Single-Cell RNAseq Data Preprocessing

Checkpoint @ 2020/03

Bastien JOB, Bioinformatics Core Facility

Reminder : Main steps of a preprocessing pipeline

From reads to normalized data :

- Reads QC [FastQC / FastqScreen]
- Reads processing (trimming, **decontamination**) [fastp / **Xenome**]
- Count matrix generation [Kallisto BUStools]
- Empty droplets filtering [DropletUtils::emptyDrops]
- Cells QC (#features, #counts) [Seurat]
- Bias evaluation (Library size, cell phase, %mito, %ribo, **%stress**) [scran/Seurat]
- Cells /genes filtering (Metrics)
- Normalization [SCTransform]
- Bias regression [Seurat]
    - **Which bias to regress ?**

Reminder : Main steps of a preprocessing pipeline

From normalized data to downstream analysis :

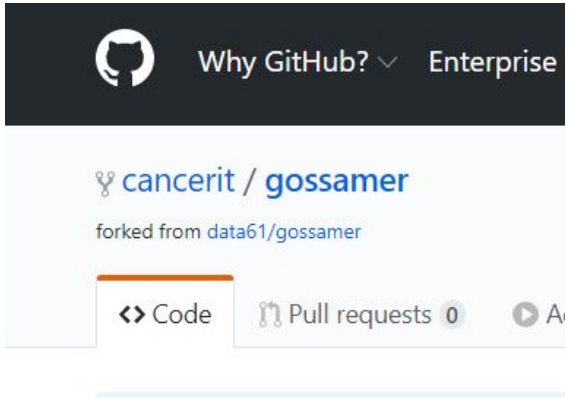
- Early dimension reduction (PCA, **MDS**, **ICA, BPCA, BFA**) [Seurat]
  - **How many dims to keep ?**
- Clustering (Louvain) [Seurat]
  - **How many clusters ?**
- Second dimension reduction for visualization (tSNE, uMAP)
- Dimension plots (metrics, given markers)
- **Quick differential analysis at high stringency**
- **Automatic cell type prediction**
- **Visualization package for project holder [Cerebro]**

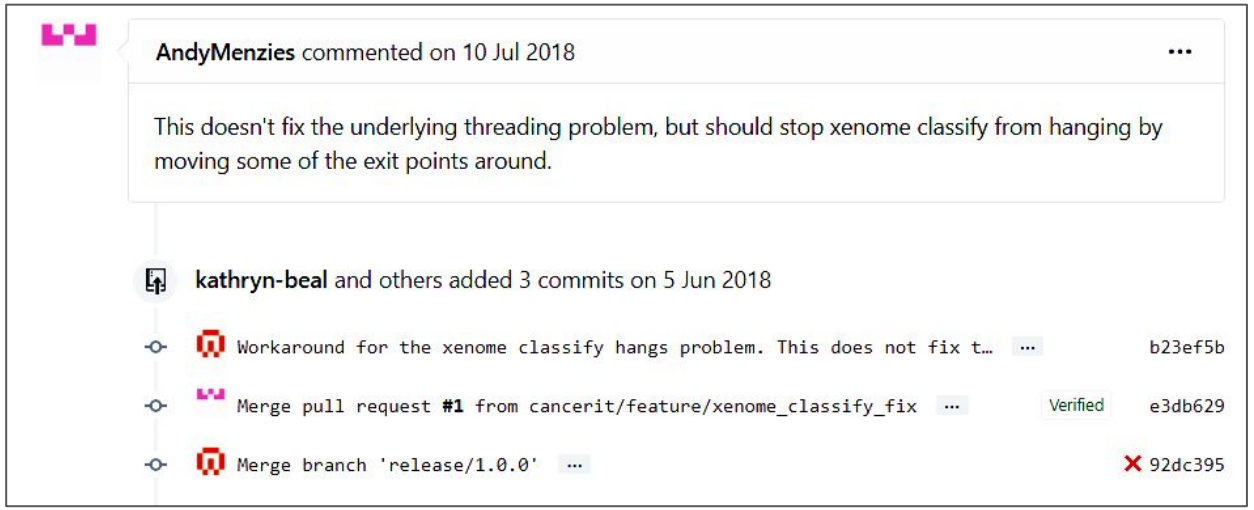
Reminder : AN ITERATIVE PROCESS

scRNAseq analysis consists in an iterative process !!

- First : analysis on a **completely unfiltered** (ie, post- empty droplets filtering) cell count matrix
  - To observe unexpected amount of cells, unexpected sub-populations (ex : ribo-less clusters), evaluate the fraction of cells that *would* be filtered), etc…
- Second : analysis on a **filtered** dataset, with **doublets kept**
  - Just to visualize them, and assess the effect of filters
- Third : same as second but with **doublets removed**, without any **covariate regression**
  - To measure effect of covariates (bias sources) on the reduced space
- Fourth, up to ∞ : same as third but **with covariate(s) regression**. Ends when/if:
  - No need for any particular regression
  - Good covariate(s) combo found (ie, reduces biases without altering true signal)
  - Death by boredom / frustration of your bioinformatician

Reads de-contamination : XENOME

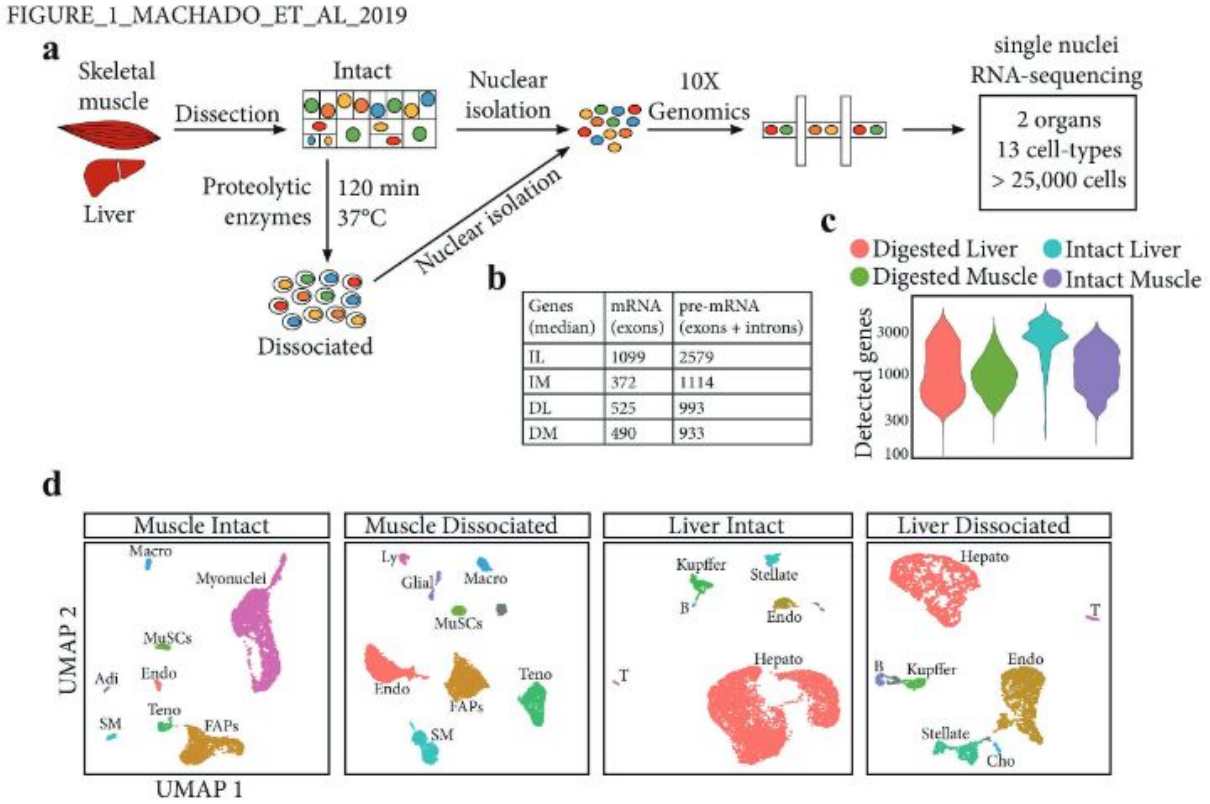




- XENOME classify : an **old** (last update in 2012), **neverending**, yet very useful tool
  - Alternatives like *bamcmp* or *BBsplit* are less efficient
  - Alternative *XenoFilter* claims to be more efficient but is a hassle (two mappings)a
- A (grossly) patched version is available @https://github.com/cancerit/gossamer
- For our pipeline : need a script to filter out non-graft R1 reads prior to Kallisto
  - (Hanane ? 😄)

Bias evaluation : Stress response genes

- Unpublished (about to be) results by Machado et al.
- Interest studying the early response to mechanical stress on cells.
- Generated single-nuclei atlas of intact, dissociated and injured organs (muscle)
- Performed time-course analysis
- Obtained a broadly conserved signature (~90 symbols)



Dimension reduction alternative to PCA : **MDS**

Multi-Dimensional Scaling

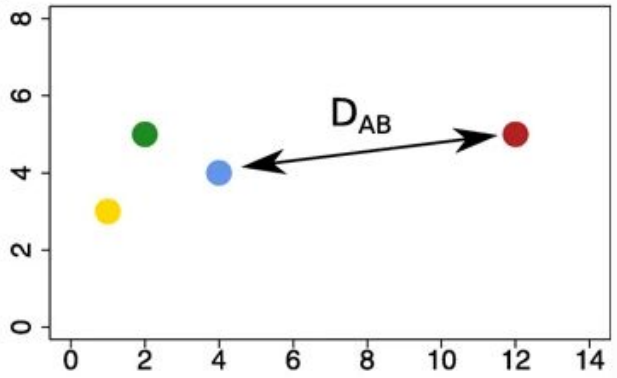
- Attempts to represent a distance matrix of M samples in a reduced N<M space (typically a map)
- Actually a FA method
- Very used by ecologists (for 2D maps)
- Among most performing and versatile methods for SC *(Sun et al, Genome Biol, 2019)*
- If data is centered + reduced = PCA !

Fig. 1

A

	A	B	C	D
A	0	d_{AB}	d_{AC}	d_{AD}
B	d_{AB}	0	d_{BC}	d_{BD}
C	d_{AC}	d_{BC}	0	d_{CD}
D	d_{AD}	d_{BD}	d_{CD}	0

B



C

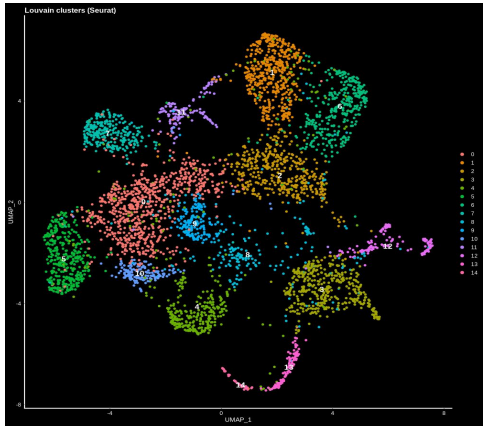
$$Stress = S = \sum (D_{ij} - d_{ij})^2$$

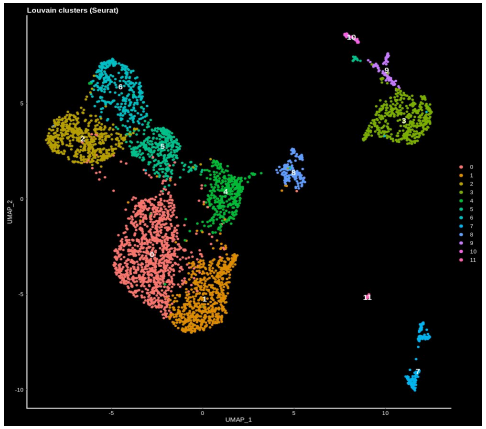
Schematic representation of the strategy for multidimensional scaling. **a** An example positive, symmetric matrix of distance values between four objects. **b** A dimension-reducing MDS representation of the distances in the matrix. **c** The stress equation for calculating the overall difference between the distances in the feature space (panel A, $d_{i,j}$) and the distances on the 2D plane (panel B, $D_{i,j}$)

Urpa & Anders, BMC Bioinformatics, 2019

MDS : Tests and observations

MDS

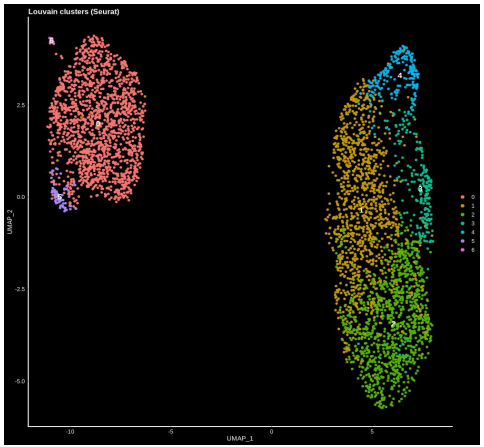


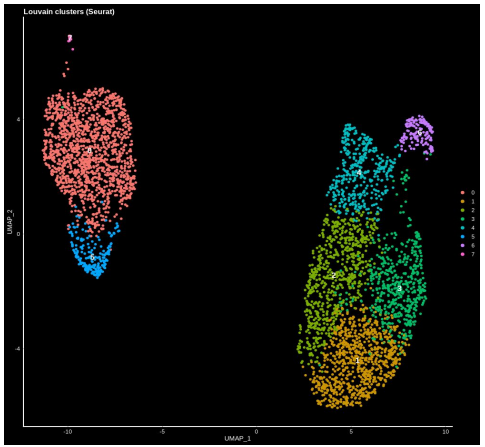


PCA

High QC sample

MDS



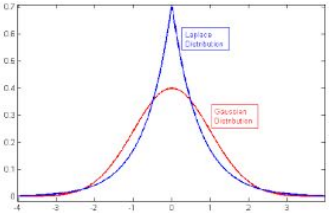


PCA

Low QC sample

Other alternatives to PCA : **ICA, scBFA, BPCA,**

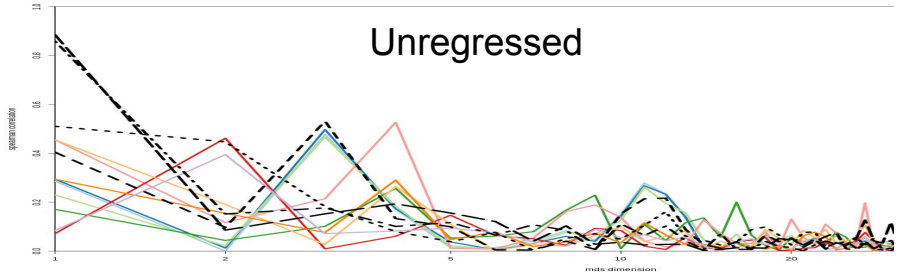
- **ICA *(fastica, through Seurat::RunICA)***
  - Generates independent components (ie, lowly correlated) through a rotation maximizing a measure of non-gaussianity, on a priorly “whitened” dataset (processed by a linear reduction method, typically SVD or EVD ==> PCA)
  - Useful In biology (Gaussianity = noise, Laplacianity = signal).
  - Fast
  - Hard to select and interpret components...

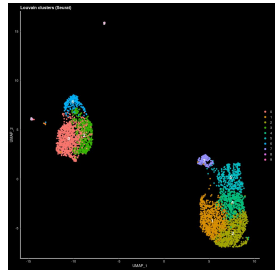


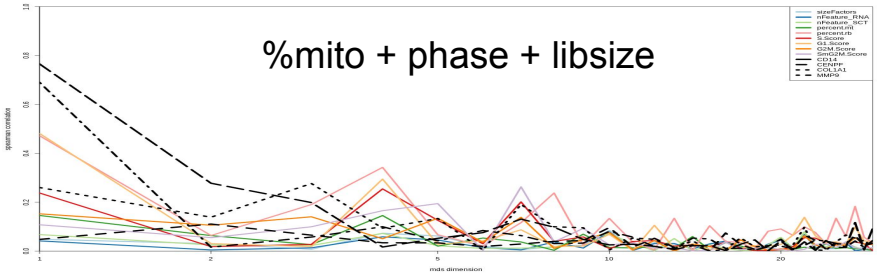
- scBFA (single cell Binary Factorial Analysis, *through the eponymous R package*)
  - From a **raw count matrix** : normalization + reduction combo
  - Simple in principle : binarization of raw counts (where any count > 0 is set to 1) then FA
  - Requires a genes selection step (HVGs)
  - Inappropriate for small datasets with high gene detection rate.
  - Claims to be much less sensitive to noise than count methods
  - Much slower than PCA or MDS (~2H / 5000 cells x), much faster than complex methods (ZinB, etc)
  - Compatible with scATAQseq
- **BPCA (Binary PCA)**
  - A faster hyper-approximation of scBFA

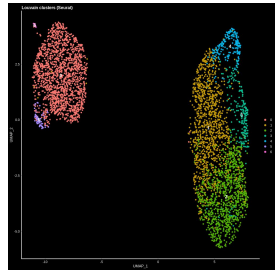
Coveriate(s) regression : correlation plot

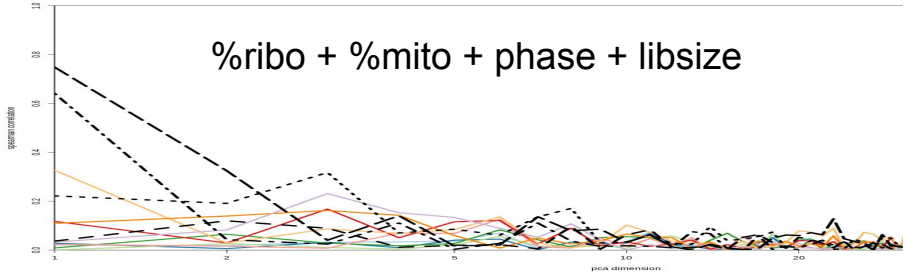
- After dimension reduction (any kind)
- Correlation of each covariate (bias, score) to each computed component
- Plot in log scale, with the help of expected markers genes (also correlated)
- Quite useful !
- But risk of **over-regression**...

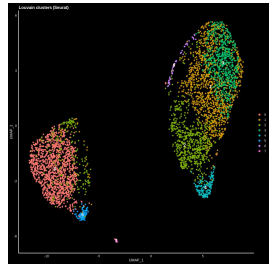






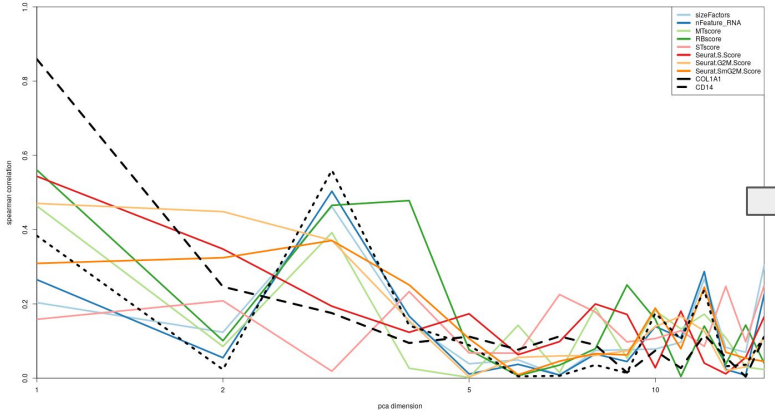




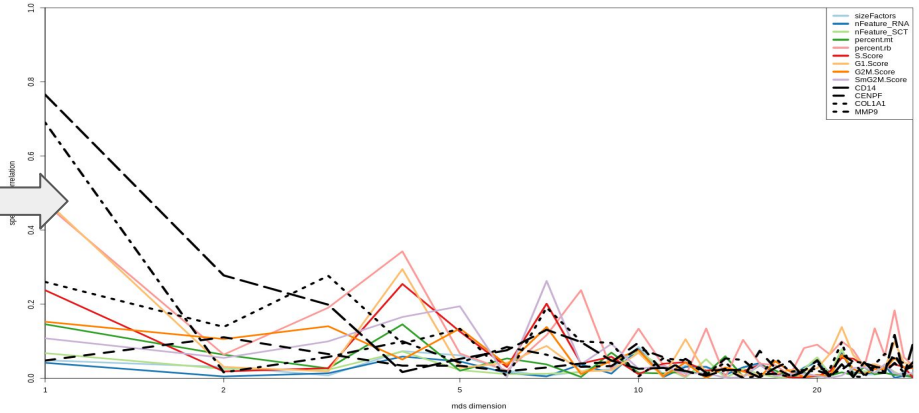


Coveriate(s) regression : correlation plot

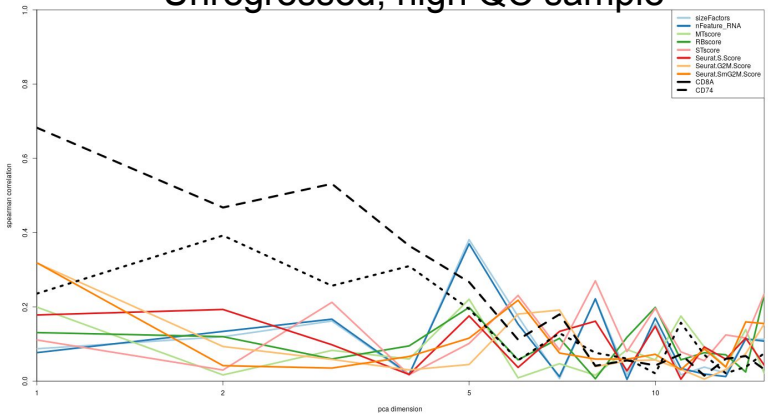
Unregressed, low QC sample (high %mt)



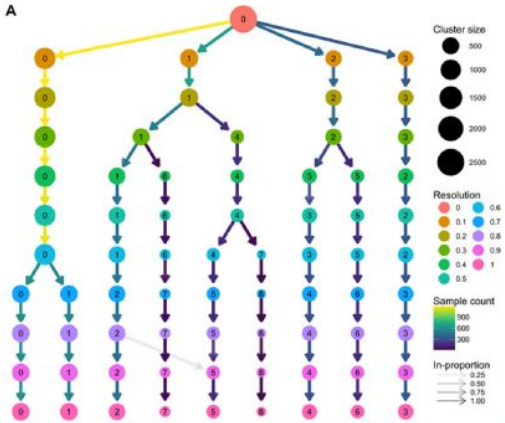
Regressed, low QC sample (high %mt)



Unregressed, high QC sample



Reduction : dimensions to keep ?
Clustering : at which resolution ?

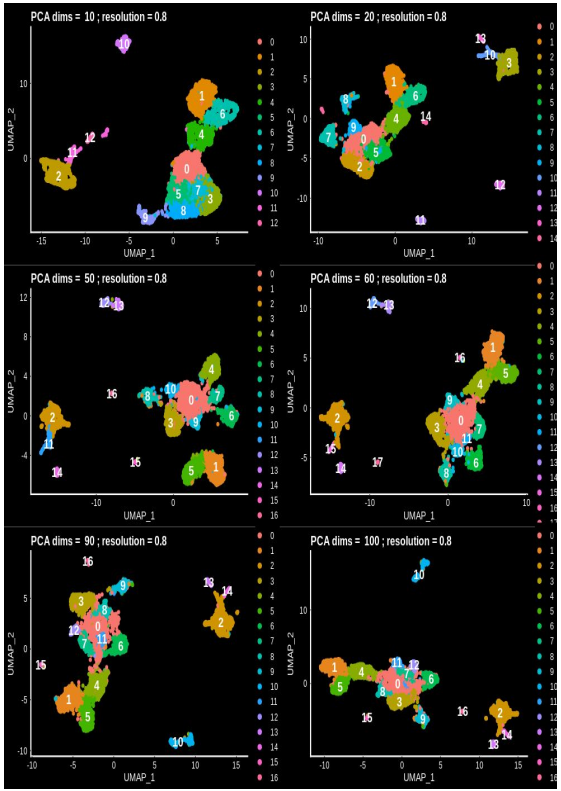


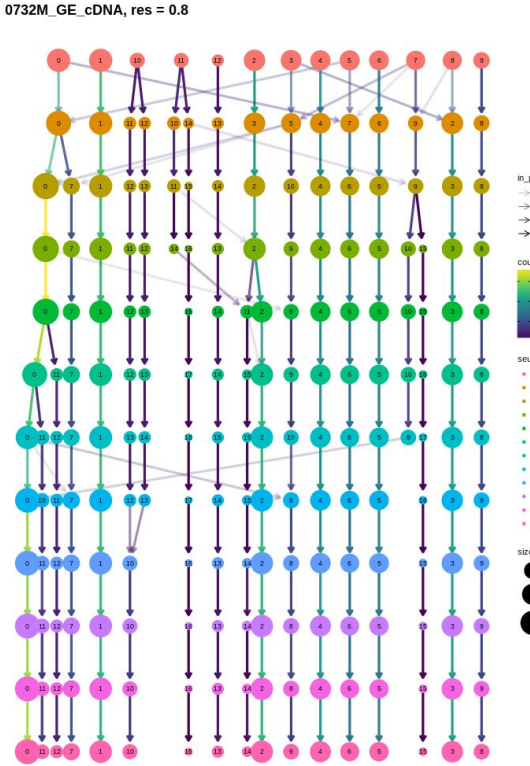
Clustering trees: a visualization for evaluating clusterings at multiple resolutions

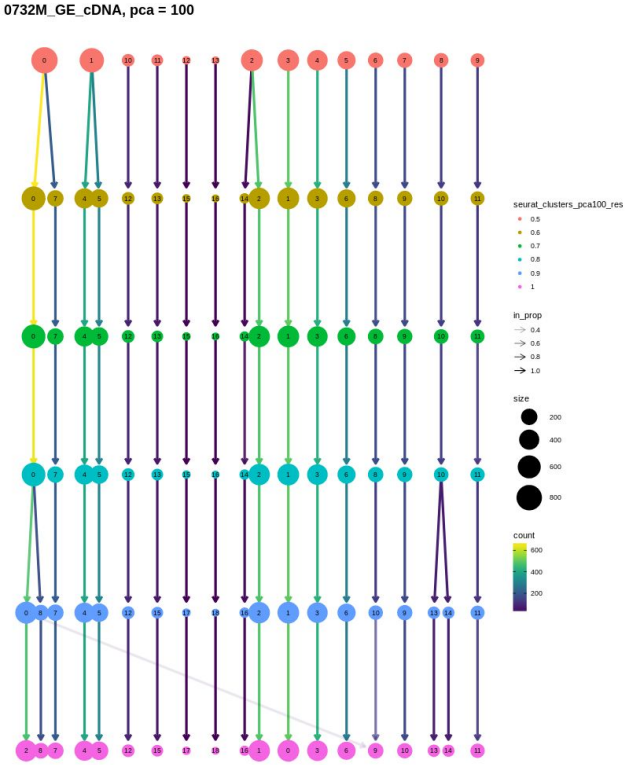
Luke Zappia [1,2] and Alicia Oshlack [1,2,*]

- Routine : double loop on both parameters. Ex :
  - dimensions <- c(seq.int(9, 25, 2), seq.int(30, 100, 10))
  - resolutions <- seq(.5, 1, .1)
- Looking for stability across contiguous iterations (mainly for dimensions)
- Also strongly relies on manual observation of resulting clusters based on uMAP reduction
  - Can converge to multiple solution (simpler @ low dim, low res ; complex @ higher)
- Alternatives :
  - Jackstraw plot (often hard to interpret, long, unstable)
  - Elbow in varplot (unreliable)

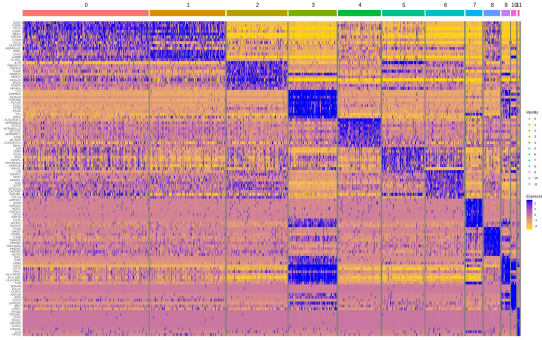
Reduction : dimensions to keep ?
Clustering : at which resolution ?

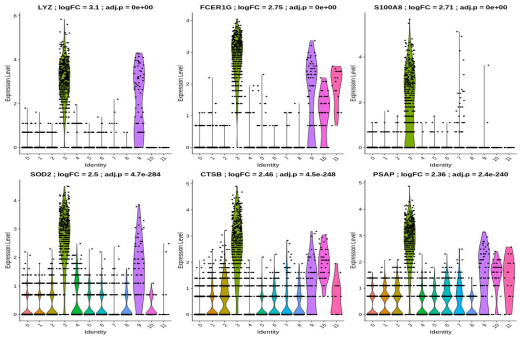


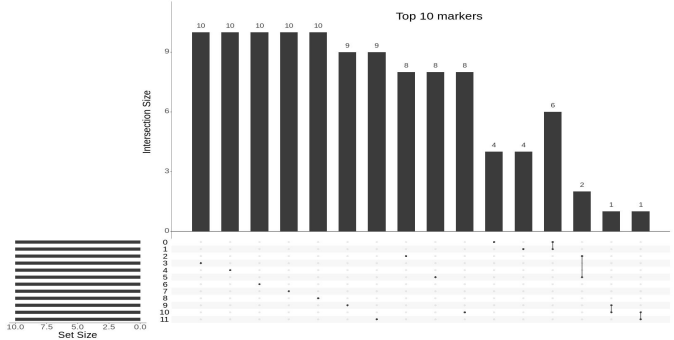


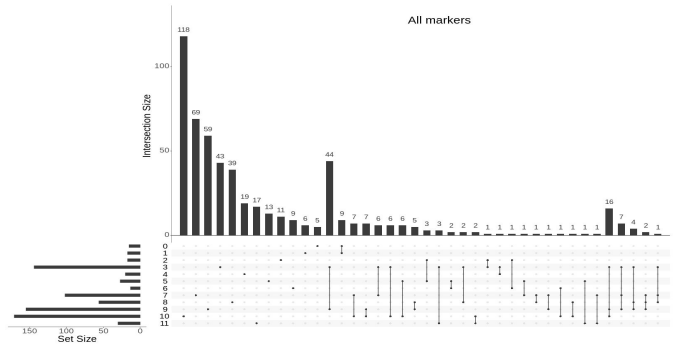


Quick differential expression analysis





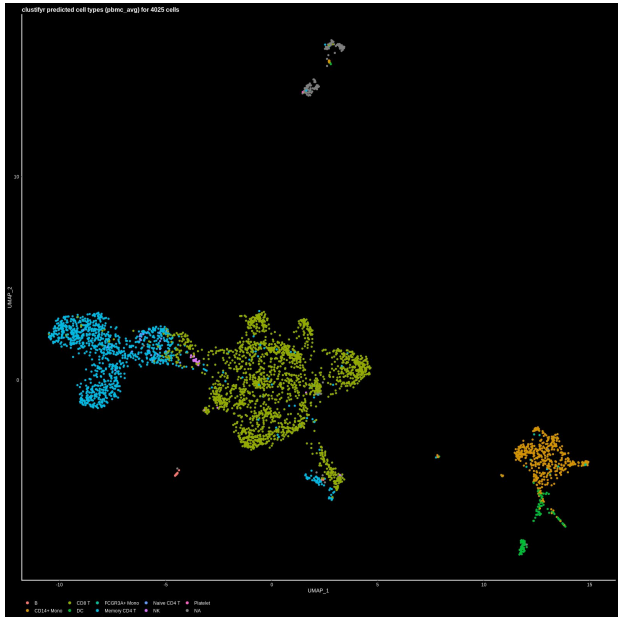


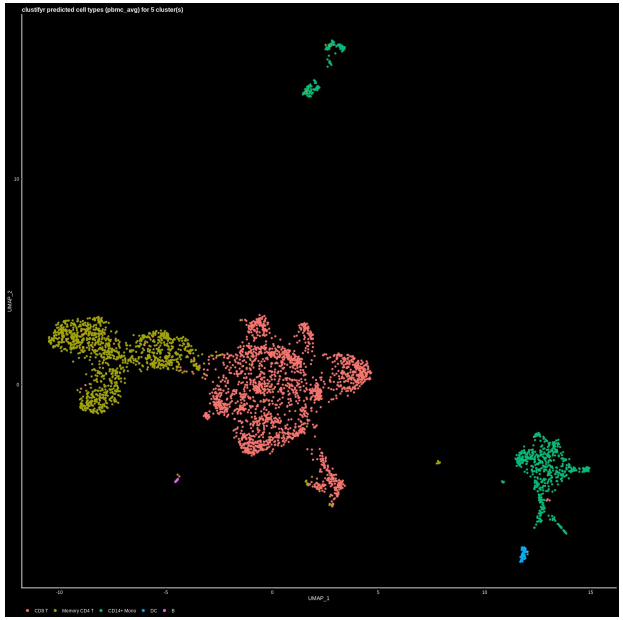


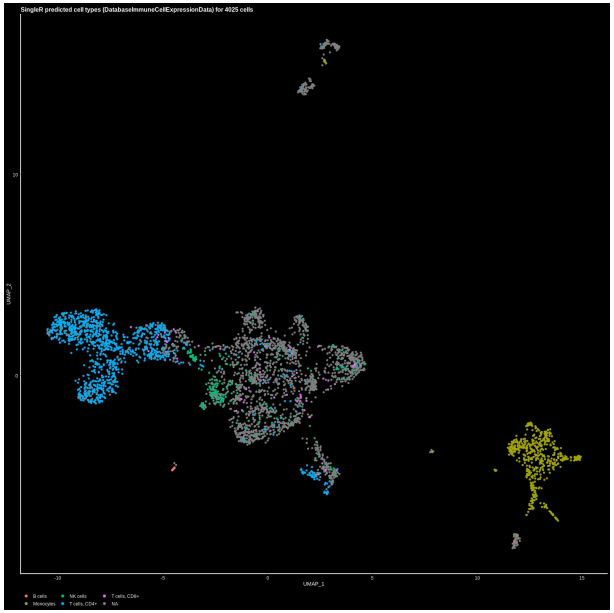
p_val	avg_logFC	pct.1	pct.2	p_val_adj	cluster	gene
2.58297337270421E-251	0.871144211245547	0.944	0.386	4.21308786821784E-247	0	CD8A
2.56177323228459E-230	0.887869679740215	0.907	0.398	4.1785083191794E-226	0	GZMH
4.93362404314396E-224	0.856845612462005	0.997	0.659	8.04723417677211E-220	0	CCL5
1.84392105905805E-201	0.758818133234689	0.927	0.479	3.00761963942958E-197	0	GZMA
1.10003408798873E-181	0.592368007448339	0.972	0.523	1.79426560091841E-177	0	NKG7
4.69253598720831E-179	1.07762545332016	0.801	0.359	7.65399544873547E-175	0	CRTAM
7.92413505426953E-169	0.607969572089822	0.897	0.432	1.2925056687019E-164	0	CTSW
1.23255351457246E-157	0.553656085000971	0.824	0.323	2.01041803761914E-153	0	KLRD1
8.42345247114342E-155	0.970126999977904	0.906	0.554	1.3739493325682E-150	0	CCL4
6.70903818248469E-141	0.543329698095384	0.946	0.533	1.09431121794508E-136	0	GZMB
5.48893111731593E-132	0.579866351703322	0.911	0.624	8.95299554545402E-128	0	CLEC2B
8.6182755235572E-103	0.531418072539514	0.999	0.986	1.40572692064742E-98	0	HSPA8
5.81644469259956E-100	0.610949368913948	0.998	0.995	9.48720293809915E-96	0	HSP90AA1
1.5961155191242E-99	0.508510864953599	0.853	0.603	2.60342402324349E-95	0	CCNH
4.14953275150606E-51	0.51456281801823	0.754	0.551	6.76830287098154E-47	0	TRBC2
2.89342529662031E-247	1.11939332271426	0.943	0.352	4.71946600131738E-243	1	PRF1
6.80893824428162E-244	1.43584341807484	0.988	0.575	1.11060591702477E-239	1	NKG7
2.0047177467359E-231	1.15105634528893	0.94	0.432	3.26989511670092E-227	1	LAG3
6.22342220275014E-211	1.26898045991644	0.993	0.575	1.01510239549057E-206	1	GZMB
9.84984669996772E-172	0.766421632529232	0.915	0.368	1.60660849523174E-167	1	KLRD1
1.74649590745848E-151	0.961722350823324	0.991	0.702	2.84870947465552E-147	1	CCL5
2.29955215763085E-142	0.780241905373246	0.924	0.485	3.75079952431168E-138	1	CTSW
6.36893733713504E-135	0.831991759345611	0.876	0.468	1.0388373690601E-130	1	CD8A
1.93701920391684E-132	0.762288049108	0.925	0.457	3.15947202350876E-128	1	GZMH
2.76765535096329E-132	0.61049668289943	0.806	0.353	4.51432264295622E-128	1	PTMS
1.87972037880779E-101	0.705419834330427	0.922	0.625	3.06601190987339E-97	1	CST7
2.21131546436679E-101	0.623876868746824	0.957	0.732	3.60687665393048E-97	1	CD7
3.66231726915558E-97	0.643776750821614	0.91	0.537	5.97360569771966E-93	1	GZMA
9.46284652502429E-93	0.740815004456164	0.901	0.662	1.54348489669671E-88	1	CLEC2B
3.55130193388166E-90	0.503821290345053	0.845	0.467	5.79252858435437E-86	1	CBLB
2.54297394646918E-73	0.767093679583648	0.776	0.418	4.14784480408589E-69	1	CRTAM
2.57101978622642E-67	0.599428411696192	0.821	0.576	4.19359037331391E-63	1	LITAF
4.59995743963475E-173	1.10775605706576	0.84	0.258	7.50299057978824E-169	2	IL7R
4.22031244159386E-155	0.797202091728955	0.792	0.292	6.88375162348374E-151	2	TMEM173
2.68295546074154E-144	1.07290454546495	0.913	0.601	4.37616865201547E-140	2	YWHAQ
1.98937794241853E-138	0.992042573093295	1	0.83	3.24487436187887E-134	2	TXNIP
2.63637561296245E-133	0.797308787200908	1	0.999	4.30019226230305E-129	2	TMSB10

- **NOT** a real, refined DE analysis, **JUST** an assessment of the retieval of expected markers
- One vs other DE test for each cluster (Seurat::FindAllMarkers)
- Extremely high stringency : Wilcoxon test, genes expressed in at least 90% of cells, over-expressed only, logFC >= 0.5, focus on top10 per cluster.
- Useful to assess over-clustering

Automatic cell type annotation

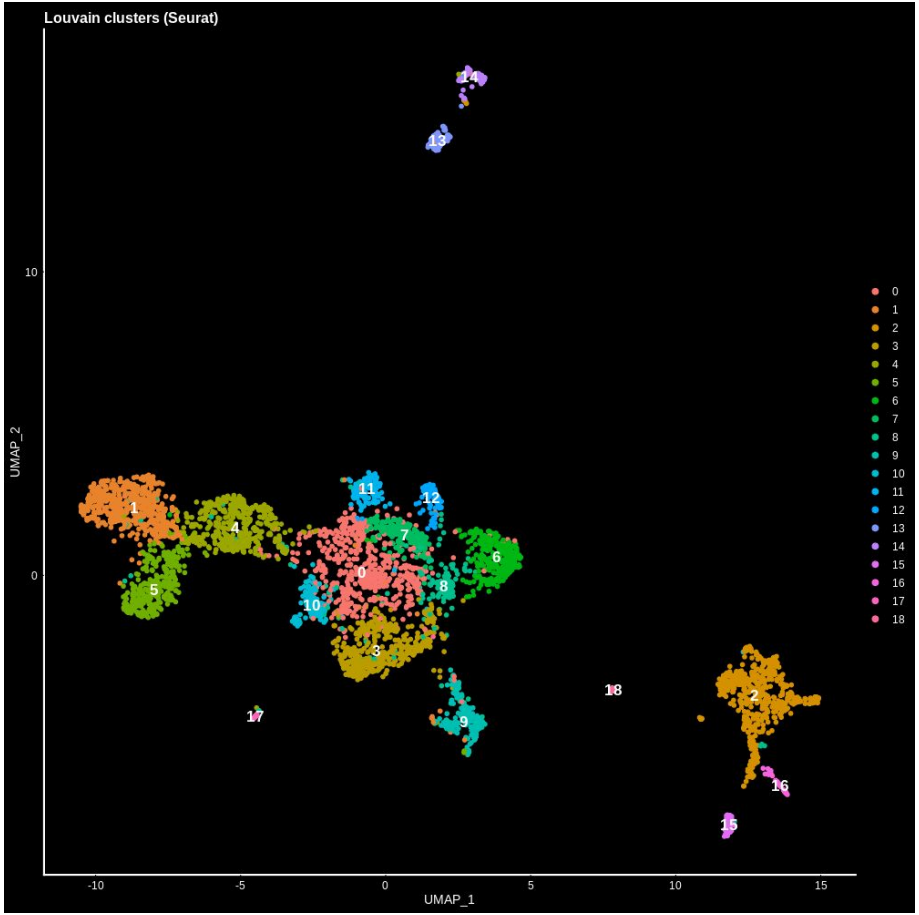






- Relies on *singleR* and *clustifyr* (and their respective databases)
- Rough estimation
- Scoring either by cell, or by cluster
- Included databases are mostly immuno / hemato…

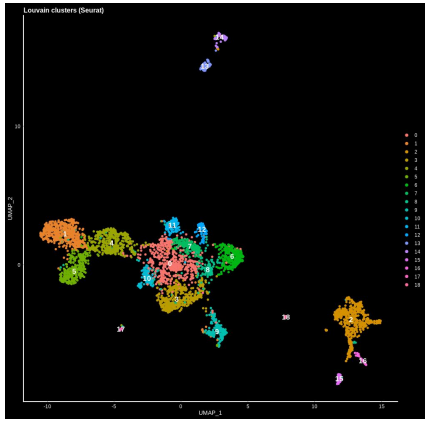
New dataset : 5’ CITE-seq + TCR immunoprofiling

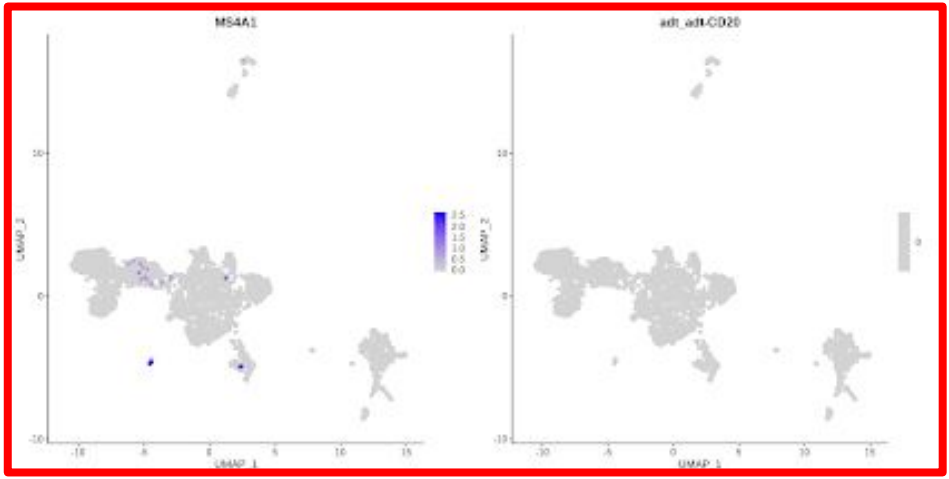


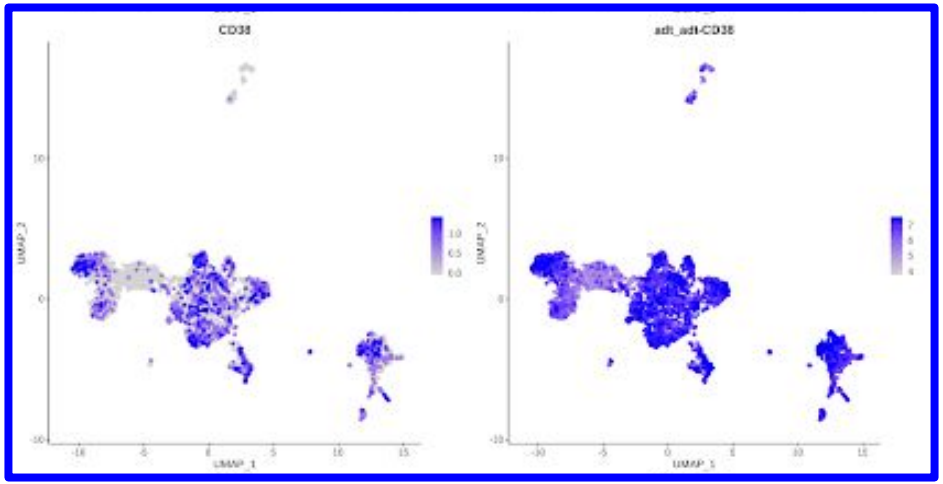
- 3 libraires :
  - RNA expression
  - Membrane proteins Ab (12)
  - TCR profiling
- Project P30_FXDA (François-Xavier Danlos)
- 2 samples
- QC : best yet @ GR !
- ~ 4,000 cells

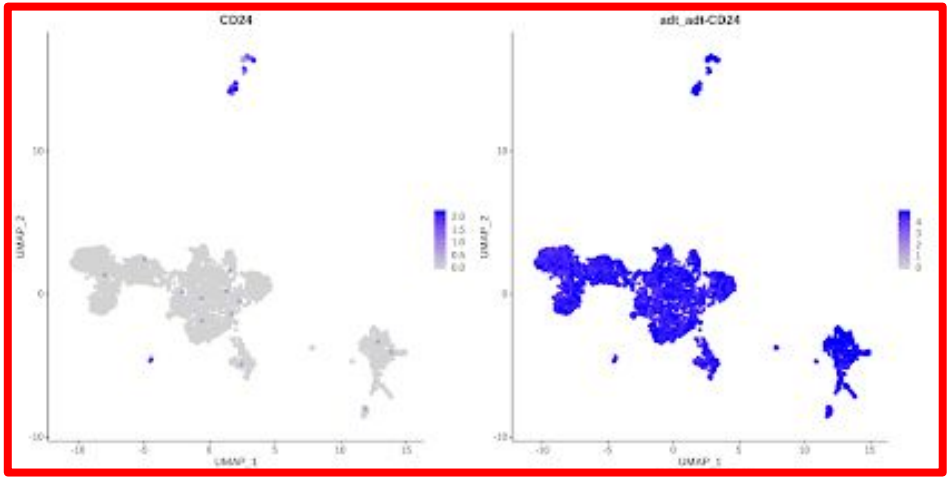
5’ CITE-seq + TCR immunoprofiling

- ~ 40% of concordance between expression and protein level
- Unexpectedly : Protein >> Exp ?!









5’ CITE-seq **+ TCR immunoprofiling**

- Performed through 10X’s CellRanger v3+, VDJ mode (no real alternative… *ImmCantation* ?)
- Generates a characterization of cells, used as new meta data

barcode	is_cell	contig_id	high_confidence	length	chain	v_gene	d_gene	j_gene	c_gene	full_length	productive	cdr3
AAACCTGAGCGATAGC-1	True	AAACCTGAGCGATAGC-1_contig_1	True	506	TRA	TRAV25	None	TRAJ56	TRAC	True	True	CAGRDGANSKLTF
AAACCTGAGCGATAGC-1	True	AAACCTGAGCGATAGC-1_contig_2	True	512	TRA	TRAV25	None	TRAJ53	TRAC	True	True	CAGRGSGGSNYKLTF
AAACCTGAGCGATAGC-1	True	AAACCTGAGCGATAGC-1_contig_3	True	485	TRB	TRBV27	None	TRBJ2-1	TRBC2	True	True	CASSFSDRREQFF
AAACCTGAGCGATAGC-1	True	AAACCTGAGCGATAGC-1_contig_4	True	378	TRB	None	None	TRBJ1-1	TRBC1	False	False	None
AAACCTGAGTACACCT-1	True	AAACCTGAGTACACCT-1_contig_1	True	467	TRA	TRAV21	None	TRAJ9	TRAC	True	True	CAVRPRDTGGFKTIF
AAACCTGAGTACACCT-1	True	AAACCTGAGTACACCT-1_contig_2	True	310	TRB	TRBV30	None	TRBJ1-2	TRBC1	True	False	CAWWGRGYTF
AAACCTGCACTCAGGC-1	True	AAACCTGCACTCAGGC-1_contig_1	True	510	TRB	TRBV7-9	None	TRBJ2-7	TRBC2	True	True	CASSPPRGPFSYEQYF
AAACCTGCACTCAGGC-1	True	AAACCTGCACTCAGGC-1_contig_2	True	498	TRA	TRAV9-2	None	TRAJ44	TRAC	True	True	CALSLLTGTASKLTF
AAACCTGCACTCAGGC-1	True	AAACCTGCACTCAGGC-1_contig_3	True	519	TRA	TRAV27	None	TRAJ34	TRAC	True	False	CAGDDRQAHLWDW
AAACCTGGTGCTTCTC-1	True	AAACCTGGTGCTTCTC-1_contig_1	True	549	TRB	TRBV30	None	TRBJ2-7	TRBC2	True	True	CAWRGITYEQYF
AAACCTGGTGCTTCTC-1	True	AAACCTGGTGCTTCTC-1_contig_2	True	690	TRA	TRAV8-3	None	TRAJ23	TRAC	True	True	CAVGTFYNQGGKLIF

cdr3	cdr3_nt	reads	umis	raw_clonotype_id	raw_consensus_id
CAGRDGANSKLTF	TGTGCAGGGCGGGATGGAGCCAATAGTAAGCTGACATTT	2282	5	clonotype41	clonotype41_consensus_3
CAGRGSGGSNYKLTF	TGTGCAGGGCGGGGGAGTGGAGGTAGCAACTATAAACTGACATTT	1300	3	clonotype41	clonotype41_consensus_1
CASSFSDRREQFF	TGTGCCAGCAGTTTCTCGGACAGGAGGGAGCAGTTCTTC	3987	10	clonotype41	clonotype41_consensus_2
None	None	5350	10	clonotype41	None
CAVRPRDTGGFKTIF	TGTGCTGTGAGACCCAGGGATACTGGAGGCTTCAAAACTATCTTT	599	2	clonotype289	clonotype289_consensus_1
CAWWGRGYTF	TGTGCCTGGTGGGGGAGGGGCTACACCTTC	707	2	clonotype289	None
CASSPPRGPFSYEQYF	TGTGCCAGCAGCCCCCCCCGGGGGCCCTTCTCCTACGAGCAGTACTTC	7466	18	clonotype3	clonotype3_consensus_1
CALSLLTGTASKLTF	TGTGCTCTGAGTCTCCTTACCGGCACTGCCAGTAAACTCACCTTT	1453	3	clonotype3	clonotype3_consensus_2
CAGDDRQAHLWDW	TGTGCAGGAGATGACCGACAAGCTCATCTTTGGGACTGG	2283	6	clonotype3	None
CAWRGITYEQYF	TGTGCCTGGCGGGGTATCACCTACGAGCAGTACTTC	4271	14	clonotype290	clonotype290_consensus_2