.com/akds/BASIC/)	v1.5.1	Python	✔	✔	✔	✔						
BraCer[16] (https://github.com/Teichlab/bracer/)	v0.2	Python		✔	✔	✔						
VDJPuzzle[17] (https://github.com/simone-rizzetto/VDJPuzzle/)	v1.0	bash	✔	✔	✔	✔						
Immcantation[19] (https://bitbucket.org/kleinstein/immcantation/)	v4.4.0	Python, R	✔	✔		✔			✔	✔		
Dandelion[24] (https://github.com/zktuong/dandelion/)	v0.3.5	Python	✔	✔		✔	✔	✔	✔			✔
scRepertoire[25] (https://github.com/ncborcherding/scRepertoire/)	v2.0.0	R	✔	✔			✔	✔	✔			✔
Scirpy[22] (https://github.com/scverse/scirpy/)	v0.16.0	Python	✔	✔			✔	✔	✔			✔
ALICE[37] (https://github.com/pogorely/ALICE/)		R	✔					✔	✔			
Platypus[53] (https://github.com/alexyermanos/Platypus/)	v3.3.6	R	✔	✔			✔	✔	✔	✔		✔
enclone[62] (https://10xgenomics.github.io/enclone/)	Beta	Rust	✔	✔				✔	✔	✔		
Tessa[65] (https://github.com/jcao89757/TESSA/)		Python, R	✔					✔			✔	
Benisse[67] (https://github.com/wooyongc/Benisse/)	v1.0.0	Python		✔				✔			✔	
CoNGA[66] (https://github.com/phbradley/conga/)	v0.1.2	Python	✔	✔				✔			✔	✔
Immunarch[54] (https://github.com/immunomind/immunarch/)	v0.9.0	R	✔	✔			✔	✔				
sciCSR[58] (https://github.com/Fraternalilab/sciCSR/)	v0.3.1	R		✔							✔	✔

'filtered_' dataset reduces potential noise from irrelevant contigs, improving the accuracy of, for example, cell-type identification, clonality assessments and mutation rates. If the 'cellranger multi' pipeline was used, contigs that do not match with any cell barcodes in the corresponding gene expression data are removed. In contrast, files with the 'all_' prefix will include any contigs associated with a droplet, regardless of whether the droplet contains a cell. It also does not assert the same filtering criteria at the contig level; therefore, all possible contigs are returned. Correspondingly, this increases the amount of noise to account for when filtering data for downstream analysis. However, if a user is analyzing developing T cells/B cells, it could be relevant to retain the nonproductive and non-fully rearranged contigs as they may be biologically relevant.

Reannotation

The Immcantation suite[19] was the first tool that implemented a strategy to parse 10x Genomics V(D)J sequencing data by reannotating the fasta files using IgBLAST[11] with IMGT[8] reference sequences. This effectively allows Immcantation to convert the output to a tabular spreadsheet format (later aligned to AIRR standards) and allows access to downstream tools that rely on IMGT references. This was necessary as 10x Genomics cellranger vdj uses an Ensembl-based reference for their annotation, which lacked the gapped information due to the IMGT unique numbering system or allele-level annotation for V(D)J loci. Dandelion adopted Immcantation's reannotation steps and included additional steps to separately annotate D/J genes, allowing Dandelion to retain contigs lacking a V gene (discarded by IgBLAST), which also led to the discovery of 'J-gene multi-mappers' (contigs with multiple sequential J-gene segments)[24]. It should be noted that 10x Genomics cellranger vdj already recovers contigs without V genes. However, for contigs with multiple J genes, where a large proportion of these also lacked a V gene[24], cellranger vdj would annotate the J genes with the highest alignment scores regardless of the positioning, instead of choosing the leftmost J gene.

As mentioned, there are alternatives to 10x Genomics cellranger vdj such as TRUST4 (ref. 13) and MiXCR[14] and they are capable of de novo assembly of TCR/BCR sequences followed by V(D)J-gene annotation from the raw 10x Genomics V(D)J sequence files. Interestingly, both TRUST4 and MiXCR can reconstruct V(D)J sequences from the 10x Genomics gene expression library, with outputs highly correlated to the V(D)J library[13]. However, as observed in Suo et al.[24] in the case of TRUST4, there may be some bias toward reconstructing V(D)J contigs by asserting the presence of a V gene; while this is fine for most use cases in studying mature T and B cells, this may pose a challenge for studying datasets containing early developing T cells/B cells where active (D)J rearrangements, and not V(D)J rearrangements, occur (Fig. 2a). It is also important to note that, by default, TRUST4 annotates the V(D)J genes using the IMGT reference, whereas MiXCR uses its own built-in set of references. Therefore, some discrepancies may arise from using different references and applications to different cell maturation stages, which users should consider.

BCR/TCR clustering and filtering

TCR/BCR clustering is important for defining clonal relationships between cells. A TCR/BCR clonotype refers to cells that share either the same (TCR) or similar (BCR) receptors. In the AIRR data standards, receptors and cells that are part of the same clonotype will be tagged with the same 'clone_id'. This allows researchers to infer that cells of the same clonotype are derived from the same ancestral cell during development and/or clonal expansion. Traditionally, TCRs/BCRs are clustered based on two major criteria: (1) using the same V and J gene and (2) CDR3 length. Cells with receptor configurations that fulfill the former but not the latter indicate that random N/P nucleotide insertions have occurred. These cells could still be considered evolutionarily part of the same ancestral lineage but cannot be strictly considered as 'clones'. Clonal relationships of BCRs are more complex because SHMs may occur throughout the recombined BCR chains, which can also introduce insertions/deletions into the junctional regions[26]. The field has devised different methods to define BCR clonotypes such as setting a cutoff based on bimodal distribution of pairwise hamming distances[27] and/or setting an empirical threshold of 85–90% similarity[28]. The Immcantation suite provides a variety of metrics to perform the clonotype grouping, including hamming distance distribution thresholding[19,29].

Another definition of clonotype that is frequently used is by grouping TCR/BCR sequences as 'functional clonotypes'. This definition is reserved for adaptive immune receptor configurations with shared specificity to the same epitope due to structural similarities and/or enrichments of certain amino acid motif patterns within the CDR3 junction, also sometimes referred to as the ligand repertoire. Several approaches have been developed to perform motif clustering based on edit distance or similarity metrics, for example, GLIPH[30], GLIPH2 (ref. 31), ClusTCR[32], GIANA[33], tcrdist[34], tcrdist3 (ref. 35) and iSMART[36]. However, they were mostly developed for, and applied to, bulk TCR-seq data (single-chain only; TRB only). While not strictly falling under the 'functional clonotype' category, clustering based on recombination frequencies for scTCR-seq has been proposed in ALICE[37] (Antigen-specific Lymphocyte Identification by Clustering of Expanded Sequences) to help identify TCRs and neighboring TCRs involved in a shared immune response using sequence similarity. ALICE constructs a single-cell TCR neighborhood space using stochastic TCR recombination model based on IGoR[38] and finds sequences that are clustered more than expected by random chance. The clustered sequences are proposed to potentially respond to the same antigens[37].

While these methods offer interesting strategies to infer the relationships between clonotypes, antigen specificity and overall cell phenotypes, caution needs to be exercised when applying them to scTCR/BCR-seq data. The single-cell data are likely to be under-sampled and not suitable for running these methods. scTCR/BCR-seq produces paired-chain data, whereas the tools above were designed for analyzing bulk TRB-only data. The results may also be challenging to experimentally validate, and thus may have limited predictive value. Although they are more challenging, wet-lab-based assays for predicting antigen specificity such as those developed for bulk TCR-seq are arguably more valuable for finding novel clonotype–phenotype associations. For example, the Multiplexed Identification of T Cell Receptor AntigenSpecificity (MIRA) assay[39] expands T cells to peptide pools followed by bulk TRB sequencing. Su et al.[40] paired data from MIRA assays with scTCR-seq data and identified T cells from patients with COVID-19 that share the same TCRβ chain configurations (CDR3 amino acid sequence, V-gene and J-gene usage) as COVID-19-associated antigen-specific TRB-clonotypes. The cellular phenotypes of these antigen-specific T cells were then inferred from the single-cell data and correlated with COVID-19 symptoms[40]. There are also emerging commercial technologies that can capture antigen specificity at the single-cell level for paired-chain TCR/BCR-seq with scRNA-seq experiments, for example, 10x Genomics Barcode-Enabled Antigen Mapping (BEAM)-T and BEAM-Ab, dCode Dextramers. There are also emerging computational techniques that can leverage these new data to predict novel antigen binding. For example, pMTnet[41] uses a transfer learning framework to learn and predict the binding associations between peptide-major histocompatibility complex (MHC)-I and TCRs, validating its utility on single-cell data from 10x Genomics single-cell immune profiling paired with dCode Dextramers reagents. Overall, development and use of these emerging technologies are poised to revolutionize how scTCR/BCR-seq analysis and interpretation may be performed in the future.

Currently, most scTCR/BCR-seq analysis tools have incorporated functions to regroup single cells into clonotype groups based on their TCR/BCR sequence and V(D)J-gene usage similarities. Importantly,

https://doi.org/10.1038/s41592-024-02243-4

Table 2 | Summary of relevant cellranger vdj output files

Prefix	Associated files	What information does it contain	Impact on analysis
filtered	filtered_contig.fasta[a] filtered_contig_annotations.csv[a]	The .fasta file contains the sequence of each reconstructed BCR/TCR contig/chain. The .csv file contains annotation information called by cellranger vdj, including V/D/J/C genes, CDR3 junctional sequence, clonotype ID and whether the contig is flagged as a productive contig. In later versions of cellranger, sequences of framework and CDR regions are also included in this annotation file. In cellranger multi mode, there is another check to retain only contigs with matching cell barcodes found in the gene expression data.	Using 'filtered' datasets as a starting point is the standard practice. The contigs are full length, productive and do not contain any premature stop codons. This format is useful for studying mature T cells/B cells with fully rearranged receptors. To reannotate the sequences with other reference databases, users will need to supply the fasta file at the minimum.
	airr_rearrangement.tsv[a]	This file contains the annotations for each 'filtered' contig as above but in AIRR rearrangement format.	This is an alternative to using the above two files. Only available from cellranger version 4 onwards.
all	all_contig.fasta[a] all_contig_annotations.csv[a] all_contig_annotations.bed all_contig_annotations.json[a]	The .fasta and .csv files contain information as above. The .bed file contains detailed annotation information on the structure of each reconstructed TCR/BCR, which can be used to explain why some contigs are filtered, for example, not full-length and/or not productive. The .json file contains a higher level of detail for the contig annotations not present in the .csv file or .bed file, such as the full reconstructed sequence, start and end positions for the framework and CDR regions, and validated/non-validated UMI sequences.	Using 'all' datasets as a starting point is similar to above with the caveat that this also includes any contig reconstructed by cellranger that can be assigned to a droplet. This may be relevant for analyzing contigs that are not fully rearranged, for example, contigs arising from D->J or V->DJ rearrangement events will be captured here. As per above, users can reannotate the contigs using the .fasta file as a minimum. Otherwise, information contained in the .csv or .json file can be converted to single-cell and AIRR formats for downstream analyses by some of the tools reviewed in this paper.
consensus	consensus.fasta consensus_annotations.csv	The .fasta and .csv files contain information as above but reports the information for the consensus sequence (for each contig) for each clonotype defined by cellranger.	These files are typically not used by current scTCR/BCR-seq analysis tool kits.

[a]Files that are used by current tool kits.

some of these tools have taken special considerations toward the quality control and filtering of scTCR/BCR-seq data before performing clonotype groupings (Fig. 2b), which can greatly influence the results from downstream analysis. For example, although it is common to sort contigs by unique molecular identifier (UMI) count and selecting chain pairs with the highest UMI counts for downstream analysis, hard filtering by contig UMI count, while not frequently practiced, could reduce the total number of contigs/chains considered for clonotype calling and result in different clonotype definitions. We note that for some tools that handle 10x Genomics scTCR/BCR-seq data, they default to using the clonotype definitions provided by the cellranger software. While this is generally suitable for TCR-seq data as cellranger defines TCR clonotypes based on identical V(D)J transcripts, this was problematic for BCR-seq data before cellranger v5.0 as the same criteria was used, effectively not accounting for SHMs. enclone is now provided as a module in cellranger v5.0 to group cells into BCR clonotypes, which respects the SHM events. Overall, regardless of which tool users choose to define the T cell or B cell clonotypes, we recommend that users should manually inspect whether the clonotype definitions are suitable for their data and reperform clonotype classification if necessary, as the quality-control steps for single-cell data greatly influence the outcome of clonotype calling.

scTCR/BCR-seq data analysis and single-cell integration

As alluded to previously, there were limited options to analyze scTCR/BCR-seq analysis before 2020. Most software was originally created for bulk immune repertoire analysis, and developers made efforts to adapt these methods for single-cell analysis. For example, the open-source Immcantation[19] suite is very widely used for both bulk BCR-seq and scBCR-seq analysis. It contains a plethora of tools such as changeo, shazam, alakazam and IgPhyML[42], allowing users to quantify SHMs in B cell contigs, define clonal groupings and reconstruct clonal phylogenetic lineage trees. Recent methods such as SCOPer[43] and Dowser[44] also implement new methods to define clonal groupings and visualize BCR lineages. Dowser's[44] unique feature is the development of three parsimony-based summary statistics that characterize migration, differentiation and isotype switching along B cell phylogenetic trees. This tool is particularly useful in understanding the migration of B cells between tissues, their differentiation into various cell types and isotype switching. Similarly, SCOPer[43,45] introduces a spectral clustering approach, which works on scBCR-seq data, to identify clones. The spectral clustering method is a nearest-neighbor approach that helps group BCR sequences based on junctional sequence similarity in different local neighborhoods[43]. They also subsequently introduced an updated approach that accounts for SHM and weights the similarity measure accordingly[45]. Overall, the immcantation suite has charted how the field analyzed bulk BCR-seq and scBCR-seq data[46]. A small caveat is that the Immcantation suite do not explicitly interact with the scRNA-seq analysis tool kits, which can be challenging for some users in determining the relevant analysis strategies for their single-cell datasets.

The data architecture for analyzing single-cell data has largely been spearheaded by the teams that develop and maintain packages such as Seurat[47] and SingleCellExperiment[48] in R or Scanpy/anndata[49] in Python (Fig. 1). There are also emerging efforts in creating tools meant for multi-omics representation, for example, muon[50]. Most single-cell immune repertoire analysis tools were designed to integrate

https://doi.org/10.1038/s41592-024-02243-4

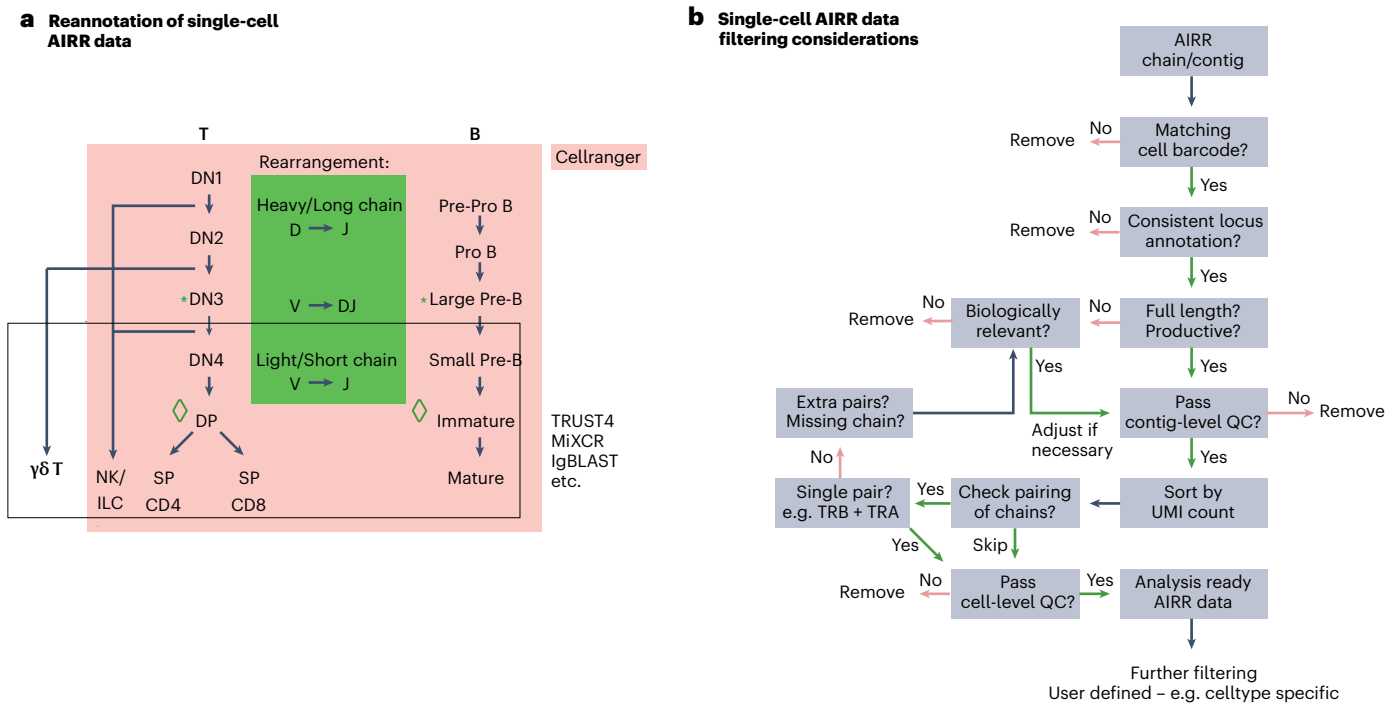


Fig. 2 | scTCR/BCR-seq data processing considerations. a, Recovery of V(D)J contigs across T cell and B cell development stages. Asterisks indicate the stage whereby productive VDJ rearrangement for BCR heavy chains or β/δ TCR chains is expected or detected. Diamond symbols indicate the stage whereby productive VJ rearrangement for BCR light chains or α/γ TCR chains is expected or detected. Cellranger can recover contigs from all stages, except for γδ-TCRs due to their biases for αβ-TCR libraries, whereas other software can recover contigs where there is a successful rearrangement containing a V gene. **b**, General quality-control strategy of scTCR/BCR-seq data. Contigs/chains are matched to cell barcodes before assessing various quality-control metrics to justify filtering or retention. Users should consider if the various filtering strategies will retain biologically relevant TCR/BCR data for their analysis. QC, quality control.

the scTCR/BCR-seq data with these single-cell data formats to enable further exploration such as performing filtering and quality-control checks, clonotyping, clonal expansion quantification and clonotype diversity estimation. Some considerations for performing scTCR/BCR-seq filtering are summarized in Fig. 2b.

Scirpy (now part of scverse[51]) was the first open-source tool that specifically dealt with the scTCR (and subsequently scBCR) data format from 10x Genomics cellranger[22]. Scirpy works as an extension of Scanpy and primarily interacts with the 'anndata' data structure. Scirpy's immune receptor model focuses on cells that primarily express single- and/or dual-immune receptor cells, following TraCeR's model[10]. Cells with more than two pairs of TCR/BCRs are flagged as multi-chain cells and filtered from downstream analyses. Scirpy uses an 'AirrCell' data structure to align with the AIRR standards, naturally populating essential contig data into the single-cell observations data frame (adata.obs). Excess chains (and other AIRR rearrangement data) are stored but not exposed to users so that they are not burdened by the high number of AIRR rearrangement data not essential for Scirpy's functions. Scirpy has ongoing support to parse the outputs from other scTCR/BCR-seq analysis tools. Recently, Scirpy has been updated to operate using a new data structure based on awkward arrays, which supports multi-omics data structure implemented in MuData[50,51]. One major difference in the new version is that the AIRR rearrangement data are no longer expanded automatically into the single-cell observations data frame like before. Scirpy implements a network-based clonotype definition strategy that is initially based on Levenshtein distance between CDR3 amino acid sequences and subsequently introduced alternative distance metrics such as Hamming distance-based metrics (to be more aligned with practices for BCR clonotype definition[19]), metrics that work on identical CDR3 nucleotide sequences to group TCR clones, and also a tcrdist[34] inspired 'alignment' metric that computes clonotypes based on BLOSUM62 (BLOcks SUbstitution Matrix) distances[52]. Scirpy leverages the plotting functionalities by Scanpy and further introduces alternative visualization methods, such as the clonotype graph where each clonotype is represented as a sub-graph that connects all single cells in a clonotype. The layout uses a rectangle packing strategy to arrange and visualize the sub-graphs from the largest to smallest clonotypes. Users can color the cells in the clonotypes to visualize cell-type-specific information associated with each clonotype, similar to how users would interact with the scRNA-seq data in Scanpy. The network visualization layout of single-cell observations was phased out in favor of a clonotype-centric layout where a single dot represents a clonotype configuration, and the size of dots scale to the number of cells in each clonotype. Scirpy also implements several useful modules to quantify clonal diversity, expansion and overlap and to visualize V(D)J-gene usage.

Similarly, scRepertoire[25] is an open-source R package and was designed to operate with the 10x Genomics cellranger vdj filtered contig outputs and works seamlessly together with Seurat (and, subsequently, SingleCellExperiment). scRepertoire appends the TCR/BCR data to the single-cell metadata and can be used for both contig-only analysis and combined analysis with gene expression data. It performs quality control and filtering of the contig data and has functions for visualizing and quantifying clonotype abundance, differential V(D)J-gene usage, clonal diversity and overlap. One of the most used features is the 'clonal space homeostasis' feature, which bins and quantifies clonotype proportions into groups, such as from rare to hyperexpanded clonotype groups. There are other R-based scTCR/BCR-seq analysis tools such as Platypus[53] and Immunarch[54] with additional features beyond scRepertoire. The Platypus package offers functions for immune repertoire simulation, repertoire classification and structural analysis of adaptive receptors. On the other hand, Immunarch has extensive

support across multiple immune repertoire sequencing platforms. As both packages require custom structures for immune repertoire data, it complicates their interoperability with other single-cell workflows. Like Immcantation, Immunarch does not directly interact with the scRNA-seq packages and thus may have limited support for analyzing combined scRNA-seq and scTCR/BCR-seq data.

Dandelion[23,24] is another open-source scTCR/BCR-seq analysis software package initially inspired by bulk BCR-seq clonotype network analyses[55–57] and has functional overlaps with the aforementioned tools, especially Scirpy. In contrast to Scirpy, Dandelion does not rely on the anndata data structure and instead focuses on retaining the contig-level AIRR rearrangement data and separately populating a cell-level data frame that can be merged with Scanpy's observation slot. Dandelion calls clonotypes based on similarities of the CDR3 amino acid/nucleotide sequences using hamming distances while asserting requirements for identical V/J-gene usages and CDR3 lengths. It constructs clonotype networks using minimum spanning trees that connect cells based on the similarity (Levenshtein distance) of the entire TCR/BCR sequences. The resulting networks enable diversity analysis based on Gini indices[23]. Dandelion's data structure is not as strict and allows for the retention of AIRR rearrangement data that do not pass normal filtering in other tools. This allows for a unique representation of incomplete/partial and nonproductive contigs (for example, those lacking a V gene). We will discuss the unique implications of this aspect in the case study section later. We note that most other tools have largely ignored these types of 'nonstandard' contigs/chains but are increasingly adopting means to include them for downstream analysis. For example, Scirpy's new data structure has removed the previous limitation of only including productive chains for downstream analysis and can now incorporate nonproductive chains. While the Immcantation suite has always enabled access to this data in a 'failed' file, this is not automatically returned during preprocessing.

Recently, another tool, sciCSR[58], introduced an analysis strategy to leverage on SHM and CSR to improve the alignment of B cell maturation. SHM was enumerated by comparing the V gene sequence to the germline V gene sequence from IMGT/HighV-Quest[59]. Based on the characteristic of how germline transcripts (also known as sterile transcripts) expression precede and can mark the onset of CSR[60], they created new features based on whether reads from the 5′ gene expression library mapped to the VDJ region, constant gene or 5′ region upstream of the constant gene. This distinguishes transcripts as productive or sterile reads, which they referred to as the 'isotype signature'. Using nonnegative matrix factorization analysis on this 'isotype signature', they defined a new metric called 'CSR potential', which underscores how naive-like or memory-like a single B cell is. By combining the Markov chain model derived from the CellRank[61] analysis of the single-cell gene expression, SHM and 'CSR potential', they implemented a transition path theory approach to the resulting transition matrix. Using data from early time points in time-course studies, they predicted the probability of CSR events in later timings, as well as CSR changes in knockout studies.

Finally, 10x Genomics has a tool for analyzing BCR and TCR data (enclone; in beta)[62]. Enclone is written in Rust and implements a unique clonotyping algorithm that does not rely on IMGT references. It solves the previous issue with 10x Genomics cellranger vdj clonotype definition that asserts identical CDR3 sequences in clonotypes, which was problematic for BCR data. It also implements a 'honeycomb' representation of the clonotypes, which were useful for visualizing light-chain coherence[63], a phenomenon observed where memory B cells tend to use the same light-chain V gene. Enclone also has a unique feature to infer donor alleles for V genes. This may have important consequences on our inference on how individual genetic variations can lead to variations of the antibody response due to genetic predispositions[64].

Overall, while each method is largely similar in their function, their approaches and results may differ. We recommend that users should consider which biological question they would like to try to address as each software has unique features, advantages and disadvantages for different contexts of lymphocyte biology when choosing methods for downstream analysis (Figs. 1 and 2 and Tables 1 and 2). A practical choice of scTCR/BCR-seq software package is by choosing tools that are compatible with the companion scRNA-seq tool kits of user's choice. For instance, users would typically pair Seurat with scRepertoire or Scanpy with Scirpy. However, it is important to consider that some tools are more adept at analyzing specific cell-type or clonotype properties. For instance, scRepertoire generates intuitive metrics for quantifying clonal expansion data; Scirpy and Dandelion are more focused on TCR and BCR data analysis, respectively; Dandelion is also more adept at analyzing developing T cells/B cells. Overall, due to AIRR standardization, there is a consistent input and output format for scTCR/BCR-seq data, which encourages interoperability between the tools. Considerable effort has also been put into developing packages that convert between scRNA-seq data formats for R and Python workflows. Therefore, users are encouraged to explore the various tool kits to fit their analysis needs, facilitating the answering of biologically relevant questions.

Advanced integrated analysis

One goal for integrated analysis of multimodal data is to strive for a common embedding or manifold. Some tools and concepts have been proposed to achieve integrated analysis of scTCR/BCR-seq and scRNA-seq in this manner.

Tessa[65] (TCR Functional Landscape Estimation Supervised with scRNA-seq Analysis) implements a Bayesian model to integrate scTCR-seq with scRNA-seq data, generating an embedding of TCR networks that reflects similarity in the TCR sequence and gene expression space. Tessa helped pull apart T cells that appear after immunotherapy treatment, suggested to be due to distinct underlying TCR profiles that separate T cells present before versus after treatment[65].

Similarly, CoNGA (Clonotype Neighborhood Graph Analysis)[66] is another tool that constructs a neighborhood graph based on the TCR sequence similarity between clonotypes and performs a graph-versus-graph correlation analysis with the gene expression neighborhood graph, highlighting common neighbors in both the gene expression and TCR neighborhood space. This approach enabled the characterization of clonotype clusters with shared features (including differentially expressed genes and TCR sequence properties).

The Benisse model[67] (BCR embedding graphical network informed by scRNA-seq) was created with a focus on generating a common latent space that incorporates the gene expression data and scBCR-seq data (encoded as Atchley factors[68], which summarizes the biochemical properties of individual amino acid sequences). The learnt embedding was proposed to reflect antigen specificity of the BCR sequences, and the core Benisse model was able to chart the BCR SHM trajectories of germinal center B cells, which correlated with maturation and memory gene signatures[67].

At the time of writing, none of the current scTCR/BCR-seq analysis tool kits described in the previous section provide native implementations or direct linkage to these methods. The input files for these methods typically are/resemble cellranger output files. To interface these methods with the scTCR/BCR-seq analysis tool kits described above, users would need to manually prepare the compatible input files (Fig. 3a). In terms of the output, the tools generally produce a file containing the new embeddings, which can then be read as a simple matrix to interact with the scRNA-seq analysis tool kit of choice. CoNGA has scripts to generate the result plots and returns an Anndata '.h5ad' file that stores the intermediate distance matrices and final scores in the relevant slots, which users can further explore with the scRNA-seq analysis packages.

Overall, these tools seek to generate a single-cell representation of an immune response by combining the AIRR and gene expression

https://doi.org/10.1038/s41592-024-02243-4

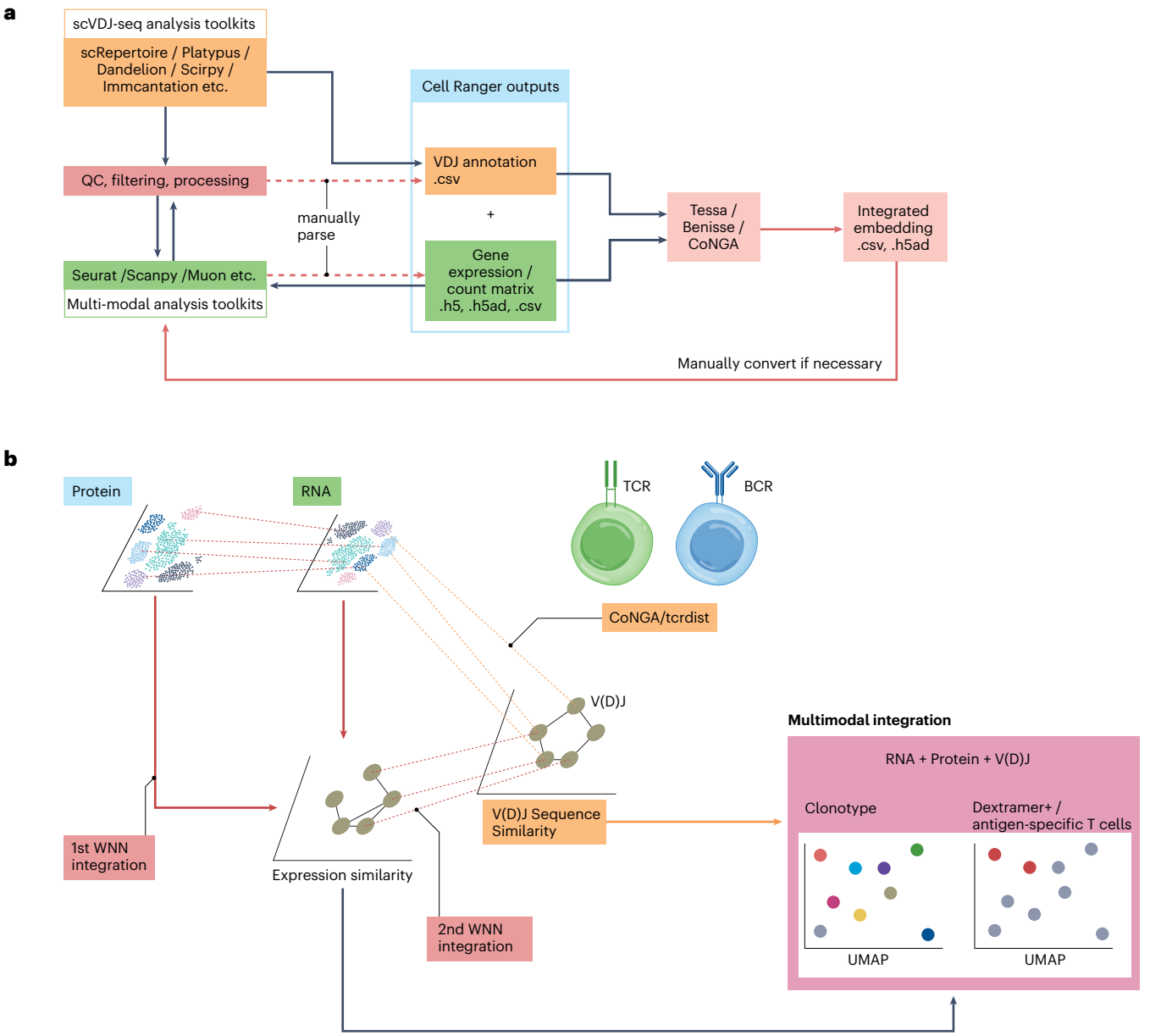


Fig. 3 | Multimodal integration of TCR with other single-cell data modes. a, Potential strategies to interface the multimodal analysis tool kits, scTCR/BCR-seq analysis tool kits and tools that perform advanced integration of TCR and multimodal data. The integration tools generally start from cellranger output files. Therefore, users will need to perform all quality control and preprocessing and prepare the corresponding input files to enable the downstream integration. **b**, Illustration of trimodal embedding strategy of RNA, protein and TCR data achieved by Zhang et al[69]. One recent example of multimodal integration of V(D)J data with other single-cell data modalities was demonstrated. The authors utilized a sequential WNN approach to integrate the neighborhood graphs of each data modality and used it to identify clonally expanded, antigen-specific $CD8^+$ T cells that responded to COVID-19 vaccination. UMAP, uniform manifold and approximation projection. Created with BioRender.com.

data. As this area is actively expanding, the inclusion of new modalities, like chromatin accessibility or protein quantification with AIRR data are being explored. Recently, trimodal embedding of single-cell gene expression, protein and TCR data was demonstrated[69]. Zhang et al. used weighted nearest-neighbor (WNN) analysis in Seurat[70] to combine the neighborhood graphs for each modality (first WNN graph combines gene expression with cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) data, followed by a second WNN graph that combines the first WNN graph with the TCR neighborhood graph derived from CoNGA[66] based on tcrdist[34,35]; Fig. 3b). Using this approach, they defined antigen-specific, 'antigen-proliferating' and 'bystander' vaccine-induced $CD8^+$ T cells and performed in-depth analysis of the cell states that accompany vaccine-induced clonal distributions.

Case studies on scTCR/BCR-seq analysis

COVID-19

For the remainder of this Review, we will describe a number of case studies to illustrate how scTCR/BCR-seq analysis led to new findings and interpretations for immunology research. As mentioned earlier, there were limited tools to perform an integrated analysis of scTCR/BCR-seq and scRNA-seq before 2020. However, with the emergence of COVID-19, several groups identified that profiling the immune repertoire during infection could be informative and thus performed scTCR/BCR-seq.

These studies analyzed scBCR/scTCR-seq data directly reported from the cellranger vdj output (Fig. 4a). For example, Ren et al.[71] explored the immune response to COVID-19 by analyzing the diversity of B cell and T cell subsets from peripheral blood mononuclear cells of infected individuals. They found that the BCR repertoire of individuals with COVID-19 exhibited biased V(D)J usage compared to healthy controls and trained a random forest classifier using the V(D)J usage frequencies, which distinguished individuals with COVID-19 with moderate or severe symptoms from healthy controls with relatively high accuracy[71]. Similarly, Liu et al.[72] performed scTCR/BCR-seq on longitudinal cohort of individuals with COVID-19 with critical disease, finding again that there was preferential V/J-gene usage in the T cells/B cells in these patients. Together, these studies suggest that specific T cell and B cell clones are preferentially expanded in response to severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection.

Dandelion was introduced in the study by Stephenson et al.[23] and incorporated a number of stringent quality-control steps to match the scBCR-seq data to the single-cell gene expression data (Fig. 4a). It also introduced a way to assess the diversity of clonotype networks across the different COVID-19 severity stages. This led to the observation of more pronounced clonal expansion in symptomatic individuals with COVID-19 compared to asymptomatic individuals and revealed sex-based disparities in clonotype size and BCR mutations, with women exhibiting higher levels of both measures than men. These differences in clonal expansion of B cells align with previous reports of worse outcomes in men with COVID-19, supporting that sex differences in the adaptive immune response may contribute to varying clinical outcomes[73]. In the same study, effector $CD8^+$ T cells were identified to be the most clonally expanded among the T cell subsets, and there was a correlation between disease severity and the ratio of effector-versus-effector memory $CD8^+$ T cells. This indicated that the balance between effector and memory T cell responses may play a role in determining the severity of COVID-19 (ref. 23).

In the study by Su et al.[74], the authors integrated the analysis of CITE-seq with scTCR-seq using Scirpy and also found marked clonal expansion of $CD8^+$ T cells in individuals with COVID-19. Hierarchical clustering of the TCRs from $CD8^+$ or $CD4^+$ T cells separated the TCRs into two groups for both cell types. For the $CD8^+$ T cells, the two TCR groups were split according to either moderate/severe disease or mild disease severity. For $CD4^+$ T cells, one of the TCR groups was associated with a cluster enriched with cytotoxic gene program; the same cluster is the only $CD4^+$ T cell cluster that presented with marked clonal expansion. These analyses help unveil unique developmental trajectories and suggest that TCRs influence cell fate choice and disease severity in COVID-19 (ref. 74).

In the study by Jaffe et al.[63], it was demonstrated that the light chains of functional antibodies are highly constrained (Fig. 4a). Using enclone's clonotyping algorithm, they analyzed paired V(D)J data from peripheral blood mononuclear cells of four unrelated individuals and discovered that B cells with similar heavy chains also tended to have similar light chains. This was shown to be the case for COVID-19, and other diseases such as multiple sclerosis and Kawasaki disease[63]. The authors concluded that this phenomenon can be generalized to memory B cells and that they represent how functional antibodies can arise in nature.

Long COVID, also known as post-acute COVID-19 syndrome/sequelae, remains an ongoing public health issue and affects an estimated ~65 million individuals worldwide[75]. Cheon et al. performed scRNA-seq and scTCR-seq on samples from older convalescent individuals with COVID-19 and found that subsets of respiratory $CD8^+$ T cells were correlated with persistent tissue conditions after COVID-19 (ref. 76). Using scRepertoire, the authors showed that clonally expanded $CD8^+$ T cells were present in blood and bronchoalveolar lavage samples of these individuals. These clonally expanded T cells also presented with higher expression of effector and/or cytotoxic molecules, consistent with chronic lung impairment[76]. This study revealed important immune features that may support the development of persistent lung abnormalities after SARS-CoV-2 infection in older individuals. In a separate study by Su et al.[40], they analyzed the scTCR-seq data with Scirpy and used the TCR sharing frequencies between cellular phenotypes and disease time points as barcodes to track the dynamics of $CD8^+$ and $CD4^+$ T cell response between acute COVID-19 and the convalescent stage (2–3 months after symptom onset). Hierarchical clustering grouped the TCR frequencies into five clusters and revealed that TCRs that predominate during the convalescent stage are different from those during acute COVID-19. By comparing the gene expression patterns of the TCR groups associated with cytotoxic or memory T cell clusters to those that were destined for clonal contraction in the convalescent stage, they showed that the cell types in each TCR group displayed distinct gene expression associated with either immune activation or suppression accordingly, suggesting that TCR patterns influence the determination of T cell memory or contraction after infection[40]. These studies are helping to shed light on the immunological heterogeneity potentially associated with long COVID development, and we anticipate that future efforts would help us fully understand the enigma of this disease[77].

The above studies provide valuable insights into the immune response to COVID-19, highlighting the importance of understanding the role of B cell and T cell subsets and their clonal expansion in the context of infection. The research underscores the importance of V/J-gene usage in shaping the immune response and highlights the immunological mechanisms underpinning long COVID.

T cell and B cell development

Other than tracking clonal expansion patterns, scTCR/BCR-seq analysis has also been useful for understanding and modeling the processes underlying the development of T cells and B cells, which occurs in the thymus[24,78–80] and bone marrow[80], respectively.

The Human Cell Atlas[81], a global collaborative effort creating comprehensive reference maps of all human cells at the single-cell level, has used scTCR/BCR-seq to understand the developing human immune system across various organs. In one of the first thymus cell atlas papers by the Human Cell Atlas, Park et al.[79] used paired scRNA-seq and scTCR-seq to explore human thymocyte development. They found a notable bias in V(D)J-gene usage from the onset of recombination to the mature T cell stage. This bias was particularly pronounced in $CD8^+$ T cells, which showed a preference for distal V–J pairs in their TRAV–TRAJ repertoire, suggesting a slower or less efficient commitment to the $CD8^+$ T lineage. In contrast, CD8αα cells displayed a slight bias toward proximal V–J pairs, more evident in postnatal thymic samples than in fetal samples. These biases were not due to the specific sequences guiding recombination but were instead strongly correlated with genomic position. This aligns with the looping structure of the locus previously observed in mice[82]. They also found that nonproductive TCRα chains were depleted in quiescent double-positive (DP(Q)) cells, suggests that cells unable to secure a productive TCRβ recombination for the first allele undergo recombination of the other allele. Overall, the study highlights the complexity of TCRα recombination and its role in shaping the T cell repertoire during human thymocyte development.

In the follow-up pan-immune cell atlas study, Conde-Domínguez et al.[78] developed a custom γδ-TCR primer set that is compatible with the 10x Genomics αβ-TCR library generation workflow (Fig. 4b). However, cellranger vdj was unable to annotate the γδ-TCR sequences despite being able to reconstruct the γδ-TCR contigs (described in detail in Suo et al.[24]). Therefore, the reconstructed contigs were processed using Dandelion[24], which successfully recovered the γδ-TCR gene annotations for each chain. While other groups may have also used scTCR-seq to study the processes underlying human T cell development, most

https://doi.org/10.1038/s41592-024-02243-4

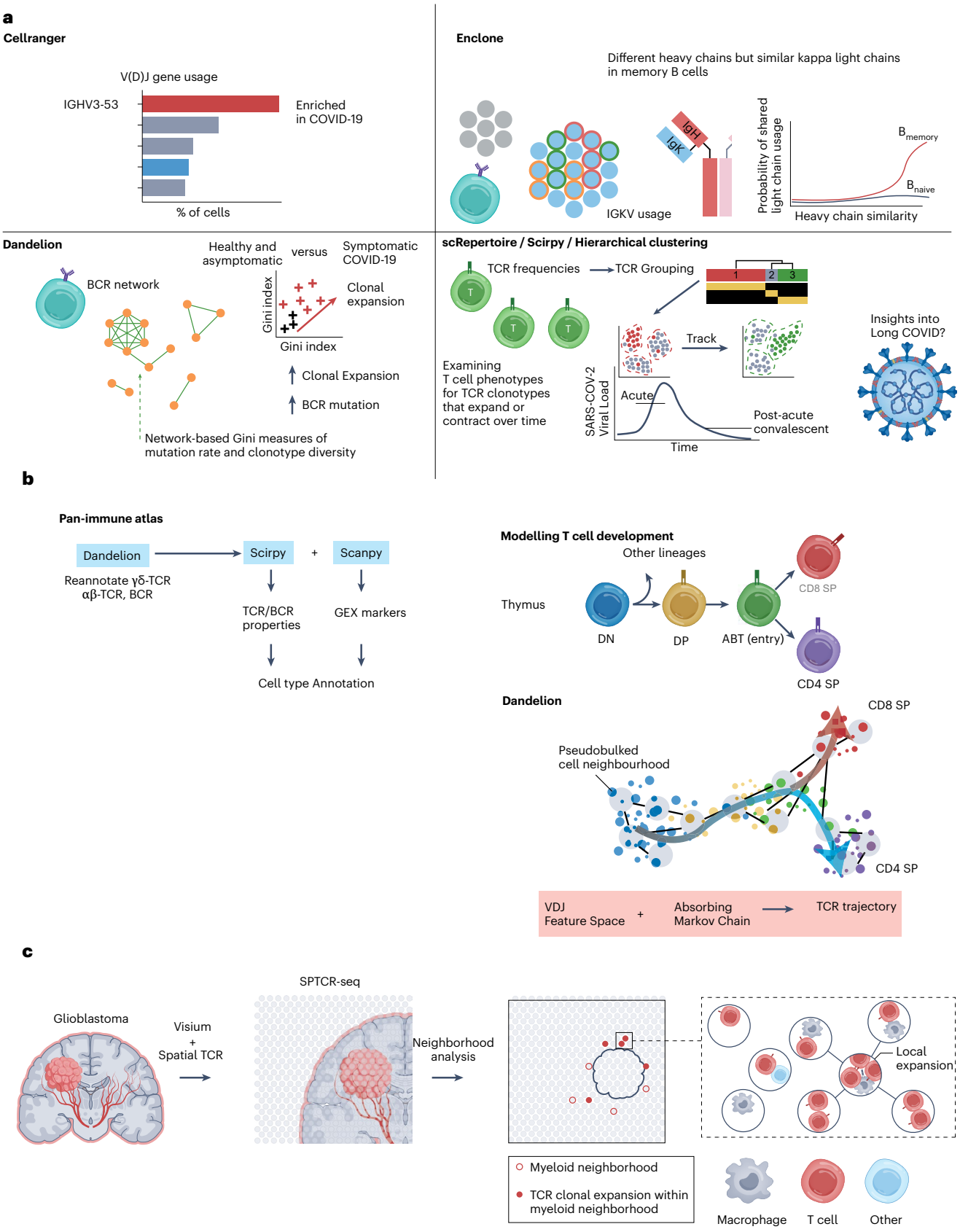


Fig. 4 | scTCR/BCR-seq analysis in COVID-19, T cell development and tumor immunology. a, Analysis strategies and key findings from scTCR/BCR-seq analysis of COVID-19 with the various tool kits. **b**, Modeling T cell/B cell development across tissues by incorporating the TCR/BCR to aid with cell-type annotations and the concept of TCR trajectory implemented in Dandelion by calculating V(D)J-gene usage frequencies across gene expression-defined pseudobulked cell neighborhoods. **c**, Utility of spatial TCR-seq with SPTCR-seq to understand the spatial TCR clonality in glioblastoma. Created with BioRender.com.

would be limited to using the presence or absence of the receptors to define cell-type annotations. For example, Cordes et al.[83] also used 10x Genomics scTCR-seq (αβ-TCR) and a custom library kit specifically designed for γδ-TCR sequencing and performed single-cell sequencing of the human thymic cells. After cell-type annotations were defined to separate αβ-TCR-expressing or γδ-TCR-expressing thymocytes and T cells, they subsequently relied only on gene expression information to define the T cell developmental routes[83]. Using Scirpy[22], Conde-Domínguez et al.[78] explored the TCR clonotype landscape within tissue-resident immune cells in adult tissues and discovered that each tissue possesses a unique 'immune neighborhood' of clonotypes reflective of adaptive immune responses tailored to specific tissue conditions. Clonal sharing among memory subtypes of $CD8^+$ T cells suggested that these subtypes may share a common precursor cell[78]. Conde-Domínguez et al.[78] processed their BCR data using the Immcantation workflow and found enrichment of isotype usage in a tissue-specific manner. For instance, the dominance of IgA2 in the gut region underscores the importance of this isotype in mucosal immunity. The detected bias toward IgA1 in memory B cells in the mesenteric lymph nodes and IgA2 in the ileum's Peyer's patches further highlights tissue-specific specialization of B cells. These findings suggest that local immune responses in the gut are tailored to counter specific pathogens and that B cells undergo class switching to produce the most appropriate isotype for a given tissue's immunological needs. The analysis of SHM levels revealed that, as expected, naive B cells had the lowest SHM frequencies, while plasma cells exhibited the highest SHM frequencies. Notably, the observed trend toward higher mutation rates in distal classes (IgG2, IgG4 and IgA2) is consistent with the accumulation of mutations during sequential class-switching events[84].

Similar efforts were conducted in the study by Suo et al.[80], but on fetal tissues. One of their efforts was focused on characterizing putative human B1 cells. B1 cells are a unique subset of B cells characterized by their self-renewal capacity, high expression of IgM and low expression of IgD[85]. These cells are believed to emerge early in development and share developmental similarities with regulatory B cells[85]. In mice, they are known for their role in the immune response, particularly in the production of natural antibodies, and can be detected using flow cytometry based on canonical markers such as CD5, CD27 and SPN (CD43), but it has been challenging to find similar human counterparts[86]. Suo et al. used Scirpy and Dandelion to identify these cells based on the distinct BCR properties compared to other mature B cells, including differences in the length of the CDR3 due to reduced random nucleotide insertions in the CDR3 junctions in both heavy and light chains compared to mature B cells[78]. These cells displayed a diverse BCR repertoire with minimal clonal expansion. They were also predominantly of the IgM isotype and spontaneously secreted IgM, a hallmark of B1 cells[87]. The adult pan-immune cell atlas[78] did not find the same putative B1 cells across the 16 tissues profiled, but they identified distinct clonal distribution patterns for the more tissue-restricted long-lived quiescent plasma cells versus the broad tissue distribution of classical memory B cell clones[78].

In a separate study, Suo et al. explored deeper into the data structure generated using 10x Genomics scTCR/BCR-seq and cellranger software and described the concept of the TCR trajectory as a natural 'time-keeper' for developing T cells. The trajectory inference method leverages data in the following three ways: (1) early developing T cells and B cells may express incomplete/nonproductive chains as part of the natural stochastic process of V(D)J rearrangement; (2) the 10x Genomics' scTCR/BCR-seq technology is capable of capturing these sequences; and (3) Dandelion's flexibility in dealing with AIRR data compared to the other scTCR/BCR-seq tools. Briefly, V(D)J usage frequencies are computed across pseudobulked cell neighborhoods (using milo[88]) based on gene expression data. The resulting matrix is used as input for trajectory inference using an absorbing Markov chain approach (implemented in Palantir[89]; Fig. 4b). This approach provides a dynamic view of the T cell maturation process in the thymus and shows that $CD4^+$ and $CD8^+$ T cell fate trajectories are influenced by the TCR V(D)J usage patterns[24], likely reflecting the selection process involving interaction with class I or class II MHC. This new technique was an improvement compared to standard trajectory inference based on gene expression alone, showing better correlation between the pseudotime ordering with $CD4^+$/$CD8^+$ defining marker genes and transcription factors[24]. It also supported the biases in TCRα V/J recombination as reported in Park et al.[79]. Secondly, Suo et al.[24] also reported on the abundance of nonproductive contigs in the 10x Genomics cellranger vdj data, a largely ignored aspect by other tools due to their limitations in only analyzing productive TCR/BCR chains. Suo et al. showed the importance of these nonproductive TCRs in lymphocyte development and selection using a TRBJ-based trajectory for innate lymphoid cell (ILC)/natural killer (NK) cell/T cell lineage analysis, which helped support the differentiation paradigms of these innate lymphocytes in the thymus[90]. They also found nonproductive TRB/TRG/TRD expression in ILC/NK cells, supporting the 'aborted' double-negative theory that suggests that ILC/NK/T cells share a common lineage trajectory in the thymus but deviate at some point before successful TCR rearrangement[91]. The nonproductive TRB expression was not organ specific but was observed consistently across various fetal organs. The expression of nonproductive TRB was also seen in early developing pre-pro B cells and the putative B1 cluster, suggesting that B1 cells have a different developmental route compared to other B cells.

Immunotherapy

In the rapidly evolving field of cancer treatment, immunotherapy is the gold standard for several advanced malignancies due to its efficacy. However, it is still not clear why some patients respond or do not respond to immunotherapy.

scTCR-seq has been used in some cancer immunotherapy studies to examine the TCR changes, including clonal expansions and diversity of T cell states, before and after treatment in responders versus nonresponders. These include, but are not limited to, studies examining combination immunotherapy with anti-PD-1 and anti-CTLA4 in melanoma[3] and anti-PD-1 with chemotherapy in breast cancer[2]. In a recent paper on pancreatic cancer, which is traditionally extremely difficult to treat, Rojas et al.[92] demonstrated that their custom neoantigen mRNA vaccines were feasible, safe and effective in 8 of 16 unselected patients. Their bulk TCR-seq data suggested that the vaccine-expanded T cells comprised up to 10% of all circulating T cells in the vaccine responders. In contrast, only 1 in 8 nonresponders showed a similar response[92]. scTCR-seq analyzed with Scirpy further demonstrated that the mRNA vaccines expanded polyclonal effector $CD8^+$ T cells after chemotherapy and vaccine booster[92]. These expanded T cells expressed lytic markers and cytokines, resembling effector T cells induced by protective viral vaccines[92].

Spatial profiling of T cell immune response is an emerging field and spatial TCR/BCR-seq methods are also being developed[93–96]. This was demonstrated in glioblastoma with SPTCR-seq[95], where the authors developed a new experimental and computational workflow to perform spatial TCR-seq (Fig. 4c). By performing spatial deconvolution of cell-type signatures and multimodal neighborhood analysis with the reconstructed TCR data, they associated spatial neighborhoods of increased myeloid cell enrichment with increased TCR diversity[95]. The technology potentially offers insights into the regional diversity of antitumor immune responses. However, the market is still relatively new and mature spatial TCR-seq technology is not commercially available at the time of writing and downstream computational analysis options for spatial TCR are very limited.

These collective studies highlight how the immune repertoire aids in decoding the immune response to both tumors and treatments. Studying the observed dynamic shifts in T cell clonality, emergence of new clones and changes in T cell diversity across various studies

provide an added window of opportunity to harness the appropriate therapeutic TCRs for improving anticancer treatment outcomes.

Outlook

Analysis of scTCR/BCR-seq data is challenging. Due to the current technologies, we are often limited by the number of immune receptors we can sample. Therefore, we often cannot generalize the findings from the single-cell level to the overall immune landscape. This coincides with a sampling issue that leads to the overrepresentation of singletons (TCR/BCR only represented once in the dataset). It is unclear if they are true singletons or part of larger clonotypes but were insufficiently sampled/sequenced. The diversity of individual TCRs/BCRs adds another layer of complexity. In addition, certain areas warrant further investigation. For instance, the presence of multiple chains or single-chain contigs can obscure potentially interesting biology. It remains unclear whether they are biological or technical challenges, for instance, ambient RNA in fluidics or index hopping, which further complicates the interpretation of the data[97].

Challenges in software

There remains several challenges and limitations in the various single-cell immune repertoire analysis tools in the handling of TCR/BCR contigs for downstream analysis. For example, in light-chain myeloma (Bence-Jones myeloma), cancer cells only express light-chain proteins and not heavy chains[98]. This poses a challenge as not every tool can call clonotypes if the heavy chain is absent. At the moment of writing, only Scirpy can define clonotypes in these types of samples as it allows users to specify whether to consider either or both chains.

Scalability poses another challenge. Current high-throughput scTCR/BCR-seq technology is dominated by 10x Genomics, and alternative technologies are emerging, for example, the recently announced 10x Genomics' GEM-X kits, Evercode TCR by Parse Biosciences, BD Rhapsody TCR/BCR Amplification Kit, BGI's DNBelab C4 single-cell sequencing for full-length transcripts, Omniscope and SeekGene. These new technologies bring with them the opportunity to sequence millions of single-cell TCR/BCR sequences, several orders of magnitude higher than what the original 10x Genomics Gel Beads-in-emulsion technology can generate for a single sample. It is important to note that while it is possible to scale up current workflows to more than a few million receptors, we may still be vastly under-sampling the repertoire. Secondly, current data structures are also not designed for dealing with large volumes of AIRR data. Better and more efficient approaches are needed to deal with the increasing volume of data so as to meaningfully assess and embed receptors other than computing pairwise distances. A possible avenue is toward featurization using deep learning/embedding techniques, which could lead to a more universal comparison of sequence similarity. The development of data standards for AIRR data is also ongoing; scverse[51] is working toward a unified data format in the single-cell space in collaboration with various stakeholders. Addressing these limitations will be key to future advancements in single-cell immune repertoire analysis.

Although there is considerable interest in spatial profiling of TCR/BCR, this field is very new. Parenthetically, downstream analysis options are limited. A common approach to spatial TCR-seq is by using MiXCR[14] to perform TCR reconstruction from the spatial sequencing data generated from full-length cDNA libraries. Engblom et al.[96] developed two protocols for spatial TCR/BCR-seq (Spatial VDJ) to capture paired-chain BCR and TCR sequences, as an extension of the 10x Genomics Visium spatial transcriptomics platform. The long-read version is compatible for capturing full-length BCR and TCR sequences, while the short-read version is only for TCR[96]. For the long-read Spatial VDJ data, relevant reads were prepared into fastq files and MiXCR was used for gene annotation. Short-read Spatial VDJ data were processed using pRESTO[99] before analyzing with MiXCR. Spatial barcodes were appended to the MiXCR outputs using custom R scripts, and downstream clonotyping analysis was primarily achieved by leveraging the Immcantation[19] workflow. Similarly, the recently described SPTCR-seq[95] method includes a bespoke probe-based library preparation protocol that leverages the fact that several commercially available spatial transcriptomics technologies generate full-length cDNA libraries at some point in their respective protocols, allowing them to sensitively and specifically capture and reconstruct full-length TCR sequences using long-read sequencing. SPTCR-seq also includes a computational pipeline to extract the TCR-seq data, which includes annotation with IgBLAST against IMGT references and their SPATAimmune R package to import and visualize the SPTCR-seq data. Overall, while we note that capturing spatial TCR/BCR information can potentially provide additional resolution of spatially relevant immunological partners and immune response, there is a general lack of dedicated computational frameworks or tool kits that can carry out quality-control checks or perform integrated analysis of spatial information, RNA and TCR/BCR data. It also remains to be determined whether the appropriate data structures are available or suitable for analyzing these new data modalities.

Emerging technologies in capturing antigen specificity

Concurrently, alternative technologies that can capture TCR/BCR specificity have emerged, for example, 10x Genomics BEAM-T and BEAM-Ab, dCode Dextramers, and so on. For the 10x BEAM-Ab technology, barcode-labeled antigens are incubated with B cells. In contrast, the BEAM-T assay is designed to take off-the-shelf antigenic peptides, load them onto MHC monomers that contain a nucleotide label, followed by incubation with T cells. The output of both BEAM assays is the pairing of the relative signal of the labeled antigen and its respective scTCR/BCR-seq data. In a related approach, LIBRA-seq (linking B cell receptor to antigen specificity through sequencing) technology uses DNA-barcoded biotinylated antigens and incorporates them into the 10x Genomics single-cell workflow, allowing the tagging of B cells that are bound to specific antigens[100]. This helped screen for virus-specific (HIV and influenza) B cells and predict reactivity to antigens for other BCRs at the single-cell level[100]. In contrast, dCode Dextramers use long chains of sugars with peptide–MHC complexes. Similarly to the BEAM-T assay, these dextramers are labeled with nucleotide tags to allow for deciphering epitope specificity downstream during sequencing. In addition, the fluorescently tagged dextramers can be used for fluorescence-activated cell sorting of antigen-specific populations before sequencing. While there are several other experimental techniques that have recently emerged aimed at quantifying single-cell TCR associations and binding with peptide–MHC complexes or BCR to their cognate antigen, the scope of this Review prevents an exhaustive discussion of them. Nevertheless, we look forward to the future uptake and development of accompanying computational methods that will be invaluable for advancing our understanding of adaptive immunity at the single-cell level.

As the scTCR/BCR-seq data analysis domain gradually improves through the emergence of more comprehensive technologies and increasing collaboration between various software communities, it will undoubtedly expand our ability to understand systems immunology processes at a single-cell level. We also hope that similar advancements will take place for spatial TCR/BCR-seq, which will help understand the adaptive immune response in tissues. The knowledge gain will help pave the way toward innovative and personalized immunotherapy and vaccine options for various immunologically relevant diseases.

References

1. Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu. Rev. Biochem.* **71**, 101–132 (2002).
2. Zhang, Y. et al. Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer. *Cancer Cell* **39**, 1578–1593 (2021).

3. Poran, A. et al. Combined TCR repertoire profiles and blood cell phenotypes predict melanoma patient response to personalized neoantigen therapy plus anti-PD-1. *Cell Rep. Med.* **1**, 100141 (2020).
4. Pilkinton, M. A. et al. In chronic infection, HIV Gag-specific $CD4^+$ T cell receptor diversity is higher than $CD8^+$ T cell receptor diversity and is associated with less HIV quasispecies diversity. *J. Virol.* **95**, e02380–20 (2021).
5. Kotagiri, P. et al. B cell receptor repertoire kinetics after SARS-CoV-2 infection and vaccination. *Cell Rep.* **38**, 110393 (2022).
6. Pai, J. A. & Satpathy, A. T. High-throughput and single-cell T cell receptor sequencing technologies. *Nat. Methods* **18**, 881–892 (2021).
7. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nat. Methods* **18**, 873–880 (2021).
8. Lefranc, M. P. et al. IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.* **27**, 209–212 (1999).
9. Tang, F. et al. mRNA-seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
10. Stubbington, M. J. T. et al. T cell fate and clonality inference from single-cell transcriptomes. *Nat. Methods* **13**, 329–332 (2016).
This study described TraCeR, the tool that reconstructed TCRs from scRNA-seq data.
11. Ye, J., Ma, N., Madden, T. L. & Ostell, J. M. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Res.* **41**, W34–W40 (2013).
12. Afik, S. et al. Targeted reconstruction of T cell receptor sequence from single cell RNA-seq links CDR3 length to T cell differentiation state. *Nucleic Acids Res.* **45**, e148 (2017).
13. Song, L. et al. TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. *Nat. Methods* **18**, 627–630 (2021).
14. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
15. Canzar, S., Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. BASIC: BCR assembly from single cells. *Bioinformatics* **33**, 425–427 (2017).
16. Lindeman, I. et al. BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat. Methods* **15**, 563–565 (2018).
17. Rizzetto, S. et al. B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics* **34**, 2846–2847 (2018).
18. Andreani, T. et al. Benchmarking computational methods for B-cell receptor reconstruction from single-cell RNA-seq data. *NAR Genom. Bioinform.* **4**, lqac049 (2022).
19. Gupta, N. T. et al. Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
This study introduced Changeo, one of the most widely used immune repertoire sequencing data analysis software as part of the *Immcantation* suite.
20. Shugay, M. et al. VDJtools: unifying post-analysis of T cell receptor repertoires. *PLoS Comput. Biol.* **11**, e1004503 (2015).
21. Rubelt, F. et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. *Nat. Immunol.* **18**, 1274–1278 (2017).
22. Sturm, G. et al. Scirpy: a Scanpy extension for analyzing single-cell T-cell receptor-sequencing data. *Bioinformatics* **36**, 4817–4818 (2020).
This study introduced Scirpy, the first and the most widely used Python package that specifically dealt with scTCR-seq data, as an extension of the Scanpy scRNA-seq package.
23. Stephenson, E. et al. Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
This study introduced the first iteration of Dandelion as a scBCR-seq analysis tool written in Python and introduced network-based diversity analysis for scBCR-seq data.
24. Suo, C. et al. Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins. *Nat. Biotechnol.* **42**, 40–51 (2023).
This study introduced an updated version of Dandelion and also introduced new concepts for analyzing scTCR/BCR-seq data, including trajectory analysis of pseudobulked cell neighborhoods using TCR usage frequencies.
25. Borcherding, N., Bormann, N. L. & Kraus, G. scRepertoire: an R-based toolkit for single-cell immune receptor analysis. *F1000Res.* **9**, 47 (2020).
This work describes scRepertoire, one of the most widely used scTCR/BCR-seq analysis software in R that integrates with Seurat and SingleCellExperiment formats.
26. Kepler, T. B. et al. Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. *Cell Host Microbe* **16**, 304–313 (2014).
27. Yaari, G. & Kleinstein, S. H. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med.* **7**, 121 (2015).
28. Chen, H. et al. BCR selection and affinity maturation in Peyer's patch germinal centres. *Nature* **582**, 421–425 (2020).
29. Yaari, G., Uduman, M. & Kleinstein, S. H. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res.* **40**, e134 (2012).
30. Glanville, J. et al. Identifying specificity groups in the T cell receptor repertoire. *Nature* **547**, 94–98 (2017).
This study uses an amino acid-based motif approach to quantify repertoire dynamics and to identify patterns in epitope specificity in the context of *Mycobacterium tuberculosis*.
31. Huang, H., Wang, C., Rubelt, F., Scriba, T. J. & Davis, M. M. Analyzing the *Mycobacterium tuberculosis* immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nat. Biotechnol.* **38**, 1194–1202 (2020).
32. Valkiers, S., Van Houcke, M., Laukens, K. & Meysman, P. ClusTCR: a Python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity. *Bioinformatics* **37**, 4865–4867 (2021).
33. Zhang, H., Zhan, X. & Li, B. GIANA allows computationally-efficient TCR clustering and multi-disease repertoire classification by isometric transformation. *Nat. Commun.* **12**, 4699 (2021).
34. Dash, P. et al. Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**, 89–93 (2017).
This study introduced the use of the edit distance of the CDR loop in grouping viral antigen-specific sequences.
35. Mayer-Blackwell, K. et al. TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *Elife* **10**, e68605 (2021).
36. Zhang, H. et al. Investigation of antigen-specific T-cell receptor clusters in human cancers. *Clin. Cancer Res.* **26**, 1359–1371 (2020).
37. Pogorelyy, M. V. et al. Detecting T cell receptors involved in immune responses from single repertoire snapshots. *PLoS Biol.* **17**, e3000314 (2019).
38. Marcou, Q., Mora, T. & Walczak, A. M. High-throughput immune repertoire analysis with IGoR. *Nat. Commun.* **9**, 561 (2018).
39. Klinger, M. et al. Multiplex identification of antigen-specific T cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS ONE* **10**, e0141561 (2015).

40. Su, Y. et al. Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell* **185**, 881–895 (2022).
41. Lu, T. et al. Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* **3**, 864–875 (2021).
42. Hoehn, K. B. et al. Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proc. Natl Acad. Sci. USA* **116**, 22664–22672 (2019).
43. Nouri, N. & Kleinstein, S. H. A spectral clustering-based method for identifying clones from high-throughput B cell repertoire sequencing data. *Bioinformatics* **34**, i341–i349 (2018).
44. Hoehn, K. B., Pybus, O. G. & Kleinstein, S. H. Phylogenetic analysis of migration, differentiation, and class switching in B cells. *PLoS Comput. Biol.* **18**, e1009885 (2022).
45. Nouri, N. & Kleinstein, S. H. Somatic hypermutation analysis for improved identification of B cell clonal families from next-generation sequencing data. *PLoS Comput. Biol.* **16**, e1007977 (2020).
46. Hoehn, K. B. & Kleinstein, S. H. B cell phylogenetics in the single cell era. *Trends Immunol.* **45**, 62–74 (2024).
47. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
48. Lun, A., Risso, D. & Korthauer, K. SingleCellExperiment: S4 classes for single cell data. *R package version 1* (2018),
49. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
50. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biol.* **23**, 42 (2022).
51. Virshup, I. et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nat. Biotechnol.* **41**, 604–606 (2023).
52. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992).
53. Yermanos, A. et al. Platypus: an open-access software for integrating lymphocyte single-cell immune repertoires with transcriptomes. *NAR Genom. Bioinform.* **3**, lqab023 (2021).
54. Samokhina, M. et al. immunomind/immunarch: Immunarch 0.9.0. Zenodo. https://doi.org/10.5281/zenodo.7446955 (2022).
55. Bashford-Rogers, R. J. M. et al. Network properties derived from deep sequencing of human B-cell receptor repertoires delineate B-cell populations. *Genome Res.* **23**, 1874–1884 (2013).
56. Bashford-Rogers, R. J. M. et al. Analysis of the B cell receptor repertoire in six immune-mediated diseases. *Nature* **574**, 122–126 (2019).
57. Fitzpatrick, Z. et al. Gut-educated IgA plasma cells defend the meningeal venous sinuses. *Nature* **587**, 472–476 (2020).
58. Ng, J. C. F. et al. sciCSR infers B cell state transition and predicts class-switch recombination dynamics using single-cell transcriptomic data. *Nat. Methods* https://doi.org/10.1038/s41592-023-02060-1 (2023).
59. Alamyar, E., Giudicelli, V., Duroux, P. & Lefranc, M. -P. IMGT/HighV-QUEST: a high-throughput system and Web portal for the analysis of rearranged nucleotide sequences of antigen receptors—high-throughput version of IMGT/V-QUEST. in *Journées Ouvertes de Biologie, Informatique et Mathématiques* **60** (2010).
60. Lorenz, M., Jung, S. & Radbruch, A. Switch transcripts in immunoglobulin class switching. *Science* **267**, 1825–1828 (1995).
61. Lange, M. et al. CellRank for directed single-cell fate mapping. *Nat. Methods* **19**, 159–170 (2022).
62. Jaffe, D. B. et al. enclone: precision clonotyping and analysis of immune receptors. Preprint at *bioRxiv* https://doi.org/10.1101/2022.04.21.489084 (2022).
63. Jaffe, D. B. et al. Functional antibodies exhibit light chain coherence. *Nature* **611**, 352–357 (2022).
64. Rodriguez, O. L. et al. Genetic variation in the immunoglobulin heavy chain locus shapes the human antibody repertoire. *Nat. Commun.* **14**, 4419 (2023).
65. Zhang, Z., Xiong, D., Wang, X., Liu, H. & Wang, T. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods* **18**, 92–99 (2021).
66. Schattgen, S. A. et al. Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA). *Nat. Biotechnol.* **40**, 54–63 (2022).
This study demonstrates how neighborhood graphs from single-cell and TCR data can be integrated to achieve integrated analysis.
67. Zhang, Z. et al. Interpreting the B-cell receptor repertoire with single-cell gene expression using Benisse. *Nat. Mach. Intell.* **4**, 596–604 (2022).
68. Atchley, W. R., Zhao, J., Fernandes, A. D. & Drüke, T. Solving the protein sequence metric problem. *Proc. Natl Acad. Sci. USA* **102**, 6395–6400 (2005).
69. Zhang, B. et al. Multimodal single-cell datasets characterize antigen-specific $CD8^+$ T cells across SARS-CoV-2 vaccination and infection. *Nat. Immunol.* **24**, 1725–1734 (2023).
70. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
71. Ren, X. et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **184**, 1895–1913 (2021).
72. Liu, C. et al. Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell* **184**, 1836–1857 (2021).
73. Takahashi, T. et al. Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* **588**, 315–320 (2020).
74. Su, Y. et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479–1495 (2020).
75. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat. Rev. Microbiol.* **21**, 133–146 (2023).
76. Cheon, I. S. et al. Immune signatures underlying post-acute COVID-19 lung sequelae. *Sci. Immunol.* **6**, eabk1741 (2021).
77. Brodin, P. et al. Studying severe long COVID to understand post-infectious disorders beyond COVID-19. *Nat. Med.* **28**, 879–882 (2022).
78. Domínguez Conde, C. et al. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* **376**, eabl5197 (2022).
79. Park, J.-E. et al. A cell atlas of human thymic development defines T cell repertoire formation. *Science* **367**, eaay3224 (2020).
80. Suo, C. et al. Mapping the developing human immune system across organs. *Science* **376**, eabo0510 (2022).
81. Lindeboom, R. G. H., Regev, A. & Teichmann, S. A. Towards a human cell atlas: taking notes from the past. *Trends Genet.* **37**, 625–630 (2021).
82. Skok, J. A. et al. Reversible contraction by looping of the Tcra and Tcrb loci in rearranging thymocytes. *Nat. Immunol.* **8**, 378–387 (2007).
83. Cordes, M. et al. Single-cell immune profiling reveals thymus-seeding populations, T cell commitment, and multilineage development in the human thymus. *Sci. Immunol.* **7**, eade0182 (2022).
84. Kitaura, K. et al. Different somatic hypermutation levels among antibody subclasses disclosed by a new next-generation sequencing-based antibody repertoire analysis. *Front. Immunol.* **8**, 389 (2017).
85. Baumgarth, N. The double life of a B-1 cell: self-reactivity selects for protective effector functions. *Nat. Rev. Immunol.* **11**, 34–46 (2011).

86. Griffin, D. O., Holodick, N. E. & Rothstein, T. L. Human B1 cells in umbilical cord and adult peripheral blood express the novel phenotype $CD20^{+}CD27^{+}CD43^{+}CD70^{-}$. *J. Exp. Med.* **208**, 67–80 (2011).
87. Holodick, N. E., Tumang, J. R. & Rothstein, T. L. Immunoglobulin secretion by B1 cells: differential intensity and IRF4-dependence of spontaneous IgM secretion by peritoneal and splenic B1 cells. *Eur. J. Immunol.* **40**, 3007–3016 (2010).
88. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using *k*-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
89. Setty, M. et al. Characterization of cell fate probabilities in single-cell data with Palantir. *Nat. Biotechnol.* **37**, 451–460 (2019).
90. Karimi, M. M. et al. The order and logic of CD4 versus CD8 lineage choice and differentiation in mouse thymus. *Nat. Commun.* **12**, 99 (2021).
91. Qian, L. et al. Suppression of ILC2 differentiation from committed T cell precursors by E protein transcription factors. *J. Exp. Med.* **216**, 884–899 (2019).
92. Rojas, L. A. et al. Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* **618**, 144–150 (2023).
93. Hudson, W. H. & Sudmeier, L. J. Localization of T cell clonotypes using the Visium spatial transcriptomics platform. *STAR Protoc.* **3**, 101391 (2022).
94. Liu, S. et al. Spatial maps of T cell receptors and transcriptomes reveal distinct immune niches and interactions in the adaptive immune response. *Immunity* **55**, 1940–1952 (2022).
95. Benotmane, J. K. et al. High-sensitive spatially resolved T cell receptor sequencing with SPTCR-seq. *Nat. Commun.* **14**, 7432 (2023).
96. Engblom, C. et al. Spatial transcriptomics of B cell and T cell receptors reveals lymphocyte clonal dynamics. *Science* **382**, eadf8486 (2023).
97. Farouni, R., Djambazian, H., Ferri, L. E., Ragoussis, J. & Najafabadi, H. S. Model-based analysis of sample index hopping reveals its widespread artifacts in multiplexed single-cell RNA-sequencing. *Nat. Commun.* **11**, 2704 (2020).
98. Kyle, R. A. et al. Clinical course of light-chain smouldering multiple myeloma (idiopathic Bence Jones proteinuria): a retrospective cohort study. *Lancet Haematol.* **1**, e28–e36 (2014).
99. Vander Heiden, J. A. et al. pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* **30**, 1930–1932 (2014).
100. Setliff, I. et al. High-throughput mapping of B cell receptor sequences to antigen specificity. *Cell* **179**, 1636–1646 (2019).

Acknowledgements

We thank G. Sturm for the useful discussions and D. Maresco-Pennisi and C. Lee for helping to proofread the document. We acknowledge Children's Hospital Foundation's philanthropic contributions awarded to the Ian Frazer Centre for Children's Immunotherapy Research.

Author contributions

S.E.I., N.B. and Z.K.T. wrote the original draft. S.E.I. and Z.K.T. synthesized the literature and designed the review structure. M.S.F.S., N.B. and Z.K.T. critically reviewed, revised and edited the manuscript. M.S.F.S. made and synthesized the figures and tables. Z.K.T. conceptualized the review, outlined the structure, provided overall direction and supervised the writing.

Competing interests

N.B. is Head of Computational Biology at Omniscope and has consulted for Starling Biosciences and Santa Ana Bio. Z.K.T. has consulted for Synteny Biotechnology in the last 3 years. All other authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Zewen Kelvin Tuong.

Peer review information *Nature Methods* thanks Caleb Lareau, Guideng Li, and Tao Wang for their contribution to the peer review of this work. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.