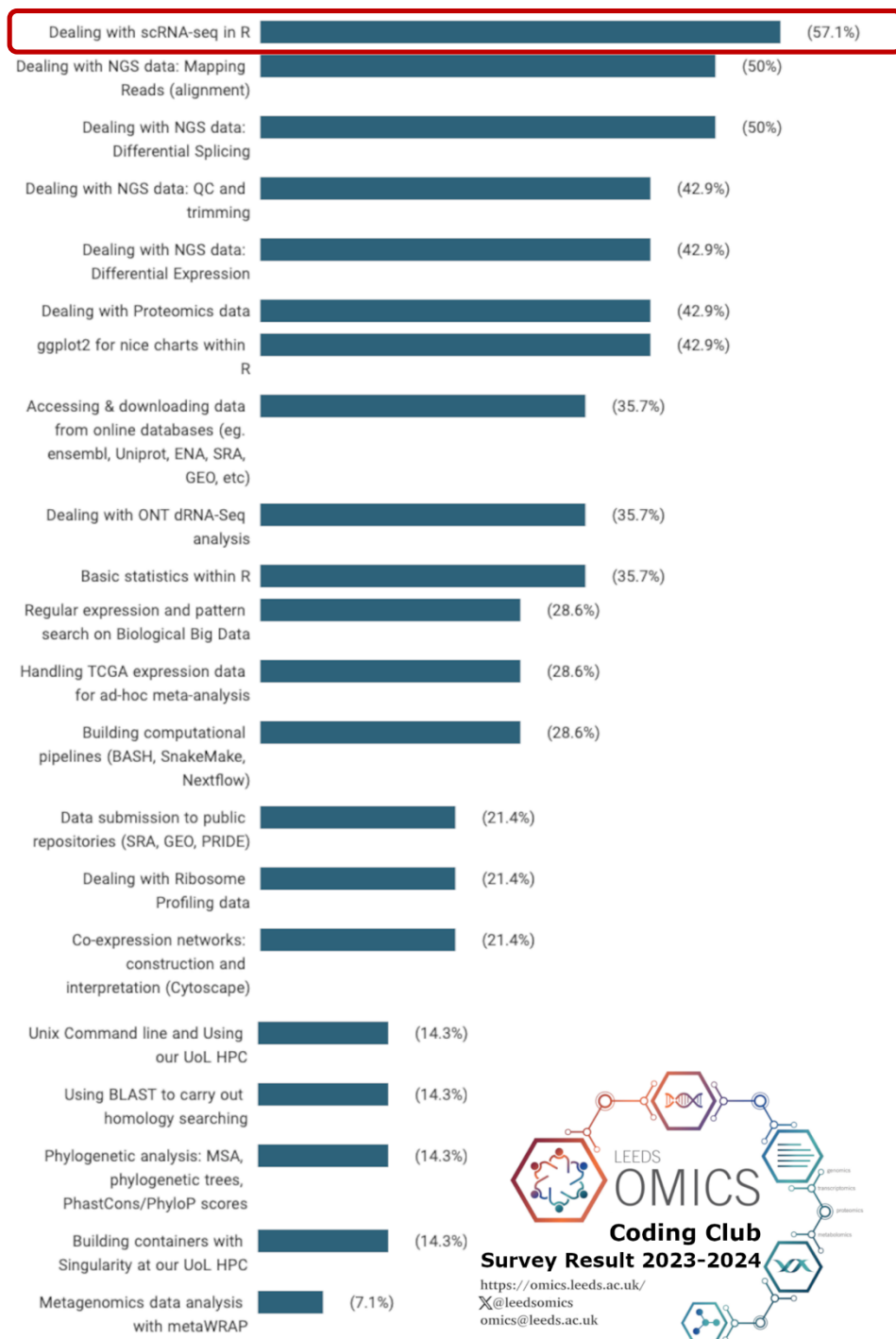


<https://omics.leeds.ac.uk/>
X@leedsomics
omics@leeds.ac.uk

Dealing with single-cell RNA-seq data in R

Club Moderator: Elton Vasconcelos

Topics to be addressed on the 2023-24 season - Survey Result



LEEDS OMICS Coding Club
Survey Result 2023-2024
<https://omics.leeds.ac.uk/>
[X@leedsomics](https://twitter.com/leedsomics)
omics@leeds.ac.uk


Seurat - R toolkit for single-cell genomics

<https://satijalab.org/seurat/>



Detail from *Circus Sideshow (Parade de Cirque)* (1889) showing pointillism and color theory

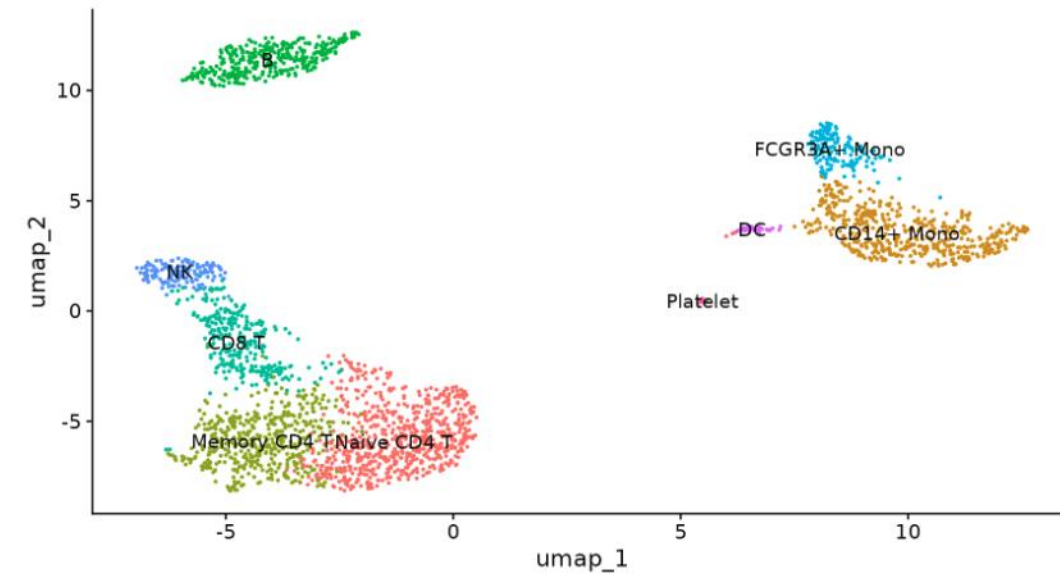
Georges Seurat



Georges SEURAT

Seurat in 1888

Born	Georges-Pierre Seurat 2 December 1859 Paris, France
Died	29 March 1891 (aged 31) Paris, France
Known for	Painting

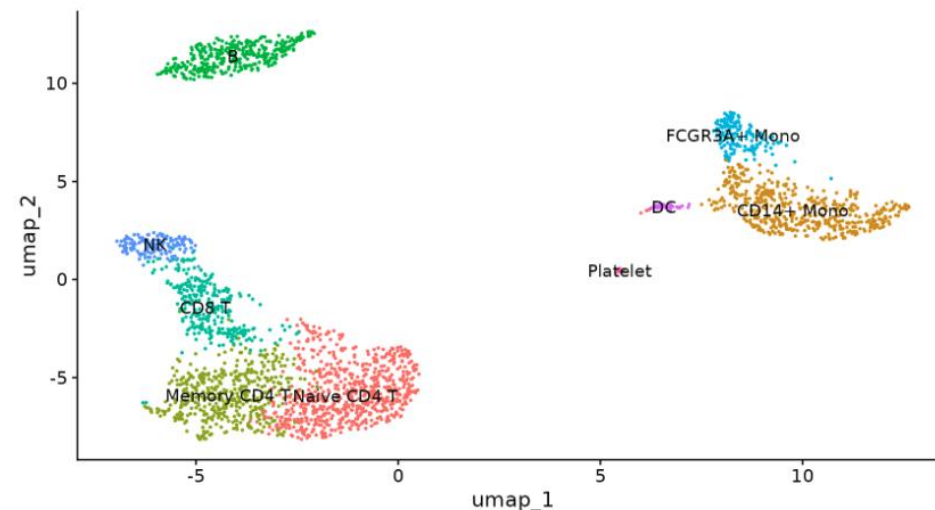


Seurat - R toolkit for single-cell genomics

https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Seurat provides a comprehensive scRNA-seq data analysis pipeline that passes through:

- (i) Quality control
- (ii) Normalization
- (iii) Selection of highly variable genes (HVGs)
- (iv) Scaling
- (v) Unsupervised linear dimensional reduction through principal component analysis (PCA)
- (vi) Clustering
- (vii) Non-linear dimensional reduction through either uniform manifold approximation and projection (UMAP) or t-distributed stochastic neighbour embedding (tSNE) methods
- (viii) Identification of differentially expressed genes (DEGs) among clusters (i.e. cluster biomarkers)
- (ix) Assignment of cell type identity to clusters



Other useful resources/material

Wellcome Trust Sanger Institute scRNA-seq course: <https://www.singlecellcourse.org/>

Single-cell Analysis in Python: <https://scanpy.readthedocs.io/en/stable/>

Single-cell Type Database: <https://sctype.app/>



3.3.1 General Considerations

Single cell RNA-seq data differ from bulk RNA seq in a number of ways (see Introduction to single cell RNA-Seq chapter above). Most modern scRNA-seq technologies generate read sequences containing three key pieces of information:

- cDNA fragment that identifies the RNA transcript;
- Cell barcode (CB) that identifies the cell where the RNA was expressed;
- Unique Molecular Identifier (UMI) that allows to collapse reads that are PCR duplicates.

In contrast to bulk RNA-seq, scRNA-seq deals with a much smaller amount of RNA, and more PCR cycles are performed. Thus, UMI barcodes become very useful and are now widely accepted in scRNAseq. Library sequencing is often done with paired-end reads, with one read containing CB + UMI (read 1 in 10x Chromium), and the other containing actual transcript sequence (read 2 in 10x Chromium).

A classical scRNA-seq workflow contains four main steps:

- Mapping the cDNA fragments to a reference;
- Assigning reads to genes;
- Assigning reads to cells (cell barcode demultiplexing);
- Counting the number of unique RNA molecules (UMI deduplication).

The outcome of this procedure is a gene/cell count matrix, which is used as an estimate of the number of RNA molecules in each cell for each gene.



3.3.4 Chromium Versions and Cell Barcode Whitelists

Cellular barcode sequences are synthetic sequences attached to the beads that identify individual cells. The library of unique sequences is called a whitelist and depends on the Chromium library preparation kit version. The whitelist files are available from the [Cell Ranger repository](#). There are three whitelists used for Chromium: `737K-april-2014_rc.txt`, `737K-august-2016.txt`, and `3M-february-2018.txt`. CBs from the first list are 14 bp long, and two others are 16 bp. The table below provides cellular barcodes and UMI lengths, as well as appropriate whitelist files, for popular 10x single cell sequencing kits:

Chemistry	CB, bp	UMI, bp	Whitelist file
10x Chromium Single Cell 3' v1	14	10	737K-april-2014_rc.txt
10x Chromium Single Cell 3' v2	16	10	737K-august-2016.txt
10x Chromium Single Cell 3' v3	16	12	3M-february-2018.txt
10x Chromium Single Cell 3' v3.1 (Next GEM)	16	12	3M-february-2018.txt
10x Chromium Single Cell 5' v1.1	16	10	737K-august-2016.txt
10x Chromium Single Cell 5' v2 (Next GEM)	16	10	737K-august-2016.txt
10x Chromium Single Cell Multiome	16	12	737K-arc-v1.txt

```

==> 3M-february-2018.txt <==
AAACCCAAGAAACACT
AAACCCAAGAAACCAT
AAACCCAAGAAACCCA
AAACCCAAGAAACCCG
AAACCCAAGAAACCTG
AAACCCAAGAAACGAA
AAACCCAAGAAACGTC
AAACCCAAGAAACTAC
AAACCCAAGAAACTCA
AAACCCAAGAAACTGC

==> 737K-april-2014_rc.txt <==
AACATACAAAACG
AACATACAAAAGC
AACATACAAACAG
AACATACAAACGA
AACATACAAAGCA
AACATACAAAGTG
AACATACAACAGA
AACATACAACCAC
AACATACAACCGT
AACATACAACCTG

==> 737K-august-2016.txt <==
AAACCTGAGAAACCAT
AAACCTGAGAAACCGC
AAACCTGAGAAACCTA
AAACCTGAGAAACGAG
AAACCTGAGAAACGCC
AAACCTGAGAAAGTGG
AAACCTGAGAACAAC
AAACCTGAGAACAATC
AAACCTGAGAACTCGG
AAACCTGAGAACTGTA

```

Aligners for scRNA-seq

- Cell ranger

<https://github.com/10XGenomics/cellranger>

- STARsolo

<https://github.com/alexdobin/STAR>

At the bottom of "STAR --help",
we'll find STARsolo parameters

```
### STARsolo (single cell RNA-seq) parameters
soloType          None
  string(s): type of single-cell RNA-seq
                  CB_UMI_Simple  ... (a.k.a. Droplet) one UMI and one Cell Barcode of fixed length in read2, e.g. Drop-seq and 10X Chromium.
                  CB_UMI_Complex ... multiple Cell Barcodes of varying length, one UMI of fixed length and one adapter sequence of fixed length are allowed in read2 only (e.g. inDrop, ddSeq).
                  CB_samTagOut   ... output Cell Barcode as CR and/or CB SAM tag. No UMI counting. --readFilesIn cDNA_read1 [cDNA_read2 if paired-end] CellBarcode_read . Requires --outSAMtype BAM Unsorted [and/or SortedByCoordinate]
                  SmartSeq      ... Smart-seq: each cell in a separate FASTQ (paired- or single-end), barcodes are corresponding read-groups, no UMI sequences, alignments deduplicated according to alignment start and end (after extending soft-clipped bases)
soloCBwhitelist   -
  string(s): file(s) with whitelist(s) of cell barcodes. Only --soloType CB_UMI_Complex allows more than one whitelist file.
                  None          ... no whitelist: all cell barcodes are allowed
soloCBstart       1
  int>0: cell barcode start base
soloCBlen         16
  int>0: cell barcode length
soloUMIstart      17
  int>0: UMI start base
soloUMIlen        10
  int>0: UMI length
soloBarcodeReadLength 1
  int: length of the barcode read
                  1 ... equal to sum of soloCBlen+soloUMIlen
                  0 ... not defined, do not check
```


- **STARsolo** HPC shell script for **10X Chromium libraries**: needs debarcoding, each barcode corresponds to a cell

```
#!/bin/bash
#$ -cwd -V
#$ -l h_rt=16:00:00,h_vmem=6G,nodes=1,ppn=20
##$ -l node_type=40core-768G

for i in *.fq; do
    base=`basename $i .fq`;
    STAR --runMode alignReads --genomeDir ../HsapSAindexedGenome/ --readFilesIn $i --quantMode GeneCounts --sjdbGTFfile ../gencode.v36.annotation.gtf --outFileNamePrefix ${base}-vsGRCh38.d1.vd1_ --soloType CB_UMI_Simple --soloBarcodeMate 1 --clip5pNbases 26 --soloStrand Reverse --soloBarcodeReadLength 0 --soloCBstart 1 --soloCBlen 16 --soloUMIstart 17 --soloUMIlen 10 --soloCBwhitelist barcodes.tsv --soloCBmatchWLtype 1MM_multi_Nbase_pseudocounts --soloUMIfiltering MultiGeneUMI_CR --soloUMIidedup 1MM_CR --outFilterScoreMin 30 --outSAMtype BAM SortedByCoordinate --outSAMattributes CR UR CY UY CB UB --limitBAMsortRAM 12000000000 --runThreadN 20
done
```

ATTENTION: You must rename your Chromium WhiteList barcode file to “**barcodes.tsv**” and place it in the directory where you’re running STAR.

- **STARsolo** HPC shell script for **SmartSeq2 libraries**: each fastq (or pairs of fastq) corresponds to a cell

```
#!/bin/bash
#$ -cwd -V
#$ -l h_rt=16:00:00,h_vmem=5G,nodes=1,ppn=24
##$ -l node_type=24core-768G
/nobackup/fbsev/bioinformatics-tools/STAR-2.7.10a_alpha_220818-intel/source/STAR --runMode alignReads --genomeDir ../HsapSAindexedGenome/ --readFilesManifest readManif.tsv --readFilesCommand zcat --soloType SmartSeq --soloUMIidedup Exact --outSAMtype BAM SortedByCoordinate --outSAMattributes RG --limitBAMsortRAM 12000000000 --runThreadN 24
```

Example of the reads manifest file (readManif.txt) provided on the cmd above

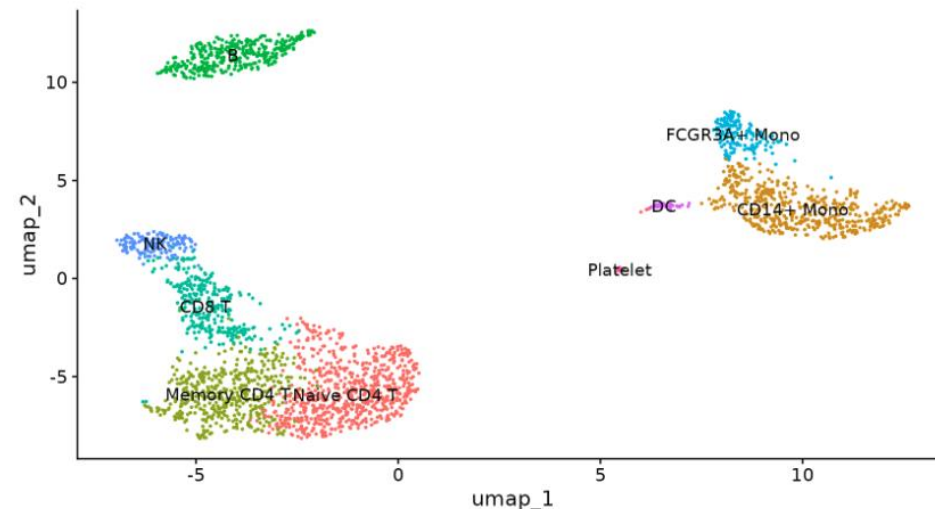
```
../../../../01-QC/GSE75688/SRR2973272_1.fastq.gz    ../../01-QC/GSE75688/SRR2973272_2.fastq.gz    SRR2973272
../../../../01-QC/GSE75688/SRR2973273_1.fastq.gz    ../../01-QC/GSE75688/SRR2973273_2.fastq.gz    SRR2973273
../../../../01-QC/GSE75688/SRR2973274_1.fastq.gz    ../../01-QC/GSE75688/SRR2973274_2.fastq.gz    SRR2973274
../../../../01-QC/GSE75688/SRR2973275_1.fastq.gz    ../../01-QC/GSE75688/SRR2973275_2.fastq.gz    SRR2973275
../../../../01-QC/GSE75688/SRR2973276_1.fastq.gz    ../../01-QC/GSE75688/SRR2973276_2.fastq.gz    SRR2973276
../../../../01-QC/GSE75688/SRR2973277_1.fastq.gz    ../../01-QC/GSE75688/SRR2973277_2.fastq.gz    SRR2973277
```

Seurat - R toolkit for single-cell genomics

Let's briefly go together through the tutorial → https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Seurat provides a comprehensive scRNA-seq data analysis pipeline that passes through:

- (i) Quality control
- (ii) Normalization
- (iii) Selection of highly variable genes (HVGs)
- (iv) Scaling
- (v) Unsupervised linear dimensional reduction through principal component analysis (PCA)
- (vi) Clustering
- (vii) Non-linear dimensional reduction through either uniform manifold approximation and projection (UMAP) or t-distributed stochastic neighbour embedding (tSNE) methods
- (viii) Identification of differentially expressed genes (DEGs) among clusters (i.e. cluster biomarkers)
- (ix) Assignment of cell type identity to clusters



Bring your issues on!