Title:

T cell receptor-centric perspective to multimodal single-cell data analysis.

Authors:

Kerry A. Mullan[1,2,3*], My Ha[2,4,5], Sebastiaan Valkiers[1,2,3], Nicky de Vrij[1,2,3,6], Benson Ogunjimi[2,4,5,7], Kris Laukens[1,2,3] and Pieter Meysman[1,2,3*]

Affiliations:

[1]Adrem Data Lab, Department of Computer Science, University of Antwerp, Antwerp, Belgium

[2]Antwerp Unit for Data Analysis and Computation in Immunology and Sequencing (AUDACIS), Antwerp, Belgium

[3]Biomedical Informatics Research Network Antwerp (Biomina), University of Antwerp, Antwerp, Belgium

[4]Antwerp Center for Translational Immunology and Virology (ACTIV), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

[5]Centre for Health Economics Research and Modelling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium

[6]Clinical Immunology Unit, Department of Clinical Sciences, Institute of Tropical Medicine, Antwerp, Belgium

[7]Department of Paediatrics, Antwerp University Hospital, Antwerp, Belgium

* Corresponding authors: Pieter Meysman, pieter.meysman@uantwerpen.be

Abstract

The T cell receptors (TCR) carry critical information regarding T cell functionality. Yet is often overlooked in single cell experiments, which commonly focus on gene expression (GEx)-first analysis. Here we propose a dedicated TCR-first analysis for single T cell data, implemented as the STEGO.R package. A curated T cell atlas was built using STEGO.R from 12 prominent human studies, containing in total 500,000 T cells spanning multiple diseases, including melanoma, head-and-neck cancer, T-cell cancer, and lung transplantation, allowing the reanalysis of these datasets. The TCR-first approach was able to highlight dynamic features in hyperexpanded clones and treatment-specific changes, which were missed in the original GEx-first analyses. This meta-analysis also revealed previously unknown public T-cell clusters with common gene signatures and antigen specificities. Overall, adopting the TCR-first approach revealed T-cell features often missed by conventional GEx-focused methods and the STEGO.R tool facilitates this paradigm shift.

Introduction

T cells are specialized cells of the adaptive immune system positioned to recognize various infected or otherwise aberrant cells. This recognition is driven by the T-cell receptor (TCR) expressed on the cellular surface. Their importance for immunology, and various underlying disease states, have driven recent innovations in the domain of single-cell sequencing. We can now capture a multimodal view of a single T cell that includes the transcription gene expression (GEx), as well as the hypervariable TCR . This detailed perspective allows distinction of the T cell population into its phenotypes with reference to the TCR. The broadest separation is a distinction between CD4+ T cells, which recognize the class II major histocompatibility complex (MHC) loaded with extracellular peptide sources, and CD8+ T cells that recognize the class I MHC loaded with intracellular peptides. CD4+ T cells exhibit diverse functional phenotypes, categorized into T helper (Th) subtypes such as Th1, Th2, Th9, Th17, follicular T cells (Tfh), and regulatory T cells (Tregs). T cells also display a variety of memory states and can transition from activation to senescence or exhaustion. In addition to their expression markers, T cells are divided into two principal lineages based on the T cell receptor (TCR), namely alpha-beta (αβ) or gamma-delta (γδ). Each of these four TCR chains is made up of a Variable (V), Diversity (D) and Junction (J) genes, with alpha/gamma having VJ and beta/delta VDJ arrangements. During development a T cell will acquire a quasi-random recombination of V(D)J genes, which results in a hypervariable sequence in the complementarity determining region 3 (CDR3), the primary driver of variable antigen specificity. Recent methodological advances have enhanced the analysis of bulk TCR repertoire data, with novel tools based on identifying antigen-specific TCR groups (e.g., ClusTCR, tcrdist3, GIANA, ismart, and anchor clustering) and epitope annotation models (TCRex , ImmuneWatch™ DETECT, ImRex , TCRGP , NetTCR), however thus far these approaches have had limited adoption for single-cell data.

A critical challenge involves the analysis of single-cell data with not one but two key data levels. Conventionally, the data analysis pipelines have been developed for bulk TCR-seq and GEx experiments alike. As a result, both data layers are viewed in parallel rather than concurrently. Nonetheless, efforts have been made to leverage both modalities, including CoNGA and mvTCR . Those approaches that do integrate both levels build on the assumption that they are alternative representations of the same information, i.e. one TCR translates to one function , which does not match current immunological knowledge . In addition, most analysis pipelines simply extend on the existing methods for GEx analysis (e.g., Seurat , scanpy , DALI), with the TCR often seen as a secondary aspect to annotate the gene expression clusters . This GEx-first approach relies on the accuracy of T cell phenotyping, which lacks robustness and is often not reproducible across studies due to subtle dynamic differences that drive the T cell phenotype . The TCR is then solely used as an index marker for clonal expansion, ignoring the information contained within, with respect to immunological convergence or antigen specificity.

Here, we compiled a T cell atlas of 12 publicly available datasets containing ~500,000 cells with both TCR and GEx and aimed to determine the utility of a TCR-first analysis. This required the development of a novel software tool coined ‘Single-Cell T-cell Receptor and Expression Grouped Ontologies’ (STEGO.R) to resolve many of the described analysis issues. Switching to a TCR-first approach, allows identification of (1) dynamic T cell states within a single clone, (2) epitope/disease-specific signatures, and (3) functional similarity between T-cells from different lineages. These findings resulted in a unique single T cell atlas, covering

49 distinct T cell phenotypes. Overall, we demonstrate that the TCR-first approach reduces biases, increases statistical power, and improves experimental design.

Results

Clones of the same T cell can have dynamic gene expression features.

To demonstrate the phenotypic heterogeneity of T cell clones with identical T cell receptor sequences, we used STEGO.R to assess the dynamics of the most prevalent clone in the 12 datasets [19-32]. Here, we identified a highly expanded clone (n = 36,430), TRAV12-2.TRAJ9_CAATTGGFKTIF & TRBV20-1.TRBD2.TRBJ2-4_CSATEGNIQYF, originating from the T-cell large granular lymphocyte (LGL) leukemia dataset [27] (**Supplementary Table S1**). While the clone featured a predominant CD8+ effector-like phenotype, its cells featured a mix of phenotypes that could be annotated with multiple T cell functions (**Figure 1A**). The gene expression patterns indicate that the cells of this TRAV12-2/TRBV20-1 clone are present in various states, from early activated to terminally differentiated senescent or even exhausted, and with a few cells undergoing cell cycling (**Figure 1B**). To assess the function of a T-cell Th1 clone, a cytokine panel of IFNγ (IFNG) and TNF would be commonly used in experimental biology assays. Therefore, we aimed to further characterize the TRAV12-2/TRBV20-1 clone through the measured gene expression of IFNG and TNF. This T cell clone had a mixture of Th1 cytokine expression with subpopulations either being IFNG+TNF+ (3.8%), IFNG+ (46.3%) or TNF+ (2.6%) (**Figure 1C**). As expected, there are distinct gene expression profiles associated with each of the Th1 cytokine phenotypes of this clone (**Figure 1D**). This illustrates how a single T cell clone can have multiple functionalities depending on which cytokines they co-express.

The original purpose of the LGL study was to block CD52 and cause apoptosis of the over-proliferated T cells. While some cases fully responded to treatment, the anti-CD52 (alemtuzumab) only caused a partial response in some patients. The proliferated TRAV12-2/TRBV20-1 T cell clone was from one of these partial responders, namely UPN4 (as described in **Figure 1**), and the clone was present both pre- and post-treatment (**Supplementary Figure S1A**). The post-treatment average phenotype of this clone displayed a reduced expression of the CD52 transcript and maintained a cytotoxic transcriptional expression (**Supplementary Figure S1A-B**). Therefore, this T cell population appeared to escape immune therapy by loss of the CD52 target and lacked the apoptosis signals found in the responders (**Extended results R1.1**) These observations, offering a possible explanation for the partial response of UPN4, were missed in the original GEx-centric analysis, as the diverse expression patterns of this critical clone diluted its overall signal.

Identification of dynamic TCR-related signatures in a hyperexpanded clone.

Despite the dynamic diversity of individual T-cell clones, a TCR-specific approach can still highlight consistent phenotypic trends across samples. To illustrate this concept, we focused on a head and neck squamous cell carcinoma (HNSCC) dataset. In the original study, the authors state a clear narrative based on all single-cell and experimental data that PD-1 may be a potential target to boost the T cell immunity in HNSCC [26].

Reanalysis of the T-cell receptor HNSCC data with STEGO.R revealed only a small number of public clones, which exhibited limited clonal expansion and lacked specificity towards disease (**Figure 2A; Supplementary Table S2**). Subsequently, we investigated the presence

of clones shared across various treatments and conditions within individual patients and their clonal expansion. Specifically, this dataset included epitope specificity for TCR sequences able to recognize one of two HPV-epitopes HLA-A*01:01–HPV $E2_{329–337}$ (peptide sequence KSAIVTLTY;KSA), and HLA-A*01:01–HPV $E2_{151–159}$ (peptide sequence QVDYYGLYY; QVD) [26]. Patient HPV34 featured KSA- or QVD-associated TCRs in their primary tumor and metastatic lymph node (metLN) samples (**Figure 2A**), although they were not necessarily the most expanded clones (**Supplementary Table S2**). Furthermore, several of these HPV-specific TCRs were in both the PD-1 positive sorts as well as the epitope-specific sorts KSA (n=1; **Figure 2B**) and QVD (n=26; **Figure 2C; Supplemental Table S3**).

The KSA-specific and PD-1-sorted TCR (n=1) presented an CD8+ effector phenotype (cytotoxic profile) (**Figure 2Bi**). This TCR's associated phenotype was marked by the expression of both CD39 (ENTPD1) and PD-1 (PDCD1), with both markers only being expressed in <50% of the population. However, there were markers that had improved expression coverage (e.g., >90% of total cells) including LINC02446, CXCL13, TRIR, HLA-class II genes, RACK1 (**Figure 2Bii; Supplementary Table S4**). Next, we pooled the 26 clones of the QVD-specific TCRs that overlapped with the PD-1 sort. Like the KAS-specific clone, this pooled QVD population was mostly comprised of CD8+ effector like phenotype (**Figure 2Ci**) and identified similar KAS transcripts including LINC02446, TRIR, HLA-class II genes, CXCL13, and RACK1 (**Figure 2Cii; Supplementary Table S5**). Therefore, at least within this individual, there was a fairly consistent transcriptional profile of HPV-specific T cells in both the primary tumor and metLN and suggesting that they are performing the same function i.e., recognition of HPV-epitopes expressed in the cancerous tissue. Overall, starting with the epitope-specific TCRs, we were able to identify a T cell signature with consistent transcriptional markers, that were not apparent from the GEx cluster-based analysis.

3.3 Sequence similarity identifies disease-specific features in a melanoma cancer dataset

Despite the added benefit of utilizing the TCR to derive the primary analysis, a common complication is that most TCRs are infrequent and private. In addition, all T cells express at least two distinct TCR chains (commonly TCRαβ or TCRγδ), which further complicates this analysis. To overcome these challenges, we propose the use of TCR clustering to group those sequences with high similarity into T-cell groups. Prior research has shown that high TCR sequence similarity within a population is often indicative of a common epitope response [33]. To demonstrate this approach, we reanalyzed a single-cell data set of colitis complication post-melanoma treatment as it was the only one with all four TCR chains (αβTCR and γδTCR) [22]. This dataset was comprised of eight colitis (C), six non-colitis (NC) patients and eight healthy controls (CT). For details on the overall TCR structure and common clusters, please see the **Extended Results R1.1**.

Interrogation of the clusters identified several potential disease-specific motifs (**Supplementary Table S6 and S7**). One TRAG cluster with TRAV13-2-TRAJ45 was present in all three conditions, with more sequences featuring in the CT and NC that had a mixture of effector CD8ab+ T cells and activation/cytotoxic features (**Figure 3A**). Interestingly, a TRGV4 gamma cluster associated with the C and NC cases, were CD8ab+ T cells with cytotoxic/activation status, with strong indications that they were melanoma-specific (**Figure 3B**). Importantly we also identify an alpha cluster with TRAV29/DV5-TRAJ40 that was over-represented in the colitis cases (**Figure 3C**), which may thus be a disease-specific signal missed

in the original analysis. The TCR beta chains contained fewer clusters of interest, which were mostly specific to singular individuals (**Supplementary Figure S2A; Supplementary Table S7**). Interestingly, due to the similar gene expression profile i.e., both CD8ab+ T cells with cytotoxic features, the gamma clusters (γδ T cells) were functionally indistinguishable from the αβ T cells. If these cells were annotated based on the GEx alone, they would have been incorrectly assigned as αβTCRs. Thus, in this instance, capturing all four TCR chains and using this information for interrogation rather than the GEx, identified novel disease specific features.

Global TCR motifs associate with invariant T-cells and high generation probability.
The last analysis was to interrogate the global patterns across all 12 datasets with STEGO.R [19-32]. We describe some of the annotation differences observed in the dataset in **Extended Results R1.2** as well as the limited public clone interrogation in **Extended Results R1.3**. This below section focuses on public clusters common to all 12 datasets.

There were three alpha clusters (**Supplementary Table S8**) and three beta clusters (**Supplementary Table S9**) common to all 12 datasets. The top alpha cluster had a junctional length of 12 amino acids and TRAV1-2 TRAJ33 gene arrangement and was present in 94 of 144 samples. This cluster had 104 unique sequences with 2304 total clones. The DETECT tool predicted 98.4% of the total cells were highly (Score >0.23) likely to interact with MR1:5-OP-RU and indicated that this cluster may be the mucosal invariant T cells (MAIT) (**Figure 4A**). When examining the average GEx profile of this cluster, identified that ~75% of cells expressed the TRAV1-2 along with KLRB1, which are commonly used to identify this population in the GEx alone (**Figure 4B**).

The two remaining alpha clusters, TRAV8-3 (14mer) and TRAV27 (15mer) were not functionally distinct from the background data (**Supplementary Figure S3A and S3B)** and there were no consistent epitopes predicted for majority of TRAV8-3 or TRAV27 cells (**Supplementary Table S8**). To understand if these sequences were also more likely to occur, we evaluated their recombination probabilities using OLGA [34]. All three sequences fell below the geometric mean of all TRA (**Figure 4C**). All three beta-clusters contained TRBV20-1, had a junction length of 13 and lacked diversity gene inserts (**Supplementary Figure S3C to E**). These TRBV20-1 had a wide variety of epitopes predicted but were inconsistent within each cluster (**Supplementary Table S9**). TRBV20-1 was the most common V gene identified ~17.1% of the total dataset. [34] These three TRBV20-1 TCR CDR3 sequences displayed a high probability of generation (**Figure 4D**) as per OLGA [34]. Therefore, these sequences have a high probability of generation, which may explain their prevalence across the different diseases.

Discussion

T cells are a complex cell type to interrogate due to their two-dimensional diversity in both the GEx and TCR. Under the current paradigm, GEx information is conventionally used to identify cell type clusters, which are subsequently labeled with features of the TCR. To shift the process to a TCR-first approach required the development of a novel tool STEGO.R to uncover actionable TCR-specific features. Here we re-interrogated 12 publicly available T cell datasets and found that a TCR-first approach was able to identify patterns within the data that were previously not identified from a GEx-centralized approach. First, we observed a dynamic T cell response across clonal populations, indicating that the phenotypic behavior of a single clonotype may vary greatly depending on the molecular context. In addition, we highlight the dynamic changes in the profile of immunodominant clones spanning multiple samples. The use

of TCR clustering analysis and epitope annotation enabled us to identify invariant or disease-specific populations, thereby prioritizing actionable targets for deeper investigation. Furthermore, we recognized the need for alternative annotations strategies for better defining T cell sub-populations (**Extended discussion**). We also noted the limited utility of public clones and the importance of sequencing all TCR chains (**Extended discussion**).

This work suggests that several of the current assumptions for scRNA-seq with TCR-seq cannot be met. Perhaps the most crucial one is the idea that a single TCR clone only has one biological function for the entire population. We showcased that this assumption is unlikely to be true, by interrogating a hyperexpanded clone from the partial anti-CD52 responder UPN4 from LGL[27](**Extended discussion**). Our results demonstrate that the assumption that one TCR linearly correlates with one function cannot be met, which results in a misinterpretation of the dynamic and disease-specific features of T cells. This is further complicated by the small GEx subtleties that drive the T-cell phenotype. Common single-cell annotation strategies rely on clustering similar profiles and contrasting each cluster with the remainder of the data set. This can often miss small perturbations that are important for the T-cell phenotype and necessitates the presence of other T cells of different phenotypes. The latter is often not the case if the cells have been sorted (**Extended discussion**). For this reason, we built a novel T-cell atlas derived from 12 independent datasets. By testing the annotation with 143 samples with 1.13M cells (sorted on either CD3+ or CD45+), we achieved high coverage of most known T cell populations. Collectively, we were able to detect and annotate 49 distinct T cell phenotypes (**Extended discussion**). To our knowledge, this is the first and largest cell atlas that uses a joint GEx and TCR approach for annotation purposes. This atlas can be freely integrated into novel studies to aid in T cell annotation, even when TCR data is unavailable.

Another common limitation when interrogating TCR in a GEx-centric approach is to assume that only the most abundant clones will be useful to interrogate. For example, in the original HNSCC dataset we reanalyzed, the authors looked at only a handful of expanded clones, which, despite being public, exhibited no discernable specificity towards the disease [26]. However, as we showed for the HNSCC data, disease-relevant clones may be hidden in the less-expanded to non-expanded portion of the repertoire and were identifiable from the sample overlap that highlighted their distinct gene expression profile (**Extended discussion**). In addition, we are not limited to studying individual clonal abundances, but can rely on the phenomenon of TCR sequence convergence following epitope stimulation, both within and across individuals. This relaxes the strict constraint of public TCRs across samples, which are very rare, to common patterns contained within the TCR, as is common for bulk TCR data analysis [2, 35].

Finally, we observe significant deficiencies in the definition of GEx-based functional annotation for T cells. Current annotation protocols for T cell subtypes rely solely on the GEx information level, e.g., MAIT cells identified by TRAV1-2 and SLC4A10 or KLRB1 [36, 37]. Yet, we identified that GEx alone is insufficient for accurate identification of innate αβTCR, adaptive γδTCR and innate γδTCR populations. To find the MAIT population, we used the sequence similarity and epitope annotation tools with almost 100% predicted to be MR1-restricted. Interrogation of this cluster's average GEx identified up to 25% were missing the expected TRAV1-2 and KLRB1 expression. Additionally, some TRAV1-2+ T cells could possibly be CD1b restricted if recombined with TRAJ9 [38]. In addition to this improved definition of MAIT populations, we also identified that γδTCR can have similar profiles to the

adaptive CD8ab+ αβTCR and suggests similarity of function (**Extended discussion**). In this case, the TCR was superior to GEx for finding these populations. We strongly suggest future single-cell T cell research to include both GEx and TCR-seq with all four T cell chains, improving the accuracy in defining T cell subpopulations.

Overall, a TCR-seq first approach has identified novel features within the reanalyzed datasets that were missed with standard GEx first approach. While GEx approaches are appropriate for single modality data, this is not the most robust way to mine T cell-based single-cell data with the dual modalities. Collectively, our analyses illustrate that by shifting the paradigm towards the TCR, one can uncover novel insights into defining T cell subtypes, as well as the identification of disease-associated TCR patterns. We believe this paradigm shift to a TCR-first approach will enhance our understanding of T cell biology and push forwards novel therapeutic opportunities.

Funding
This work has been made possible by grant number 2022-249472 from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation. In addition, this work was supported by the Research Foundation Flanders [FWO: 1S40321N to S.V. 1S71721N to N.d.V.].

Conflict of interest
BO, KL and PM hold shares in ImmuneWatch BV, an immunoinformatics company, which developed IMW DETECT.

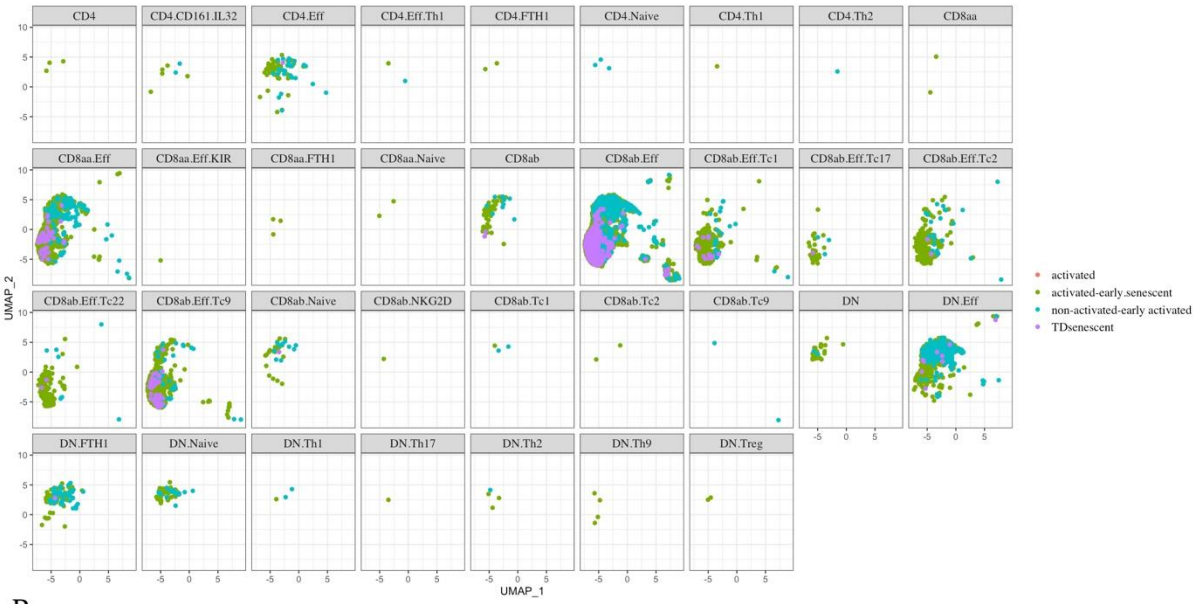
Acknowledgements
KAM would like to thank: Dr. Nicole A Mifsud, Dr. Martin Davey, and Dr. Alexandra Sharland for providing access to their BD Rhapsody datasets to refine QC pipeline. The authors would also like to thank Romi Vandoren, Sofie Gielis, Fabio Affaticati, and Briek van den Born for testing the app and aiding in the development of the STEGO.R tutorial.

References
1. Valkiers, S., et al., *Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing.* ImmunoInformatics, 2022. **5**.
2. Valkiers, S., et al., *ClusTCR: a python interface for rapid clustering of large sets of CDR3 sequences with unknown antigen specificity.* Bioinformatics, 2021. **37**(24): p. 4865-4867.
3. 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, 2021. **10**: p. e68605.
4. Hudson, D., et al., *A comparison of clustering models for inference of T cell receptor antigen specificity.* ImmunoInformatics, 2024. **13**.
5. Chang, H., et al., *Anchor Clustering for million-scale immune repertoire sequencing data.* BMC Bioinformatics, 2024. **25**(1): p. 42.
6. Gielis, S., et al., *Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires.* Front Immunol, 2019. **10**: p. 2820.
7. Moris, P., et al., *Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification.* Brief Bioinform, 2021. **22**(4).
8. Jokinen, E., et al., *Predicting recognition between T cell receptors and epitopes with TCRGP.* PLoS Comput Biol, 2021. **17**(3): p. e1008814.

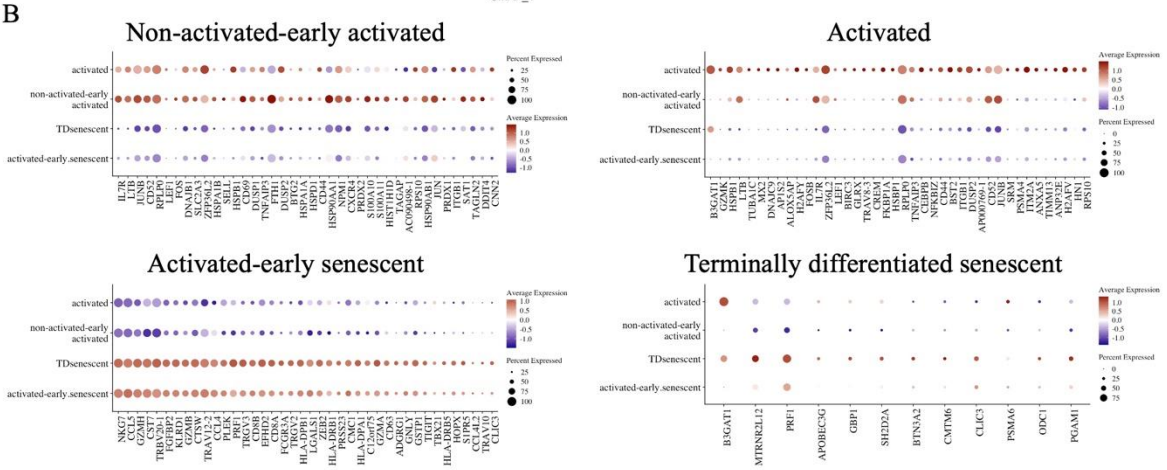
9. Montemurro, A., et al., *NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRalpha and beta sequence data.* Commun Biol, 2021. **4**(1): p. 1060.
10. Rosati, E., et al., *Overview of methodologies for T-cell receptor repertoire analysis.* BMC Biotechnol, 2017. **17**(1): p. 61.
11. Schattgen, S.A., et al., *Integrating T cell receptor sequences and transcriptional profiles by clonotype neighbor graph analysis (CoNGA).* Nat Biotechnol, 2022. **40**(1): p. 54-63.
12. Drost, F., et al., *Integrating T-cell receptor and transcriptome for 4 large-scale single-cell immune profiling analysis.* BioRxiv, 2022.
13. Becattini, S., et al., *Functional heterogeneity of human memory CD4+ T cell clones primed by pathogens or vaccines.* Science, 2015. **347**(6220): p. 400-406.
14. Hao, Y., et al., *Dictionary learning for integrative, multimodal and scalable single-cell analysis.* Nat Biotechnol, 2024. **42**(2): p. 293-304.
15. Wolf, F.A., P. Angerer, and F.J. Theis, *SCANPY: large-scale single-cell gene expression data analysis.* Genome Biol, 2018. **19**(1): p. 15.
16. Verstaen, K., et al., *DALI (Diversity AnaLysis Interface): a novel tool for the integrated analysis of multimodal DALI (Diversity AnaLysis Interface): a novel tool for the integrated analysis of multimodal.* BioRxiv, 2022.
17. Borcherding, N., N.L. Bormann, and G. Kraus, *scRepertoire: An R-based toolkit for single-cell immune receptor analysis.* F1000Res, 2020. **9**: p. 47.
18. Mullan, K.A., et al., *Current annotation strategies for T cell phenotyping of single-cell RNA-seq data.* Front Immunol, 2023. **14**: p. 1306169.
19. Azizi, E., et al., *Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment.* Cell, 2018. **174**(5): p. 1293-1308 e36.
20. Wu, T.D., et al., *Peripheral T cell expansion predicts tumour infiltration and clinical response.* Nature, 2020. **579**(7798): p. 274-278.
21. Banta, K.L., et al., *Mechanistic convergence of the TIGIT and PD-1 inhibitory pathways necessitates co-blockade to optimize anti-tumor CD8(+) T cell responses.* Immunity, 2022. **55**(3): p. 512-526 e9.
22. Luoma, A.M., et al., *Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy.* Cell, 2020. **182**(3): p. 655-671 e22.
23. Zheng, Y., et al., *Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment.* Nat Commun, 2020. **11**(1): p. 6268.
24. Han, L., et al., *Interleukin 32 Promotes Foxp3(+) Treg Cell Development and CD8(+) T Cell Function in Human Esophageal Squamous Cell Carcinoma Microenvironment.* Front Cell Dev Biol, 2021. **9**: p. 704853.
25. Mahuron, K.M., et al., *Layilin augments integrin activation to promote antitumor immunity.* J Exp Med, 2020. **217**(9).
26. Eberhardt, C.S., et al., *Functional HPV-specific PD-1(+) stem-like CD8 T cells in head and neck cancer.* Nature, 2021. **597**(7875): p. 279-284.
27. Gao, S., et al., *Single-cell RNA sequencing coupled to TCR profiling of large granular lymphocyte leukemia T cells.* Nat Commun, 2022. **13**(1): p. 1982.
28. Snyder, M.E., et al., *Modulation of tissue resident memory T cells by glucocorticoids after acute cellular rejection in lung transplantation.* J Exp Med, 2022. **219**(4).
29. Shibata, T., et al., *Expansion of Human Papillomavirus-Specific T Cells in Periphery and Cervix in a Therapeutic Vaccine Recipient Whose Cervical High-Grade Squamous Intraepithelial Lesion Regressed.* Front Immunol, 2021. **12**: p. 645299.
30. Gate, D., et al., *CD4(+) T cells contribute to neurodegeneration in Lewy body dementia.* Science, 2021. **374**(6569): p. 868-874.

31. Cheon, I.S., et al., *Immune signatures underlying post-acute COVID-19 lung sequelae.* Sci Immunol, 2021. **6**(65): p. eabk1741.
32. Penter, L., et al., *Coevolving JAK2V617F+relapsed AML and donor T cells with PD-1 blockade after stem cell transplantation: an index case.* Blood Adv, 2021. **5**(22): p. 4701-4709.
33. Meysman, P., et al., *On the viability of unsupervised T-cell receptor sequence clustering for epitope preference.* Bioinformatics, 2019. **35**(9): p. 1461-1468.
34. Sethna, Z., et al., *OLGA: fast computation of generation probabilities of B-and T-cell receptor amino acid sequences and motifs.* Bioinformatics, 2019. **35**(17): p. 2974-2981.
35. Hudson, D., et al., *Can we predict T cell specificity with digital biology and machine learning?* Nature Reviews Immunology, 2023. **23**(8): p. 511-521.
36. Andreatta, M., A.J. Berenstein, and S.J. Carmona, *scGate: marker-based purification of cell types from heterogeneous single-cell RNA-seq datasets.* Bioinformatics, 2022. **38**(9): p. 2642-2644.
37. Garner, L.C., et al., *Single-cell analysis of human MAIT cell transcriptional, functional and clonal diversity.* Nat Immunol, 2023. **24**(9): p. 1565-1578.
38. Gras, S., et al., *T cell receptor recognition of CD1b presenting a mycobacterial glycolipid.* Nat Commun, 2016. **7**: p. 13257.

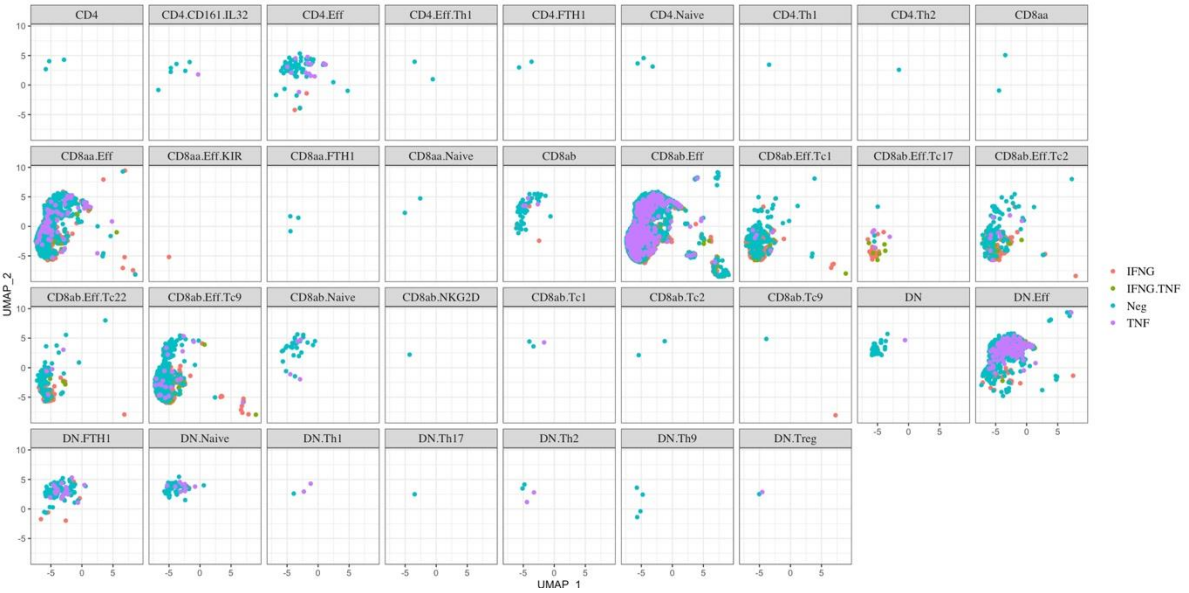
A


CD4
CD4.CD161.IL32
CD4.Eff
CD4.Eff.Th1
CD4.FTH1
CD4.Naive
CD4.Th1
CD4.Th2
CD8aa
CD8aa.Eff
CD8aa.Eff.KIR
CD8aa.FTH1
CD8aa.Naive
CD8ab
CD8ab.Eff
CD8ab.Eff.Tc1
CD8ab.Eff.Tc17
CD8ab.Eff.Tc2
CD8ab.Eff.Tc22
CD8ab.Eff.Tc9
CD8ab.Naive
CD8ab.NKG2D
CD8ab.Tc1
CD8ab.Tc2
CD8ab.Tc9
DN
DN.Eff
DN.FTH1
DN.Naive
DN.Th1
DN.Th17
DN.Th2
DN.Th9
DN.Treg
UMAP_1
UMAP_2
activated
activated-early.senescent
non-activated-early activated
TDsenescent

B


Non-activated-early activated
Activated
Activated-early senescent
Terminally differentiated senescent
activated
non-activated-early activated
TDsenescent
activated-early.senescent
Percent Expressed
Average Expression

C


CD4
CD4.CD161.IL32
CD4.Eff
CD4.Eff.Th1
CD4.FTH1
CD4.Naive
CD4.Th1
CD4.Th2
CD8aa
CD8aa.Eff
CD8aa.Eff.KIR
CD8aa.FTH1
CD8aa.Naive
CD8ab
CD8ab.Eff
CD8ab.Eff.Tc1
CD8ab.Eff.Tc17
CD8ab.Eff.Tc2
CD8ab.Eff.Tc22
CD8ab.Eff.Tc9
CD8ab.Naive
CD8ab.NKG2D
CD8ab.Tc1
CD8ab.Tc2
CD8ab.Tc9
DN
DN.Eff
DN.FTH1
DN.Naive
DN.Th1
DN.Th17
DN.Th2
DN.Th9
DN.Treg
UMAP_1
UMAP_2
IFNG
IFNG.TNF
Neg
TNF

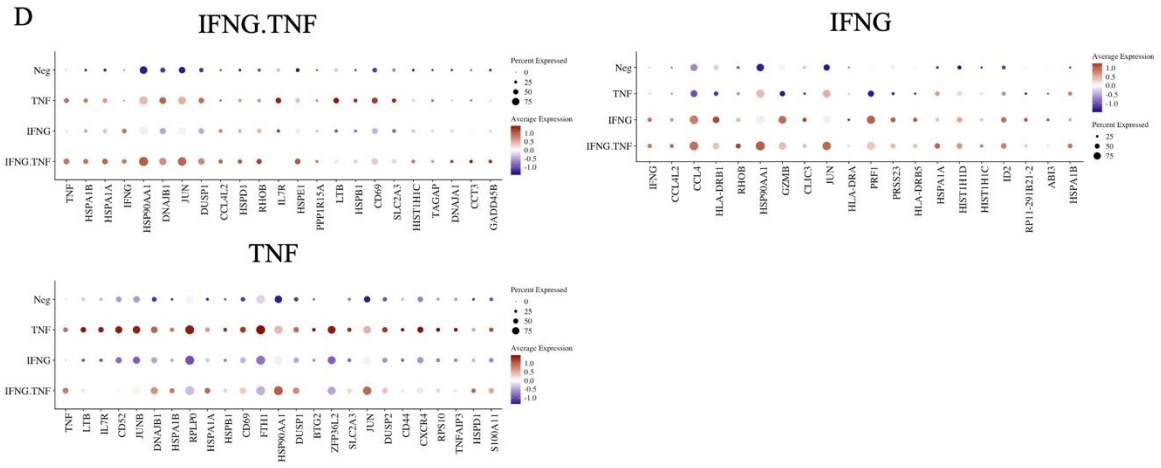


Figure 1. phenotype of TRAV12-2 & TRB20-1 top clone.

(**A**) UMAP of the cluster split by the T cell function annotations and color by the senescence annotation from scGate modelling. (**B**) Dot plots of the only clone data for each senescence label. (**C-D**) Th1_cytokines (IFNG and/or TNF) presented as the (**C**) UMAP and (**D**) dot plots of the average expression of IFNG.TNF, IFNG and TNF.

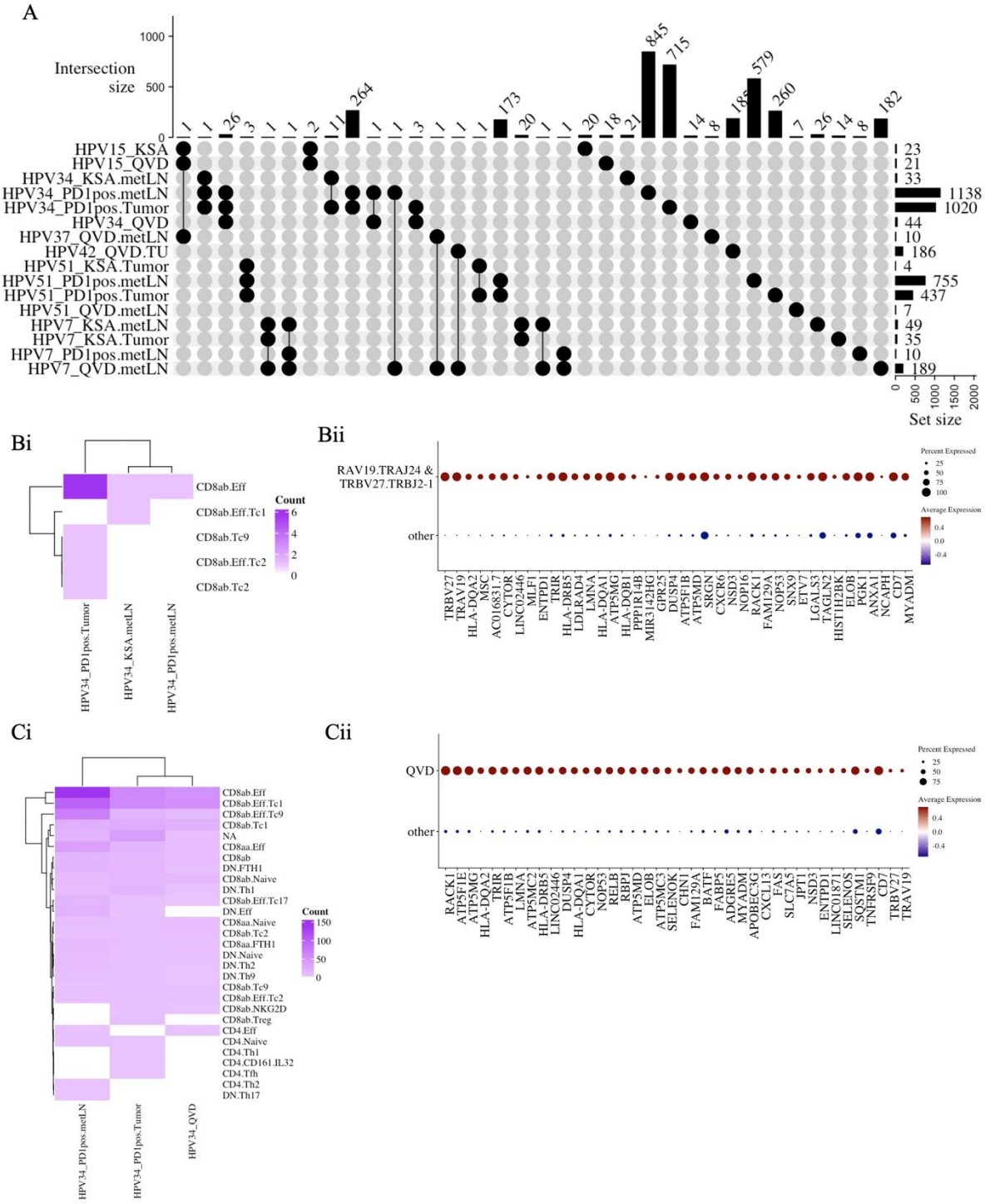


Figure 2. Re-examining the HNSCC identifies novel transcriptional signature of the HPV-epitope and PD-1 specific clones from HPV34.

(**A**) Upset plot (**B**) KAS clone (TRAV19.TRAJ24 CALSGTDSWGKLQF & TRBV27.TRBJ2-1 CASSLSGTLGNEQFF) and (**C**) pooled QVD clones. (**i**) scGate-based T cell function annotation vs samples in a heatmap and (**ii**) average expression for the top 40 genes enriched for this TCR.

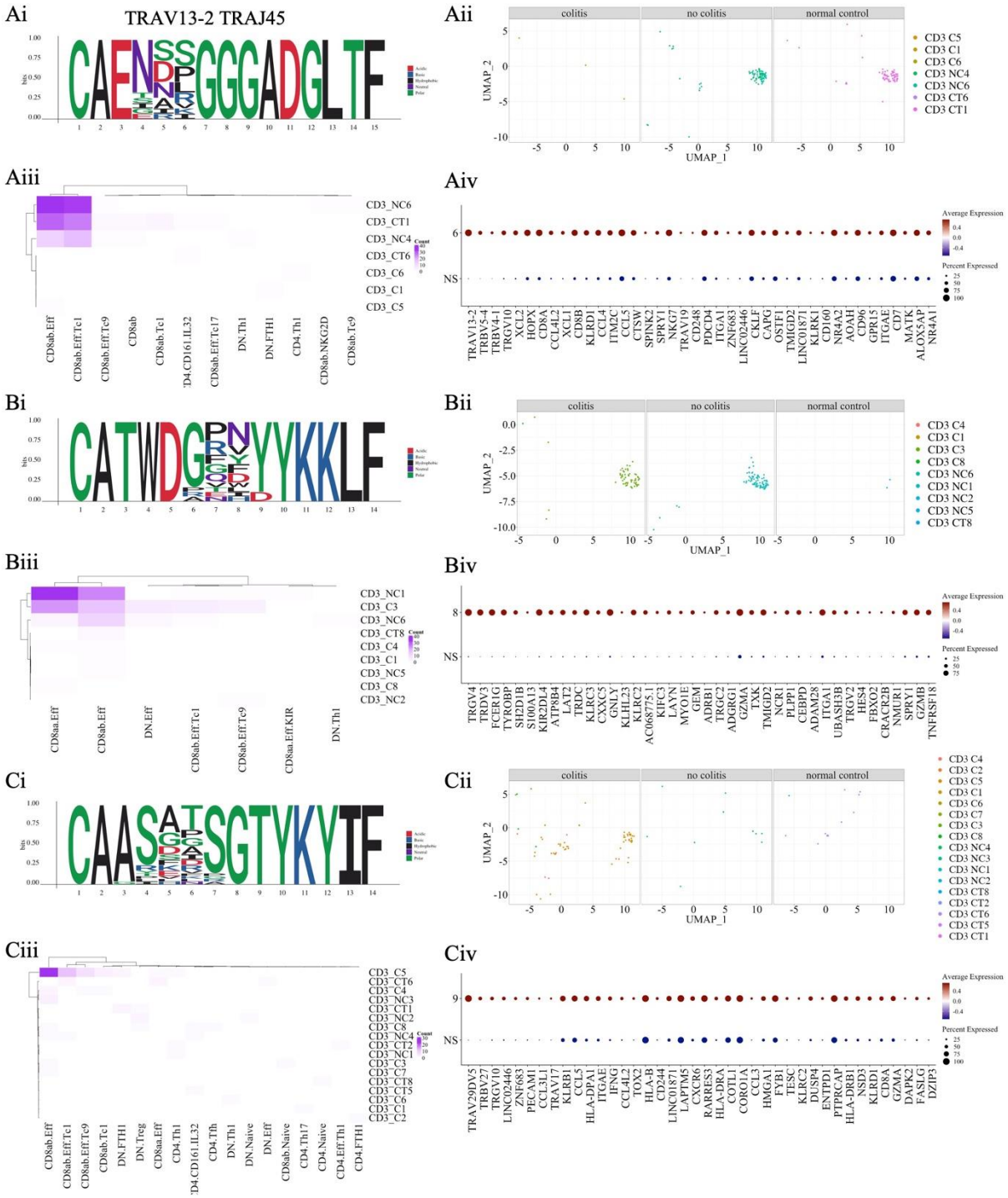


Figure 3. Clustering analysis of the colitis dataset highlighting disease specific clusters. (**i**) Motif of the cluster, (**ii**) Cluster on the UMAP plot coloured by individual, (**iii**) heatmap of the T cell phenotype vs the individual samples, and (**iv**) the top 40 transcripts enriched on this cluster. C: colitis; NC: non-colitis; CT: normal controls. (**A**) Cluster 6 (TRAV13-2 TRAJ45) more common in NC and TC. (**B**) Cluster 8 (TRGV4 TRGJ2) more common in melanoma cases. (**C**) Cluster 9 (TRAV29/DV5 TRAJ40) more common in colitis cases. C, colitis; NC, non-colitis; CT, normal controls.

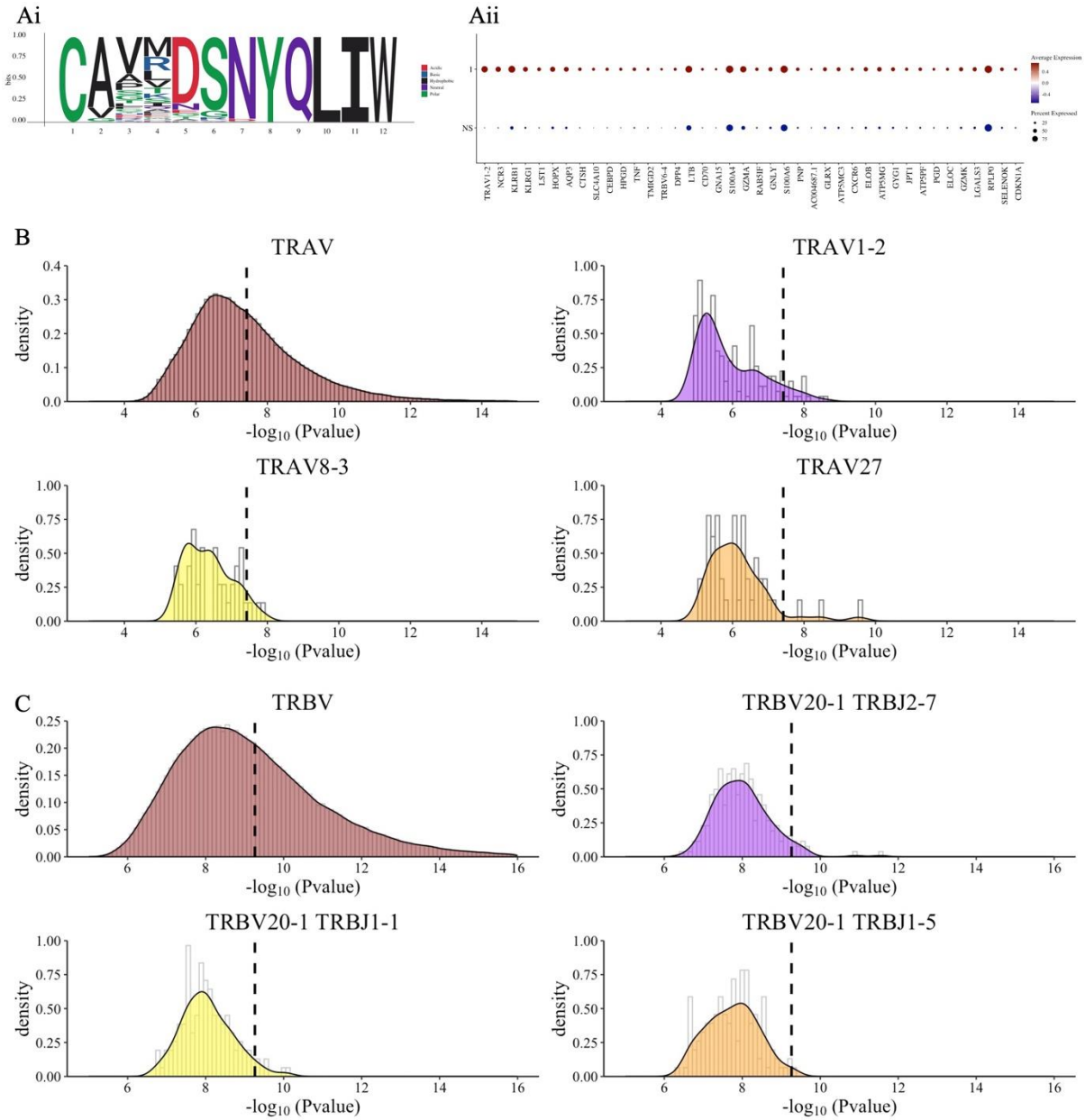
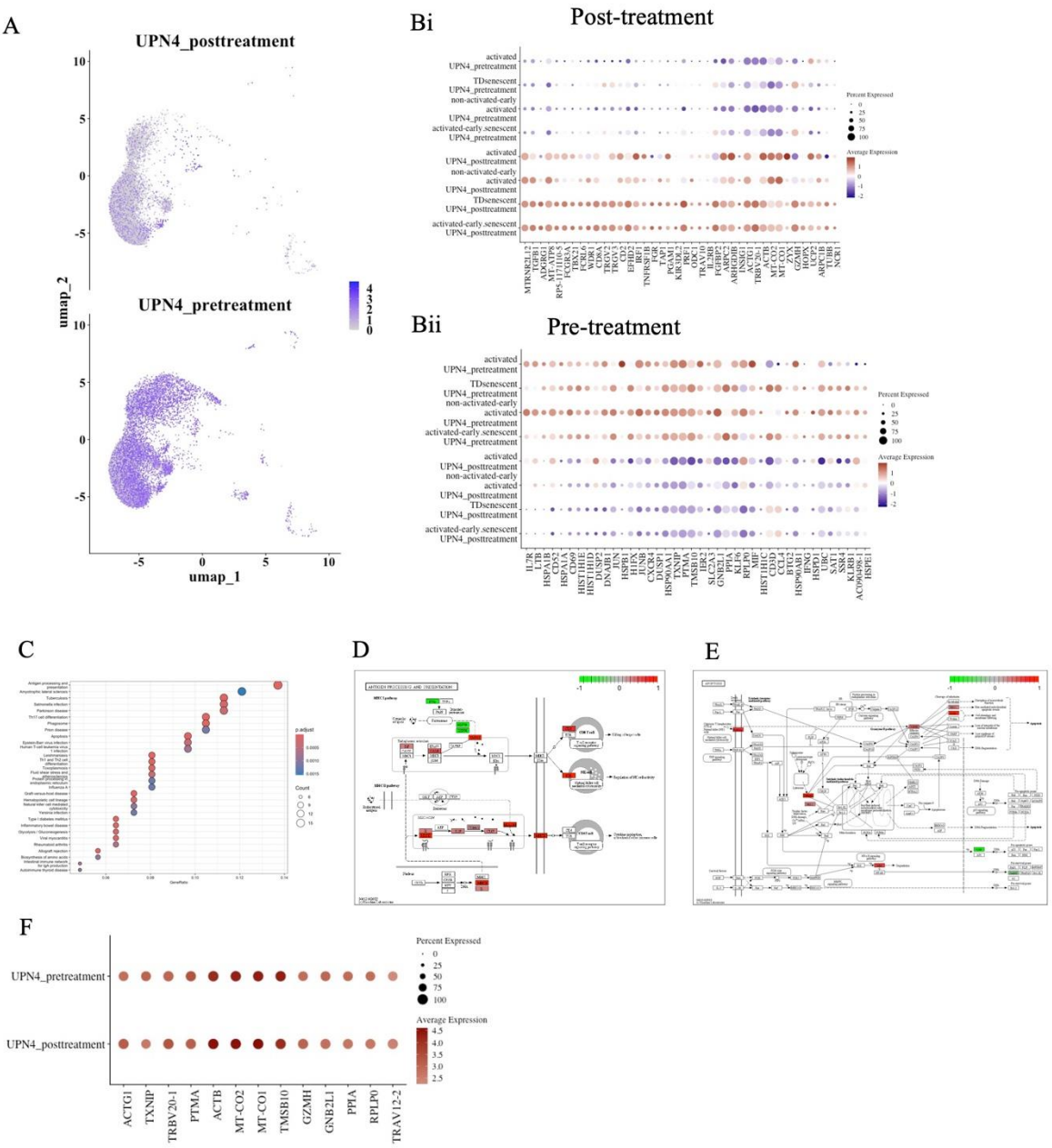
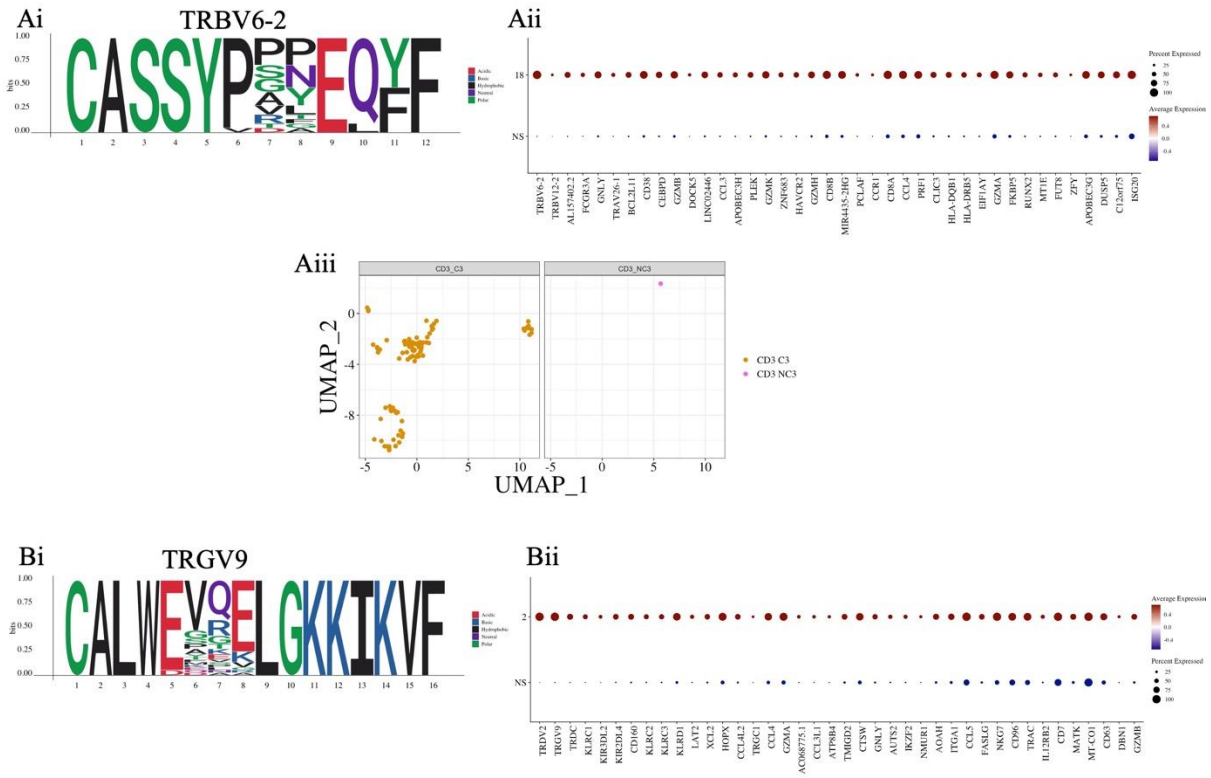


Figure 4. Global clustering analysis TRAV and TRBV clusters present in all 12 samples. TRAV1-2 TRAJ33 (**A**) Motif and (**B**) corresponding dot plot of the average relative expression and coverage. (**C**) TRAV probability of generation curve of the all sequences, TRAV1-2 TRAJ33, TRAV8-3 and TRAV27. (**D**) TRBV probability of generation curve of all sequences, TRBV20-1 TRBVJ1-1, TRBVJ1-5 and TRBV J2-7.

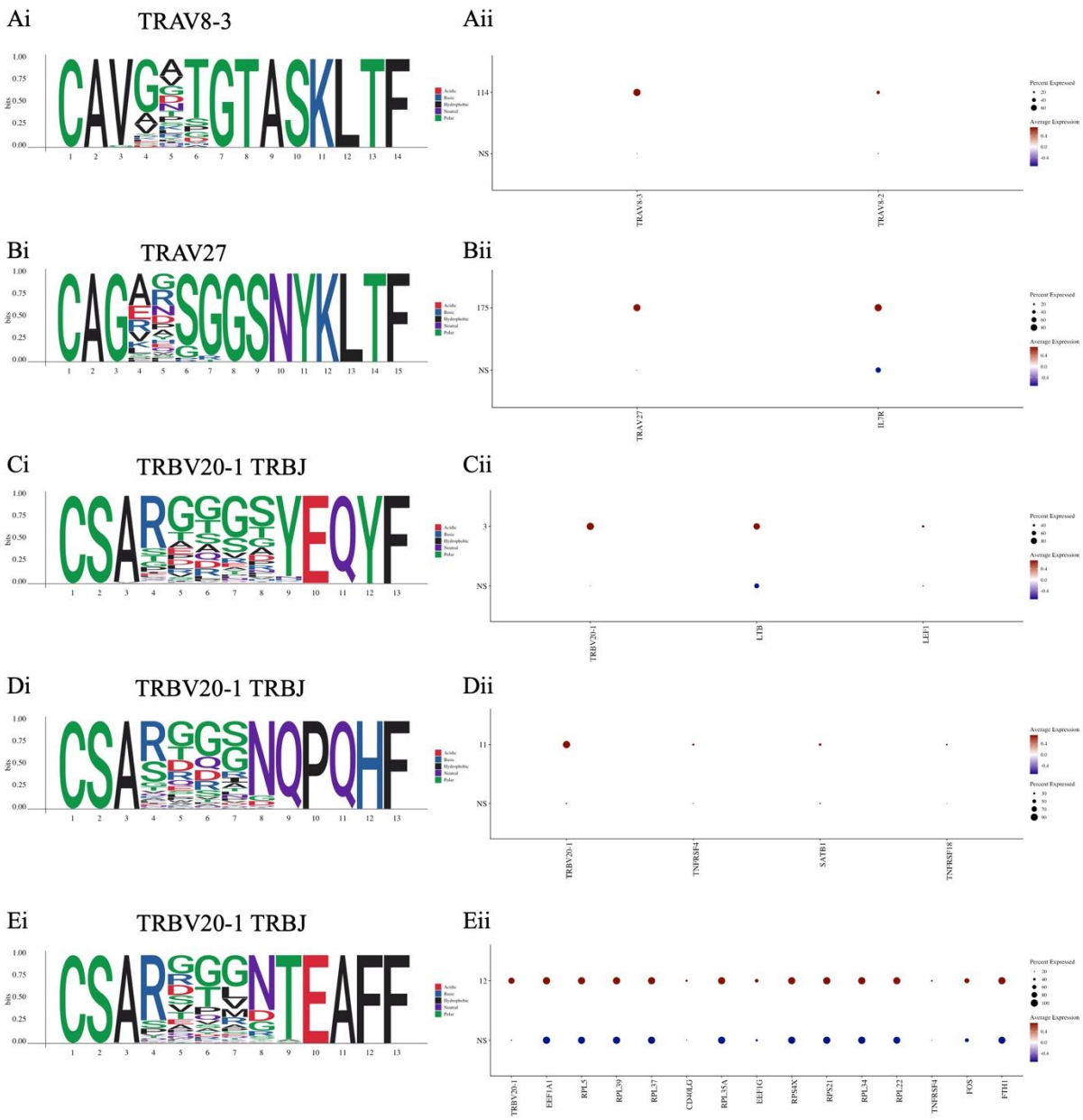


Supplementary Figure 1. Additional features of UPN4 top clone TRAV12-2 & TRB20-1 stratified by pre- and post-treatment.

(**A**) Feature plot expression of CD52 in post- and pre- treatment. (**B**) average expression of (**i**) post- and (**ii**) pre-treatment. (**C-E**) is the pathway analysis looking at the genes associated with the post-treatment (regardless of senescence status) looking at the (**C**) top 30 pathways, (**D**) top enriched pathway with antigen presentation and (**E**) the apoptosis KEGG pathway. (**F**) dot plot of markers un-altered by therapy and with >90% expression across the population.



Supplementary Figure 2. Colitis dataset showcasing two clusters(**A**) represents the TRGV9 invariant γδTCR cluster and (**B**) a TRBV6-2 cluster overrepresented in the C3 individual. (**i**) motif plot, (**ii**) top 40 associated genes and (**iii**) UMAP plot split and coloured by individual. C = c colitis. NC = non-colitis.



Supplementary Figure 3. Global clustering analysis of two TRAV and three TRBV common clusters. (**A**) Represent the TRAV8-3, (**B**) TRAV27, and TRBV20-1 (**C**) TRBJ (**D**) TRBJ and (**E**) TRBJ. (i) Motif and (ii) genes associated with the cluster.