DALI (Diversity AnaLysis Interface): a novel tool for the integrated analysis of multimodal single cell RNAseq data and immune receptor profiling.

Kevin Verstaen[1,2], Inés Lammens[3,4], Jana Roels[1,2], Yvan Saeys[1,2], Bart N Lambrecht[3,4], Niels Vandamme[1,2,5*] and Stijn Vanhee[3,4*#]

[1] Data Mining and Modeling for Biomedicine, VIB Center for Inflammation Research, Ghent, BE
[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, BE
[3] Department of Internal Medicine and Pediatrics, Ghent University, BE
[4] Laboratory for Immunoregulation and Mucosal Immunology, VIB Center for Inflammation Research, Ghent, BE
[5] VIB Single Cell Core Facility, BE

*equal contribution
corresponding author stijn.vanhee@ugent.be

Abstract

Single-cell RNA sequencing is instrumental to unravel the cellular and transcriptomic heterogeneity of T and B cells in health and disease. Recent technological advances add additional layers of information allowing researchers to simultaneously explore the transcriptomic, surface protein and immune receptor diversity during adaptive immune responses. The increasing data complexicity poses a burden on the workload for bioinformaticians, who are often not familiar with the specificities and biology of immune receptor profiling. The wet-lab modalities and sequencing capabilities currently have outpaced bioinformatics solutions, which forms an ever-increasing barrier for many biologists to analyze their datasets. Here, we present DALI (Diversity AnaLysis Interface), a software package to identify and analyze T cell and B cell receptor diversity in high-throughput single-cell sequencing data. DALI aims to support bioinformaticians with a functional toolbox, allowing seamless integration of multimodel scRNAseq and immune receptor profiling data. The R-based package builds further on workflows using the Seurat package and other existing tools for BCR/TCR analyses. In addition, DALI is designed to engage immunologists having limited coding experience with their data, using a browser-based interactive graphical user interface. The implementation of DALI can effectively lead to a two-way communication between wet-lab scientists and bioinformaticians to advance the analysis of complex datasets.

Introduction

Dissecting heterogeneity of immune cell responses in disease is at the heart of immunological research. With the development of single cell sequencing methods, the ability to distinguish cellular heterogeneity based on RNA expression became a popular tool for immunological research. Recent developments in these methods have added additional layers of information, including surface protein expression (CITE-Seq) and immune receptor profiling (VDJseq). Understanding the evolution of immune receptor profiles in adaptive immune responses is essential to fully appreciate the heterogeneity of adaptive immune responses in the fields of infectiology, allergy, autoimmunity, cancer and vaccinology.

Both B lymphocytes and T lymphocytes express immune receptors with an astonishing diversity, forming an anticipatory repertoire to fend off any antigenic exposure. Throughout development of B- and T- lymphocytes, diversity is induced in the B cell receptor (BCR) heavy and light chains and the T cell receptor (TCR) alpha and beta chains, respectively. During this rearrangement process genomic segments of the variable (V), diversity (D) and joining (J) regions are recombined in a random fashion, which are then combined with constant (C) region segment. In addition, at the junctions of these recombined fragments, random nucleotides are inserted, increasing the diversity even more. The anticipatory repertoire of B and T cells is estimated to contain over 10^{13} unique receptor sequences (1). Upon antigen exposure, B cells will develop into either memory cells or antibody secreting plasma cells. The main goal of the B cell response is to produce antibodies with high specificity towards the antigen. This is promoted by introducing random mutations into the recombined BCR sequences, increasing antigen affinity in a stochastic manner. To this end, B cells undergo somatic hypermutation (SHM) and subsequent selection for receptor affinity in germinal center reactions. Hence, these processes further increase the complexity of the immune receptor repertoire.

While we have seen rapid technological advances now allowing us to sequence the immune receptor repertoires of hundreds to thousands of cells, the tools that allow easy and meaningful analysis of these datasets are not developing at the same speed. To help the field in accommodating this need, we have developed the Diversity AnaLysis Interface (DALI), which interacts with the Seurat R package (2) and is aimed to support the advanced bioinformatician with a set of novel methods and an easier integration of existing tools for BCR and TCR analysis in their single cell workflow. With this, we aim to produce meaningful output from complex datasets. On the other hand, DALI aids immunologists with a limited background in bioinformatics through an interactive R Shiny app based graphical user interface (GUI), allowing easy data exploration. The implementation of DALI can effectively lead to a two-way communication between wet lab scientists and bioinformaticians, and advance the analysis of complex datasets

Results

Tool availability

DALI is available as an open-source tool to the community and is available on the VIB Single Cell Core Facility Github page https://github.com/vibscc/DALI .

Supported platforms

DALI can be used on any single-cell dataset processed with the R Seurat package, and for which there is corresponding immune profiling data. The package has been built and tested based on the 10x Genomics cellranger vdj and cellranger multi output. However, any platform that generates AIRR compliant output can be integrated into an immune receptor annotated R Seurat object (3).

DALI analysis pipeline

DALI can be integrated in an existing single-cell workflow in two different ways, as show in figure 1. First, an existing R Seurat object, containing single cell sequencing data obtained through any platform, can be linked to immune receptor profiling data in one step into the interactive DALI Shiny app (Fig 1A, step A) (given that the immune receptor data is either generated by the cellranger vdj/multi pipeline or compliant with the AIRR specifications (3) (Fig 1A, step A). This can be done by simple selection of the source R object and the folder containing the immune receptor sequencing output (Figure 1A). DALI allows to link both BCR and TCR data in the same object, and allows for seamless switching between the 2 data modalities. Alternatively, an extended R eurat object can be generated in a separate step, in which the immune receptor data is linked to its corresponding cell. This extended R Seurat object can then be loaded or be used in the R environment for further analysis using DALI or any other R tool.

Overview of the Shiny application

Most of the functionalities of DALI are bundled in a user-friendly Shiny application, which can be run in the browser. Here, the user has the option to go through different steps adding up to an in-depth analysis of the integrated (multimodal) single cell dataset. The user can navigate through the **General View** to view immune receptor gene segment usage plotted on the dimensionality reductions available in the object (Figure 2A). Next, the **Clone View** allows for an instantaneous assessment of clonotype or clone expansion and potential clonal overlap between clusters using the clonotype connection plots (Figure 2B). Additionally, in-depth information on the top clonotypes within each of the metadata factors can be obtained. Next, using the **Population comparison view**, users can directly compare immune receptor properties between patient, condition, treatment or any other available metadata variable (Figure 2C). Once a clonotype of interest is identified, the **Clonotype view** tab can be used to obtain full length information on these clonotypes (Figure 2D). This information can be used to generate recombinant monoclonal antibodies, perform receptor modeling or to generate immune receptor transgenic cells/animals. Lastly, the clonotype view allows for interactive generation of hierarchical clusters of BCRs, denoting somatic hypermutation of one

clonotype. The nodes can be colored based on gene expression data or any meta data parameter of interest. In the **Transcriptomics view**, expression of the genes of interest can be plotted on the dimensionality reduction plots and linked to the position of the clonotype of interest (Figure 2D). In the **Differential Gene Expression view**, differential gene expression can be calculated between cells using any metadata parameter, including immune receptor properties (Figure 2F).

A test case: analysis of house dust mite driven B cell responses in mediastinal lymph nodes

As a test case, we have performed single cell immune profiling of both B and T cell receptors of enriched subsets of B and T cells from mediastinal lymph node. These cells were isolated from mice sensitized to house dust mite extract (HDM) and subsequently challenged with HDM for 3 weeks. During the adaptive immune response in the lymph node in response to allergen exposure, a germinal center response will be initiated (Figure 3). During this process, naïve B cells and T cells will be activated and will differentiate into germinal center B cells (GCB) and follicular helper T cells (TFH), respectively. Germinal center B cells will increase the affinity of their B cell receptors and undergo receptor class switching in the dark zone of the germinal center. Upon exit of the dark zone, B cells enter the germinal center light zone where there newly obtained receptor affinity is tested. If successful in competing for limiting amounts of antigen, the light zone B cell will enter the dark zone again for another round of affinity maturation. From these germinal center responses, dedicated antibody producing cells (plasma cells) or memory B cells can arise, the latter effectively forming a memory of previous antigen exposure. We will here define dark zone B cells (DZB) and light zone B cells (LZB), memory B cells (memB), as well as plasma cells (PC). In addition, we zoom in at clonal evolution of a clone of interest and we will define antigen specific B cells and analyze their immune receptor repertoires.

<u>Definition of DZB, LZB, memB and PC</u>

We set out to define different B cell subsets using the DALI interactive mode (Figure 4A), we defined B cells through Cd79a expression through the transcriptomics view (Figure 4B). To define germinal center B cells, we defined cells expressing Bcl6 and Mki67 (Figure 4C). These cells can be traced back to cluster 7 (Figure 4A). We defined LZB within cluster 7 by CD86 and Fcer2a expression (Figure 4D), while DZB were defined using Hmmr and Aicda expression (Figure 4E) (4). Plasma cells were defined by Xbp1 and Jchain expression (Figure 4F), while memory B cells were defined by high Bach2 and Klf2 expression (Figure 4G)(5). Using the general view, we can next determine heavy and light chain constant region usage, where we see evidence for class switching towards IgG1 and IgG2b in cluster 7 (Figure 4H). Using the clone view, we can see that there is clonal expansion within cluster 7 (Figure 4I) and that some clones are shared between cluster 7, cluster 0 and cluster 15 (Figure 4J).

<u>Understanding clonal evolution</u>

Next, we aim to understand the clonal evolution of some of the expanded clones in cluster 7. To this end, we define an expanded clonotype in the clone view, and define this clone further in the clonotype view. Here we zoom in on clonotype 2 (Figure 5A, B). Using our tree analysis in the clonotype view, we can appreciate the clonal evolution of clonotype 2 (Figure 5C,D).

171 Here we highlight the evolution of class switching within this tree view (Figure 5C) and the
172 evolution of AID expression (Figure 5D).
173
174 Analysis of antigen specific B cells
175
176 Finally, we analyze the repertoire of Der p2 specific B cells, a major allergen in house dust
177 mite. Using fluorescence activated cell sorting (FACS), we sorted Der p2 reactive cells using
178 Der p2 protein tetramers (Figure 6A), and upon sample hashing, combined these for single
179 cell sequencing. Demultiplexing of the samples shows the location of the two Der p2+ samples
180 on the UMAP. Using the clone view, we can obtain information on the Der p2 reactive BCR
181 sequences (Figure 6C) and obtain detailed information on the clonotype in the clonotype view
182 (Figure 6D).
183
184 **Conclusion and future direction**
185
186 We here provide a comprehensive R based tool for the integrated analysis of single cell gene
187 expression data and immune receptor profiling in the Seurat ecosystem. One of the unique
188 aspects of this tool is both the availability of a toolset in the R environment for the
189 experienced bioinformatician, as well as a GUI based interface for the biologist with limited
190 bioinformatics background. While we strongly believe that this will aid the field and reduce
191 the generation of data black boxes, we do realize that no tool is complete, and that the GUI
192 analysis might be limited to obtain complete understanding of the dataset. Nonetheless we
193 believe that this way of analyzing data provides a first step, upon which the biologist can
194 interact with experienced bioinformaticians. DALI is also fully open-source, which allows the
195 community to aid in the further development of this tool. Lastly, DALI will receive regular
196 updates to accommodate the rapid evolutions in the field of single cell transcriptomics.

Methods

Scripting and computational analysis

All scripting was performed using R (6). All analyses were performed using R version 4.1.1 "Kick Things" and DALI version 1.0.0.

Mice and treatments

All mice were obtained through Janvier (France) and were of the C57BL6 strain. All procedures were approved by the VIB-UGent ethical review board. Mice were sensitized with 1ug house dust mite extract (ALK, Denmark) in 80ul of PBS instilled intratracheally and challenged 3 times a week for 3 weeks with 50ug of house dust mite extract in 40ul PBS intranasally. These methods are described in greater detail in (7).

Single cell sequencing

Single-cell RNA and BCR/TCR sequencing experiments

Upon isolation of mediastinal lymph nodes from 2 biological replicates, cells were stained with CD19 Pe-Cy5 (clone 1D3), Fixable viability dye Live/Dead (L/D) Efluor 506, homemade Der p2 protein tetramers coupled to streptavidin-PE and biotin-streptavidin-PE-AF647 decoys. This staining was performed on ice for 30' in PBS containing 3% of FCS. After sorting Der p2 reactive and non-reactive cells, the cells were stained with TotalSeq-C hashing antibodies (BioLegend) diluted 1:500. 4 samples were multiplexed per lane using TotalSeq-C Cell Hashing Antibodies. Sorted cells were mixed according to the following ratio: non B cells/B cells/Der p2 reactive B cells in a 45:45:10 ratio. Samples from 2 biological replicates were then mixed in a 1:1 ratio. Sorted single-cell suspensions were resuspended at an estimated final concentration of 2000 cells/µl and loaded on a Chromium GemCode Single Cell Instrument (10x Genomics) to generate single-cell gel beads-in- emulsion (GEM). The scRNA/Feature Barcoding/BCR/TCR libraries were prepared using the Chromium Next GEM Single Cell 5' Kit v2 (PN-1000263) according to the manufacturer's instructions. The cDNA content of pre-fragmentation and post-sample index PCR samples was analyzed using the 2100 BioAnalyzer (Agilent). Sequencing libraries were loaded on an Illumina NovaSeq flow cell at VIB Nucleomics core with sequencing settings according to the recommendations of 10x Genomics, pooled in a 75:20:5 ratio for the gene expression, TCR/BCR and antibody-derived libraries, respectively.

Single-cell data processing and analysis

For the demultiplexing and FASTQ generation, bcl2fastq2 (Illumina, v2.20) was used. This resulted in FASTQ files for read 1, read 2, i5 and i7 sample index for both the gene expression and cell surface protein libraries. Cellranger multi (10X Genomics, v6.1.1) was used to map gene expression reads to the reference genome (Mus musculus, Ensembl release 99), assemble VDJ contigs (GRCm38 v5.0.0 provided by 10X Genomics), process cell barcodes and count UMIs. The raw gene expression matrix was then further processed in R, using a combination of the workflow proposed by the Marioni and Theis lab (8) and the Seurat

pipeline (2). Outlier cells were identified based on 3 metrics (library size, number of expressed genes and mitochondrial proportion) and cells were tagged as outliers when they were 5 median absolute deviation (MADs) away from the median value of each metric across all cells. This removes low quality cells (low UMI count and/or high percent mitochondrial reads) and potential doublets (cells with very high UMI count). Data normalization was performed using the NormalizeData function of Seurat (v4.0.5) with default parameters. Scaling, detection of higlhy variable features, clustering and dimensionality reduction were all done using the Seurat package. Both clustering and dimensionality reduction were performed using the first 30 principal components.

Data availability

A test data R Seurat object is available on the VIB-UGent Inflammation Research Center cloud https://cloud.irc.ugent.be/public/index.php/s/9ys5czsaNtNQtSd

Acknowledgements

We would like to thank Dr. Nincy Debeuf, Dr. Andrew Brown and Sam Dupont for testing beta versions of DALI.

We would like to thank the VIB Flow Core for training, support and access to the instrument park. The authors would also like to thank the VIB Single Cell Core facility for support and sample processing. We also thank the VIB Protein Service Facility for in house generation of tetramers.

SV is supported by a Research Foundation – Flanders (FWO) senior postdoctoral fellowship FWO-1244321N

Author contributions

KV conceived the project, wrote the code and wrote the manuscript, IL performed experiments, JR wrote the code, YS, BNL, NV conceived the project and wrote the manuscript, SV conceived the project, wrote the code, wrote the manuscript, performed experiments, analyzed data and supervised the study.

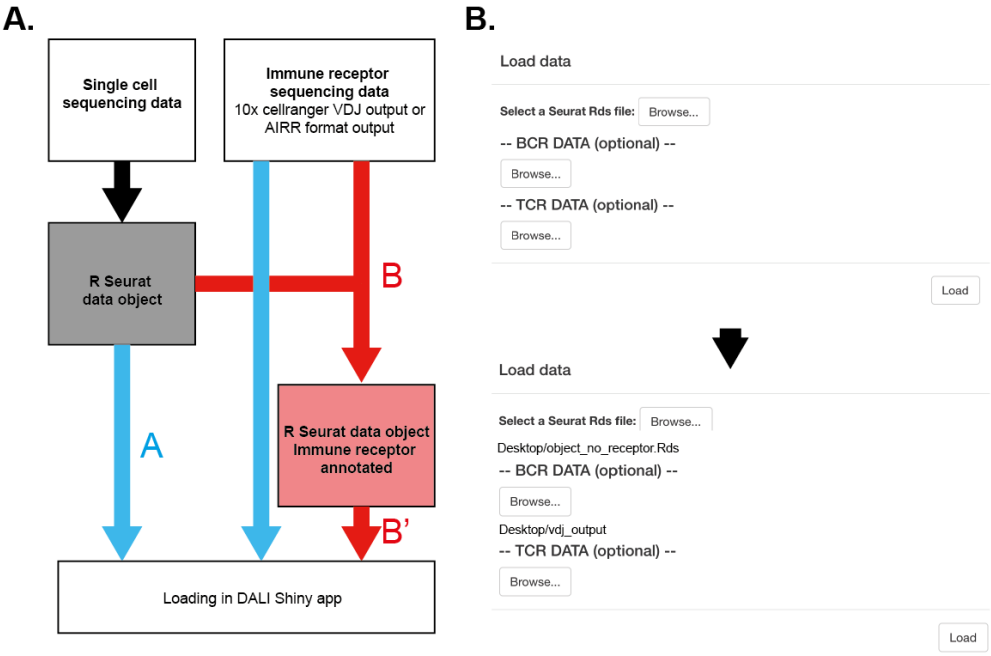

***Figure 1**: Overview of DALI pipelines using either **(A)** loading of the R Seurat object and immune profiling data directly into the interactive Shiny app **(B)** an extended R Seurat object containing immune receptor profiling data, which can be loaded into the interactive Shiny app (B').*

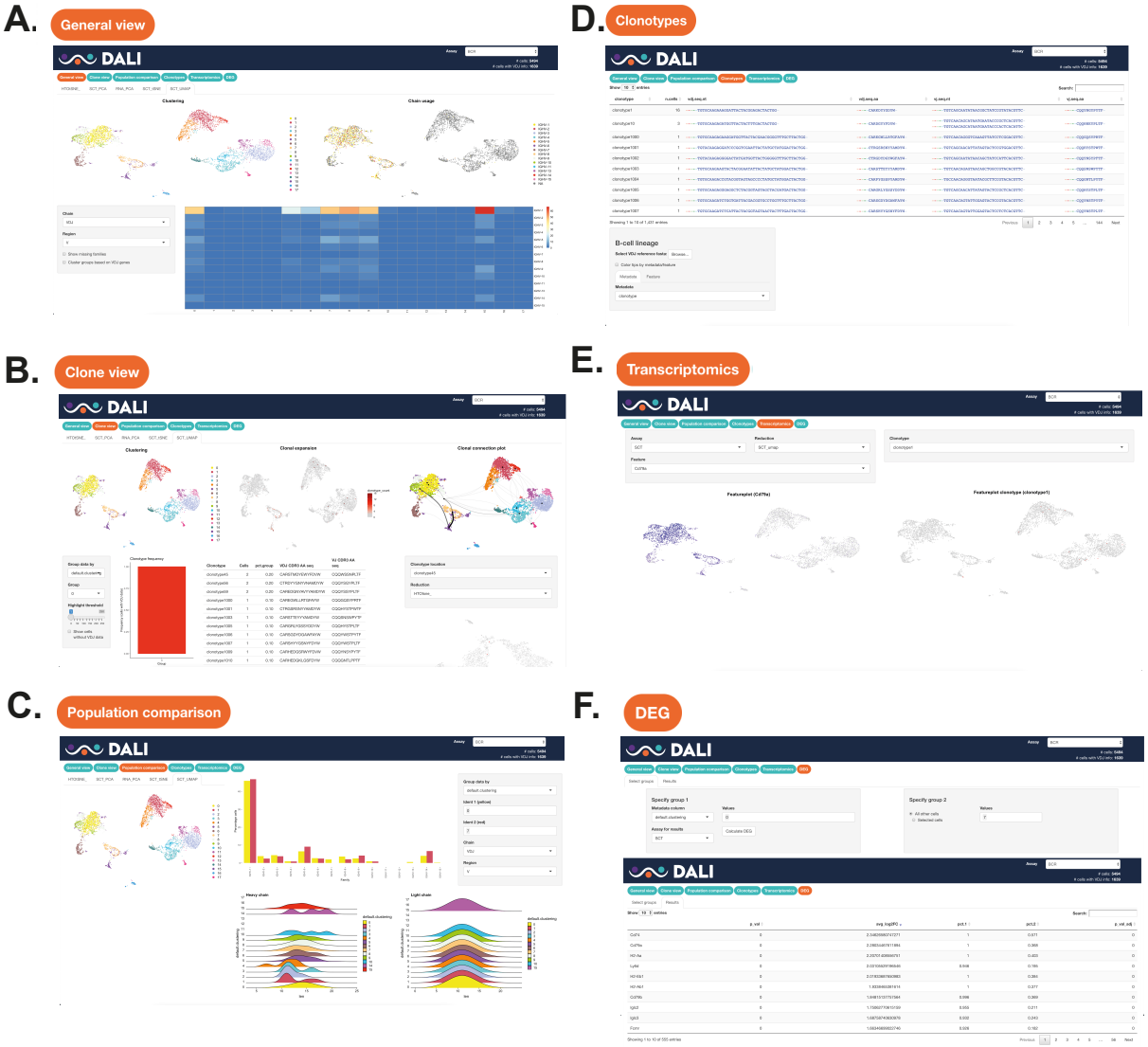

Figure 2: *Overview of the different DALI Shiny app tabs.* ***(A)*** *General View,* ***(B)*** *Clone View,* ***(C)*** *Population comparison,* ***(D)*** *Clonotype view,* ***(E)*** *Transcriptomics view,* ***(F)*** *Differential Gene Expression view.*

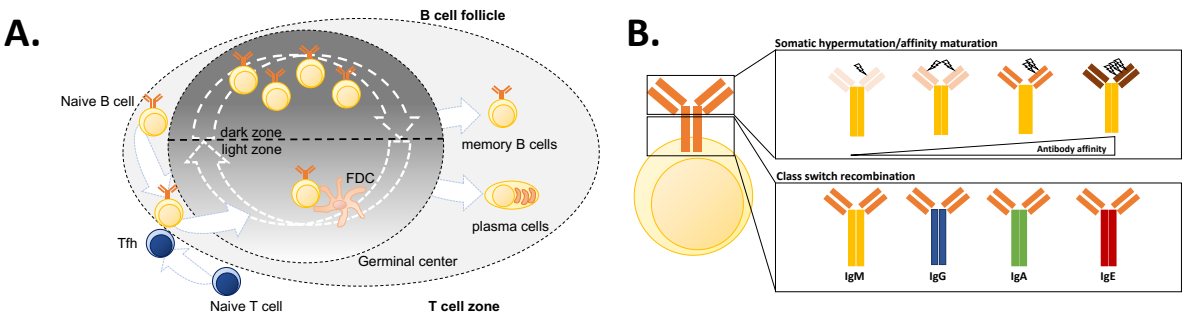

Figure 3: *Overview of B cell responses* ***(A)*** *B cell germinal center (GC) responses. An activated B cell will interact with its cognate follicular helper T cell (Tfh). Upon this interaction, the B cell will differentiate into a germinal center B cell (GCB) and migrate into the GC dark zone. Here, the B cell will proliferate and can undergo class switch recombination and somatic hypermutation directed by the AID protein. The edited receptor is then tested for its potential to bind antigen presented by follicular dendritic cells (FDC). The GCB can undergo multiple iterations of the cycle. From the GC response, memory B cells and plasma cells can arise. The former will quickly respond and re-enter a GC reaction upon future antigenic exposure, while plasma cells are dedicated antibody producing cells.* ***(B)*** *Schematic overview of B cell receptor (BCR) affinity maturation and class switch recombination. With increasing somatic hypermutation, the binding affinity of the BCR/secreted antibody might be increased. Upon class switch recombination, the BCR/antibody constant region is switched, providing the receptor with different signaling capacities and the secreted antibody with different functional properties.*

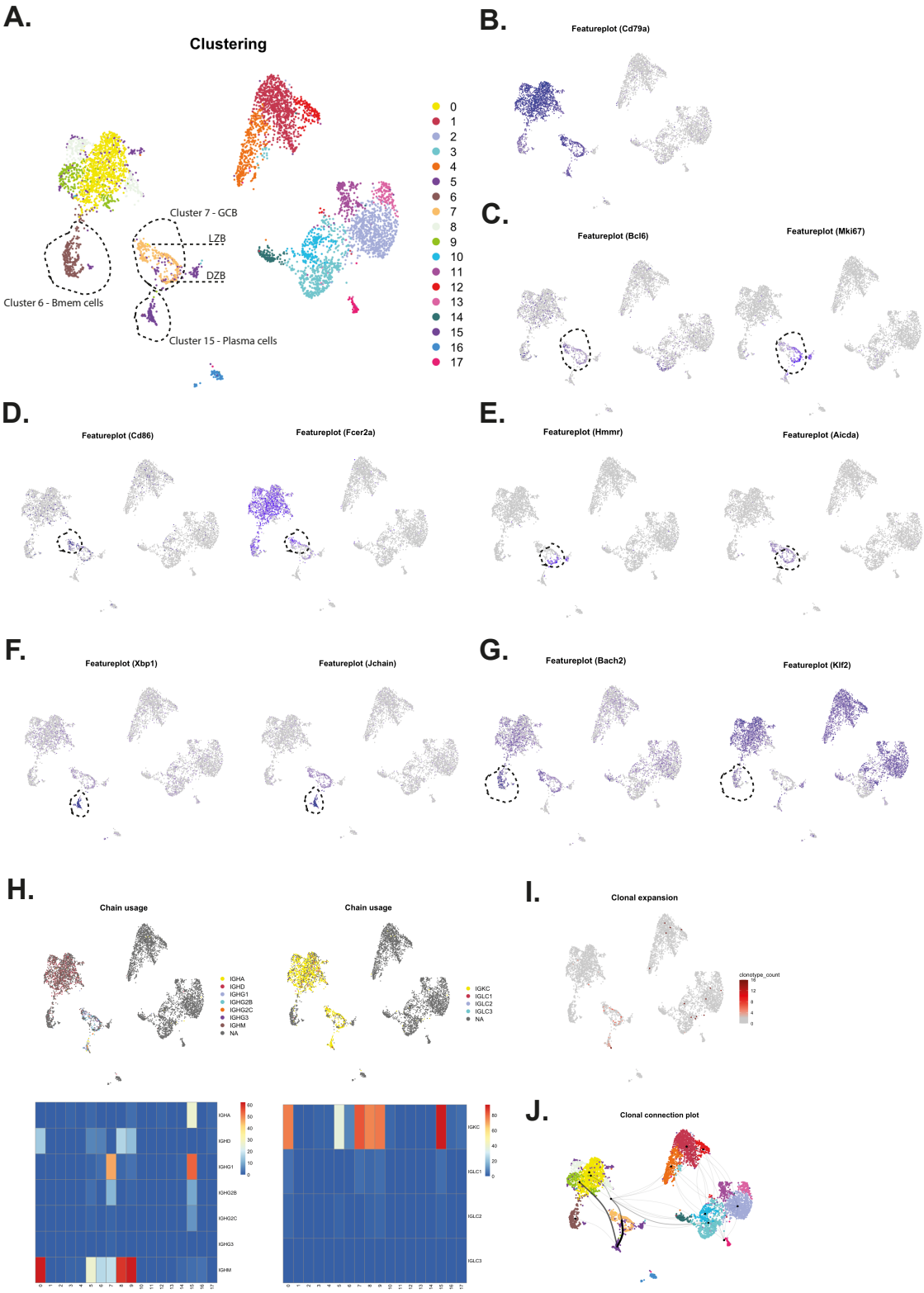

Figure 4: *Definition of B cell subsets using DALI.* ***(A)*** *Definition of the cluster containing different B cell subsets.* ***(B)*** *Definition of CD79a expression B cells* ***(C)*** *Definition of germinal centers B cells by Aicda and Mki67.* ***(D)*** *Definition of light zone B cells (LZB) by CD86 and Fcer2a expression.* ***(E)*** *Definition of dark zone B cells (DZB) by Hmmr and Aicda expression.* ***(F)*** *Definition of plasma cells by Xbp1 and Jchain expression.* ***(G)*** *Definition of memory B cells by Bach2 and Klf2 expression. All populations were defined using the transcriptomics view,* ***(H)*** *DimPlots and heatmaps depicting B cell receptor heavy and light chain constant region usage from the general view,* ***(I)*** *Clonotype expansion plot from the clone view,* ***(J)*** *Clonotype connection plot from the clone view.*

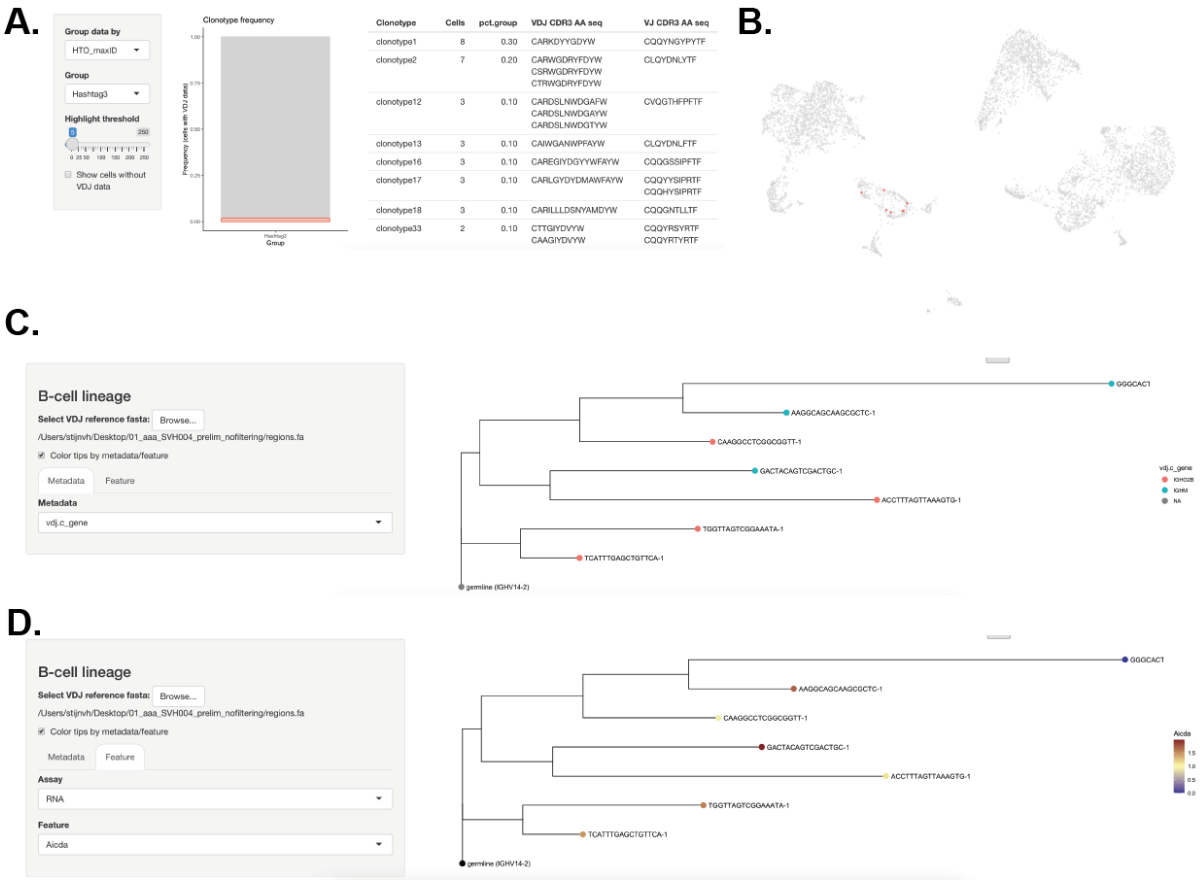

Figure 5: *Analysis of clonal evolution.* ***(A)*** *Definition of expanded clones using the clone view,* ***(B)*** *Localisation of selected clones on the UMAP plot (clonotype2 is highlighted),* ***(C)*** *Hierarchical clustering of clonotype2, with node ends highlighting the isotype class,* ***(D)*** *Hierarchical clustering of clonotype2, with node ends highlighting the Aicda expression levels.*

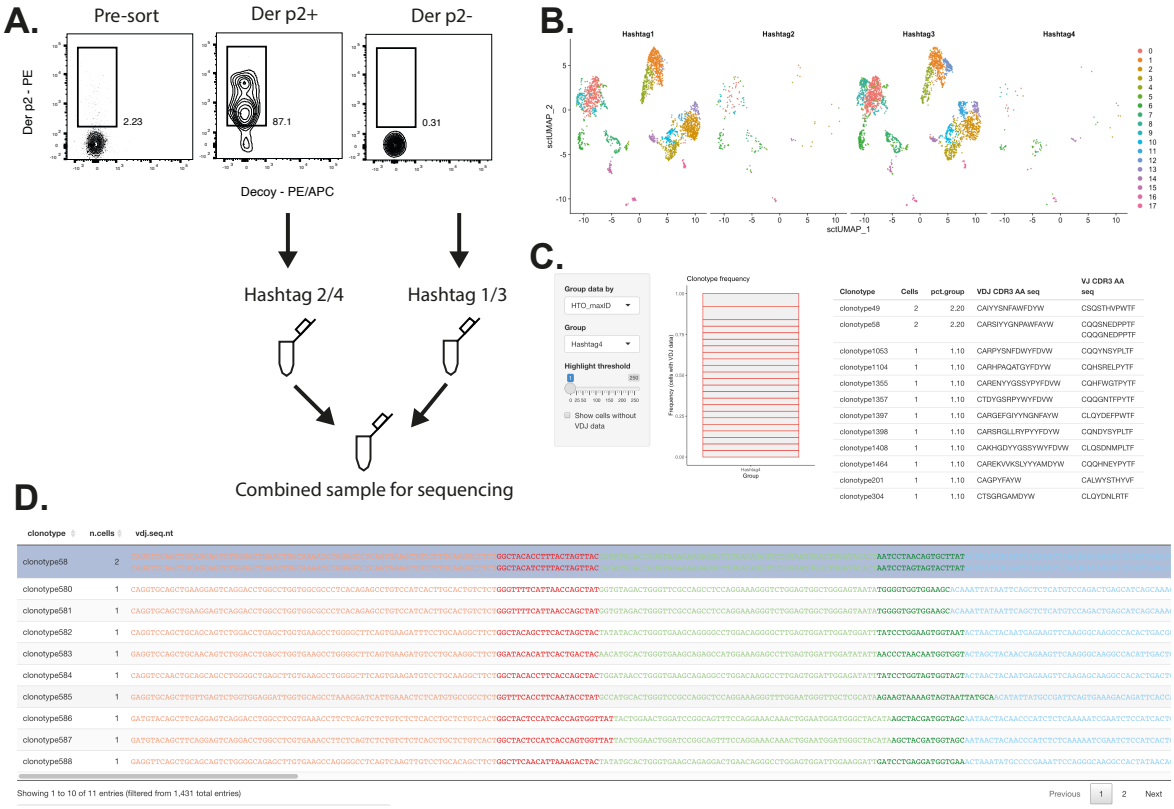

Figure 6: *Analysis of antigen reactive B cells.* ***(A)*** *Isolation of Der p2 reactive B cells from mediastinal lymph node by FACS using a Der p2 protein tetramer (FACS plots show CD19+ pregated cells), the sample hashing strategy is shown* ***(B)*** *Demultiplexing of the sample based on hashtag expression,* ***(C)*** *Definition of Der p2 reactive clones in sample 4,* ***(D)*** *Detailed information on one of the clonotypes shown in C.*

References

1. Miho E, Yermanos A, Weber CR, Berger CT, Reddy ST, Greiff V. Computational Strategies for Dissecting the High-Dimensional Complexity of Adaptive Immune Repertoires. Front Immunol. 2018;9:224.
2. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29.
3. Rubelt F, Busse CE, Bukhari SAC, Burckert JP, Mariotti-Ferrandiz E, Cowell LG, et al. Adaptive Immune Receptor Repertoire Community recommendations for sharing immune-repertoire sequencing data. Nat Immunol. 2017;18(12):1274-8.
4. Victora GD, Dominguez-Sola D, Holmes AB, Deroubaix S, Dalla-Favera R, Nussenzweig MC. Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas. Blood. 2012;120(11):2240-8.
5. Laidlaw BJ, Cyster JG. Transcriptional regulation of memory B cell differentiation. Nature reviews Immunology. 2021;21(4):209-20.
6. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2021.
7. Debeuf N, Haspeslagh E, van Helden M, Hammad H, Lambrecht BN. Mouse Models of Asthma. Curr Protoc Mouse Biol. 2016;6(2):169-84.
8. Lun AT, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Research. 2016;5:2122.