STEGO.R : an application to aid in scRNA-seq and scTCR-seq processing and analysis

Title:
STEGO.R: an application to aid in scRNA-seq and scTCR-seq processing and analysis.

Authors:
Kerry A. Mullan[1,2,3*], My Ha[2,4,5], Sebastiaan Valkiers[1,2,3], Benson Ogunjimi[2,4,5,6], 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]Department of Paediatrics, Antwerp University Hospital, Antwerp, Belgium

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

Abstract
Introduction. The hypervariable T cell receptor (TCR), created through somatic recombination, allows for recognition of a diverse array of antigens. Single sequencing technologies allow capture of both the single cell expression data (scRNA-seq) with the paired single cell TCR sequencing (scTCR-seq). However, the current analytical pipelines have limited capacity to integrate both data levels. To overcome these limitations, we developed STEGO (Single cell TCR and Expression Grouped Ontologies) Shiny R application to facilitate the complex analysis required for understanding T cells role in various conditions.
Program parameters. STEGO.R application includes the Seurat quality control (QC) process, merging with Harmony, followed by semi-supervised cellular annotations with scGate. The scRNA-seq with scTCR-seq is broken down into four sections: top clonotype, expanded clonotypes, clustering (ClusTCR^2) and target epitopes from TCRex predictions. The Shiny R interface also facilitates the program's accessibility to novice R coders. The application can be found at https://github.com/KerryAM-R/STEGO.R.
Preliminary analysis. Out of 22 selected public datasets, 12 could be processed with STEGO.R. We re-interrogated the dataset concerning colon inflammations following melanoma therapies, as original studies did not integrate the scRNA-seq with scTCR-seq analysis. From one study, our novel process identified that the colitis expanded T cells were cytotoxic CD8+ T cells with over-represented transcripts including IFNG, GNLY, PFR1, GZMB, NKG7, HLA-DR, KLRD1 transcripts relative to both the non-expanded clonotypes, non-colitis cases and healthy colon donors. The analysis also identified a TRGV4 cluster associated with melanoma cases as well as two TRBV6-2 clusters specific to colitis.
Discussion. STEGO.R facilitates fast and reproducible analysis of complex scRNA-seq with TCR repertoire data. We have demonstrated its utility by extracting novel biologically relevant insights into T-cells. We anticipate this program will facilitate the identification of subtle T population differences and if these are specific to a TCR clone and/or the expanded repertoire.

1. Introduction

T cells, part of the adaptive arm of the immune system, have a critical function in maintaining human health. T cells are specialised to recognise a diverse array of epitope/antigens (e.g., peptides, lipids, small molecules) to identify and remove dysfunctional cells. To ensure coverage of all 'foreign' antigens, T cells have distinct functional profiles (e.g., cytotoxic CD8+ T cell are responsible for monitoring intracellular pathogens)[1]. Moreover, identification of the variable epitopes also relies on the T cells hyper variable receptor (TCR), created through recombination of Variable (V), Diversity (D) and Junctional (J) gene segments [2]. There are variable estimates on the total number of possible TCR recombination (~10^{15}), with ~10^{6} functional clones [3]. The collection of the total sum of the total variation is referred to as the TCR repertoire. Overall to determine the T cells function with epitope specificity requires interrogating both the gene expression phenotype and TCR repertoire.

With current next generation sequencing (NGS) technologies, we have access to single cell resolution of the genes expressed in the T cell subsets for both the gene expression (scRNA-seq) and TCR paired chain sequencing (scTCR-seq). With this enhanced resolution, it has become clear that the protein-based classifications are not all inclusive. For instance, CD4+ cytotoxic T cells have now been observed [4]. Thus, combining scRNA-seq with scTCR-seq is a powerful tool for identifying the dynamic nature of T cell function. These technological advancements mean we can identify disease specific markers and identify the TCRs to functionally validate. The experimental workflow has come a long way to gain this breadth of information.

The current approaches, reviewed in [5], identify many ways to process single cell gene expression data, with few programs taking into consideration the TCR-seq. Several approaches look for correlations in gene expression patterns and TCR similarity features (ConGA [6]). The target audience for most of these tools are computational biology experts, and therefore can require a steep learning curve for many experimental immunologists. Programs are either written in python (ConGA[6], pyTCR[7], Dandelion[8]) or R (scRepertoire[9]). Data generated through the 10x Genomics technologies can be analysed using their own software platforms, including CellRanger and Loupe Browser v6.4. While the BD rhapsody data requires the proprietary program SeqGeq®. The depth of the TCR repertoire analysis focuses on clonally expanded cells with similar expression pattern. Yet, it may be worthwhile interrogating if CDR3 sequence similarity could have similar functions (clustering), identify features of TCRs with predicted epitope specificity, and determine if a TCR may have multiple roles (*i.e.,* multiple distinct expression profiles). However, these aspects are not considered in the current analysis pipelines, potentially missing crucial insights buried in single cell data.

There are many approaches to annotating single cell RNA-seq dataset (reviewed in [5] and [Mullan et al. unpublished review]). To overcome some of these barriers to the annotation of single cell data, members of the Chan-Zuckerberg Initiative (CZI) have been developing a user-friendly python program (cellxgene)[10]. These label transfer models (e.g., CellTypist[11], cellxgene) perform reasonable well for labelling cellular populations with distinct profiles. Additionally, many of the programs cannot distinguish certain T cell populations (e.g., CD8+ γδ T cells from CD8+ αβ T cells). However, distinguishing the T cells sub-populations should include the TCR repertoire, and alternative approach are needed to add this distinct dataset.

Here we present out new tool STEGO.R to aid in scRNA-seq and scTCR-seq analysis. The program is a shiny R package for ease of installation and accessibility to novice R coders. We included extensive documentation (written and videos) to aid in the installation process, as well as the processing steps of the data analysis.

2. Material and Methods

We selected 22 publicly available scRNA-seq with scTCR-seq for STEGO.R benchmarking based on a literature search (**Table 1**). Only 12 of the 22 datasets could be processed with STEGO.R (**Table 1**). The main issues for not being able to process the remaining ten datasets were due to missing information in the public repositories (i.e., no available gene expression [n=2], no TCR-seq[n=2]) or data format issue (e.g., summarised TCR data[n=1], cannot separate cases in merged files [n=2], incompatible format[n=3]). These 12 datasets were used to create a 10x Genomics workflow. In addition to the 10x Genomics datasets, we also used one Array-based dataset and unpublished BD Rhapsody for the respective workflows.. The array pipeline was developed for re-analysis of a COVID-19 dataset and published in [12]. The 10x Genomics data was formatted as either raw filtered files (barcode, features, and matrix), .h5, .h5ad, rds.gz or csv.gz. STEGO.R can currently process the raw files, .h5 and csv.gz, but not the other formats. There is currently no process to make the cloupe and/or vloupe as the original matrix or TCR file and therefore were not analysable with STEGO.R. The direct 10x outputs of the barcode, features and matrix with the filtered_contig were the most accessible to processing in STEGO.R. The csv.gz required some manual manipulation to process the files, including add a group to the barcode file (e.g., S2 to S32).

Table 1. Publicly available datasets to test STEGO (10x Genomics based data).

GEO	Name	Species	Condition	STEGO.R compatible	# of individuals	# of samples	Cells included	Total functional TCR sequences	% of captured TCR sequenced	Ref.
GSE114724	Immune Phenotypes in the Breast Tumour Microenvironment	*Homo sapiens*	Breast Cancer	yes	3	5	28341	24039	85%	[13]
GSE121637	Peripheral blood and tumour-infiltrating immune cells in renal clear cell carcinoma	*Homo sapiens*	Renal Cancer	scTCR-seq only	-	-	-	-	-	[14]
GSE139555	Peripheral clonal expansion of T lymphocytes associates with tumour infiltration and response to cancer immunotherapy	*Homo sapiens*	Anti-PD1 therapy	yes	13	32	194519	67700	35%	[15, 16]
GSE144469	Colon Inflammation Induced by Cancer Immunotherapy	*Homo sapiens*	Melanoma and therapy	yes	22	22	75569	68760	91%	[17]
GSE145370	Immune suppressive landscape in a human oesophageal squamous cell carcinoma microenvironment	*Homo sapiens*	Oesophageal cancer	yes	7	14	108226	35449	33%	[18, 19]
GSE148190	Single cell RNA and TCR sequencing of tumor-infiltrating lymphocytes from human melanoma	*Homo sapiens*	Skin cancer	yes	2	2	8794	4904	56%	[20]
GSE184330	Single cell RNA sequencing and TCR repertoire analysis of MIS-C affected patients versus healthy controls and severe adult COVID-19	*Homo sapiens*	COVID-19 infection	Cannot separate cases in merged files	16	-	-	-	-	[21]
GSE160173	Induction of T cell dysfunction and NK-like T cell differentiation *in vitro* and in patients after CAR T cell treatment [scTCR-seq]	*Homo sapiens*	CAR T cell treatment	scTCR-seq only	2	-	-	-	-	[22]
GSE180268	Functional HPV-specific PD-1+ stem-like CD8 T cells in head and neck cancer	*Homo sapiens*	HPV and Head and Neck cancer	yes	6	19	53303	26844	50%	[23]

GEO	Name	Species	Condition	STEGO.R compatible	# of individuals	# of samples	Cells included	Total functional TCR sequences	% of captured TCR sequenced	Ref.
GSE168859	Coupled Single Cell RNA Sequencing and TCR Profiling of T Cells in Large Granular Lymphocytosis to Infer Pathophysiology and Mechanism of Drug Action	*Homo sapiens*	LGL and Response to Drugs[&]	yes	20	32	558795	276796	50%	[24]
GSE168163	Single-cell profiling of T lymphocytes in deficiency of adenosine deaminase 2	*Homo sapiens*	DADA2 deficiency	Summarised TCR file	15	-	-	-	-	[25]
GSE185659	Human lung tissue resident memory T cells are re-programmed but not eradicated with systemic glucocorticoids after acute cellular rejection	*Homo sapiens*	Lung transplant rejection	yes	3	7	22397	8227	37%	[26]
GSE185058	The Single Cell Sequencing of Immune Cells in malignant pleural effusion	*Homo sapiens*	malignant pleural effusion	scRNA-seq only	5	-	-	-	-	[27]
GSE179994	Temporal single cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer	*Homo sapiens*	anti-PD-1 therapy in lung cancer	Cannot process due to rds.gz format (incompatible format)	38	-	-	-	-	[28]
GSE184703	Expansion of Human Papillomavirus-Specific T Cells in Periphery and Cervix in a Therapeutic Vaccine Recipient Whose Cervical High-Grade Squamous Intraepithelial Lesion Regressed	*Homo sapiens*	HPV and cervical cancer	yes	1	1	14946	10222	68%	[29]
GSE161192	CD4+ T cells contribute to neurodegeneration in Lewy Body dementia	*Homo sapiens*	LB dementia	yes	4	4	6438	5642	88%	[30]

GEO	Name	Species	Condition	STEGO.R compatible	# of individuals	# of samples	Cells included	Total functional TCR sequences	% of captured TCR sequenced	Ref.
GSE178991	10X genomics single cell GEX and VDJ 5' sequencing of PBMC from Type 1 Diabetes patients treated with Treg therapy alone or plus low dose IL-2	*Homo sapiens*	Type 1 Diabetes	Cannot process due to .h5ad format (incompatible format)	28	-	-	-	-	[31]
GSE176201	Immune signatures underlying post-acute COVID-19 lung sequelae	*Homo sapiens*	COVID-19 lung tissue	yes	6	6	34781	23081	66%	[32]
GSE172158	Single-cell RNA-seq of T cells in B-ALL patients reveals an exhausted subset with remarkably heterogeneity	*Homo sapiens*	Leukemia (B-ALL)	Cannot separate cases in merged files	4	-	-	-	-	[33]
GSE182536	Single-cell RNA-sequencing on human naïve and memory CD4+ T cells during Plasmodium falciparum infection	*Homo sapiens*	Malaria infection	scTCR-seq in Vlope format (incompatible format)	6	-	-	-	-	-
GSE165499	Co-evolving JAK2V617F+ relapsed AML and donor T cells with PD-1 blockade after stem cell transplantation: an index case	*Homo sapiens*	Anti-PD-1 therapy in AML after transplantation	yes	1	6	32714	3081^	9%	[34]
GSE181279	Single-cell RNA sequencing of peripheral blood reveals immune cell signatures in Alzheimer's disease	*Homo sapiens*	Alzheimer's	scRNA-seq only	5	-	-	-	-	[35]

- Not applicable due to missing files or incompatible file format.
^ BCR & TCR sequence
[&] Out of memory issues reached with >500,000 cells in the annotation process. To annotate all models with the Posttreatment and Pre-treatment files annotation (~8-12h). We recommend in this case to annotate <200,000 cells and then merge the files together before the analysis.

2.1 Data pre-processing

10x Genomics: Each of the datasets included four files including the Barcode, feature annotations, matrix, and filtered contig annotations. As there were many file formats uploaded for 10x Genomics data, the process needed to include adding in h5, and csv.gz pipelines. To decrease compatibility and accessibility issue, we recommend storing the raw filtered files as either barcode, feature annotations, matrix, csv.gz or h5 matrix object with the TCR contigs in a separate file. The user needs to add the group and treatment label. This will be included in the "Sample_Name" column of the meta-data file. This will be the unique identify of the file.

Similarly, BD Rhapsody included a labelled matrix or raw files (barcode, features, matrix) with a paired or unpaired TCR and sample tags file. For experiments missing the sample tags file, a mock file needs can be produced by the program. The unique identifier is already included in the "Sample_Name" column or the file name added to the "orig.ident" column.

For the TCR/BCR sequence files are paired from the AIRR format. Due to the meta-data mering issue only the dominant sequence was maintained. However, there is some evidence the most dominant sequences may not be epitope specific, we added in a filtering pipeline from the AIRR format to identify sequences with multiple chains and non-standard pairing (*e.g.,* alpha/delta). However, this is currently implemented only for BD Rhapsody datasets, as the 10x Genomics usually only had the filtered TCR files available. This will be merged at the end of the Seurat quality control process, if needed.

Regardless of the platform of the data, the user will need to download several files which includes: ClusTCR^2 (ID_ClusTCR2_date.csv), TCRex (ID_TCRex_date.tsv), SeuratQC (matrix[raw only]:ID_count-matrix_date.csv, meta data: ID_metadata_platform_date.csv), TCR_Explore (ID_TCR_Explore_date.csv).

2.2 Clustering

ClusTCR^2 was based on the ClusTCR python package[36]. This clustering is based on V family matching and sequence similarity CDR3 amino acid sequences of the same length. ClusTCR uses a hashing function to determine all pairs of sequences with a hamming distance of 1 and builds a graph from the edge list obtained through hashing. Next, it uses the Markov clustering algorithm (MCL) to identify similarity groups in the graph. The MCL R package originally only numerically labelled one node away connection relative to the row, and that part was looped to ensure that the whole cluster received the same numeric label (https://github.com/KerryAM-R/ClusTCR2) . However, unlike the ClusTCR python package, ClusTCR^2 does not include the biochemical properties needed for the K-means clustering before the MCL step. As current scRNA-seq experiments have fewer clonotypes (e.g., <100,000 unique clonotypes), the K-means pre-clustering resulted in limited speedup.

The user can upload multiple ID_ClusTCR2_date.csv for merging prior to the clustering if more than one individual is present or directly to the clustering section. After uploading, the user will need to download the clustering table (ClusTCR2_output_date.csv).

2.3 TCRex processing

If multiple individuals are used, the individual TCRex files can be merged in "TCRex merge" section. This process removes the duplicate CDR3 sequences and downloaded as the corresponding .tsv file (TCRex_merged_date.tsv). The user uploads the file to TCRex

webpage https://tcrex.biodatamining.be and selects the epitopes of interest[37]. This can be all viral and/or cancer, or focus on a specific subset (*e.g.,* COVID-19, Melanoma *etc*.). The processed file is downloaded and will be uploaded to STEGO.R Analysis section.

2.4 Gene expression quality control

This followed the Seurat pipeline version 4[38]. Briefly, each individual dataset is processed to remove low quality cells (<200 features) or possible doublets (>6000 features). Cells with high mitochondrial (*e.g.,* 20%) and low ribosomal gene can be removed (*e.g.,* 5-10%). The data is scaled on the whole dataset, normalised, feature identification (n=2000), use the principal components (PC) for dimensional reduction (UMAP) and resolution for identifying Seurat clustering. The number of PCs can be interrogated using the Elbow method (*e.g.,* usually between 10-15). The files are downloaded as the .h5Seruat object, which contains all the Seurat embeddings.

2.5 Merging multiple samples.

The harmony package was used to combine multiple .h5Seurat Seurat objects. This process requires the combined datasets to be re-scaled using the top 2000 transcripts and pre-selected genes for phenotyping (*e.g.,* CD4, CD8). The process uses 30 PCs for the harmony reduction, which removes the batch/individual biases. The user can check that they annotated the files correctly with the presented UMAP plot. The user is required to then download the processed merged file (ExperimentID_merged_date.h5Seurat).

If there are more than 200,000 total cells, it is best to batch merge in lots of ~200,000 cell for each annotation (see 2.5 for details), and divide the samples as evenly as possible between the batches. This was needed for the GSE168859, which had ~560,000 cells. After annotating, the files will then be merged into one .h5Seurat file.

2.6 Single cell annotation

The gold standard approach to single cell annotation involves (1) automated annotations, (2) manual annotation/inspection of the annotations and (3) expert review [39]. The common strategies for the automated annotation include transfer models (CellTypist[11], cellxgene). However, the annotations for the T cell sub-populations are incomplete. For instance, modelling approaches cannot distinguish the γδ T cells and NKT cells within the cytotoxic CD8+ T cell population (**Figure 1A-C**). Additionally, upon integration of the TCR-seq with the scRNA-seq, we could identify that there was mixing of the TRAC and TRDC expression and therefore do not have distinct expression patterns. In this manner, γδTCR can be missed due to reduced TRDC expression (**Figure 1D**).

Some of the label transfer models do not include some common T cell population (e.g., Th2) [Mullan et al. unpublished review]. Additionally, none of the modelling investigates functional aspects of T cells including activation markers (*e.g.,* IFNγ) that may not be subset specific. The other issue with these models is that they often do not include how the datasets were originally labelled. Until we have a robust list of markers for consistent labelling that fit the subtle differences of T cell subpopulations, they are not the best strategy for annotating T cells.

The unsupervised Seurat clustering was not an ideal annotation strategy due to the dynamic nature of gene expression within clusters and between the clustering. Thereby this may have missed some important dynamic features. While these annotation strategies are good for

labelling cells with distinct transcriptional differences, an alternative annotation approach for identification of the subtle and dynamic differences within T cell population was used. The semi-supervised scGate [40] annotation methodology was integrated with custom databases to cover the missing models. Additionally, we also used the FindMarker function from the Seurat package within each of the "TCR and GEX" sub-section to identify markers that were over-represented for the individual clones, clonal expansion, clustering, and predicted epitopes. To check the expression the users could visualise the average expression (dot plot) or scaled expression (violin plot). Thereby, there were two strategies for annotating the T cells.

Human Annotations

Human specific markers were chosen based on several sources. The program includes the standard generic, CD4_TIL and CD8_TIL models from scGate [40], with the latter two models requiring sorted CD4 or CD8 T cell populations. To overcome this pre-sorting issue, there was a need to add the CD8A and CD4 transcriptional expression to the databases. Additionally, the models did not include double-negative (DN) annotations. From observation across the 12 usable datasets, the CD4 expression had poorer expression than CD8 markers (**Figure 2**). Often the DN population clustered with the CD4 population.

However, as the three scGate models did not capture some of the more subtle T cell features. Therefore, these databases of annotation started with the general T cell markers identified in our recent review [Mullan et al. unpublished review]. Each of the annotation strategies was tested and refined and viewed against the raw expression to the final list in **Table 2**. Based on the above considerations, the models did not distinguish the αβTCR from γδTCR, as this required the TCR-seq level for certainty. Additionally, due to the CD4 coverage issue, this marker was not included in the T cell Function marker list and added the -like to each description. The lower expression of CD4 could be driven by activation or due to the 10x Genomic chemistry.

Table 2. Literature based annotation strategies.

Classification	Sub-classification	Transcriptional markers
Activation	Early	CD69[§]
	Late	IL2RA
	Very late	CD38, HLA-DRA
	Resting	IL4R
COVID^	Effector	GZMB, GNLY, PRF1, PRDM1, KLRD1, SLC9A3R1
	Exhausted	TIGIT, PDCD1
	Memory	GZMK
	Naïve	LEF1, TCF7, CCR7
	Proliferative	MKI67, TYMS
	Senescence	B3GAT1
Cell cycling	Cycling	MKI67, TOP2A
ESCC [18]	B cells	MS4A1, CD19, CD3D-, CD3E-, CD2-
	CD4 T cells	CD3D, CD3E, CD2, CD14-,CD4, CD8A-
	CD8 T cells	CD3D, CD3E, CD2, CD14-, CD4-, CD8A
	Double negative (DN) T cells	CD3D, CD3E, CD2, CD4-, CD8A-
	mDC	CD1C, FCER1A
	Mast cell	TPSB2, CPA3
	Basal cells/fibroblasts	KRT19, IGFBP4, CTSB
	Monocyte/macrophages	CD14, VCAN, FCGR2A, CSF1R
	Natural Killer (NK)	CD3D-, CD3E-, CD2-, KLRD1,KLRC1
	pDC	CLEC4C

	Plasma	LAMF7, IGKC
Exhausted and Senescence	Exhausted	PDCD1, TIGIT
	Senescence	B3GAT1
T cell Function	MAIT-like	CD3E, TRAV1-2
	NKT-like (inhibitory receptor)	CD3E, KLRC1, KLRD1
	Tfh-like[%]	CD3E, CXCR5
	Th1-like [%]	CD3E, CXCR3, TBX21
	Th2-like [%]	CD3E, CCR4
	Th9-like	CD3E, IL9
	Th17-like [%]	CD3E, RORC
	Th22-like	CD3E, IL22
	Tregs	CD3E, FOXP3
Cytotoxic	GNLY.GZMB.PFR1	GZMB, GNLY, PFR1
	GZMB.PFR1	GZMB, GNLY-, PFR1
	GNLY	GZMB-, GNLY, PFR1-
IFNγ and TNFα	IFNγ and TNFα	IFNG, TNF
	IFNγ	IFNG, TNF-
	TNF	IFNG-, TNF
Interleukin[%]	IL-2	IL2
	IL-4	IL4
	IL-6	IL6
	IL-8	IL8
	IL-9	IL9
	IL-10	IL10
	IL-17	IL17A, IL17F

[%] Markers identified based on transcriptional interrogation. The literature identified that interleukin markers are more poorly captured at the transcriptional level. It is unknown which markers can replace IL-9 or IL-22.
* Based on COVID-19 x publication [ref].
§ when CD69 is expressed with CD103 (transcript ID: ITGAE) they may represent resident memory T cells[41]. Other's report that CD69 alone represents tissue resident T cells[42].

The current database requires some additions. Memory and tissue resident markers as well as including an NK-like annotation will be included in a later version of STEGO.R. Additionally, the cut-offs within the scGate thresholds depended on the origin of the dataset that includes: 10x Human, 10x Mouse (underdevelopment), BD Rhapsody (Human Immune panel), BD Rhapsody (Mouse; underdevelopment).

User instruction for annotating
The user uploads the merged or single .h5Seurat object for annotation. They can choose from a range of annotation strategies for mouse, human or other. For other, the user will download the database and alter the gene names as needed. However, the folder names must not be altered so that R can recognise each database correctly. Once annotated, the user will download this file (ExperimentID_Annotated_date.h5Seurat).

2.7 Analysis

The user will upload the annotated .h5Seurat file as well as the ClusTCR_output.csv and TCRex_ID.tsv (Beta chain only). If needed, the user can also upload a custom annotation file. The first column is labelled "ID" and will match the "Sample_Name" column. This allows for greater flexibility if new parameters need to be added to the dataset.

The program will do all the necessary summarisation for identification of clonal expansion and overlapping TCR sequences. For a more in-depth interrogation of the structure of the TCR repertoire, can be done with TCR_Explore [43], which can also aid in reformatting to TCRdist3[44] for more in-depth distance statistics.

The analysis was split into several types of analyses: TCR or gene expression overview analysis and the TCR-seq with gene expression (TCR with GEx) split into single TCR interrogation, expanded TCR, TCR clustering and epitope prediction. Every section includes the capacity to download a UMAP figure with the distinct annotations. The clustering, epitope, clonal expansion, and repertoire overlap highlight a subset of data relevant to each section. The user can change which groups of markers to interrogate as per the annotation present **Table 1**. The section also includes the "FindMarker" to better understand how the population of interest differs from the rest of the population. The user can view the scaled data in violin plots or the average relative expression in the dot plot. The statistics table can be downloaded.

2.8 Testing and requirements

STEGO.R was installed by the intended user on all major operating systems (Unix, Microsoft Windows and Linux). We recommend having 32Gb of RAM, as the merging and annotation steps can be quite memory intensive. The user needs to download the latest version of R and RStudio. The user needs to make sure devtools and BiocManager before install STEGO.R. For a detailed installation and running instruction can be found at https://stego.readthedocs.io/en/latest/. STEGO.R functionality was tested in-house (MMR, VZV), University of Sydney (Sharland group), Monash University (Mifsud and Purcell lab's) and other collaborators.

2.9 Code availability

STEGO code is available on the GitHub repository (https://github.com/KerryAM-R/STEGO.R).

2.10 Data availability

The publicly available data was sourced based on the GEO numbers (**Table 1**).

3. Results

3.1 Common practices for interrogating and presenting the scRNA-seq with scTCR-seq prior to STEGO.R

12 of the 22 selected datasets could be processed through STEGO.R and were summarised to showcase how the authors annotated and displayed their scRNA-seq with scTCR-seq (**Table 3**). These 12 datasets represent 88 individuals across 148 scRNA-seq with scTCR-seq datasets. There were ~1.14 million high quality cells with ~0.55 million functional TCR sequence. The TCR coverage for each dataset averaged 55% (range: 9% to 91%). All datasets, apart from [13], were processed in R 'Seurat' package. The majority of the datasets were annotated based on the Seurat clusters. However, few articles listed their markers and often referred to a previous publication or assumed that this was common knowledge. The ESCA dataset[18] had clear listing of each population, and was added to the annotation database in STEGO.R application. Some studies opted to not annotate the clusters and performed single marker interrogations, as the studies were focused on biomarker identification and validation[20, 23, 26].

TCR-seq data was most frequently analysed only in the context of the gene expression. The most common analysis was a dot plot that looked at clonal frequency across two conditions to see if they expanded or contracted [15, 26, 29, 34]. When combined with the scRNA-seq transcriptional expression, the most common approach to understanding the role of the scTCR-seq was to colour the UMAP plot by TCR repertoire clonal frequency.

Table 3. Summary of analysis strategies and findings in the original twelve articles

GEO	Study Goal	Sort	Program and packages	Annotation strategy	TCR-analysis (figures/tables)	scRNA-seq and scTCR figures and/or findings
GSE114724 [13]	Interrogating T cells from breast carcinoma tissue, LN and PBMC microenvironment	CD45+ DAPI–	Charlotte Python package	Manual checking clusters on top ranked genes. Markers not listed		Overlayed TCR onto the cluster. Described multiple phenotypes of T cells within the samples.
GSE139555 [15]	Compare paired tumor and non-tumor adjacent tissue and blood in NSCLC	CD3+ (PBMC) or CD45+ (Tumor)	Seurat in R	Cluster based. Markers not listed	Compared T cell expansion using dot plots. Use dominant phenotype within expanded TCR	PBMCs were somewhat representative of tumor T cell. Therefore, could be used for monitoring and direct treatment.
GSE139555 [16]	Interrogating CD28+CD226+ CD8+ T cells in NSCLC to assess the impact of treatment	CD3+ (PBMC) or CD45+ (Tumor)	Seurat in R	Cluster based on published gene signatures. Focused on CD28 and CD226 Markers not listed.	Previous article	Previous article
GSE144469 [17]	Interrogating cellular and molecular mechanisms of colitis side effect of CTLA-4 or PD-1/PD-L1 therapy (melanoma)	CD3+ CD45+ from colon biopsies	Seurat in R	SingleR projection model Manual checking clusters on top ranked genes CD4 and CD8 based on normalized expression Markers not listed	No main scTCR figures	No overlap analysis
GSE145370 [18]	Understanding the tumor microenvironment in ESCC	CD45+ CD235- from Tumor and adjacent tissues	Seurat in R Monocle in R (trajectory analysis)	Manual checking clusters on top ranked genes Markers listed (no MAIT, NKT, γδTCR included) Exhausted listed in Sup. Data 3	Upset plots, bar graphs of % clonal	Overlayed clonal expansion on UMAP and compared within clusters. Distinct and overlapping T cells in both tumor and adjacent tissue.

GSE145370 [19]	Understanding the role of IL-32 in ESCC	Previous publication	Seurat in R	Manual checking clusters on top ranked genes Markers listed (no MAIT, NKT, γδTCR included)	Previous publication only	Previous publication only
GSE148190 [20]	Understanding LAYN role in human melanoma	CD3+ CD8+ (melanoma tissue)	Seurat in R	Focused on LAYN expression only. No cell annotation	Single chain analysis of top 20 expanded clonotypes ±LAYN	Overlayed clonal expansion on UMAP and focused on LAYN being expressed in more expanded T cells
GSE168859 [24]	T-cell large granular lymphocyte leukemia (T-LGLL)	CD3+	Seurat in R	Used raw data from GSE93777[45] and used gene sets to define cells Markers not listed.	3D bar graphs, diversity calculation (Gini), length distributions, Clustering	Overlayed clonal expansion on UMAP. Limited phenotypic analysis on expanded clones.
GSE185659 [26]	Understanding the effect of glucocorticoid therapy for treating Acute rejection in lung transplant patients	CD3+	Seurat in R	Some markers listed (e.g. Naïve: S1PR1 and SELL) or based on CD4 vs CD8 expression.	Dot plot frequency graph.	Overlayed clonal expansion on UMAP. Expression (volcano plot) of top 4 clonotypes vs the rest.
GSE184703 [29]	Understanding HPV-Specific T Cells	Peptide derived T cells and IFNγ+	Seurat in R Loupe V(D)J Browser	Single marker interrogation by cluster with UMAP and violin plots	Dot plot frequency graph. Single chain frequency	No overlap analysis
GSE176201 [32]	Understanding acute lung sequelae T cell signature from COVID-19	CD3+ from BAL and PBMC	Seurat in R	Used clusterProfiler and AddModuleScore based on three published datasets. Markers not listed	Supplementary figures bar graph (%)	Separate CD4+ and CD8+ analysis. Overlayed clonal expansion on UMAP (sup. Figure). Comparing expression of expanded (>5) clonotypes to non-expanded (≤5).

GSE161192 [30]	Understanding T cells in Lewy body dementia	No sort, From CSF	Seurat in R MAST in R topGO for gene ontology	Used Findmarker in clusters and MAST package (filter out low quality cells). Markers not listed	No main scTCR figures.	Overlayed clonal expansion on UMAP and single marker analysis of expanded clonotypes. Gene ontology analysis.
GSE165499 [34]	Demonstrating how single-cell technology can understand the relapses GVL	-	Seurat in R	FindMarkers within clusters used. Populations based on published articles. CITE-seq and scRNA-seq expression. Some markers in heatmap, but not listed in a table.	Dot plot frequency graph (two conditions).	Overlayed clonal expansion on UMAP.
GSE180268 [23]	Understanding HPV+ T cells in head and neck cancer	tetramer-sorted HPV-specific PD-1+ CD8+ T cells	Seurat in R VISION in R Monocle3 in R (trajectory analysis)	FindMarkers within clusters used. Single marker and observed across the three clusters.	No main scTCR figures.	Overlayed clonal expansion on UMAP.

LN, Lymph node. PBMC, Peripheral mononuclear cells. UMAP, uniform manifold approximation and projection. BAL, bronchoalveolar lavage fluid. CSF, cerebral spinal fluid. GVL, Graft-vs-leukemia. HPV, Human papilloma virus. NSCLC, non-small cell lung carcinoma. ESCC, esophageal squamous cell carcinoma.

3.2 Re-analysing the colitis dataset.

While all datasets were processed for re-analysis, the GSE144469 will be used as the primary example as the original study did not contain scTCR-seq analysis. Additionally, this dataset included both αβ and γδ T-cells. The purpose of the GSE144469 study was to understand why some individuals that were treated for melanoma developed the gastrointestinal inflammation i.e., colitis[17]. The researchers analysed CD3+ CD45+ T cells from colon biopsies that represented eight melanoma patients with colitis (C), eight melanoma patients without colitis (NC) and six healthy controls (CT).

Firstly, we used STEGO.R to analyse the TCR-seq and gene expression independent of each other. There was evidence of clonal expansion (**Fig. 3A**) with the vast majority of clonotypes were unique to the individual (**Fig. 3B**). The clonal expansion on the UMAP plot (**Fig. 3C**) appeared to be mostly from CD8+ T cell population (**Fig. 3D**). Additionally, the clonal expansion represented both αβTCR and γδTCR, which was based on the TCR sequencing rather than the gene expression (**Fig. 3E**).

The upset plot, an alternative of a Venn diagram for representing overlap when four or more groups are listed, identified that there were few public clonotypes (**Fig. 3B**). Therefore, interrogating the top clonotypes was not performed. Instead, analysing the phenotype of the clonally expanded population (Ex; ≥3 clonotypes) vs non-expanded (NEx; <3 clonotypes) was completed (**Fig. 4A**). Comparing the phenotype of the colitis Ex vs NEx identified 55 markers associated with the expanded colitis populations include CD8A and CD8B, as well as class II HLA genes, (e.g., HLA-DRB1, HLA-DPA1, HLA-DRA, HLA-DPB1) cytotoxic markers (e.g.,PFR1, GNLY, GZMB, IFNG), NK receptors (CD160, KLRD1) and tissue homing integrins (ITGA1, ITGAE) (**Fig. 4B**). Additionally, there were 124 markers over-represented in inflamed colitis Ex compared to the non-colitis Ex (**Fig. 4C**). These markers include GBP genes that are interferon induced proteins, CD38 (late activation marker), cytotoxic markers (e.g., GZMB, GZMK, GZMH, GNLY, IFNG, PRF1) as well as antigen processing (TAP1) and effector-linked (NKG7) transcripts.

The next step of the analysis was to identify clusters on CDR3 sequence similarity. Interestingly, there was a preference for the TRBV6-2 11mer and 12mer in the colitis cases that had a greater degree of GNLY, GZMB and PFR1 expression (**Fig. 5A and 5B**). There were also TRGV10 (pseudogene) and TRGV2 clusters specific to the NC and CT including a (**Fig. 5C and 5D**). Additionally, an alternative gamma genes TRGV4 and TRGV8 clusters were over-represented in the melanoma (C and NC) with the (**Fig. 5E and 5F**). All these clonotypes appeared to express cytotoxic markers.

4. Discussion

As part of testing STEGO.R we searched for datasets that contained scRNA-seq with scTCR-seq and identified 22 datasets to interrogate for re-analysis purposes. 55% of these publicly stored datasets were able to be processed through our novel STEGO.R. The remaining 45% could not be re-interrogated due to formatting or missing data issues. This highlights a broader issue with how single cell RNA-seq data is stored in public repositories. Based on our experience, the easiest data to reanalyses was the output of the cellranger pipeline containing the barcode, features and matrix file as well as storing both the filtered and unfiltered contig (TCR) file in the AIRR format. This will allow other to readily re-interrogate previous data and/or add published dataset to help improve statistical power. We emphasize the storing of public data into common file formats that apply to the current standards of the field [39].

Many studies failed to list the main markers used to annotate the clusters, subjected to researcher expertise when interpreting the data and are unlikely reproducible. To improve this, based on our recent review and refinement with the 12 datasets, we included several databases of common T cell specific markers. This list is not yet inclusive of all T cell subtypes (e.g., memory, NKT), and needs further input from T cell immunologists. The memory markers have not been included in part due to the issue of CD45RO and CD45RA proteins being translated from the same CD45 transcript. Those two variants of CD45 proteins distinguish naïve from memory. However, we need additional markers that are transcriptional appropriate for memory populations.

The other strategy STEGO employed was to use the findMarker function focusing on the TCR, expansion or cluster of interest. This highlighted the markers that were over-represented in the cell population of interest. For instance, the extended interrogation of the colitis clonal expansion identified a more cytotoxic profile (IFNγ, GNLY, PFR1, GZM's, NKG7), class II HLA genes) with NK receptors (KLRD1) from CD8+ T cells relative non-expanded colitis T cells. Similarly, these cytotoxic genes were more prevalent in the T cells derived from the colitis cases compared to both the non-colitis cases and healthy controls. Thereby, this analysis was able to give additional insights into the bystander cells and could be excluded as a cause of the colitis after melanoma treatment. Importantly, key biomarkers segregated out the expanded T cell population, some of which are not commonly used with focused experiments that relies on IFNγ/TNFα capture. We would recommend companies developing triplet capture protocol for intracellular proteins granulysin, granzyme B and perforin, as the triplicate expression is closely linked to the expanded populating. The sorting process would also need to include the NK receptors.

Unlike the previous tools that focus on public and private clonotypes, STEGO.R was able to identify public clusters and determine if they had similar expression profiles. Our interrogation was able to identify two colitis specific clusters that expressed TRBV6-2 (11mer and 12mer). This may indicate a preference for certain beta-chains irrespective of the alpha pairing. Additionally, the clustering analysis also identified T cells that may protect against the inflammatory, as they were absent or rarely expressed in the colitis cases. One cluster expressed the TRGV10 pseudogene, indicating it would not express a functional TCR, yet it appears to have utility in tracking distinguishing colitis from non-colitis cases. Lastly, the clustering identified several gamma specific clusters of TRGV4 and TRGV8 that were over-represented in the melanoma (C and NC) cases. A recent study showed that Vγ4+ T cells can be HLA-A*02:01 restricted to melanoma peptides [46]. Thereby, the TRGV4

cluster, as it also expressed CD8, could be a class I antigen reactive γδ T-cell that may recognise melanoma specific peptides. In this manner, STEGO.R provided additional evidence that there might be more similarity between αβ and γδ T-cells than has been assumed so far. This finding also reinforces the need to classify the T cell based on function expression in an independent manner from the TCR sequence.

5. Conclusion

Here we present our STEGO.R for processing and analysing scRNA-seq with scTCR-seq data. Through optimising STEGO pipeline we identified shortcomings with how the published literature is stored as well as annotation reporting. We showcased some of our program's functionality by extending the analysis of colitis complication to melanoma therapy. This enabled the identification of colitis specific T cells expression signature, that were specific to the expanded populations. Additionally, we identified two TRBV6-2 clusters specific to the colitis complication. Overall, STEGO.R program has improved our capacity to understand the novelties of scRNA-seq with scTCR-seq data.

Funding

This work has been made possible by grant number 2022-249472 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation.

References

1. Neefjes, J., et al., *Towards a systems understanding of MHC class I and MHC class II antigen presentation.* Nature reviews immunology, 2011. **11**(12): p. 823-836.
2. Ma, L., et al., *Analyzing the CDR3 Repertoire with respect to TCR-Beta Chain V-D-J and V-J Rearrangements in Peripheral T Cells using HTS.* Sci Rep, 2016. **6**: p. 29544.
3. Sun, X., et al., *Longitudinal analysis reveals age-related changes in the T cell receptor repertoire of human T cell subsets.* J Clin Invest, 2022. **132**(17).
4. Oh, D.Y. and L. Fong, *Cytotoxic CD4(+) T cells in cancer: Expanding the immune effector toolbox.* Immunity, 2021. **54**(12): p. 2701-2711.
5. Valkiers, S., et al., *Recent advances in T-cell receptor repertoire analysis: Bridging the gap with multimodal single-cell RNA sequencing.* ImmunoInformatics, 2022. **5**.
6. 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.
7. Peng, K., et al., *pyTCR: A comprehensive and scalable solution for TCR-Seq data analysis to facilitate reproducibility and rigor of immunogenomics research.* Frontiers in Immunology, 2022. **13**: p. 954078.
8. Suo, C., et al., *Dandelion uses the single-cell adaptive immune receptor repertoire to explore lymphocyte developmental origins.* Nature Biotechnology, 2023: p. 1-12.
9. Borcherding, N. and N.L. Bormann, *scRepertoire: An R-based toolkit for single-cell immune receptor analysis [version 1; peer review: 2 approved with.* 2020.
10. Megill, C., et al., *Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices.* bioRxiv, 2021: p. 2021.04. 05.438318.
11. Xu, C., et al., *Automatic cell type harmonization and integration across Human Cell Atlas datasets*. bioRxiv, 2023: p. 2023.05. 01.538994.
12. Postovskaya, A., et al., *Leveraging T-cell receptor - epitope recognition models to disentangle unique and cross-reactive T-cell response to SARS-CoV-2 during COVID-19 progression/resolution.* Front Immunol, 2023. **14**: p. 1130876.
13. Azizi, E., et al., *Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment.* Cell, 2018. **174**(5): p. 1293-1308 e36.

14. Borcherding, N., et al., *Mapping the immune environment in clear cell renal carcinoma by single-cell genomics.* Commun Biol, 2021. **4**(1): p. 122.
15. Wu, T.D., et al., *Peripheral T cell expansion predicts tumour infiltration and clinical response.* Nature, 2020. **579**(7798): p. 274-278.
16. 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.
17. Luoma, A.M., et al., *Molecular Pathways of Colon Inflammation Induced by Cancer Immunotherapy.* Cell, 2020. **182**(3): p. 655-671 e22.
18. Zheng, Y., et al., *Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment.* Nat Commun, 2020. **11**(1): p. 6268.
19. 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.
20. Mahuron, K.M., et al., *Layilin augments integrin activation to promote antitumor immunity.* J Exp Med, 2020. **217**(9).
21. Hoste, L., et al., *TIM3+ TRBV11-2 T cells and IFNgamma signature in patrolling monocytes and CD16+ NK cells delineate MIS-C.* J Exp Med, 2022. **219**(2): p. e20211381.
22. Good, C.R., et al., *An NK-like CAR T cell transition in CAR T cell dysfunction.* Cell, 2021. **184**(25): p. 6081-6100 e26.
23. 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.
24. 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.
25. Wu, Z., et al., *Single-cell profiling of T lymphocytes in deficiency of adenosine deaminase 2.* J Leukoc Biol, 2022. **111**(2): p. 301-312.
26. 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).
27. Huang, Z.Y., et al., *Single-cell analysis of diverse immune phenotypes in malignant pleural effusion.* Nat Commun, 2021. **12**(1): p. 6690.
28. Liu, B., et al., *Temporal single-cell tracing reveals clonal revival and expansion of precursor exhausted T cells during anti-PD-1 therapy in lung cancer.* Nat Cancer, 2022. **3**(1): p. 108-121.
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. Dong, S., et al., *The effect of low-dose IL-2 and Treg adoptive cell therapy in patients with type 1 diabetes.* JCI Insight, 2021. **6**(18).
32. Cheon, I.S., et al., *Immune signatures underlying post-acute COVID-19 lung sequelae.* Sci Immunol, 2021. **6**(65): p. eabk1741.
33. Wang, X., et al., *Single-Cell RNA-Seq of T Cells in B-ALL Patients Reveals an Exhausted Subset with Remarkable Heterogeneity.* Adv Sci (Weinh), 2021. **8**(19): p. e2101447.
34. 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.

35. Xu, H. and J. Jia, *Single-Cell RNA Sequencing of Peripheral Blood Reveals Immune Cell Signatures in Alzheimer's Disease.* Front Immunol, 2021. **12**: p. 645666.
36. 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.
37. Gielis, S., et al., *Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor Sequence Repertoires.* Front Immunol, 2019. **10**: p. 2820.
38. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587. e29.
39. Heumos, L., et al., *Best practices for single-cell analysis across modalities.* Nat Rev Genet, 2023: p. 1-23.
40. 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.
41. Kim, H.D., et al., *Implication of CD69(+) CD103(+) tissue-resident-like CD8(+) T cells as a potential immunotherapeutic target for cholangiocarcinoma.* Liver Int, 2021. **41**(4): p. 764-776.
42. Weiner, J., et al., *CD69+ resident memory T cells are associated with graft-versus-host disease in intestinal transplantation.* Am J Transplant, 2021. **21**(5): p. 1878-1892.
43. Mullan, K.A., et al., *TCR_Explore: A novel webtool for T cell receptor repertoire analysis.* Comput Struct Biotechnol J, 2023. **21**: p. 1272-1282.
44. Mayer-Blackwell, K., et al., *TCR meta-clonotypes for biomarker discovery with tcrdist3: identification of public, HLA-restricted SARS-CoV-2 associated TCR features.* bioRxiv, 2021.
45. Tasaki, S., et al., *Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission.* Nat Commun, 2018. **9**(1): p. 2755.
46. Benveniste, P.M., et al., *Generation and molecular recognition of melanoma-associated antigen-specific human gammadelta T cells.* Sci Immunol, 2018. **3**(30): p. eaav4036.

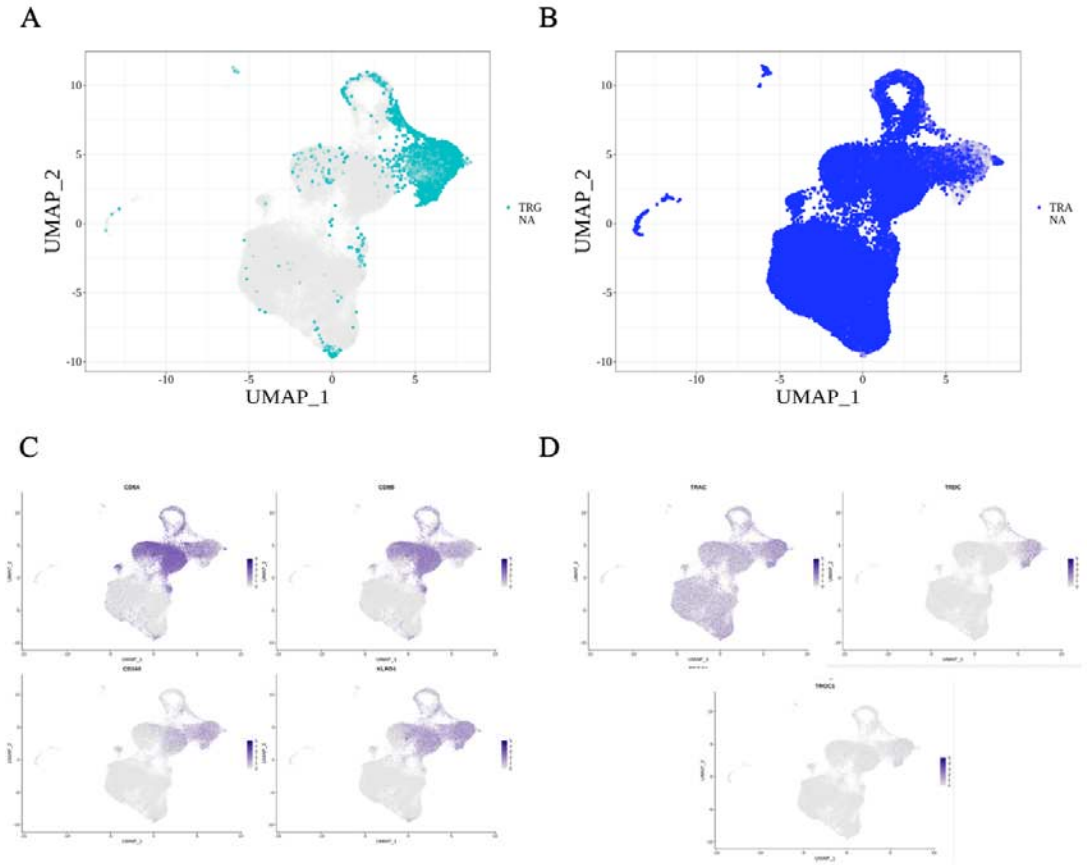

Figure 1. Understanding dynamic and variability of TCR expression. (**A-B**) UMAP plot representing the chain expression of the (A) TRA gene and (B) TRG that was derived from the TCR-seq data-level. (C) CD8 transcripts (CD8A and CD8B) and NK receptors (CD160 and KLRD1) expression. (**D**) Constant T cell gene markers (TRAC, TRDC, TRGC1).

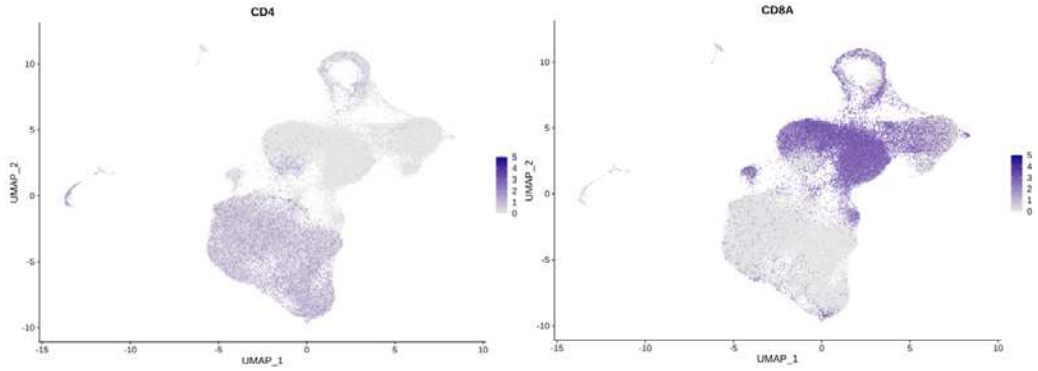

Figure 2. Limitation in CD4 expression. (**left**) CD4 and (**right**) CD8A expression.

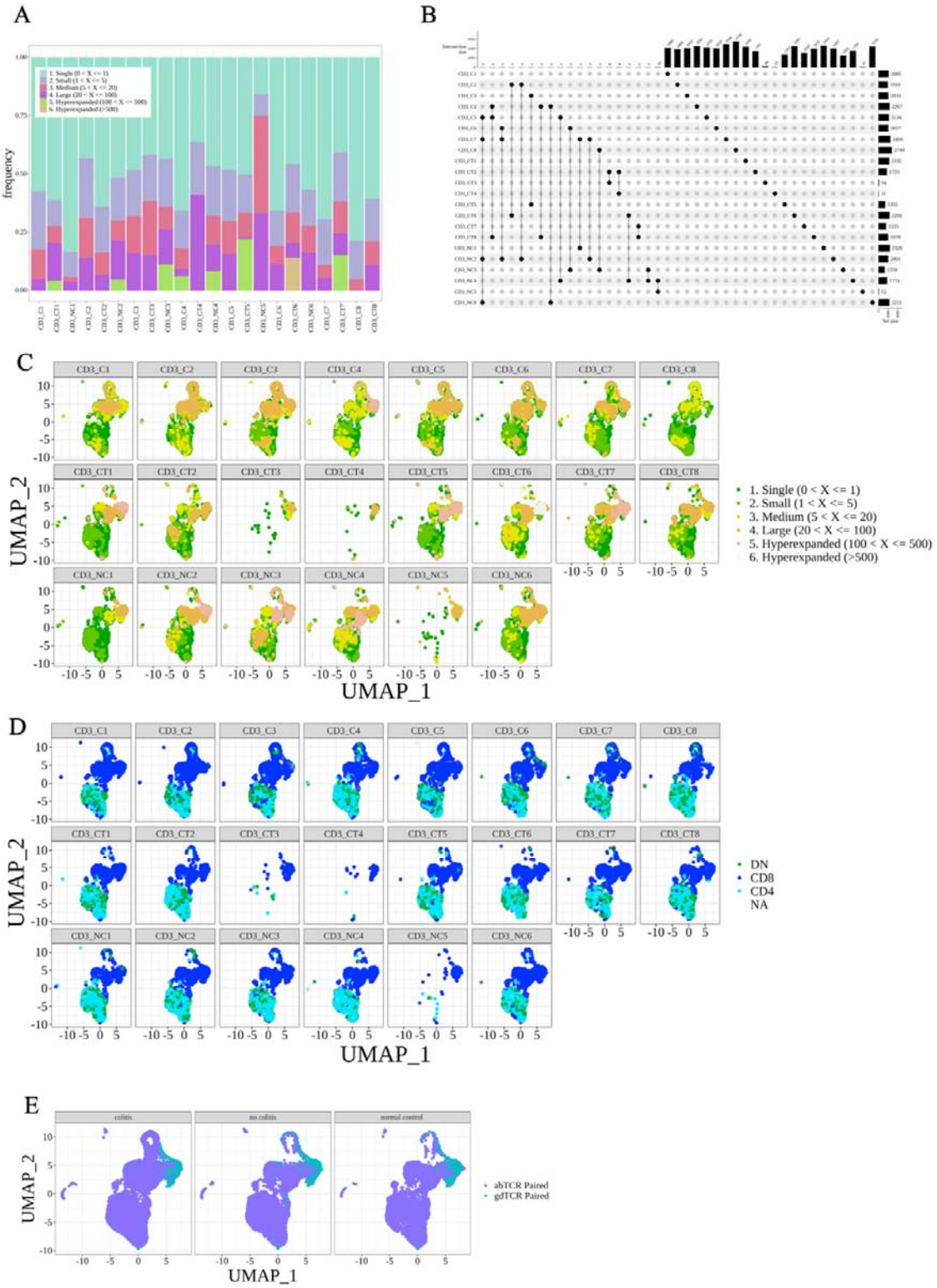

Figure 3. Exploring the gene expression the TCR expansion of the colitis dataset. (**A**) Clonal expansion based on frequency of clones across. (**B**) Upset plot of the paired TCR including both the V(D)J genes and CDR3 sequence. An upset plot similar to Venn diagrams represent numbers of overlapping samples. The black dot represents if a sequence was present in any given sample. The lines indicate if the sequence was present in multiple samples. The bar graphs on the top and right represent the number of total unique clones. (**C-D**) UMAP plot of all samples coloured by (**C**) clonal frequency or (**D**) generic T cell markers. (**E**) UMAP plot of the three groups coloured by the TCR pairing. C = colitis, CT = normal control and NC = no-colitis.

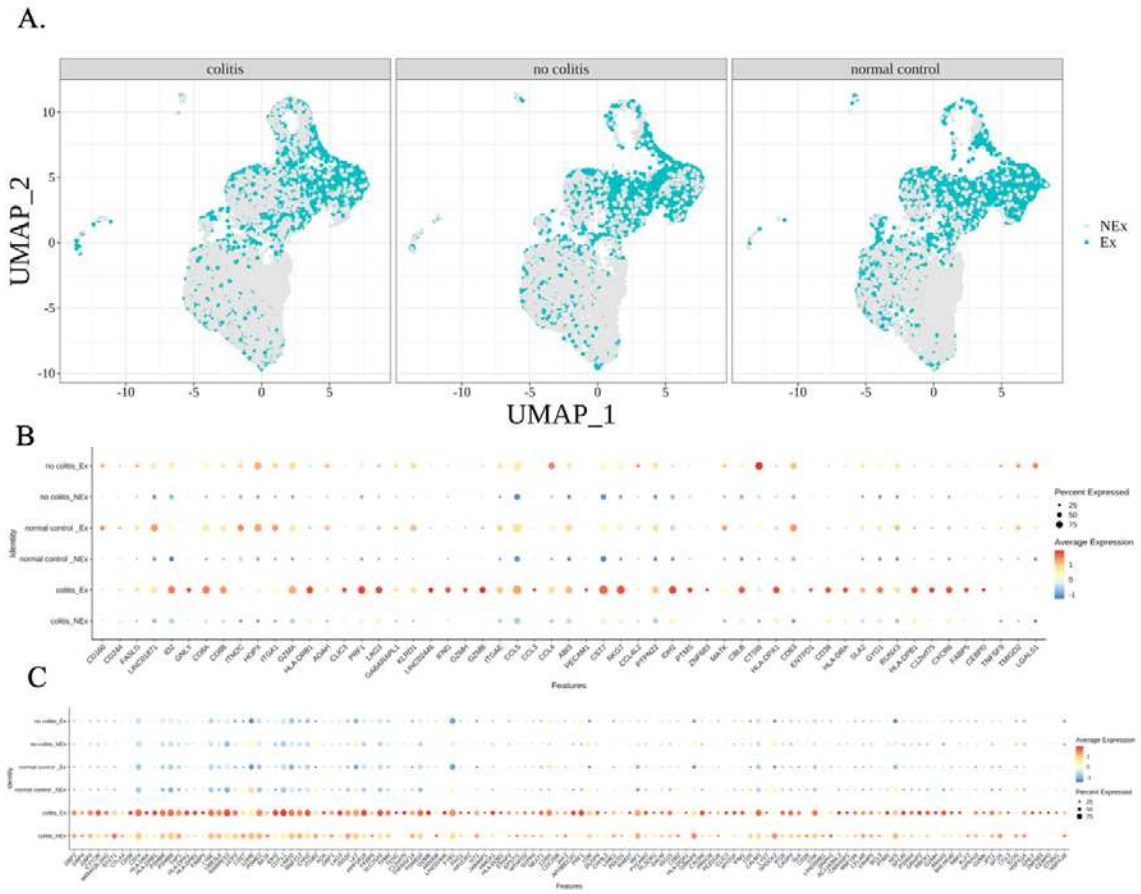

Figure 4. Interrogation of the expanded colitis expanded clones. (**A**) UMAP plot of the three groups coloured by the TCR pairing coloured by expansion (n≥3). (**B-C**) Dot plots showcasing the relative significant expression (p<0.001) for (**B**) Expanded colitis vs non-expanded colitis and (**C**) colitis compared to both the normal controls and non-colitis T cells. C = colitis, CT = normal control and NC = no-colitis.

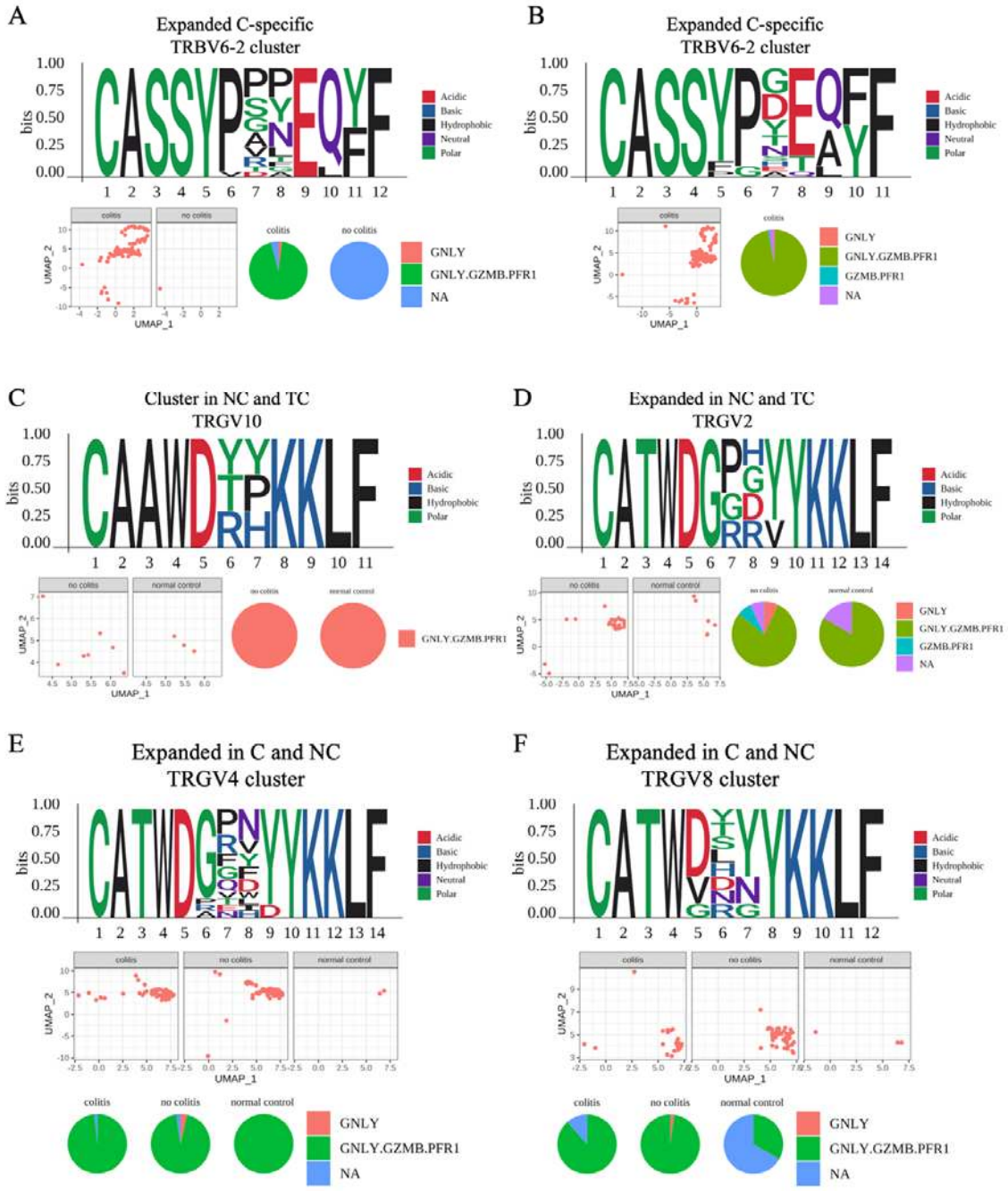

Figure 5. Interrogating the TCR CDR3 clustering to identify conditions specific sequences. (A-F) Each of the plots contains a (**top**) motif consensus sequence, (**middle**) UMAP location and (**bottom**) cytotoxic expression of GNLY, GZMB and PFR1. (**A-B**) represent two TRBV6-2 clusters that were specific to colitis. (**C-D**) represent two gamma clusters identified in non-colitis and normal controls. (**E-F**) Cluster over-represented in the melanoma patients. C = colitis, CT = normal control and NC = no-colitis.