Title Page

Single cell multi-modal analysis with VDJPuzzle identifies clonality and gene expression profiles of T and B cells

AUTHORS:

Jerome Samir*[1], Simone Rizzetto*[1], Money Gupta*[1], Mandeep Singh[2], Katherine JL Jackson[2], Curtis Cai[1], Auda Eltahla[1], Rowena Bull[1], Joanne H Reed[2], Chris C Goodnow, [1,2] Fabio Luciani[1,2].

* Equally contributed

[1] School of Medical Sciences and Kirby Institute for Infection and Immunity, UNSW Sydney, Australia

[2] Garvan Institute of Medical Research, Sydney, Australia.

Correspondence to: Fabio Luciani, luciani@unsw.edu.au

Keywords: immune cells, scRNAseq, T cell receptor, B cell receptor, multi-omics

The tools are available on BitBucket:

https://bitbucket.org/kirbyvisp/vdjpuzzle

https://bitbucket.org/kirbyvisp/vdjview

ABSTRACT

VDJPuzzle and VDJView form a tool for simultaneous analysis and visualisation of gene expression and immune receptors from the scRNAseq data of both T and B cells, that allows accurate detection of secreted/membrane bound B cell receptors. Using VDJPuzzle on scRNAseq data from breast cancer and metastatic lymph-nodes, we identified T and B cell subsets with distinct patterns of hypermutation, isotype, and gene signature associated with immune cell differentiation and activation.

Immunological studies have revealed a surprisingly high level of heterogeneity between immune cells, even in those with same clonotype and surface phenotype suggesting that lymphocyte populations of apparently similar phenotype could have different functions[2]. With the advent of single cell RNA-sequencing (scRNA-seq), it is now possible to unravel the heterogeneity of T and B cells and link receptor clonotype diversity to the gene expression profile of the cell and to clinical or other meta-data.

Here, we report a novel bioinformatics pipeline, based on the well verified tool, VDJPuzzle[3], furnished with a new graphical user interface called VDJView. This framework delivers an integrated set of novel and publicly available tools to assemble T and B cell receptors (BCR and TCR), quantify gene expression from scRNA-seq data, and integrate and visualise the resulting clonal and transcriptomic data with clinical and other data such as sample origin, molecular definition of disease and cell surface phenotype (e.g. index sorting flow-cytometric data) (Fig. 1). VDJView can be utilised by researchers that do not have advanced bioinformatic skills to test hypotheses and explore the relationship between multi-modal single cell data-sets.

For B cells, we have developed a novel algorithm to identify the splice variants associated with secreted and membrane isoforms of the B cell receptor (BCR) (Supplementary Fig 1.1). This information can reveal unknown heterogeneity in B cell phenotype, for instance identifying plasmablast cells and novel B cell subsets if often unclear with current approaches based on flow-cytometric gating of B cells with intensity of expression of markers CD27, CD38 and surface Ig. VDJPuzzle also allows for accurate identification of isotypes and estimates the somatic hypermutation (SHM) level with respect to germline for each receptor sequence.

The shiny app, VDJView, accepts the output of VDJPuzzle, or other available gene profile and immune receptor formats, including those derived from 10X scRNA-seq 5' kit with TCR detection, and utilises statistical analysis and visualisation R packages, allowing the user to integrate the multi-omics data with cell metadata to facilitate the testing of specific hypotheses and the discovery and exploration of novel subpopulations (Fig. 1). A detailed list of the features and an overview of the pipeline is reported in Supplementary Note 1.

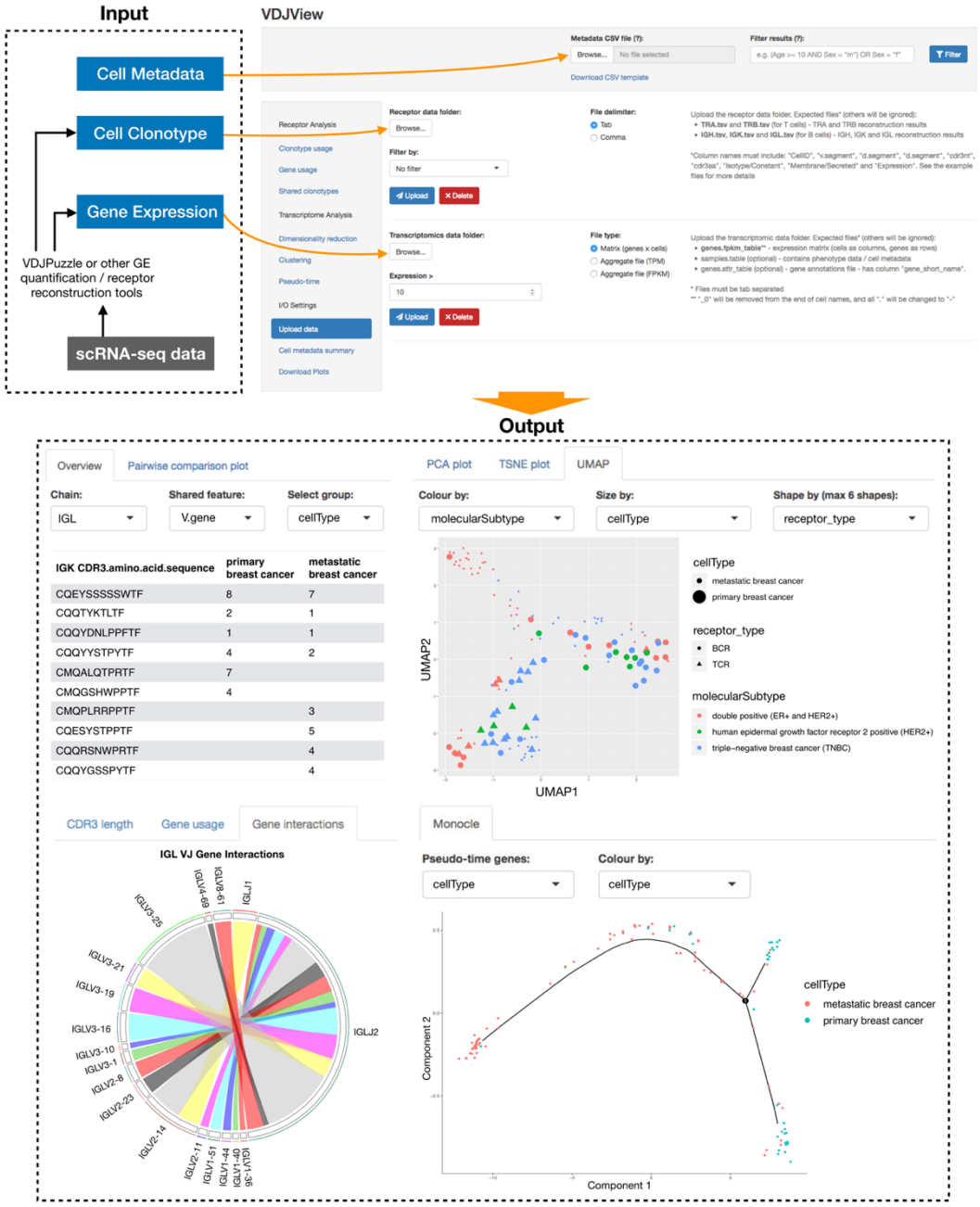

Figure 1 | Overview of VDJView. Top: VDJView upload page, showing where required (immune receptor sequences and gene expression matrix) and optional inputs (metadata) can be uploaded. **Bottom**: examples of analysis modules. The table (top left) shows a clonal expansion of IGL chains across primary breast tissue and metastatic lymph-node. The Circos plot (bottom left) shows the IgL V and J gene pairings identified[1]. Dimensionality reduction using UMAP (top right) shows a cluster of B cells derived from metastatic lymph-node in two patients with ER^+ $HER2^+$ breast cancer, while T and B cells from the primary breast cancer tissue had similar gene signature regardless of molecular subtype. Pseudo-time plot (bottom right) shows the inferred evolutionary trajectory between all immune cells determined by genes that differentiate primary from metastatic tissues in two subjects with matched samples.

VDJPuzzle identifies both membrane-bound and secreted isoforms of the BCR. The identification of these splice variants is performed by utilising a profile Hidden Markov Model (HMM) through the HMMER package[4] to map BCR constant region identified from scRNA-seq to a database of sequences corresponding to membrane isoform (see Supplementary Note 1 for further details). The advantage of this approach over the standard BLAST is the sensitivity with which sequences are searched and mapped using formal probabilistic framework that allows accurate identification of position-specific residue and statistically estimated scoring gaps based on a query profile. VDJPuzzle successfully identified the membrane and secreted isoform for each cell with a successful full-length BCR reconstruction in 117 human B cells with known phenotype[3] (Supplementary Table 2.7). These results were consistent with the phenotype of each cell, with transitional and naive B cells predominantly carrying a membrane variant, plasmablast carrying both variants, and memory with both forms identified, in line with the gating strategy which did not exclusively separate switched memory B cells (CD19+ IgD- CD27+ IgG+ CD38 +/-) from plasmablasts (CD19+ CD27+ IgD- IgG- CD38++). Similar success rates were observed with PW2 and mouse B cells[5] (Supplementary Tables 2.8 and 2.9). VDJPuzzle has been thoroughly tested and compared against other methods (BASIC[6] and BraCeR[7]), showing comparable results for the success rate in BCR reconstruction and also improved detection of isotype on human and mouse B cells[3] (Supplementary Tables 2.1 and 2.2). Notably, VDJPuzzle reports both transcripts in a cell when IgM and IgD isotypes are detected with identical Complementary Determining Region 3 (CDR3), as in naïve and transitional cell populations. When tested on a second set of human cells (89 plasmablast B cells from patient PW2[6]) and a set of 200 mouse cells, VDJPuzzle revealed a broader set of isotypes than those reported by BraCeR, identifying IgM and IgG2 isotypes (Supplementary Tables 2.3 – 2.6). In this second set of human cells, we identified 43% of the BCRs as being

membrane bound, which is consistent with the results found in the plasmablast cells from the first datasets.

To demonstrate the utility and novelty of VDJView, we analysed scRNA-seq data from the breast cancer tissue of 11 subjects, two of which had samples from associated metastatic lymph nodes[1]. We found 170 single B cells with at least one full-length H, L or K chain (hereby known as partially reconstructed BCR), of which 101 had a full-length heavy and light chain (hereby known as full-length BCR). Similarly, we found 42 single T cells with at least one full-length α or β TCR chain (partially reconstructed TCR), of which 30 had full-length TCR). We also found 33 cells with TCR and BCR chains, suggesting that they are likely contaminated or doublets, as VDJPuzzle has a low false positive identification rate (see Supplementary Note 2). The analysis of these immune cells revealed a highly diverse clonal repertoire (Supplementary Figure 2.1). The TCRαβ clone identified in subject BC03 (CAVGNNAGNNRKLIW_CASRSRDSSTGELFF, with germline TRAV8-3, TRAJ38, TRBV10-2 TRBD2, TRBJ2-2) was found in one cell from primary and one from metastatic tissues.
We identified 31 clones across the 101 full-length BCRs, with shared clonotypes between primary and metastatic tissues and also between subjects (Fig. 1 and Supplementary Table 2.10), suggesting that a common repertoire of B cells is involved in breast cancer with a migratory pattern between primary and metastatic tissues. B cells from the lymph nodes were more clonally diverse than those found in breast cancer, with 27 clones in 64 cells from the lymph node and 7 in 37 cells from the breast tissue. Notably, there was a significant overlap in gene usage between B cells across tissue types (Supplementary Table 2.11); for instance, IgHV4–59 was expressed in 6 B cells from breast cancer tissues, and 10 lymph node derived cells. This analysis also revealed a full-length BCR clone (CDR3

CACEELDVVW_CQEYSSSSSWTF, IgHV3-33, D2-15, J1, and IgKV1-5, J1) was found in 2 cells of BC09 breast tissue and 1 cell in metastatic tissue of BC03. This K chain (V1-5, J1) was expressed in 15 cells across subjects BC04, BC07, BC09 and BC03LN.

Through VDJView, we integrated immune receptor information with the gene expression profile and performed unsupervised clustering as well as a model-based differential expression analyses focussing on comparing B and T cells between tissues as well as between cancer molecular subtypes. Unsupervised clustering analysis using SC3 on all the B and T cells revealed evidence of 8 clusters, clearly demonstrating that immune cells form distinct clusters based on identity (B and T cells), isotype, tissue of origin and molecular subtype (Supplementary Figure 2.4).

To further elucidate the molecular signatures of these immune cells, we performed a model-based differential expression and clustering analyses utilizing MAST in VDJView. Specifically, we analysed gene profiles of immune cells by molecular subtypes of breast cancer (Fig. 2), tissue of origin (primary and metastatic, Supplementary Figure 2.6), and finally by cell identity (B and T cells, Supplementary Figure 2.7). Differential expression analysis comparing T and B cells between double positive (ER+ and HER2+) and triple negative breast cancers revealed distinct clusters of B cells by isotype and SHM levels (Fig. 2). We discovered B cells found in metastatic lymph nodes from the subject BC03 affected by double positive breast cancer tissues were consistent with plasmablast cells, as these cells expressed a secreted form of IgG1, elevated SHM, and expression of genes such as MZB1 and EAF2, and ELL3. This cluster was also confirmed in the analysis comparing immune cells between tissue of origin. In contrast, B cells found in the primary tissue in subjects with triple negative breast cancer were

mostly resting naïve or memory cells, with IgG1, IgG3 and IgM isotypes, expressing HLA DR markers, as well as CD99[8], thus implying that they function as antigen presenting cells (Fig. 2).

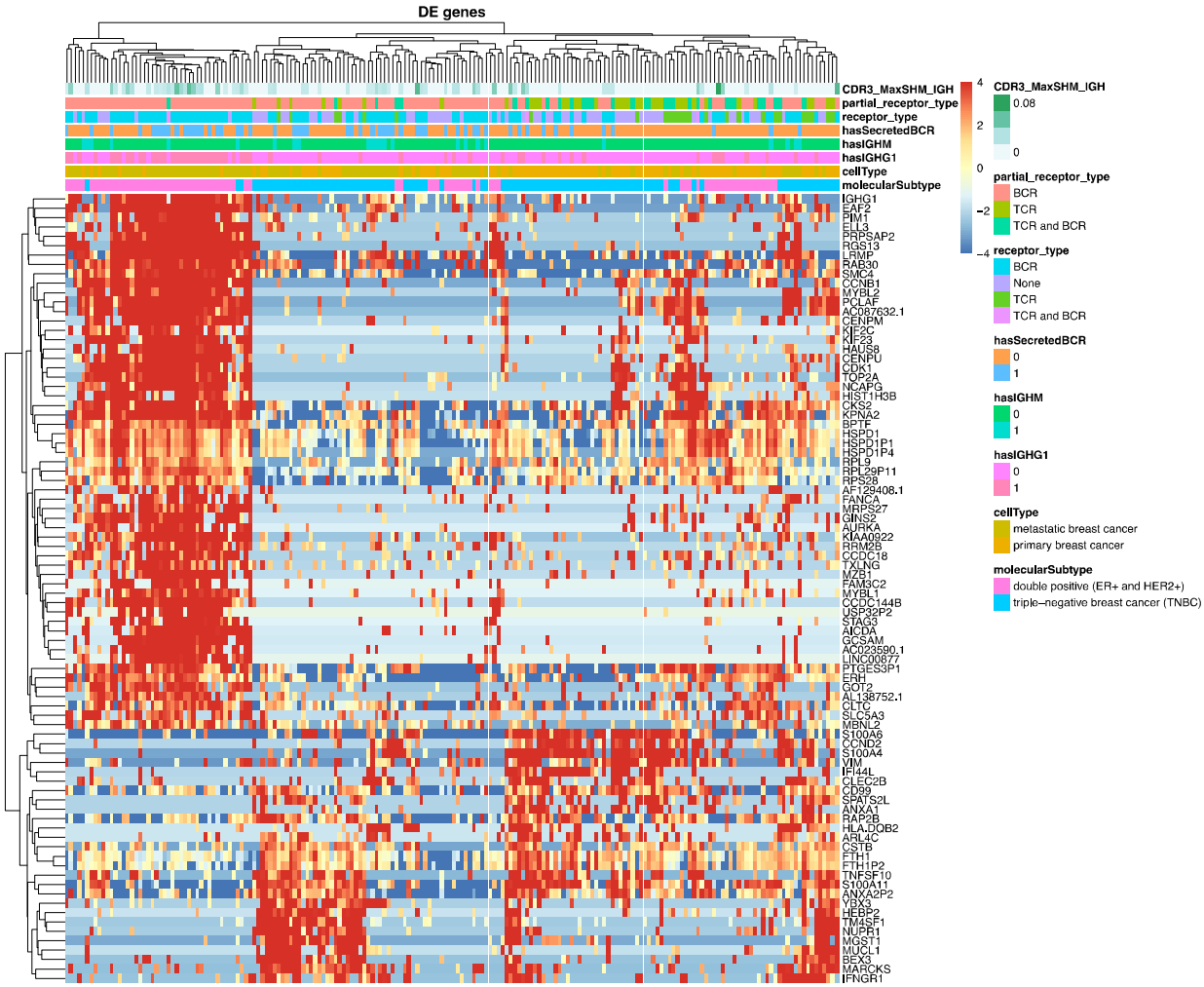

Figure 2 | **Heatmap of immune cells clustered by cancer molecular subtypes**. Supervised clustering depicting differentially expressed genes (fold change > 2) between ER$^+$HER2$^+$ and triple-negative breast cancer subtypes. The cells in the ER$^+$HER2$^+$ cluster (left) are primarily IgG1$^+$ cells with an elevated SHM, secreted BCR from the metastatic lymph node tissue, and have gene expression profiles consistent with plasmablast B cells.

We have shown that integrating immune receptor and gene expression data with clinical information is useful to discover novel, biologically relevant findings from published data that

do not emerge through previous analyses and to further understand and discover medically relevant mechanisms. VDJPuzzle and VDJView form an integrated set of known and novel tools that have a flexible design, expanding other tools and providing a robust quantitative framework to generate and study multi-omic immune cell data at the single cell level. The proposed framework can be utilised by bioinformatics experts to develop and integrate new tools, as well as by clinical scientists and immunologists without profound knowledge of bioinformatics tools.

References

1 Chung, W. *et al.* Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat Commun* **8**, 15081, doi:10.1038/ncomms15081 (2017).

2 Kaech, S. M. & Cui, W. Transcriptional control of effector and memory CD8+ T cell differentiation. *Nat Rev Immunol* **12**, 749-761, doi:10.1038/nri3307 (2012).

3 Rizzetto, S. *et al.* B-cell receptor reconstruction from single-cell RNA-seq with VDJPuzzle. *Bioinformatics* **34**, 2846-2847, doi:10.1093/bioinformatics/bty203 (2018).

4 Wheeler, T. J. & Eddy, S. R. nhmmer: DNA homology search with profile HMMs. *Bioinformatics* **29**, 2487-2489, doi:10.1093/bioinformatics/btt403 (2013).

5 Wu, Y. L., Stubbington, M. J., Daly, M., Teichmann, S. A. & Rada, C. Intrinsic transcriptional heterogeneity in B cells controls early class switching to IgE. *J Exp Med* **214**, 183-196, doi:10.1084/jem.20161056 (2017).

6 Canzar, S., Neu, K. E., Tang, Q., Wilson, P. C. & Khan, A. A. BASIC: BCR assembly from single cells. *Bioinformatics* **33**, 425-427, doi:10.1093/bioinformatics/btw631 (2017).

7 Lindeman, I. *et al.* BraCeR: B-cell-receptor reconstruction and clonality inference from single-cell RNA-seq. *Nat Methods* **15**, 563-565, doi:10.1038/s41592-018-0082-3 (2018).

8 Park, C. K., Shin, Y. K., Kim, T. J., Park, S. H. & Ahn, G. H. High CD99 expression in memory T and B cells in reactive lymph nodes. *J Korean Med Sci* **14**, 600-606, doi:10.3346/jkms.1999.14.6.600 (1999).

Acknowledgement

FL acknowledges funding from NHMRC (APP1121643 project grant and Career Development fellowship APP1128416). JS, SR and MG have PHD Scholarships from UNSW. NHMRC Project Grant 1142186 to JHR and KJLJ, NSW Health Early to Mid Career Fellowship to JHR.

Author contribution

FL and SR designed the research. SR, JS, and MG analyzed the data and wrote the code. AE, RB provided reagents. AE, MS, JHR generated the experimental scRNA-seq data. KJLJ, CCG, CC, JHR contributed to the analysis of the data. FL, SR, MG and JS wrote the manuscript. All authors have read the manuscript and provided feedback.