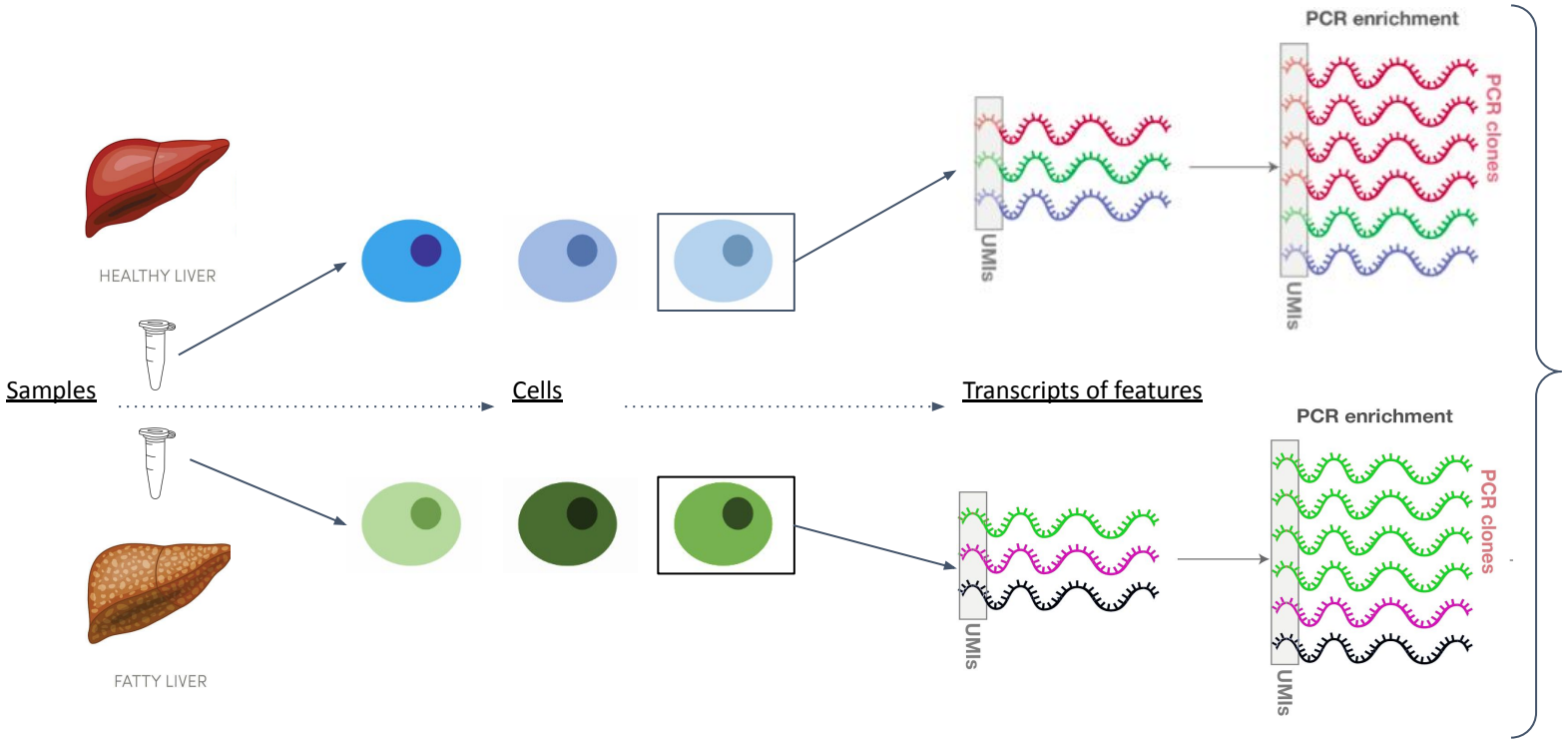
From Raw Read Data to Cell Clustering: Cell ranger and Seurat

David Carbonez

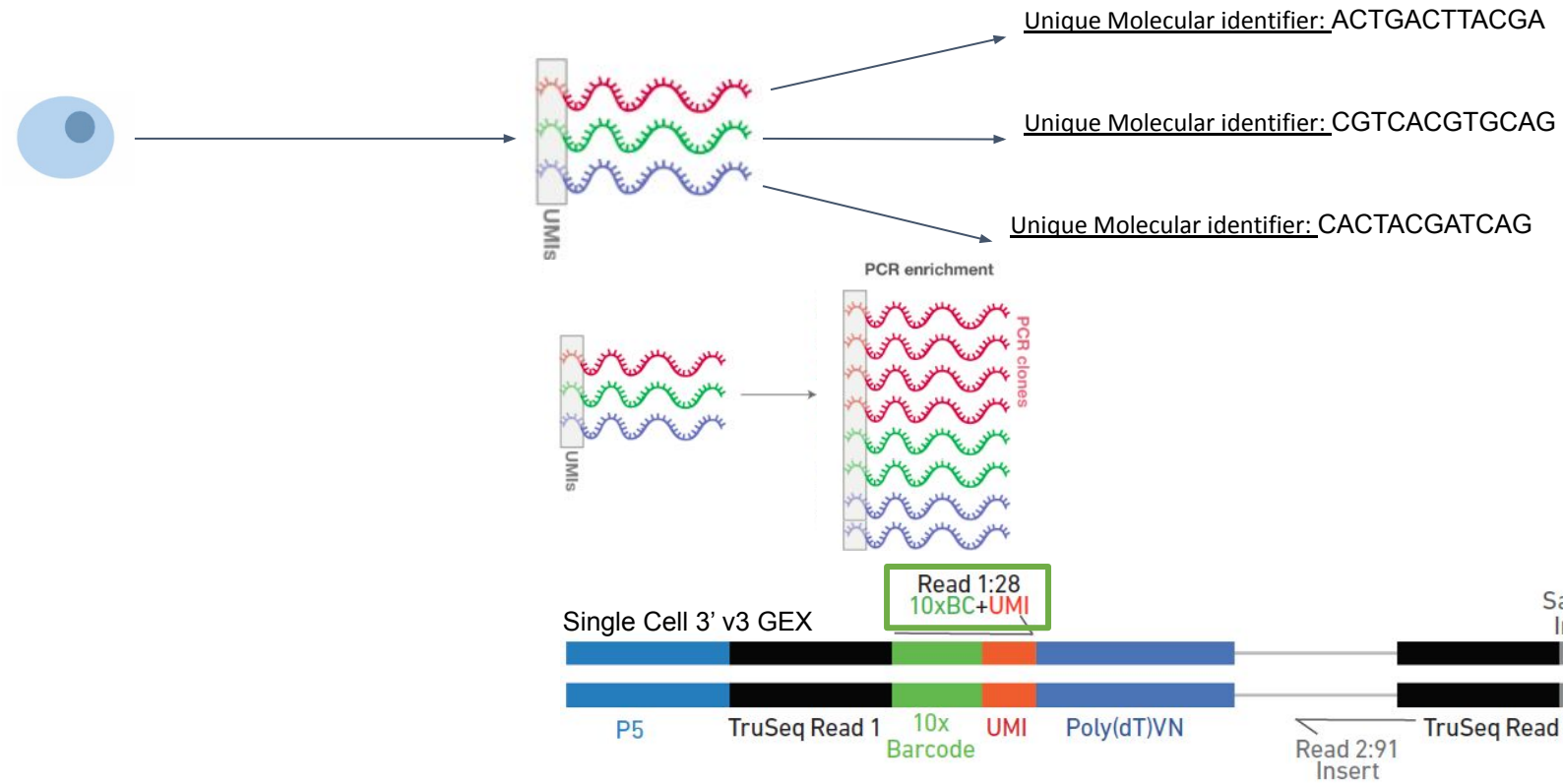
From Raw read to Cell Clustering

1. Part 1: Preparation of the data
    a. Introduction & Terminology
    b. The anatomy of a 10x Construct
    c. Sequencing
    d. Demultiplexing & Counting
2. Part 2: Analysis of the data

Introduction



UMI

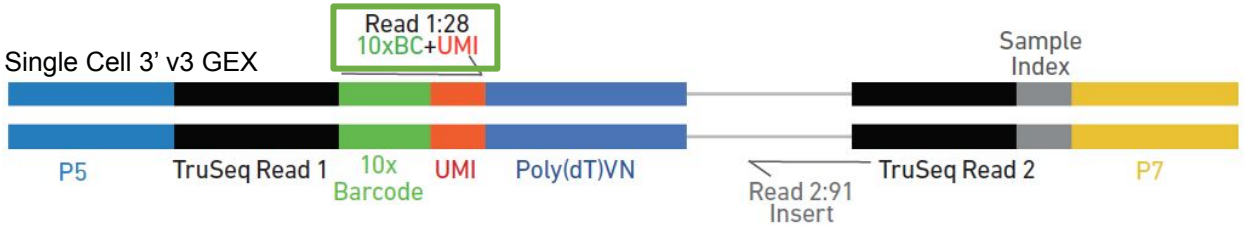

Unique Molecular identifier: ACTGACTTACGA
Unique Molecular identifier: CGTCACGTGCAG
Unique Molecular identifier: CACTACGATCAG
UMIs
PCR enrichment
PCR clones
UMIs
Single Cell 3' v3 GEX
Read 1:28
10xBC+UMI
Sample
Index
P5
TruSeq Read 1
10x
Barcode
UMI
Poly(dT)VN
Read 2:91
Insert
TruSeq Read 2
P7

Cell Barcode

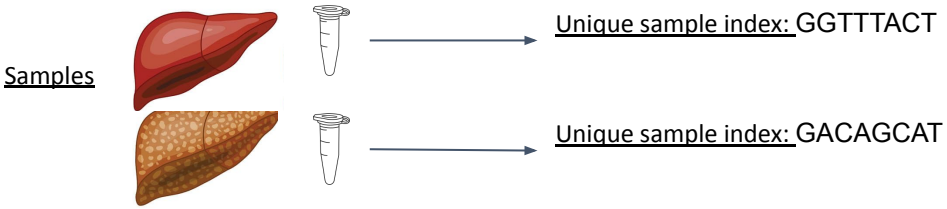
Unique Cell barcode: AACCTTGGAATATCTA

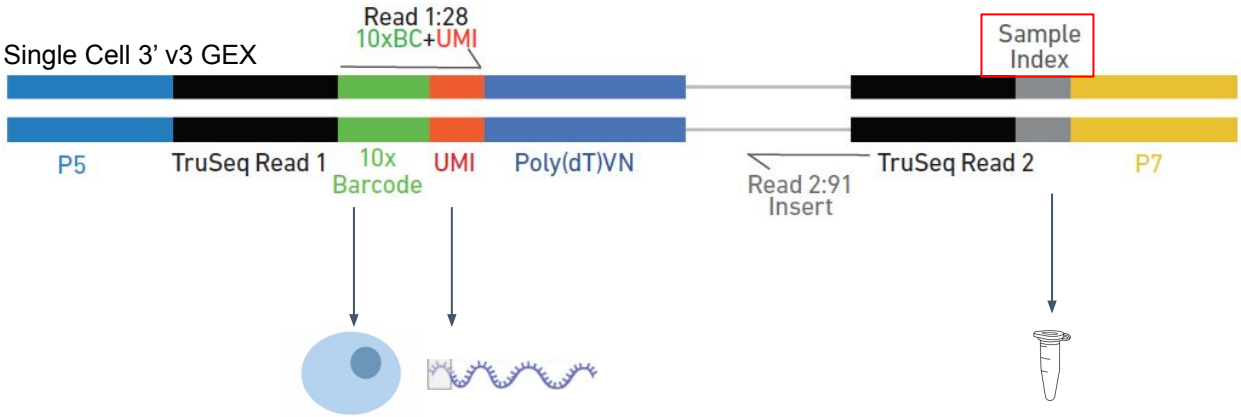
Unique Cell barcode: CTAACTAAGGCACGAC

Unique Cell barcode: GAACTCGATTCTTACG



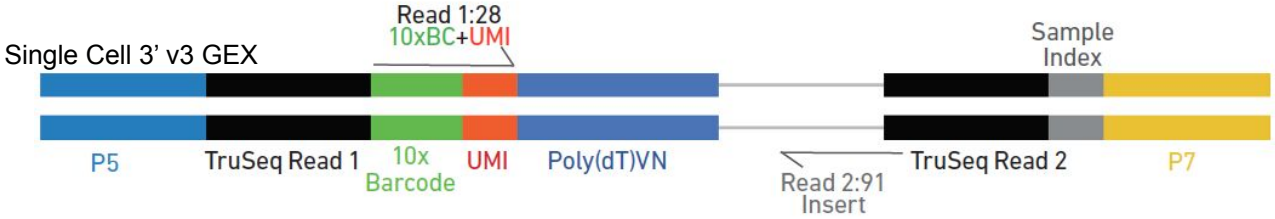
Sample index

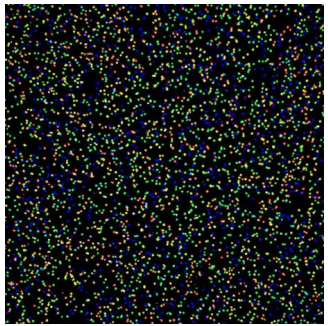


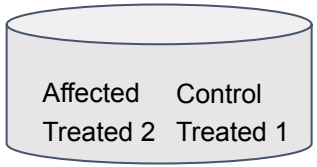


Sequencing

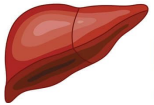
BioIT analysis: Demultiplexing







Index.fastq.gz
Read1.fastq.gz
Read2.fastq.gz

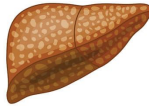


Unique sample index: GGTTTACT

Healthy data

Fatty data

Unique sample index: GACAGCAT



BioIT analysis: Counting

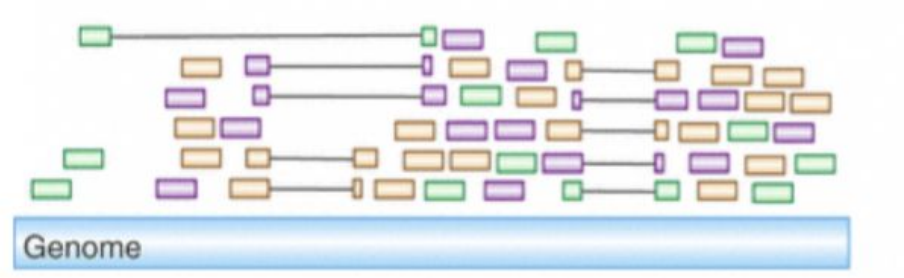
Healthy_I1.fastq.gz
Healthy_R1.fastq.gz
Healthy_R2.fastq.gz

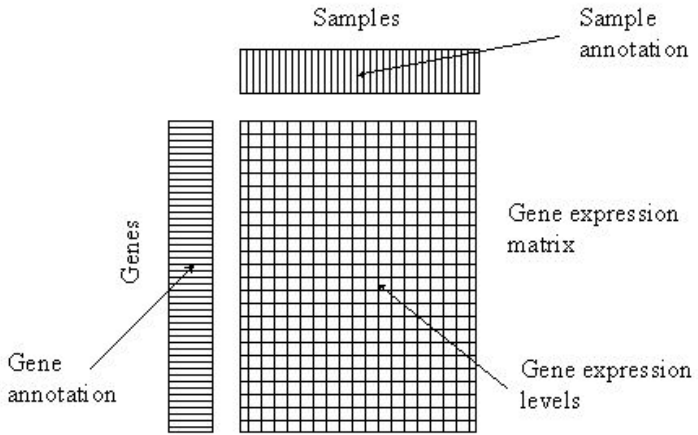
→

Sample indices (8bp)
10x Barcode (16bp) + **UMI** (12 bp)
mRNA (91 bp)

→

Identify sample
Identify cell + Transcript
Identify gene



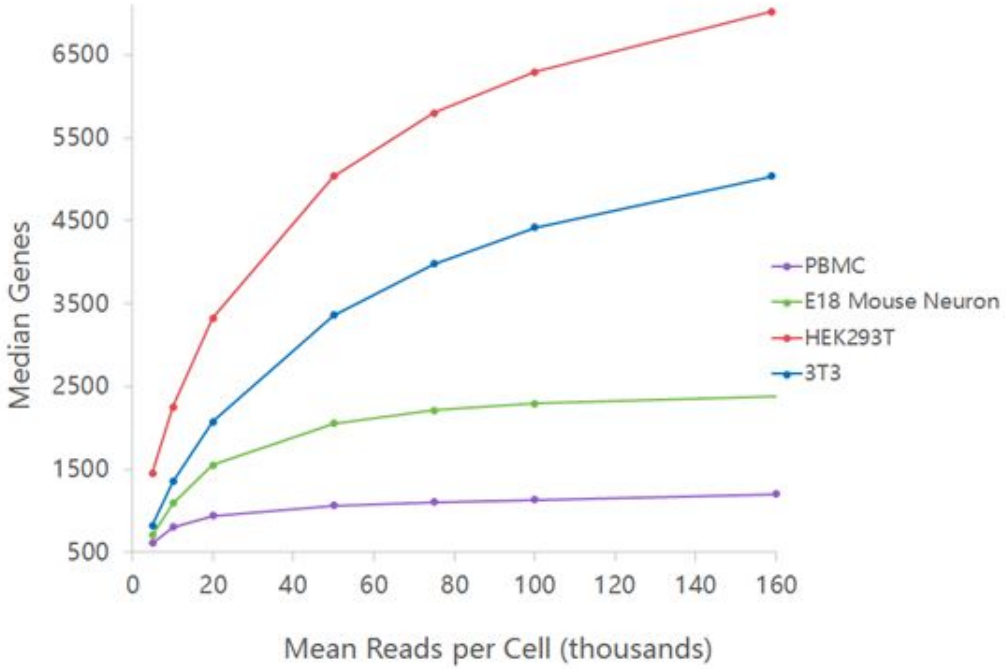


Sequencing requirements

How many reads/cell?

Saturation =

'a measure of the fraction of library complexity that was sequenced in a given experiment'



Sequencing requirements

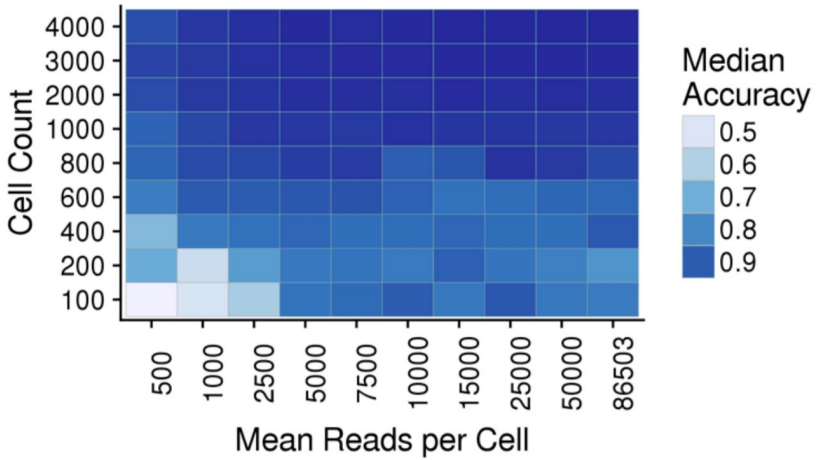
1. Overview of the cell types

mRNA content ↑ ⟶ required reads/cell ↑

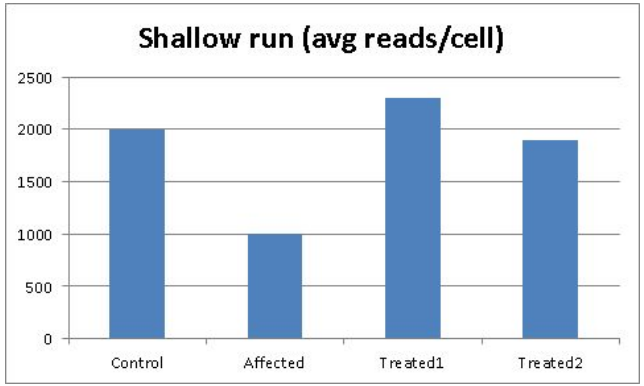
RARE cell types: Cell count ↓ ⟶ required reads/cell ↑

⟶ **Usually ~25k reads/cell**

2. Detect as many genes as possible



BioIT analysis: Shallow sequencing



Estimated Number of Cells

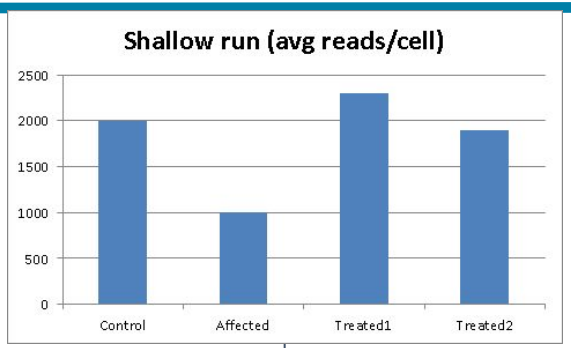
3,868

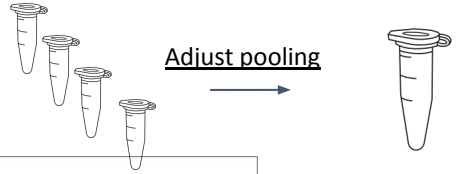
Mean Reads per Cell	Median Genes per Cell
4,939	722

Sequencing

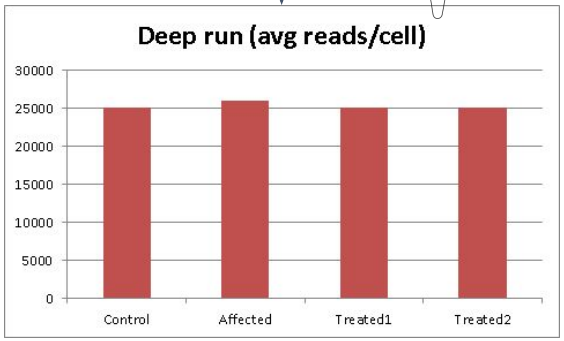
Number of Reads	19,107,635
Valid Barcodes	97.7%
Sequencing Saturation	4.7%
Q30 Bases in Barcode	97.2%
Q30 Bases in RNA Read	92.5%
Q30 Bases in Sample Index	96.3%
Q30 Bases in UMI	96.9%

BioIT analysis: Deep sequencing





Adjust pooling



Estimated Number of Cells

4,824

Mean Reads per Cell	Median Genes per Cell
57,119	3,432

Sequencing

Number of Reads	275,543,740
Valid Barcodes	97.5%
Sequencing Saturation	30.7%
Q30 Bases in Barcode	95.6%
Q30 Bases in RNA Read	95.1%
Q30 Bases in Sample Index	92.1%
Q30 Bases in UMI	95.4%

GENOMICS CORELEUVEN

KU LEUVEN

UZ LEUVEN

Web summary

10x GENOMICS™ Cell Ranger

SUMMARY ANALYSIS

Estimated Number of Cells

4,824

Mean Reads per Cell	Median Genes per Cell
57,119	3,432

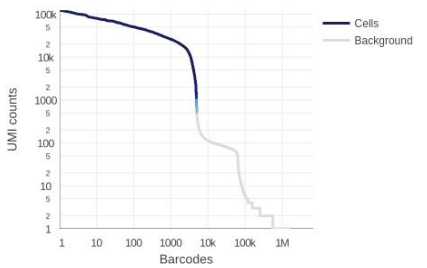
Sequencing

Number of Reads	275,543,740
Valid Barcodes	97.5%
Sequencing Saturation	30.7%
Q30 Bases in Barcode	95.6%
Q30 Bases in RNA Read	95.1%
Q30 Bases in Sample Index	92.1%
Q30 Bases in UMI	95.4%

Mapping

Reads Mapped to Genome	93.4%
Reads Mapped Confidently to Genome	87.5%
Reads Mapped Confidently to Intergenic Regions	7.7%
Reads Mapped Confidently to Intronic Regions	25.6%
Reads Mapped Confidently to Exonic Regions	54.2%
Reads Mapped Confidently to Transcriptome	50.4%
Reads Mapped Antisense to Gene	1.5%

Cells



Estimated Number of Cells	4,824
Fraction Reads in Cells	92.3%
Mean Reads per Cell	57,119
Median Genes per Cell	3,432
Total Genes Detected	22,525
Median UMI Counts per Cell	16,774

Sample

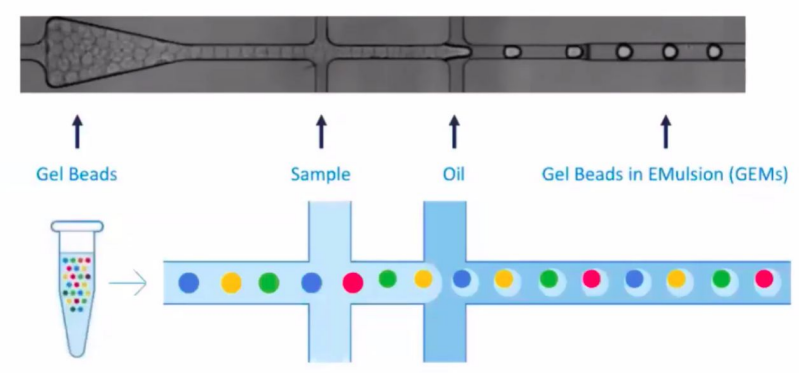
Name	GC1001745_E11
Description	
Transcriptome	GRCh38
Chemistry	Single Cell 3' v3
Cell Ranger Version	3.0.2

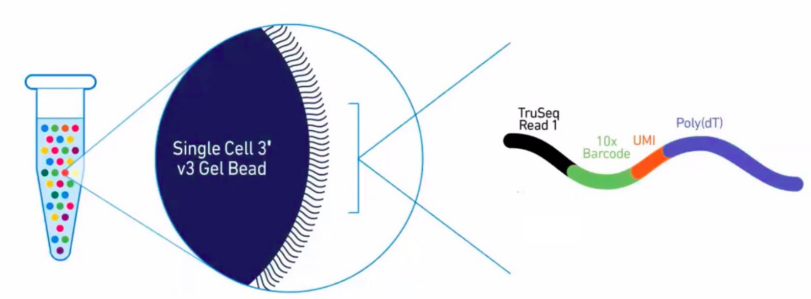
Analysing the data itself through Seurat

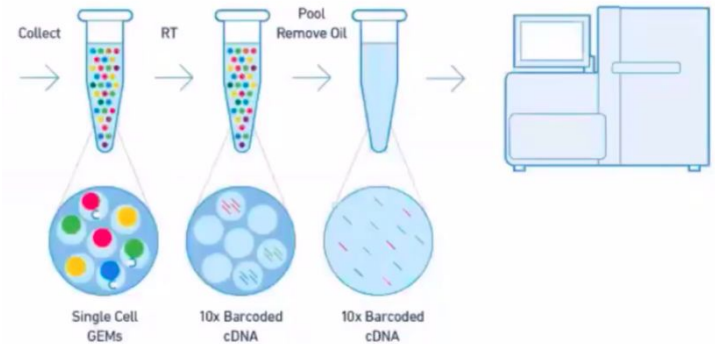
Analysing the data itself through Seurat

1. Part 1: Preparation of the data
2. Part 2: Analysis of the data
    a. Filtering
    b. Feature selection & Dimension reduction
    c. Clustering
    d. Identification
    e. Subsetting
    f. Integration
    g. Further analysis

Recap: Library Prep 10x





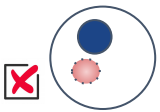


Empty GEM

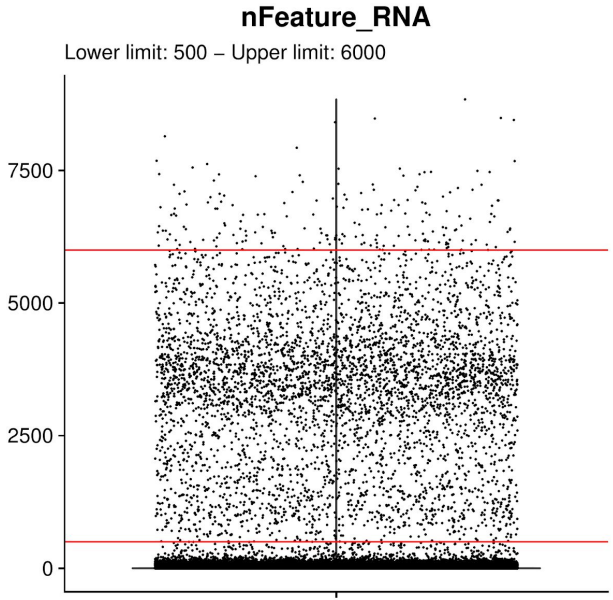
Multiplet

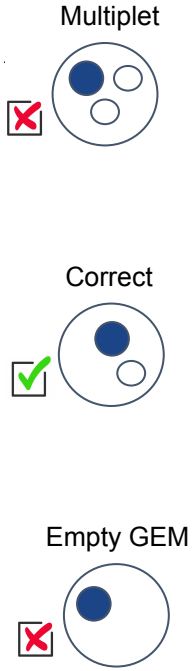
Correct

Dying/Stressed



BioIT analysis: Filtering

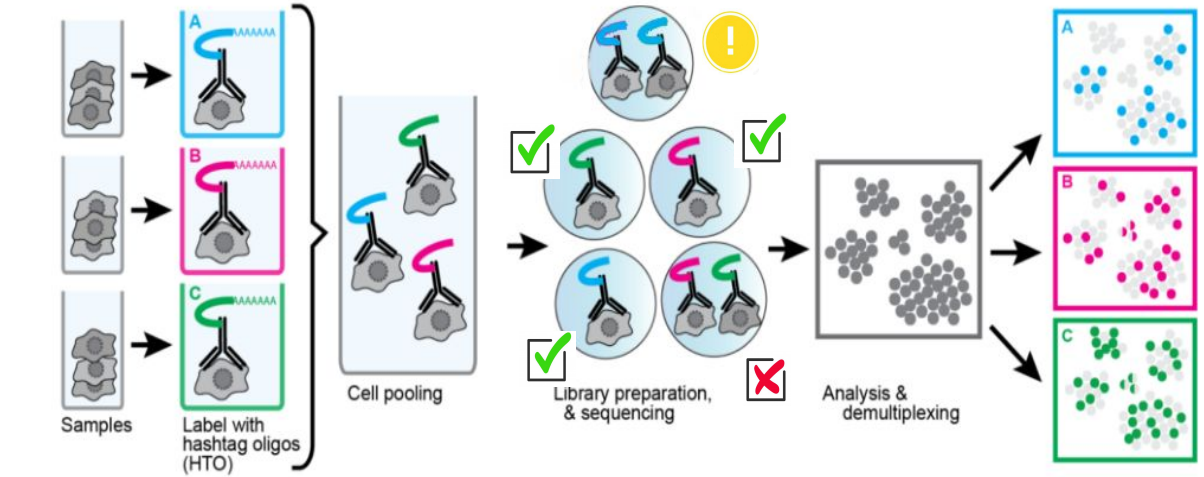




Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~870	~500
~0.8%	~1700	~1000
~1.6%	~3500	~2000
~2.3%	~5300	~3000
~3.1%	~7000	~4000
~3.9%	~8700	~5000
~4.6%	~10500	~6000
~5.4%	~12200	~7000
~6.1%	~14000	~8000
~6.9%	~15700	~9000
~7.6%	~17400	~10000

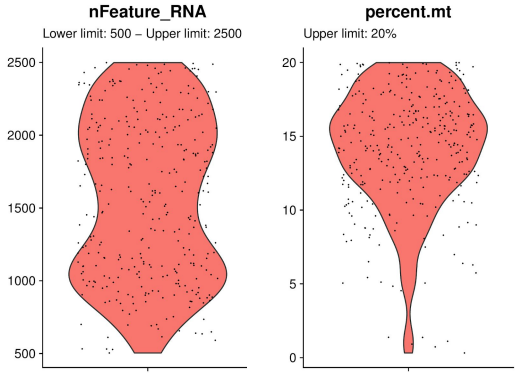
BioIT analysis: Multiplet detection

Cell hashing

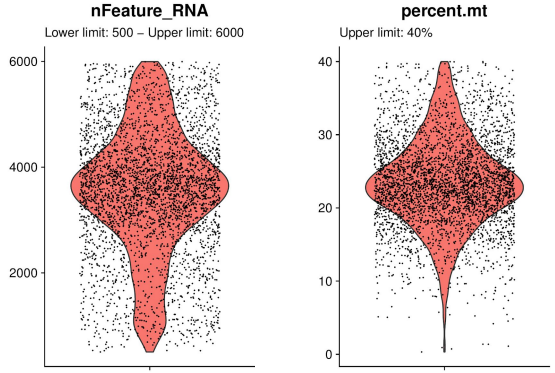


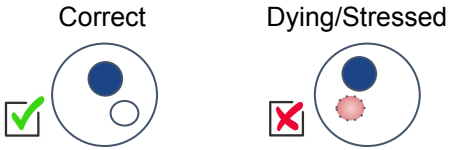
BioIT analysis: Filtering

First pass

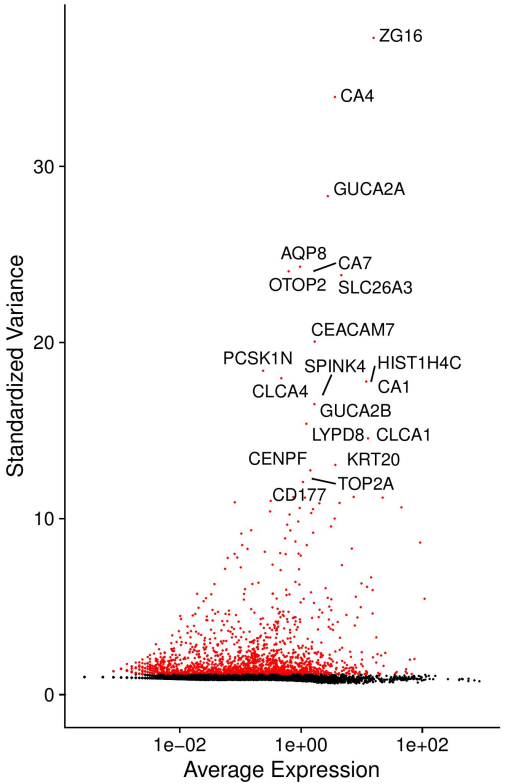


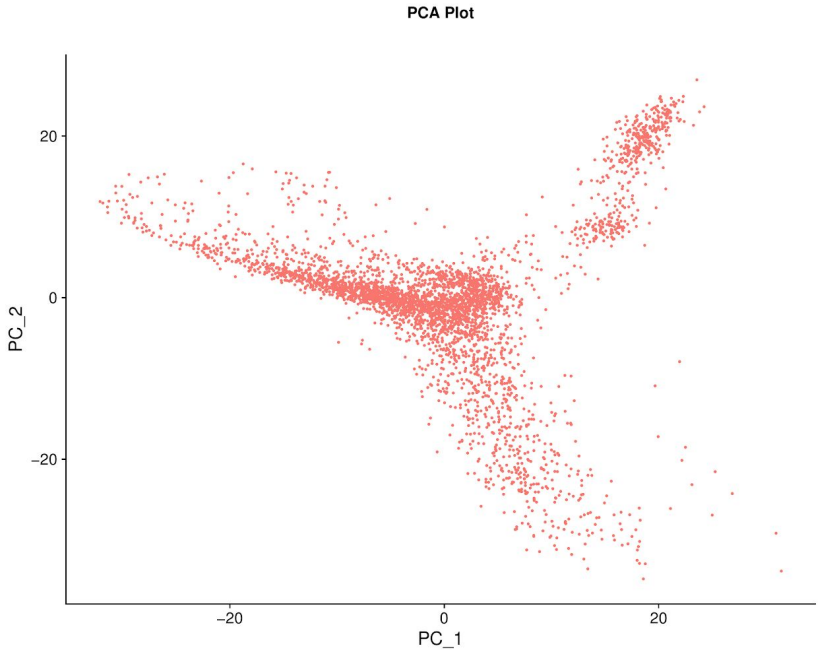
Second pass



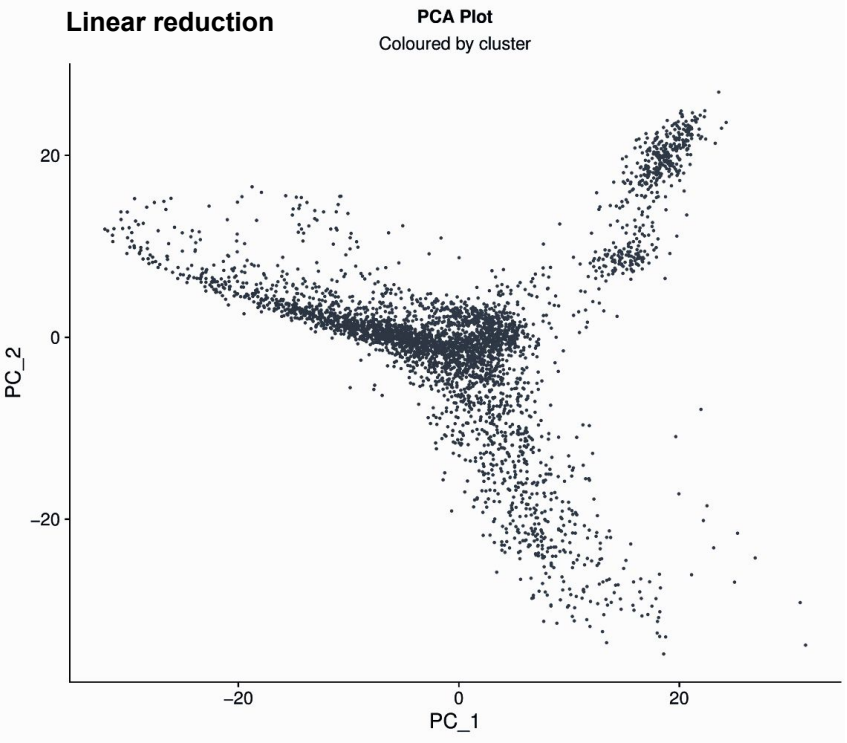


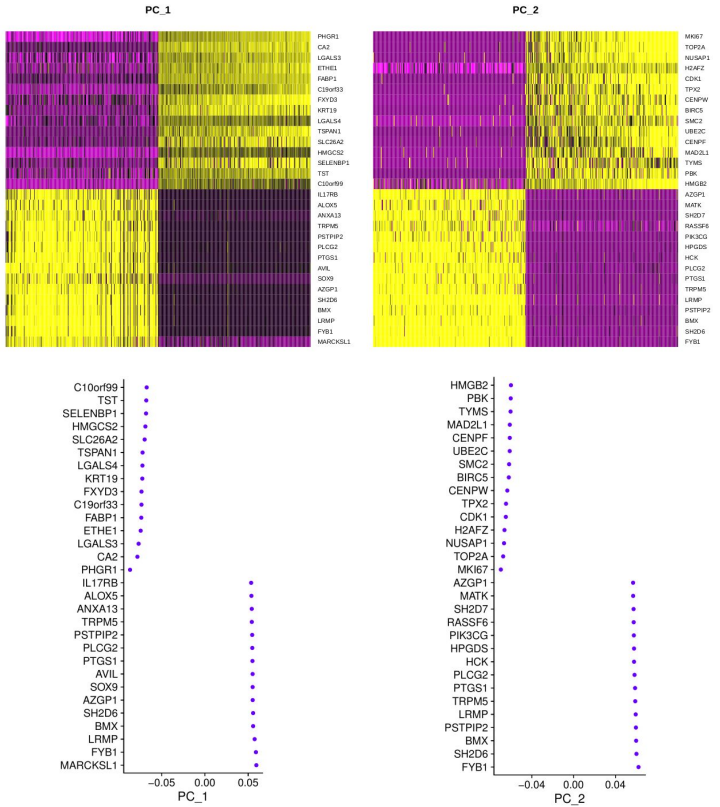
BioIT analysis: Feature selection & PCA



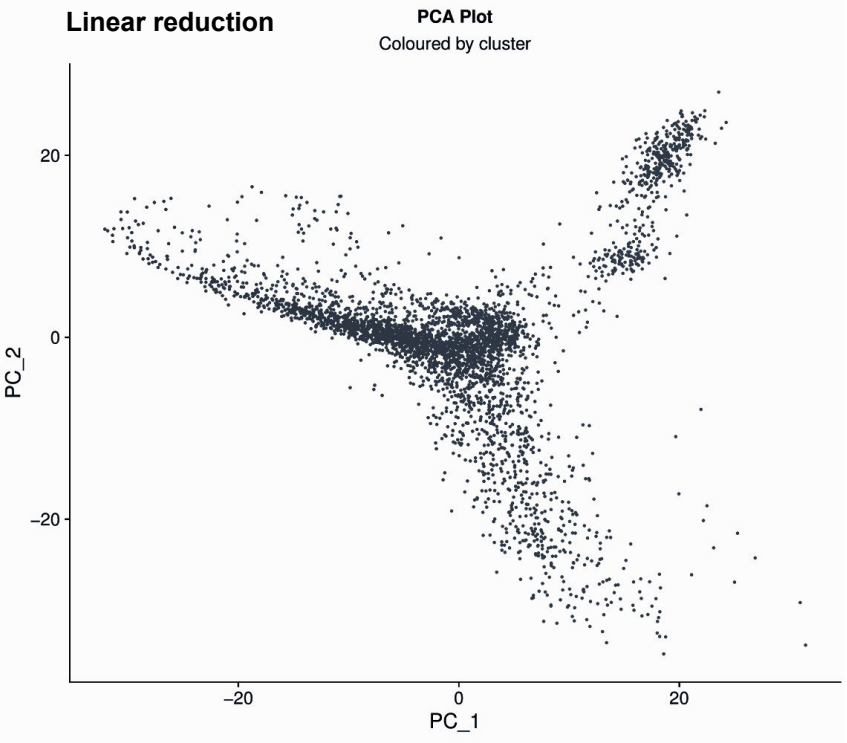


BioIT analysis: PCA



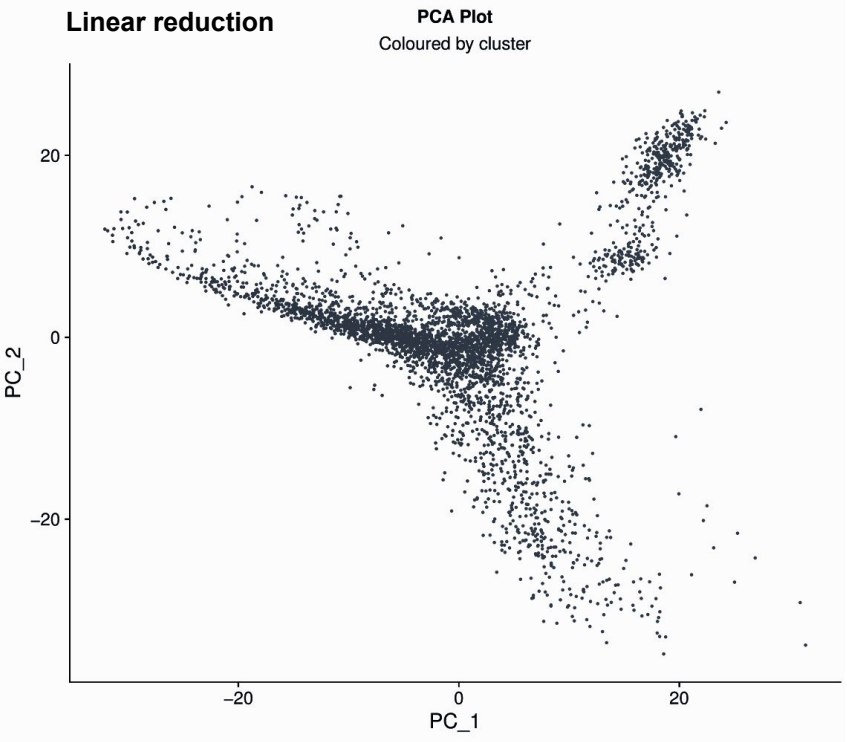


BioIT analysis: Clustering

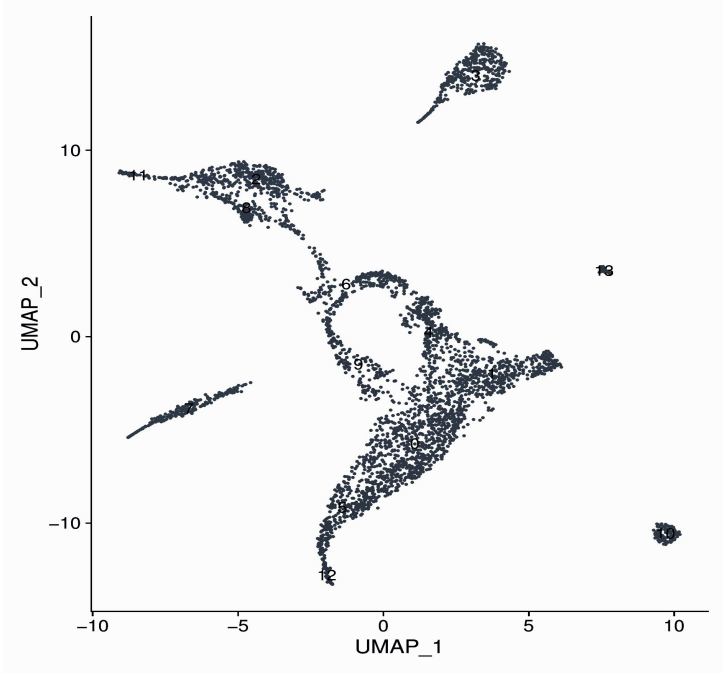


BioIT analysis: Clustering

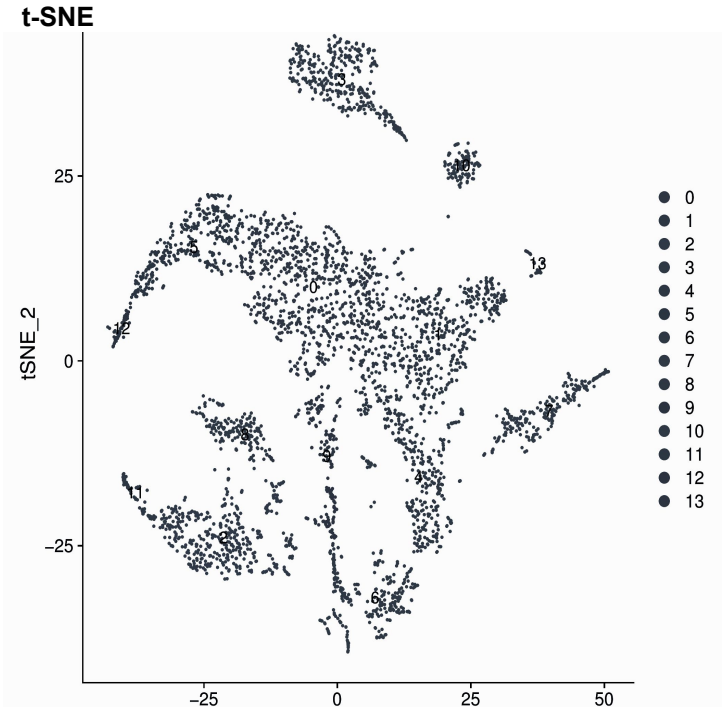
Linear reduction

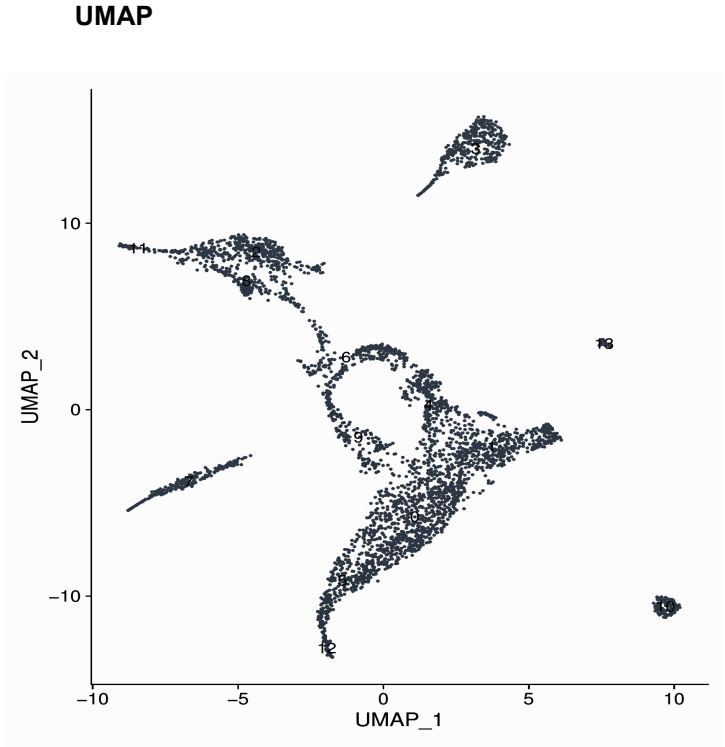


Nonlinear reduction



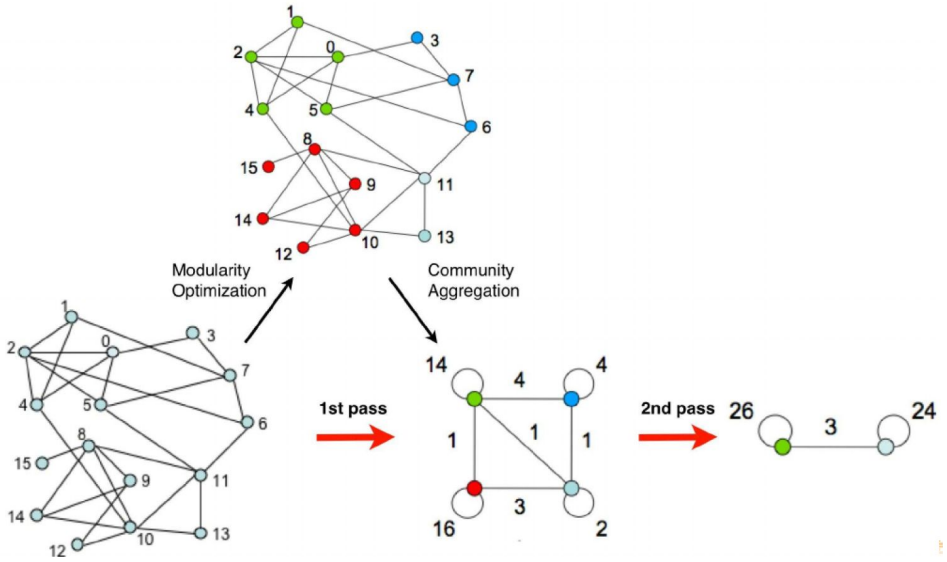
BioIT analysis: Clustering



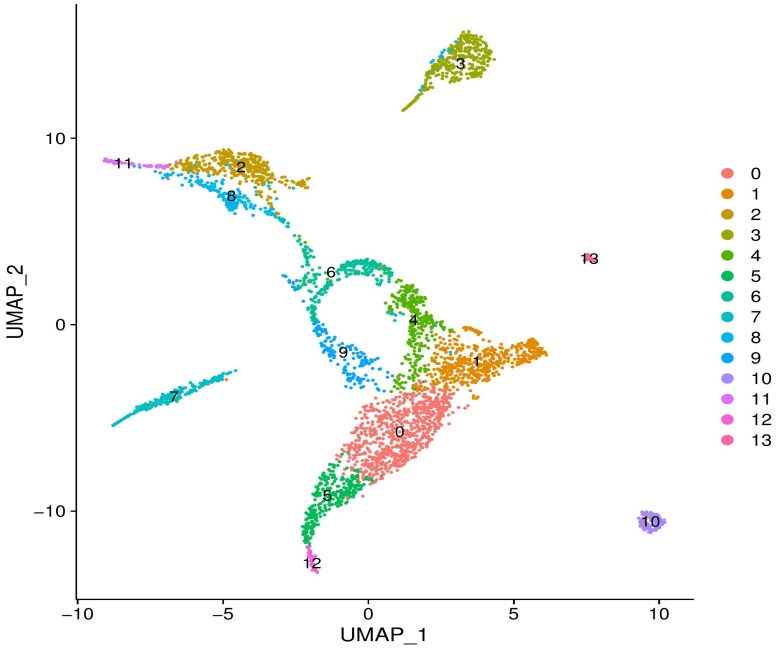


BioIT analysis: Clustering

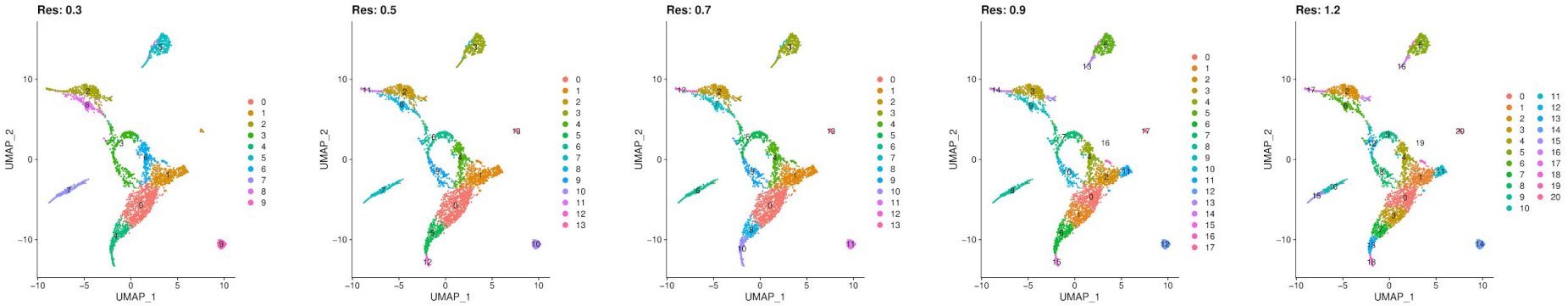
Graph-based



Nonlinear reduction

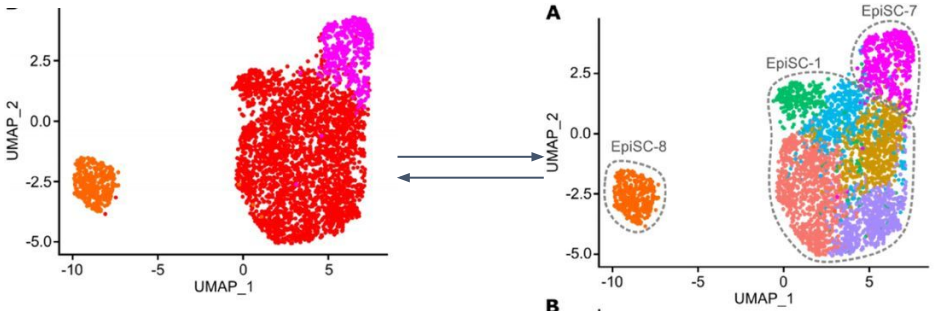


BioIT analysis: Clustering



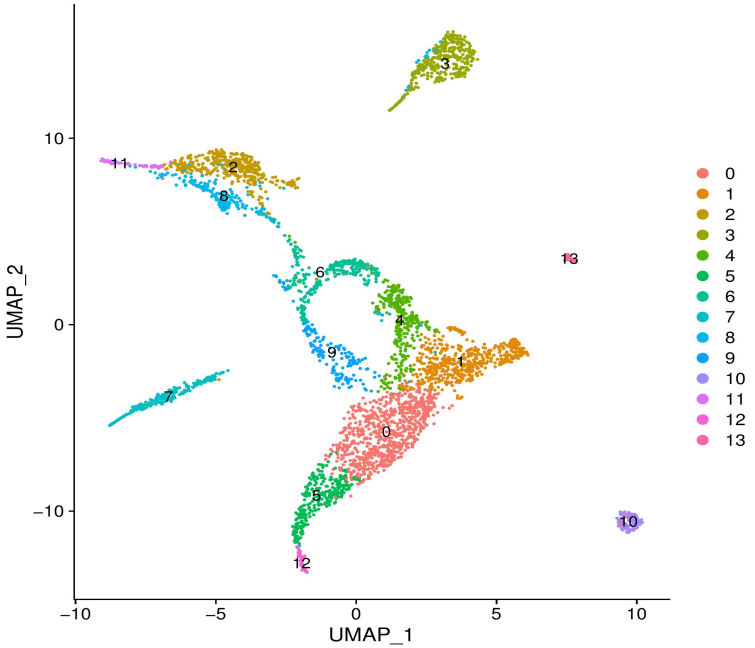
Low resolution

High resolution

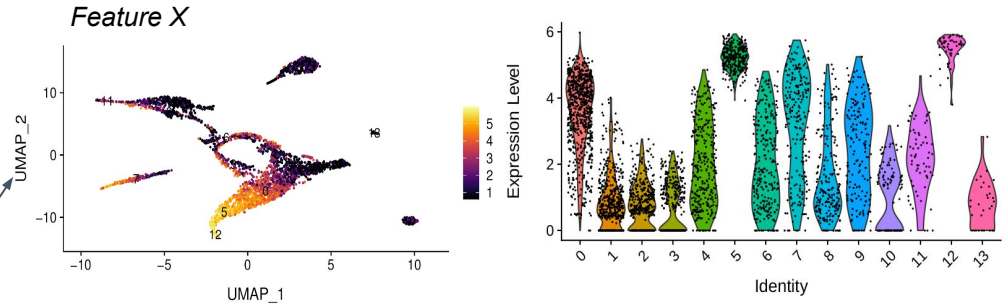


Adapted from: Single-cell transcriptomics defines heterogeneity of epicardial cells and fibroblasts within the infarcted heart, Hesse et al.

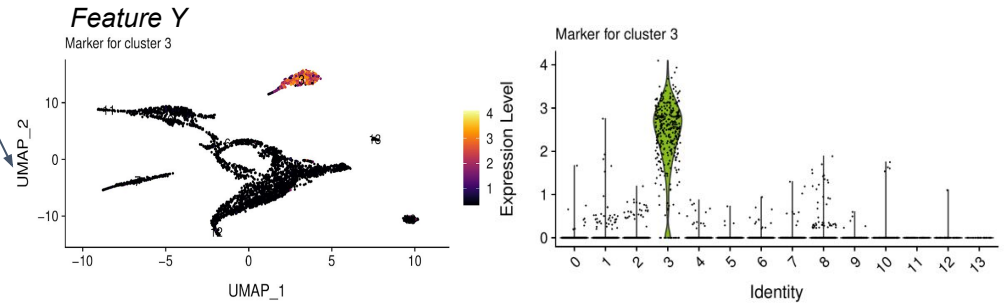
BioIT analysis: Identification



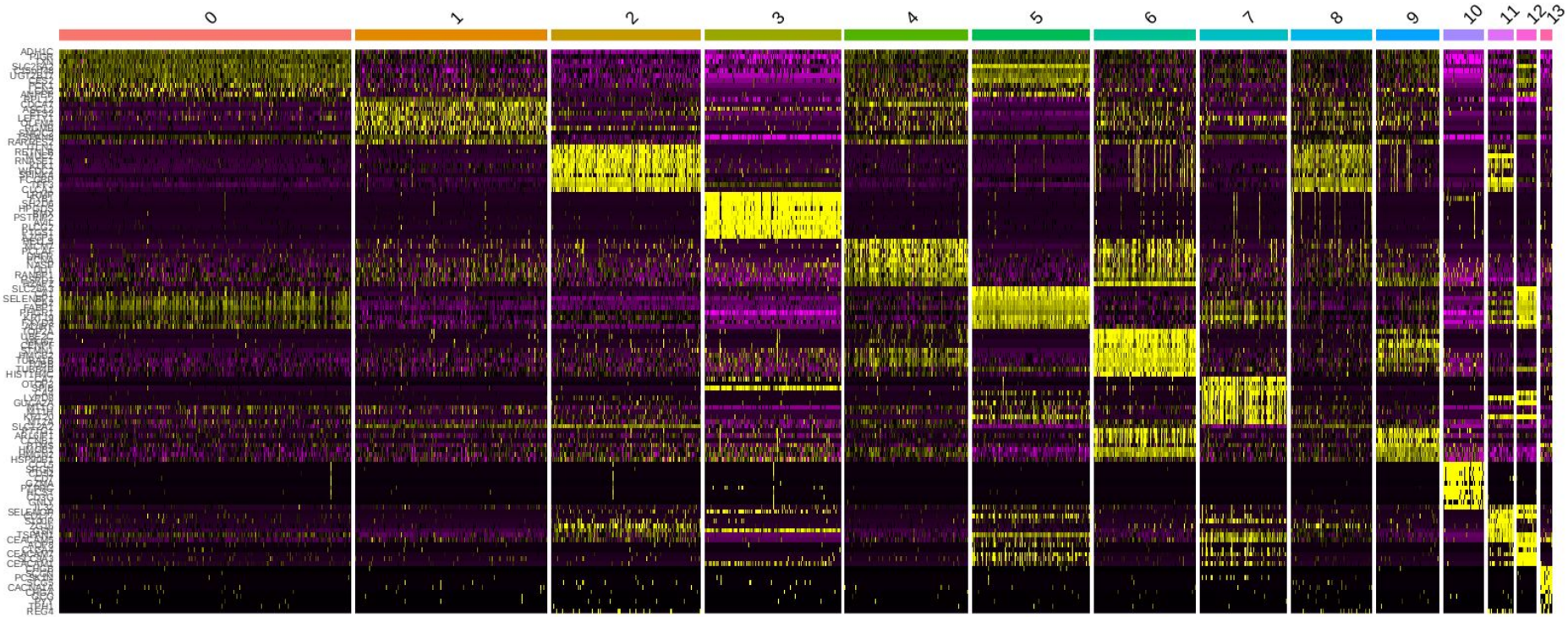
Visualize genes typically expressed in cell type



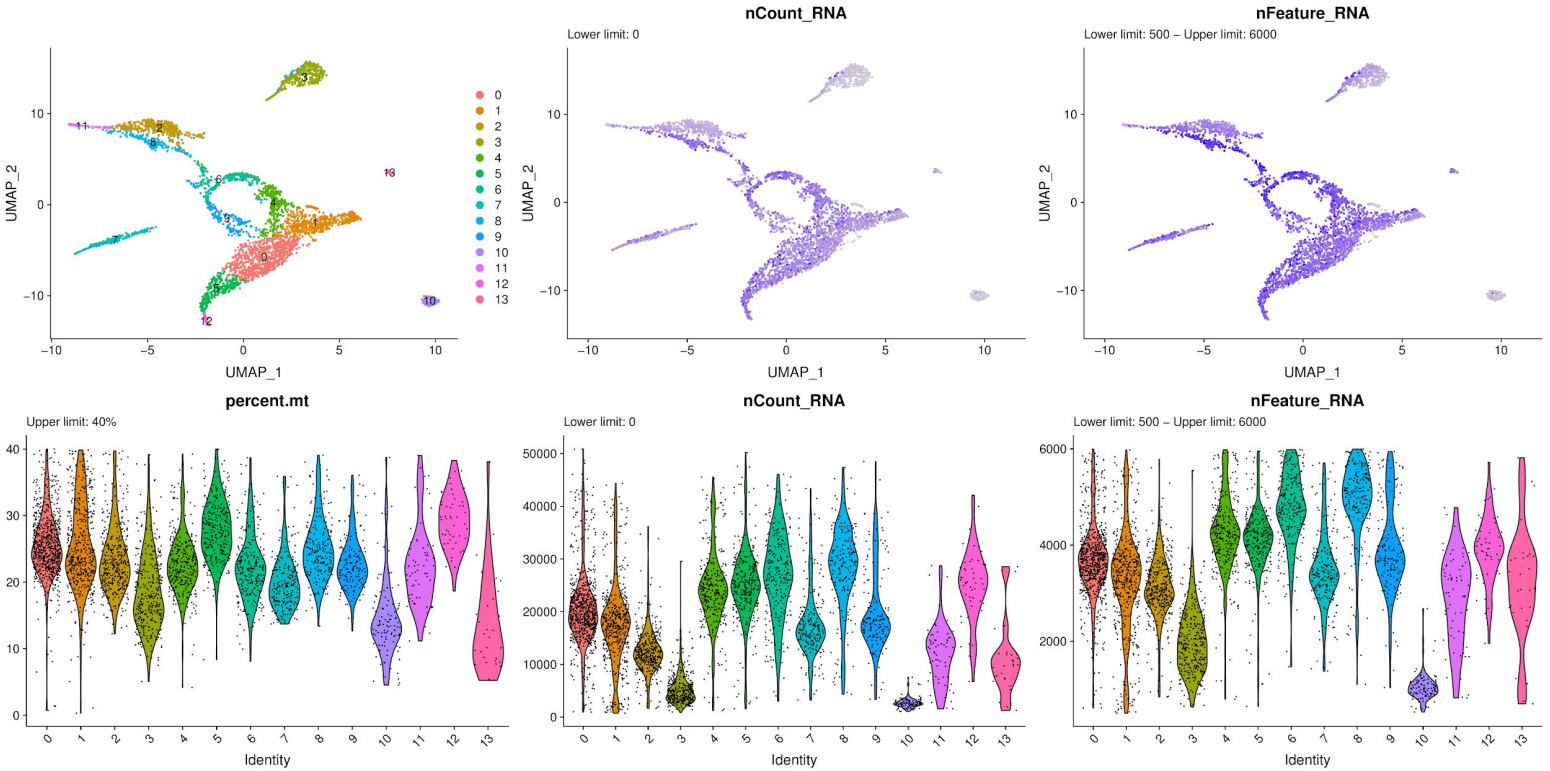
Find markers expressed for each cluster



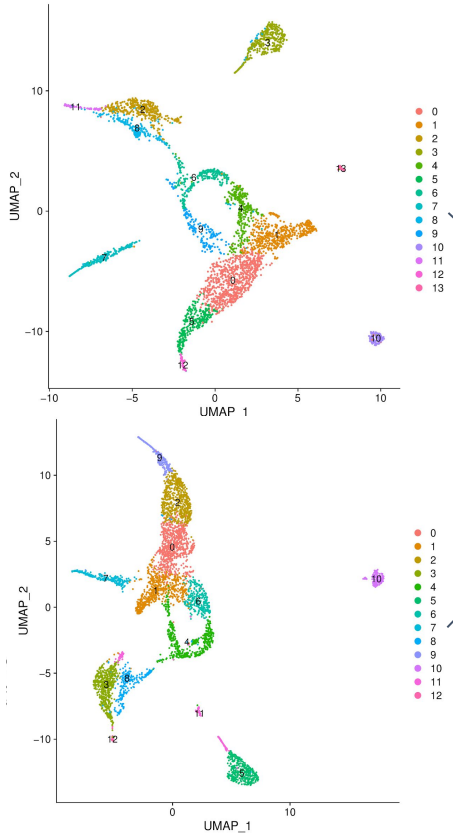
BioIT analysis: Identification



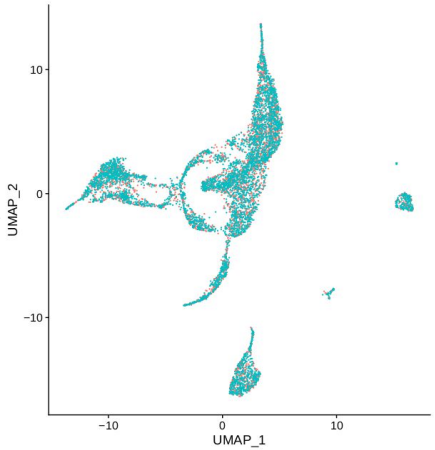
BioIT analysis: Subsetting

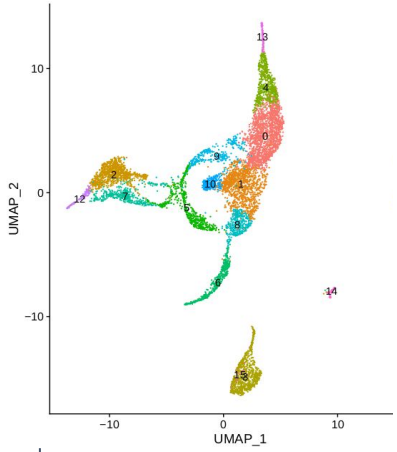


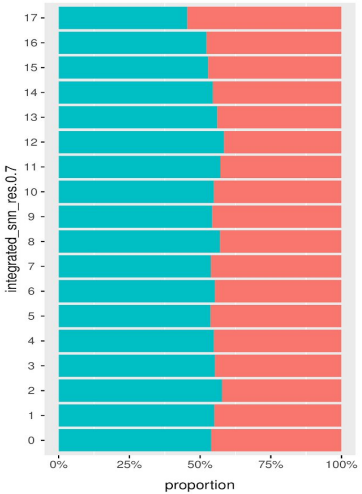
BioIT analysis: Integration



Integration

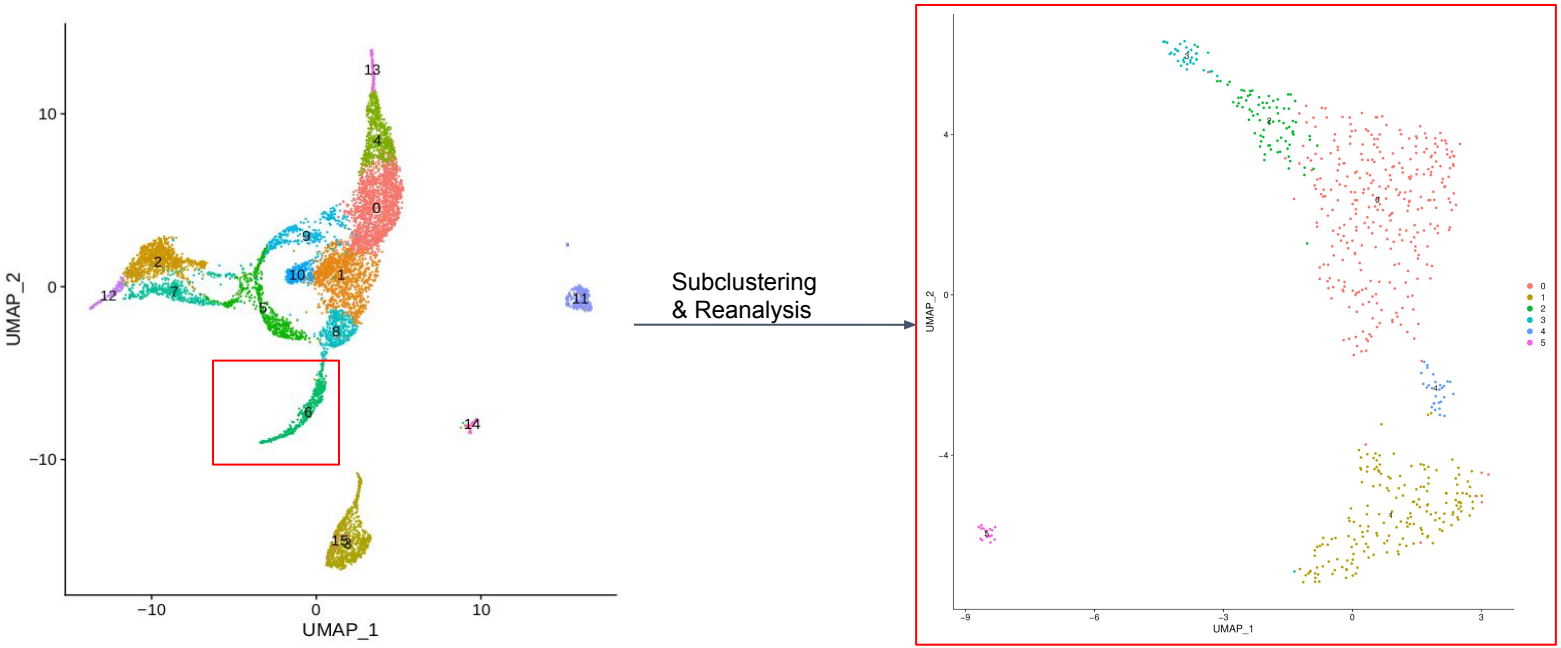




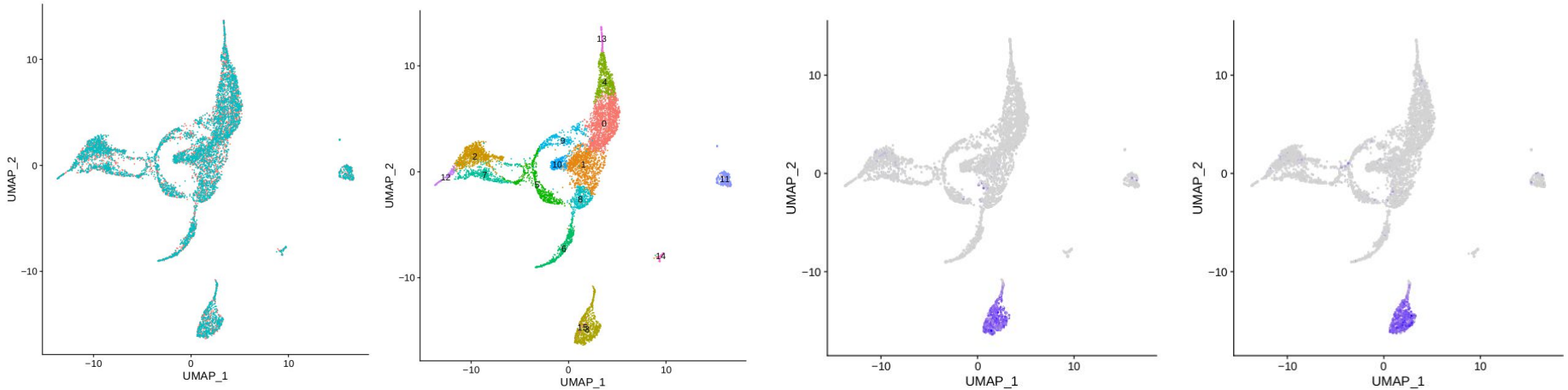


- Confirm cell types
- Expression patterns
- Differential genes
- Integration with other layers of data

BioIT analysis: Subsetting

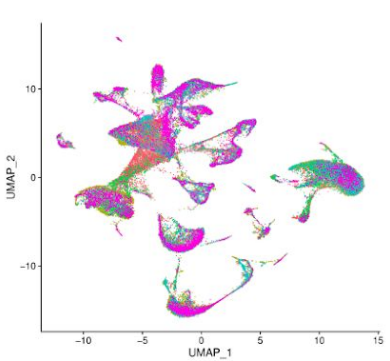


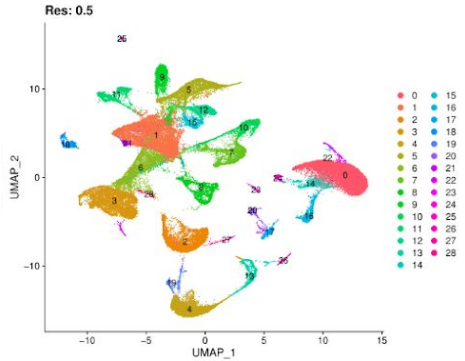
BioIT analysis: Integration

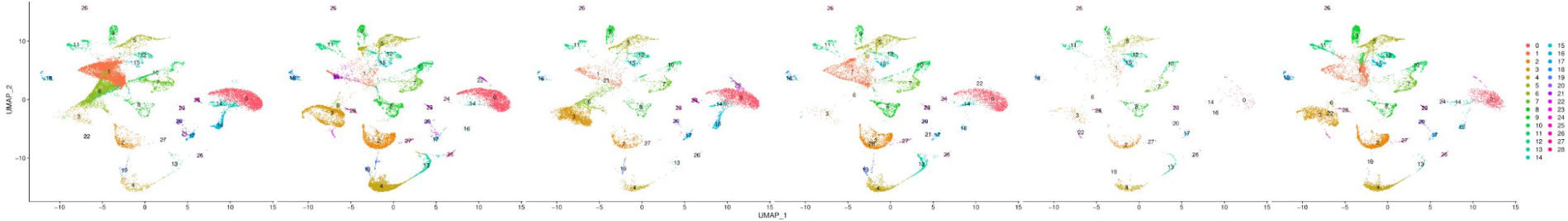


BioIT analysis: Scalability

~78k cells

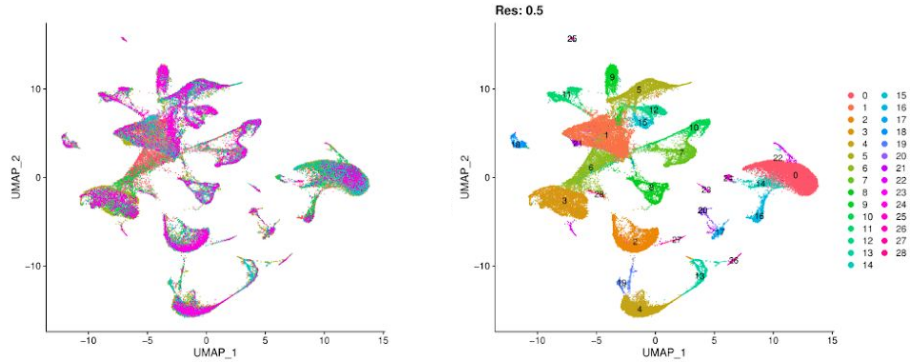


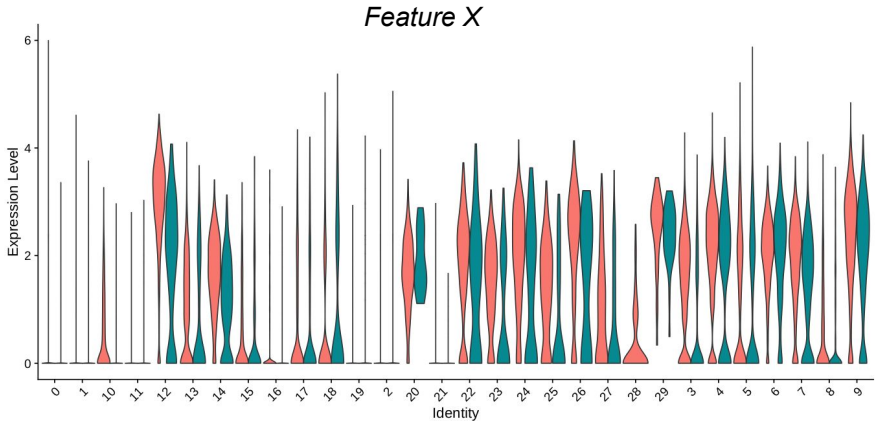




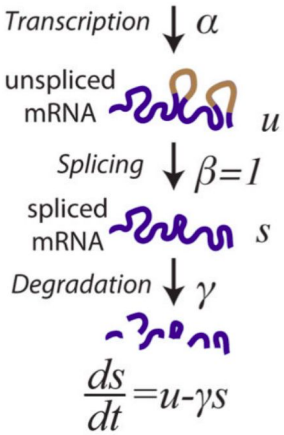
BioIT analysis: Scalability

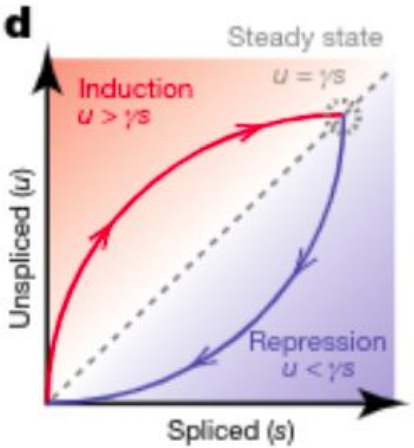
~78k cells



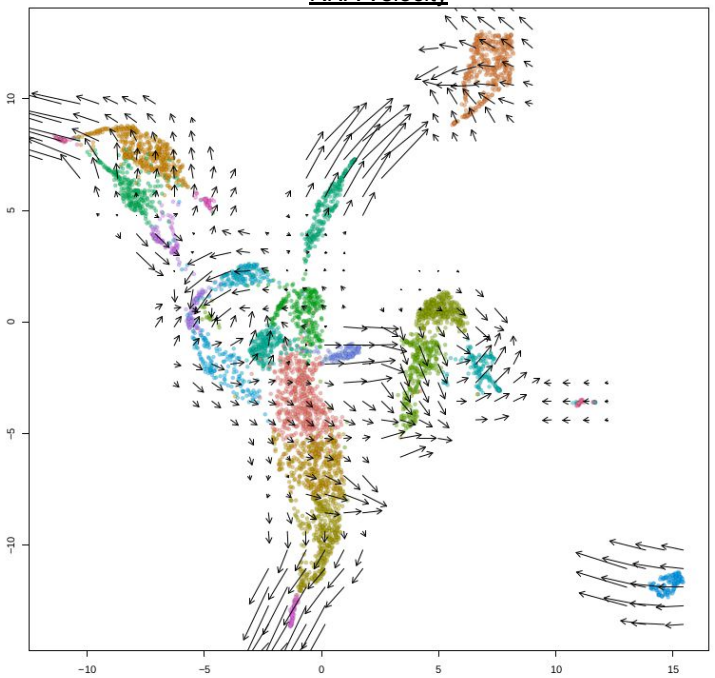


BioIT analysis: Further analysis





RNA Velocity

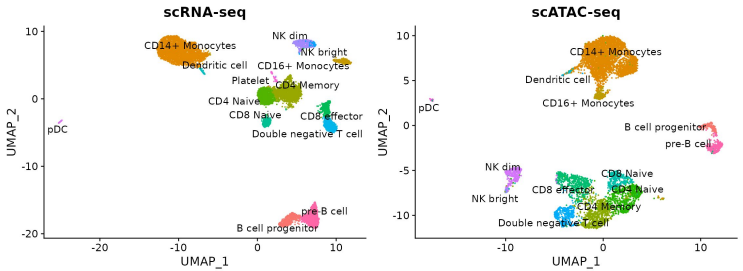


BioIT analysis: Further analysis

Multimodal analyses

RNA-Seq Antibodies ATAC-Seq Protein Spatial

Integration of multiple data types & datasets

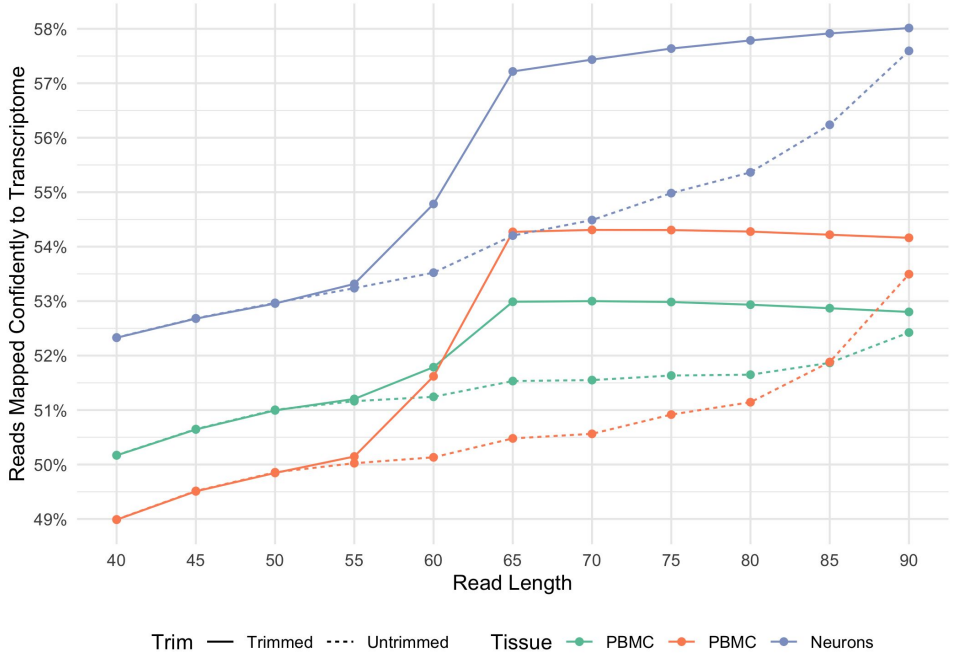


Thank you for your attention!

Questions?

info@genomicscore.be

Additional figures:



Cell calling algorithm

1) First pass: (High RNA content cells)
   a) Cutoff based on UMI count
      i) Take 99th percentile of UMI counts as m
      ii) All GEMs with nUMI > (m/10) → mark as cells
2) Second pass (Lower RNA content cells)
   a) Take set of GEMs which appear empty (low nUMI)
   b) Create a model of their RNA profile
   c) RNA profile of all GEM's uncalled in the first pass is compared to this RNA profile
   d) If strong disagreement with the profile → mark as cells