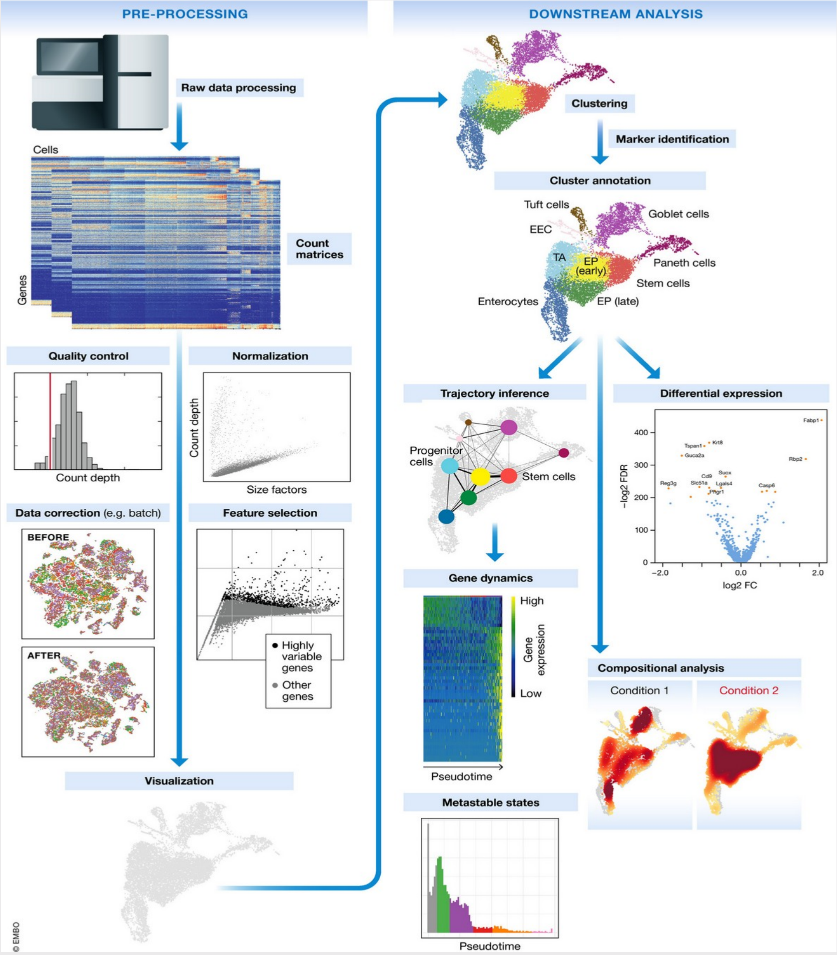
SincellTE 2024

Aziza CAIDI

Mapping, quality control and quantification

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

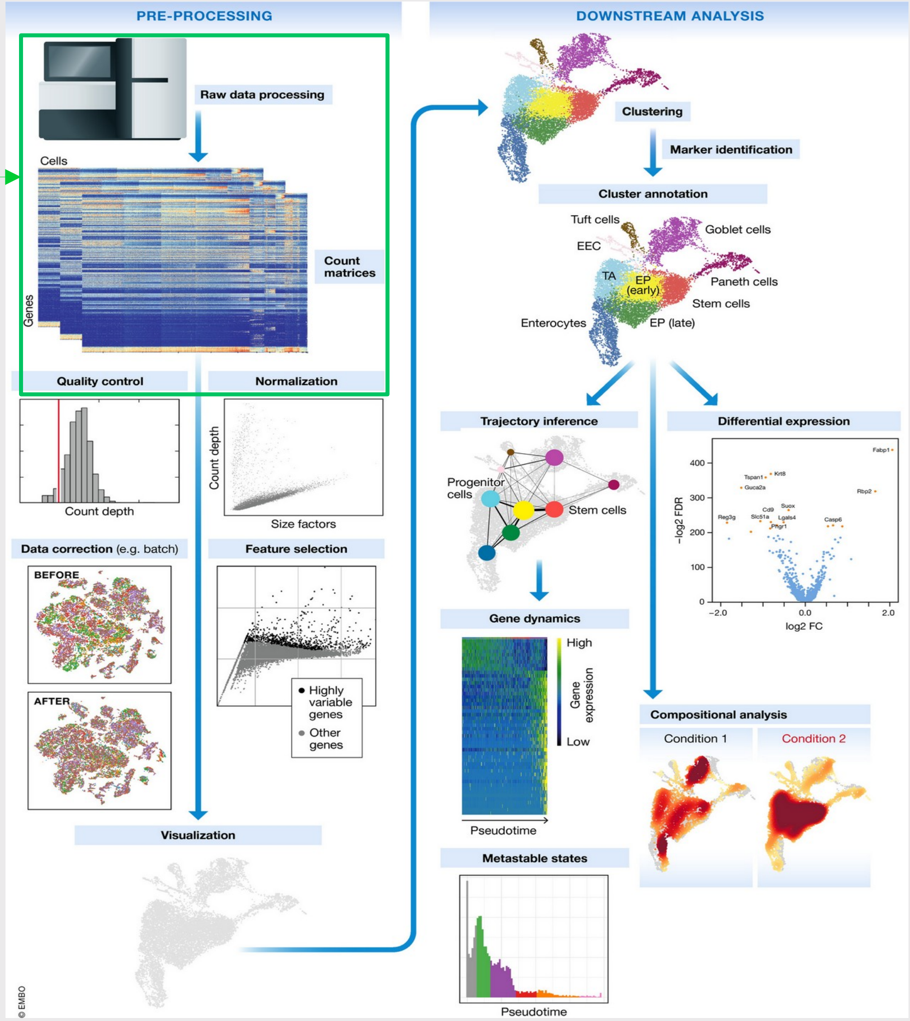
Main steps of single cell data processing



From Luecken and Theis, Mol Systems Biology 2019

Main steps of single cell data processing

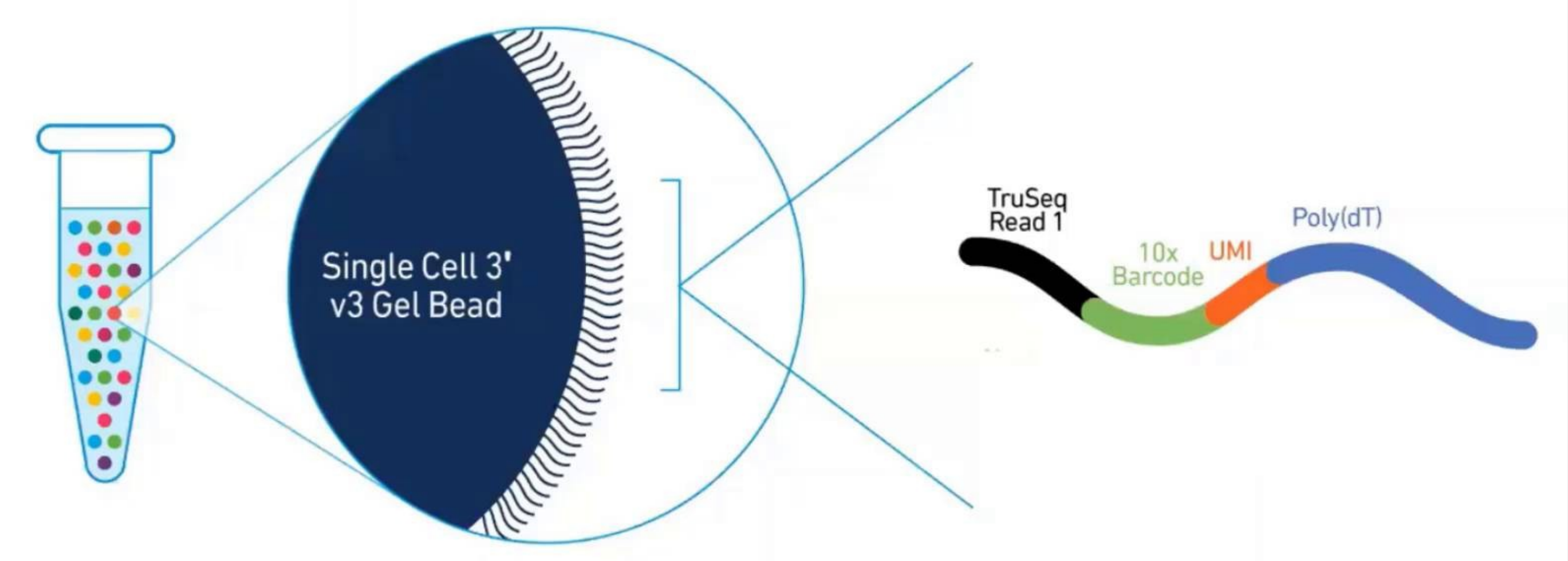
this course



From Luecken and Theis, Mol Systems Biology 2019

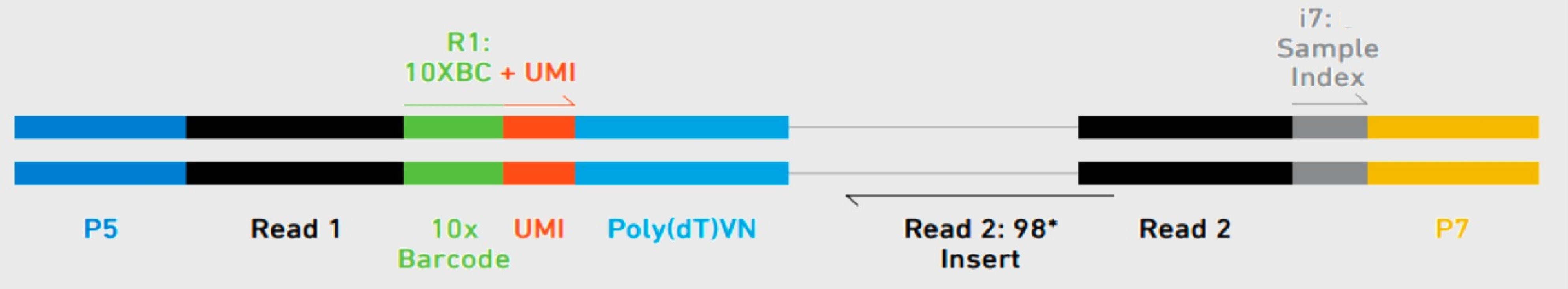
The starting library

We will use a droplet-based library as an example.



The starting library

We will use a droplet-based library as an example.



Read1: unique cell barcode (x nt) + UMI (y nt)

Read2: RNA 3’ sequence

I7: sample index: determines which sample the read originated from

Cellular barcode: determines which cell the read originated from
Unique molecular identifier (UMI): determines which transcript molecule the read originated from

Introduction

Plan

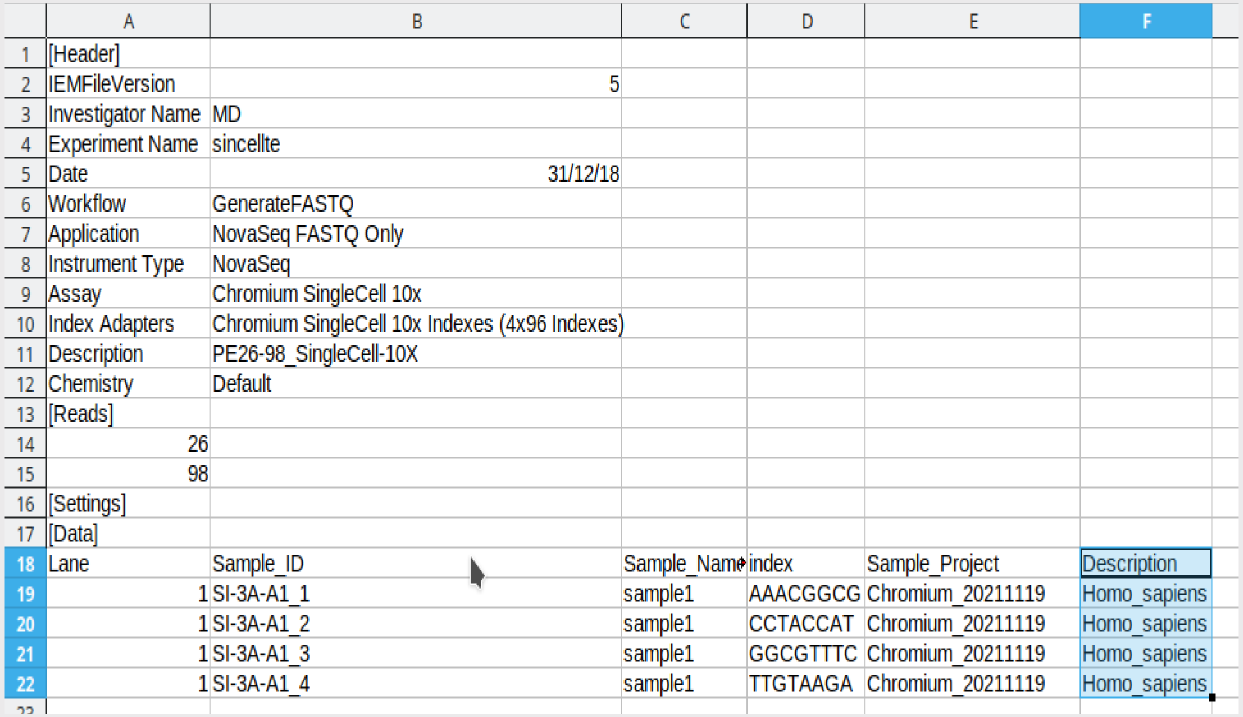
- Demultiplexing: generating fastqs from bcl
- Quality Check
- Generating a gene x cell count matrix

Demultiplexing

Convert BCL files (sequencer output) to fastq files
Most used tool : 10X's cellranger mkfastq a wrapper around bcl2fastq

- Usual sample sheet

- You must know :
  - i7 (i5) index sequence
  - R1 and R2 lengths
  - (depends on technology, version…)



	A	B	C	D	E	F
1	[Header]					
2	IEMFileVersion	5				
3	Investigator Name	MD				
4	Experiment Name	sincellte				
5	Date	31/12/18				
6	Workflow	GenerateFASTQ				
7	Application	NovaSeq FASTQ Only				
8	Instrument Type	NovaSeq				
9	Assay	Chromium SingleCell 10x				
10	Index Adapters	Chromium SingleCell 10x Indexes (4x96 Indexes)				
11	Description	PE26-98_SingleCell-10X				
12	Chemistry	Default				
13	[Reads]					
14	26					
15	98					
16	[Settings]					
17	[Data]					
18	Lane	Sample_ID	Sample_Name	index	Sample_Project	Description
19	1	SI-3A-A1_1	sample1	AAACGGCG	Chromium_20211119	Homo_sapiens
20	1	SI-3A-A1_2	sample1	CCTACCAT	Chromium_20211119	Homo_sapiens
21	1	SI-3A-A1_3	sample1	GGCGTTTC	Chromium_20211119	Homo_sapiens
22	1	SI-3A-A1_4	sample1	TTGTAAGA	Chromium_20211119	Homo_sapiens

https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/using/bcl2fastq-direct

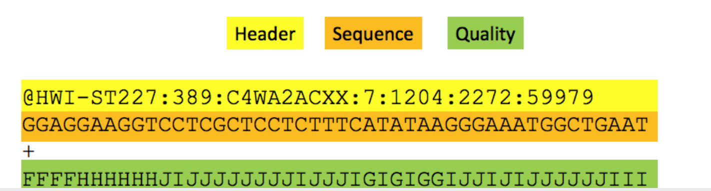
Demultiplexing

Convert BCL files (sequencer output) to fastq files
Most used tool : 10X’s cellranger mkfastq a wrapper around bcl2fastq

- Usual sample sheet

- You must know :
  - i7 (i5) index sequence
  - R1 and R2 lengths
  - (depends on technology, version…)

- 10X: 1 index = 4 sequences ⇒ 4 lines



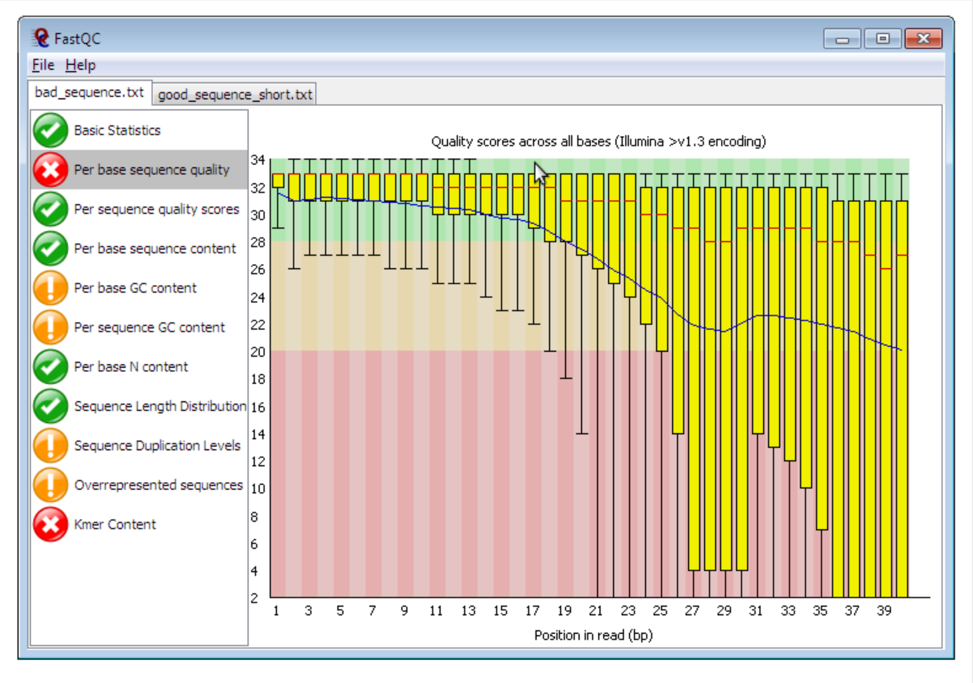
```
@HWI-ST227:389:C4WA2ACXX:7:1204:2272:59979
GGAGGAAGGTCCTCGCTCCTCTTTCATATAAGGGAAATGGCTGAAT
+
FFFFHHHHHJIJJJJJJJJIJJJIGIGIGGIJJIJIJJJJJJIII
```

Quality Check

Check reads quality : fastqc

- Performs various basic QC on reads

- For 10X scRNA datasets :
  - R1 (BC + UMI) : QC is mandatory. Watch out for Ns and highly repeated sequences
  - R2 : do as usual


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Trimming

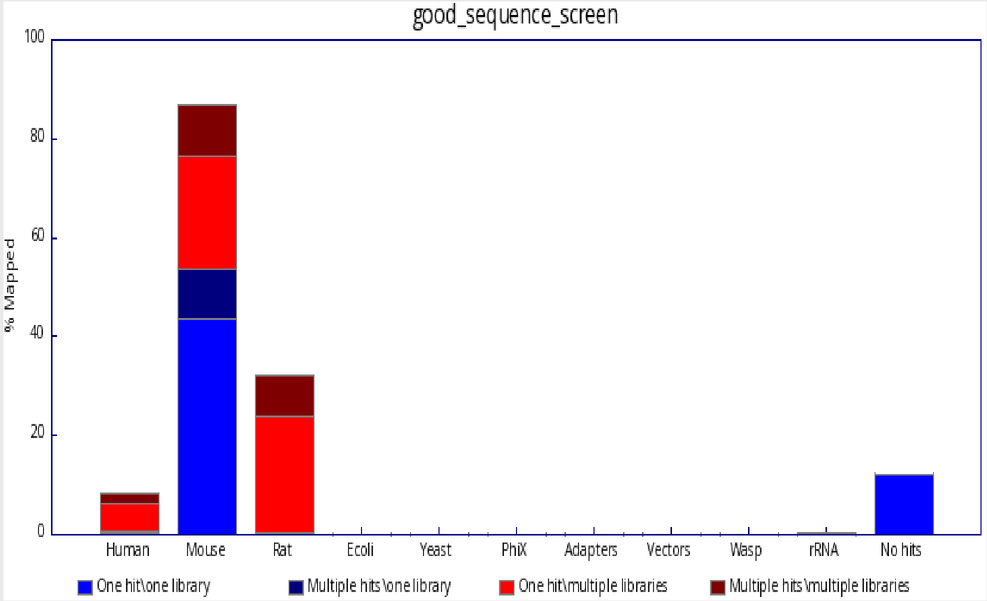
- If QC is not good:
  - Low base quality
  - Remaining adapter sequence
  - Homopolymer tailing
  - Low complexity

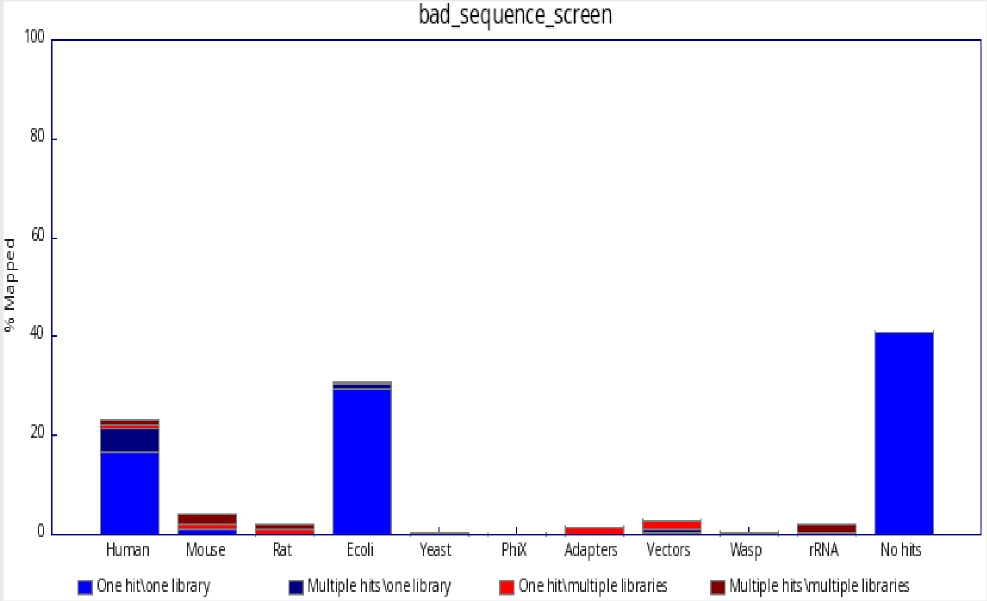
- Many tools to trim reads:
  - Trimmomatic (Bolger A.M. *et al.,* Bioinformatics (2014).
  - TrimGalore (Krueger F., https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, unpublished 2012).
  - Cutadapt (Martin M., EMBnet.journal 2011)
  - Fastp (Chen *et al.,* Bioinformatics 2018).

- For single cell, like with xenome, apply to R2 file, then sync the R1 file.

Quality Check

Check cross-species contaminations: FastQ Screen





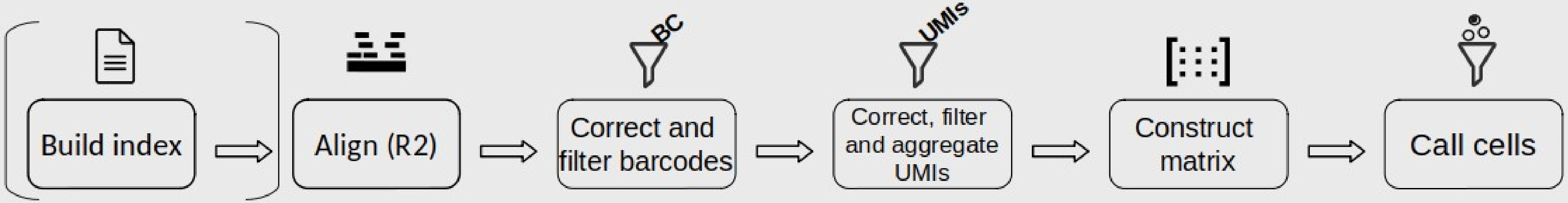
https://www.bioinformatics.babraham.ac.uk/projects/fastq_screen/

- Quick mapping (bowtie2) of a subset of reads across multiple genomes and common contaminants: human, mouse, rat, E. coli, adapters, vectors…

- Identifies 1hit-1library, multi hits-1library, 1hit-multi libraries and multi hits-multi libraries

Building the count matrix

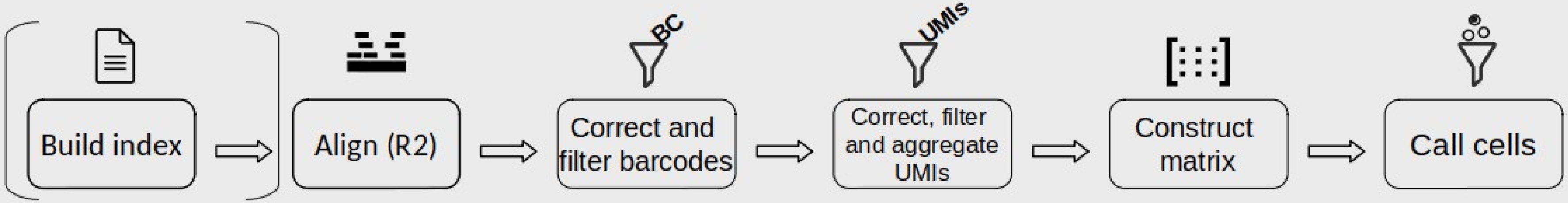
Principle


Build index
Align (R2)
BC
Correct and filter barcodes
UMIs
Correct, filter and aggregate UMIs
Construct matrix
Call cells

- 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).

Building the count matrix

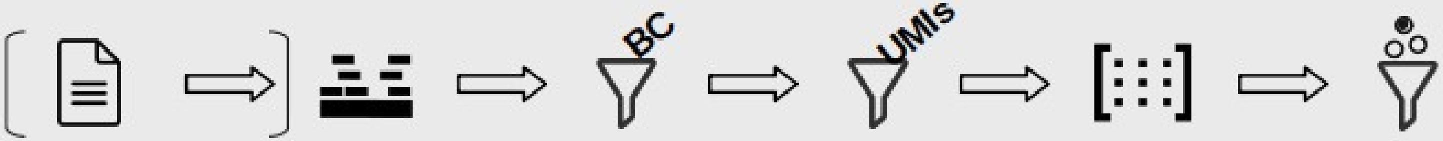
Principle


BC
UMIs
Build index
Align (R2)
Correct and filter barcodes
Correct, filter and aggregate UMIs
Construct matrix
Call cells

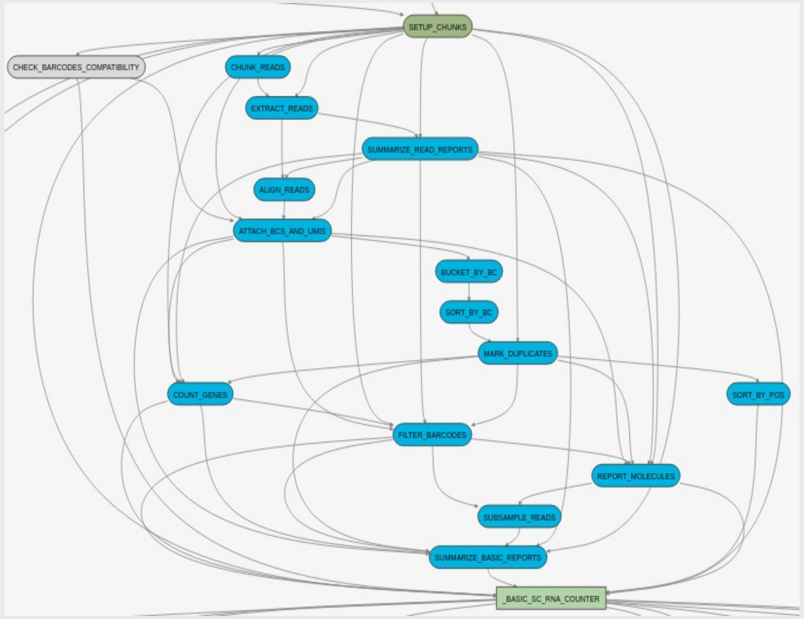
- Various tools have been developed:
  - **Cellranger**: 10X solution for 10X libraries only
  - **STARsolo**: an open source alternative to cellranger
  - **kallisto+bustools:** a pseudomapper and tool suite needing very little resources
  - (**Alevin**: a pseudomapper integrated with the salmon software)

Building the count matrix

Cellranger



- A set of pipelines for single cell analysis
- Many languages + task scheduler Martian
- Aligner: STAR
- single cell gene expression: *cellranger count*



https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest

Building the count matrix

Reference Preparation

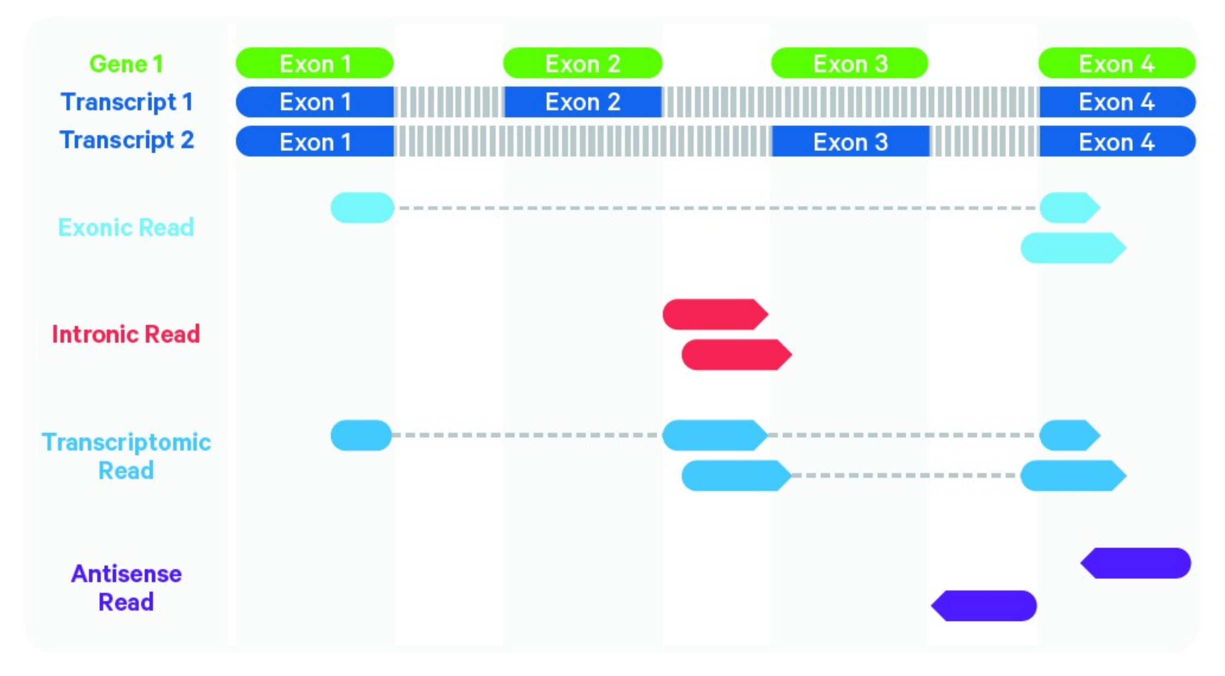
- Human/mouse retained biotypes :
- Protein coding
- Long noncoding RNA
- Antisense
- All biotypes belonging to BCR/TCR (i.e. V/D/J) Genes
- All pseudogenes and small noncoding rnas are removed.

(note that older Cell Ranger reference versions do not include BCR/TCR Genes)

Building the count matrix

Alignement

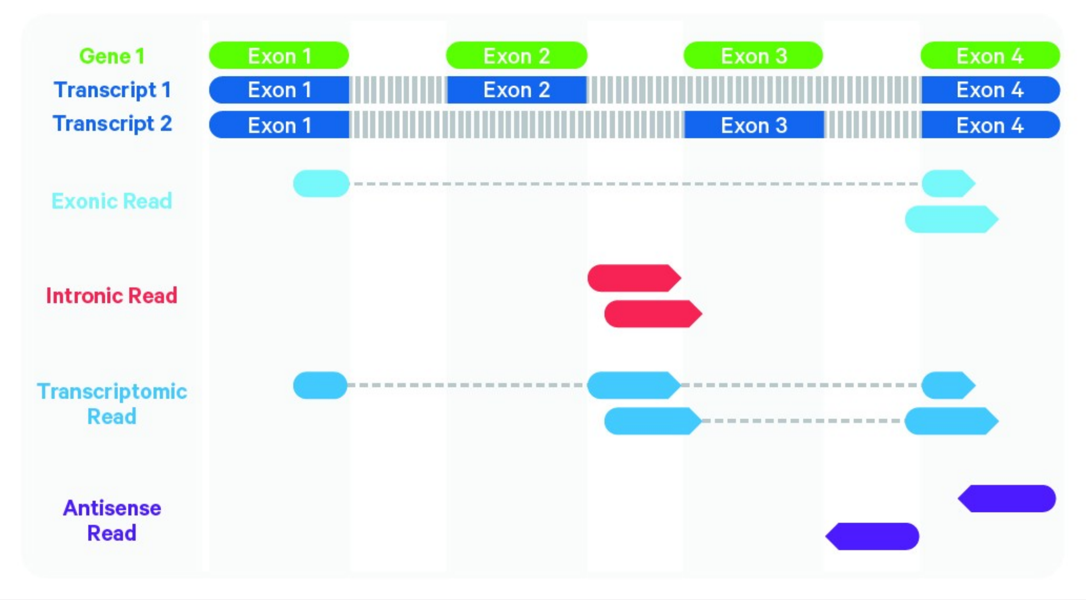
- Cell Ranger further aligns confidently mapped exonic and intronic reads to annotated transcripts by examining their compatibility with the transcriptome
- Reads are classified based on whether they are exonic (light blue) or intronic (red) and whether they are sense or antisense (purple).



Building the count matrix

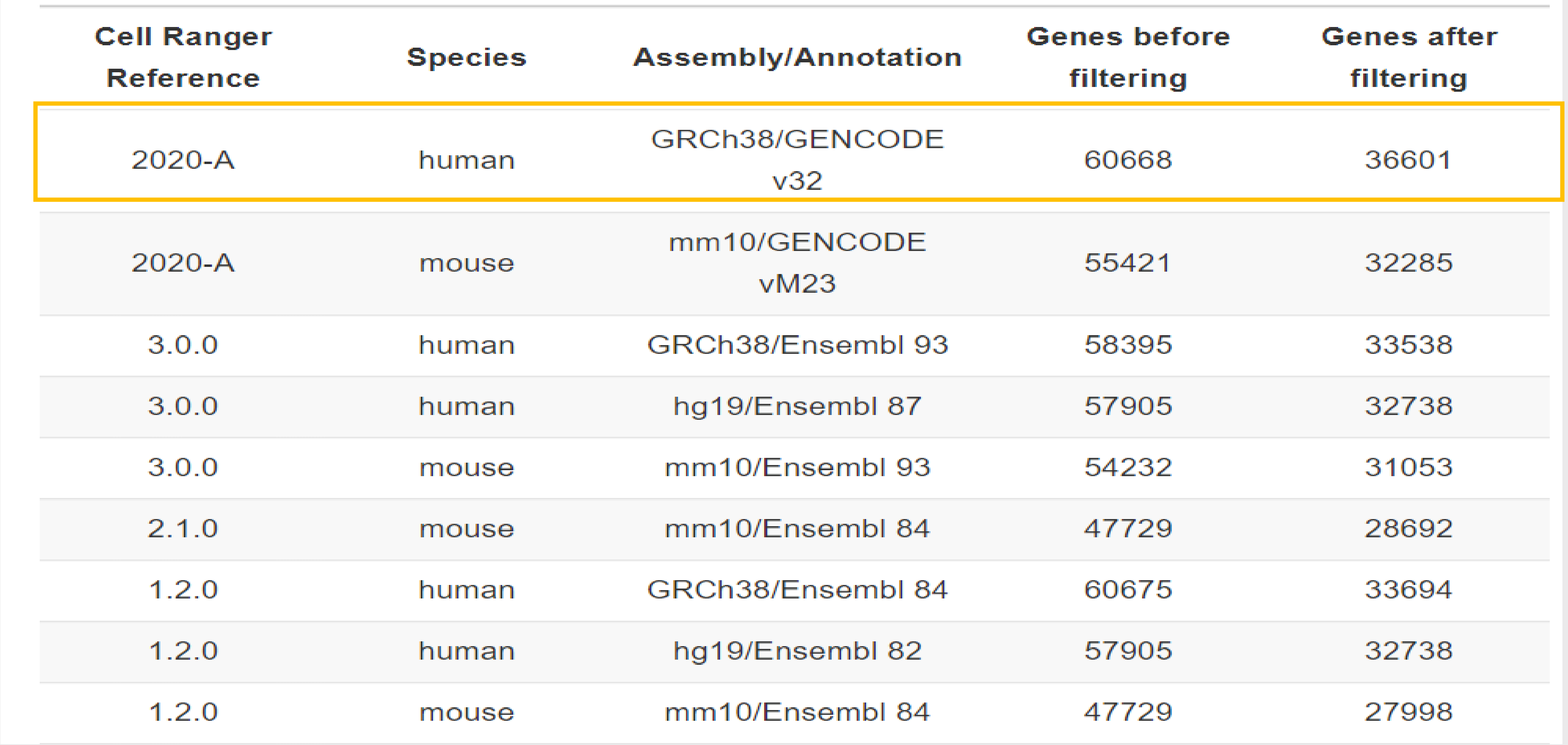
Alignement

- In Cell Ranger 7.0, by default, the cellranger-count and cellranger-multi pipelines will include intronic reads for whole transcriptome gene expression analysis –> recommended to maximize sensitivity
- Any reads that map in the sense orientation to a single are carried forward to UMI counting.
- Cell Ranger ignores antisense reads (purple).



Building the count matrix

Reference Preparation

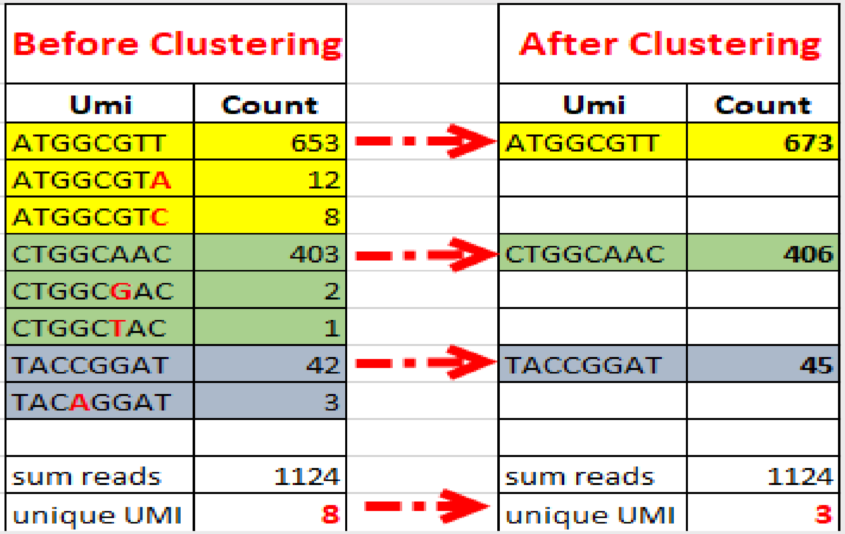


Cell Ranger Reference	Species	Assembly/Annotation	Genes before filtering	Genes after filtering
2020-A	human	GRCh38/GENCODE v32	60668	36601
2020-A	mouse	mm10/GENCODE vM23	55421	32285
3.0.0	human	GRCh38/Ensembl 93	58395	33538
3.0.0	human	hg19/Ensembl 87	57905	32738
3.0.0	mouse	mm10/Ensembl 93	54232	31053
2.1.0	mouse	mm10/Ensembl 84	47729	28692
1.2.0	human	GRCh38/Ensembl 84	60675	33694
1.2.0	human	hg19/Ensembl 82	57905	32738
1.2.0	mouse	mm10/Ensembl 84	47729	27998

Building the count matrix

UMI Counting

- Cell Ranger attempts to correct for sequencing errors in the UMI sequences by association (Group confidently mapped reads -> Associate UMIs differ by a single base (less confidently mapped) to their assigned group)
- Gene annotation with the most supporting reads is kept for UMI counting, and the other read groups are discarded



Before Clustering			After Clustering	
Umi	Count		Umi	Count
ATGGCGTT	653	→	ATGGCGTT	673
ATGGCGTA	12			
ATGGCGTC	8			
CTGGCAAC	403	→	CTGGCAAC	406
CTGGCGAC	2			
CTGGCTAC	1			
TACCGGAT	42	→	TACCGGAT	45
TACAGGAT	3			
sum reads	1124		sum reads	1124
unique UMI	8	→	unique UMI	3

Building the count matrix

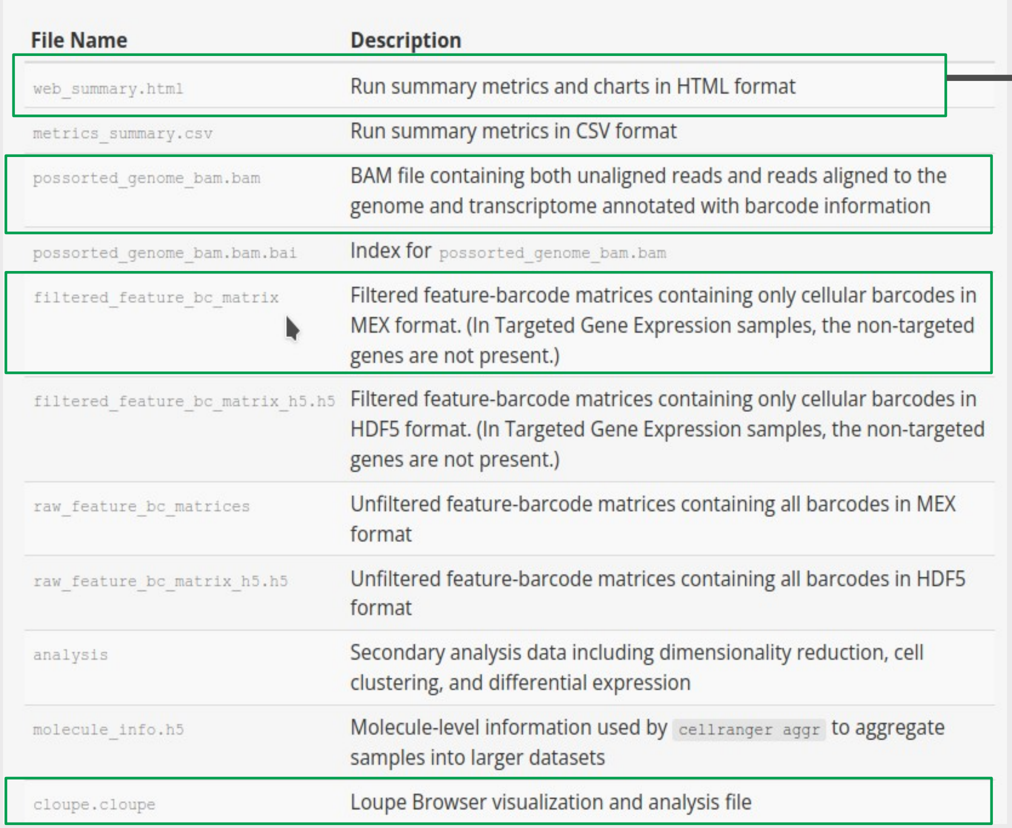
UMI Counting

- Cell Ranger attempts to correct for sequencing errors in the UMI sequences by association (Group confidently mapped reads -> Associate UMIs differ by a single base (less confidently mapped) to their assigned group)
- Gene annotation with the most supporting reads is kept for UMI counting, and the other read groups are discarded
- Aggregation: 1 BC+UMIs = 1 unique RNA molecule (filter excess)
- Finally, construct matrix with selected reads: *genes x barcodes*

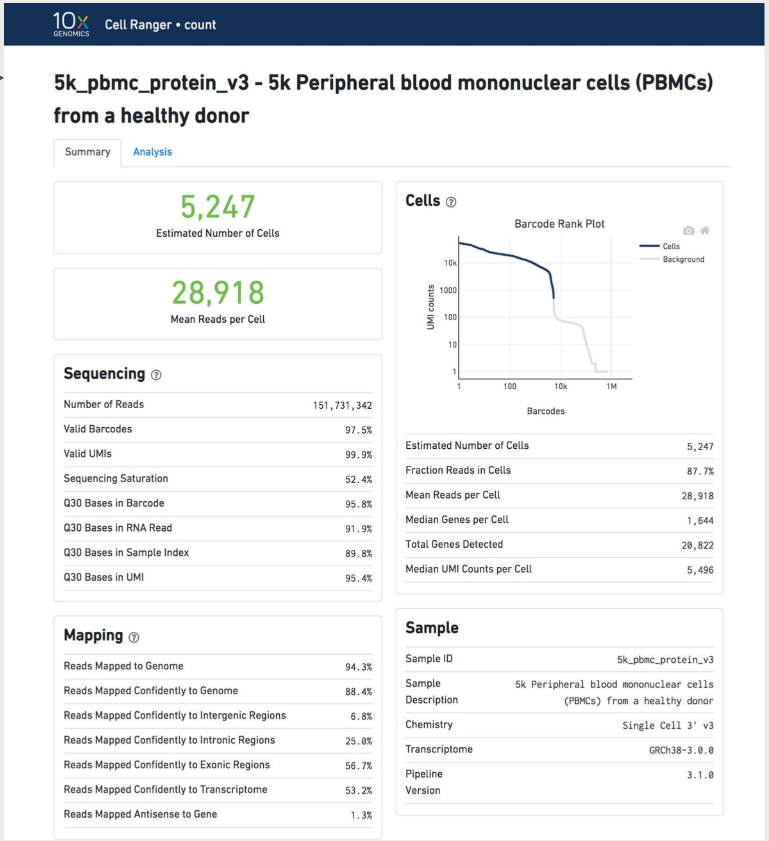
Building the count matrix

Cellranger

Outputs



File Name	Description
`web_summary.html`	Run summary metrics and charts in HTML format
`metrics_summary.csv`	Run summary metrics in CSV format
`possorted_genome_bam.bam`	BAM file containing both unaligned reads and reads aligned to the genome and transcriptome annotated with barcode information
`possorted_genome_bam.bam.bai`	Index for `possorted_genome_bam.bam`
`filtered_feature_bc_matrix`	Filtered feature-barcode matrices containing only cellular barcodes in MEX format. (In Targeted Gene Expression samples, the non-targeted genes are not present.)
`filtered_feature_bc_matrix_h5.h5`	Filtered feature-barcode matrices containing only cellular barcodes in HDF5 format. (In Targeted Gene Expression samples, the non-targeted genes are not present.)
`raw_feature_bc_matrices`	Unfiltered feature-barcode matrices containing all barcodes in MEX format
`raw_feature_bc_matrix_h5.h5`	Unfiltered feature-barcode matrices containing all barcodes in HDF5 format
`analysis`	Secondary analysis data including dimensionality reduction, cell clustering, and differential expression
`molecule_info.h5`	Molecule-level information used by `cellranger aggr` to aggregate samples into larger datasets
`cloupe.cloupe`	Loupe Browser visualization and analysis file



10x GENOMICS Cell Ranger • count

5k_pbmc_protein_v3 - 5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor

Summary Analysis

5,247
Estimated Number of Cells

28,918
Mean Reads per Cell

Sequencing

Number of Reads	151,731,342
Valid Barcodes	97.5%
Valid UMIs	99.9%
Sequencing Saturation	52.4%
Q30 Bases in Barcode	95.8%
Q30 Bases in RNA Read	91.9%
Q30 Bases in Sample Index	89.8%
Q30 Bases in UMI	95.4%

Mapping

Reads Mapped to Genome	94.3%
Reads Mapped Confidently to Genome	88.4%
Reads Mapped Confidently to Intergenic Regions	6.8%
Reads Mapped Confidently to Intronic Regions	25.0%
Reads Mapped Confidently to Exonic Regions	56.7%
Reads Mapped Confidently to Transcriptome	53.2%
Reads Mapped Antisense to Gene	1.3%

Cells

Estimated Number of Cells	5,247
Fraction Reads in Cells	87.7%
Mean Reads per Cell	28,918
Median Genes per Cell	1,644
Total Genes Detected	20,822
Median UMI Counts per Cell	5,496

Sample

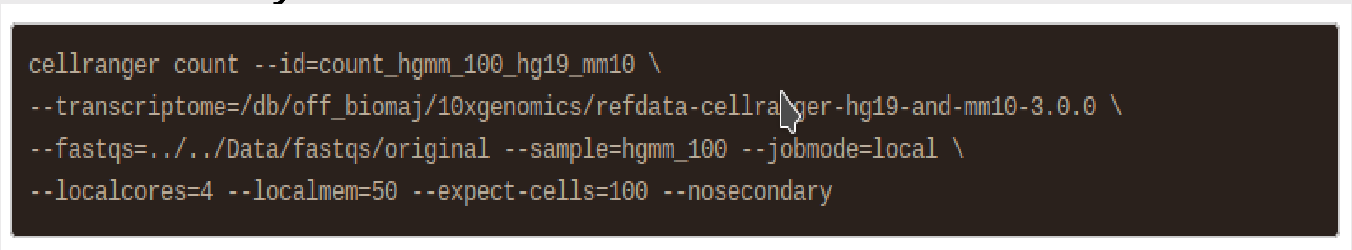
Sample ID	5k_pbmc_protein_v3
Sample Description	5k Peripheral blood mononuclear cells (PBMCs) from a healthy donor
Chemistry	Single Cell 3' v3
Transcriptome	GRCh38-3.0.0
Pipeline Version	3.1.0

Building the count matrix

Cellranger

🙂

- Turnkey solution



```
cellranger count --id=count_hgmm_100_hg19_mm10 \
--transcriptome=/db/off_biomaj/10xgenomics/refdata-cellranger-hg19-and-mm10-3.0.0 \
--fastqs=../../Data/fastqs/original --sample=hgmm_100 --jobmode=local \
--localcores=4 --localmem=50 --expect-cells=100 --nosecondary
```

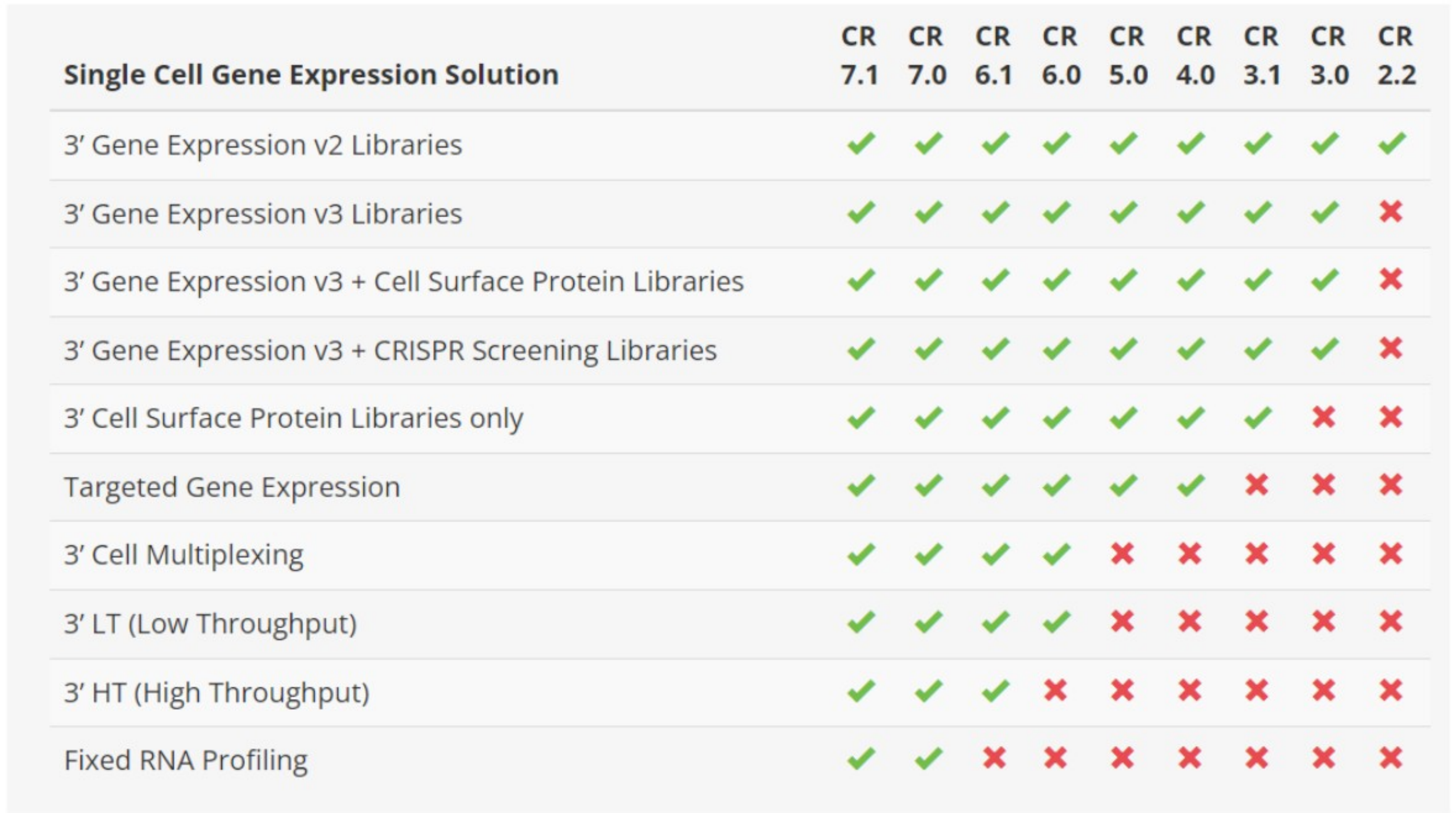
- Many QC-metrics, results summarized in 1 html.
- Some secondary analysis
- More complex experiences: VDJ analysis, feature-barcoding
- Versions for ATAC-Seq, TCR-seq and BCR-seq

☹

- Proprietary
- Analyze only 10X product (cannot customize BC and UMI)
- Has its own scheduler: hard to include in another pipeline
- Compatibility not guaranteed with all HPC managers

Building the count matrix

Cellranger

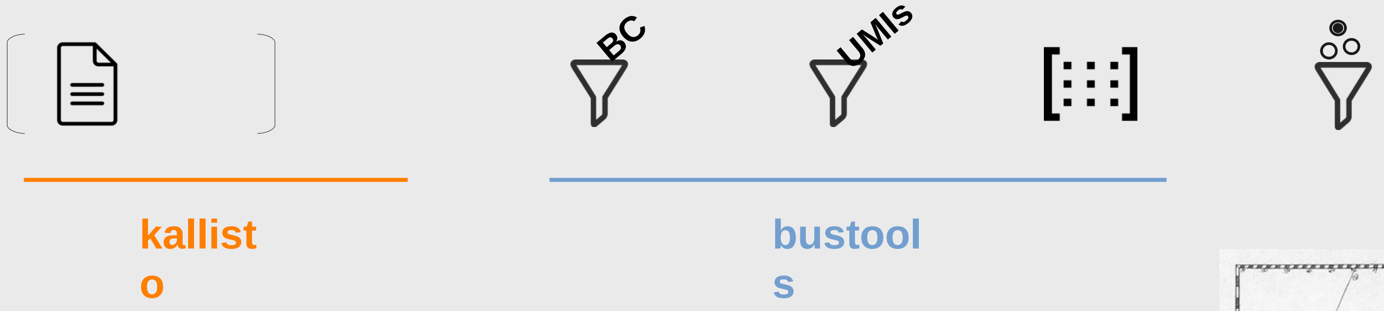


	CR	CR	CR	CR	CR	CR	CR	CR	CR
Single Cell Gene Expression Solution	**7.1**	**7.0**	**6.1**	**6.0**	**5.0**	**4.0**	**3.1**	**3.0**	**2.2**
3′ Gene Expression v2 Libraries	✓	✓	✓	✓	✓	✓	✓	✓	✓
3′ Gene Expression v3 Libraries	✓	✓	✓	✓	✓	✓	✓	✓	✗
3′ Gene Expression v3 + Cell Surface Protein Libraries	✓	✓	✓	✓	✓	✓	✓	✓	✗
3′ Gene Expression v3 + CRISPR Screening Libraries	✓	✓	✓	✓	✓	✓	✓	✓	✗
3′ Cell Surface Protein Libraries only	✓	✓	✓	✓	✓	✓	✓	✗	✗
Targeted Gene Expression	✓	✓	✓	✓	✓	✓	✗	✗	✗
3′ Cell Multiplexing	✓	✓	✓	✓	✗	✗	✗	✗	✗
3′ LT (Low Throughput)	✓	✓	✓	✓	✗	✗	✗	✗	✗
3′ HT (High Throughput)	✓	✓	✓	✗	✗	✗	✗	✗	✗
Fixed RNA Profiling	✓	✓	✗	✗	✗	✗	✗	✗	✗

Cell Ranger v8.0 introduces support for the analysis of GEM-X libraries.
Cell Ranger v7.2 is the last version to support the analysis of LT (low throughput) libraries.

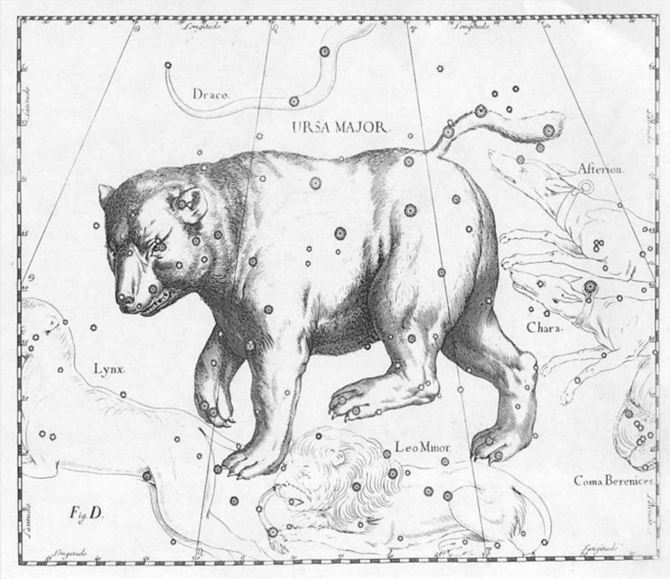
Building the count matrix

Kallisto/bustools



- Make use of the pseudo-aligner kallisto and the toolsuite bustools

- Very good time and memory performance.



https://pachterlab.github.io/kallisto/download

Building the count matrix

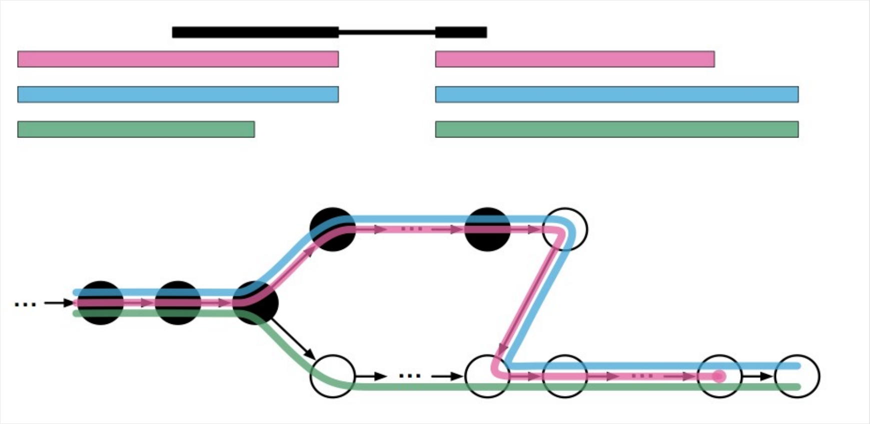
Kallisto/bustools

- Kallisto is a pseudo aligner: fast, low memory
- Working with a reference transcriptome, not genome

Building the count matrix

Kallisto/bustools

- Kallisto is a pseudo aligner: fast, low memory
- Working with a reference transcriptome, not genome
- Principle:
  - reference chunked into k-mers ==> de Bruijn Graph
  - Reads chunked into k-mers and assigned to the
  - transcript(s) they overlap with
  - 1 read generally compatible with several transcripts:
  - proportion of transcripts computed by
  - Expectation Maximization from all reads



From Bray *et al.*, *Nat Biointechno* 2016

A very nice explanation of kallisto: https://bioinfo.iric.ca/fr/comprendre-comment-kallisto-fonctionne

Building the count matrix

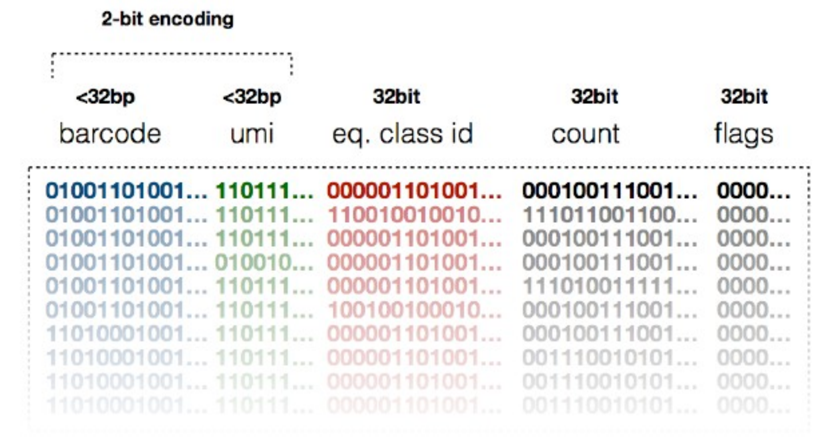
Kallisto/bustools

- Many technologies already accepted, the CB + UMI geometry is configurable

Allows analysis of non 10X technologies

- Gives relative abundance, not absolute counts

- Output format in a specific, compressed format: bus instead of sam or bam files.


2-bit encoding

<32bp barcode	<32bp umi	32bit eq. class id	32bit count	32bit flags
01001101001...	110111...	000001101001...	000100111001...	0000...
01001101001...	110111...	110010010010...	111011001100...	0000...
01001101001...	110111...	000001101001...	000100111001...	0000...
01001101001...	010010...	000001101001...	000100111001...	0000...
01001101001...	110111...	000001101001...	111010011111...	0000...
01001101001...	110111...	100100100010...	000100111001...	0000...
11010001001...	110111...	000001101001...	000100111001...	0000...
11010001001...	110111...	000001101001...	001110010101...	0000...
11010001001...	110111...	000001101001...	001110010101...	0000...
11010001001...	110111...	000001101001...	001110010101...	0000...

From Melsted et al., Bioinformatics 2019

Building the count matrix

Kallisto/bustools

- Next steps: bustools

Bus file + BC whitelist

bustools correct: correct and filter BC

bustools sort: sort results by BC, UMIs and gene

bustools count: correct and filter UMIs, construct matrix

raw gene x barcodes matrix

Building the count matrix

Kallisto/bustools

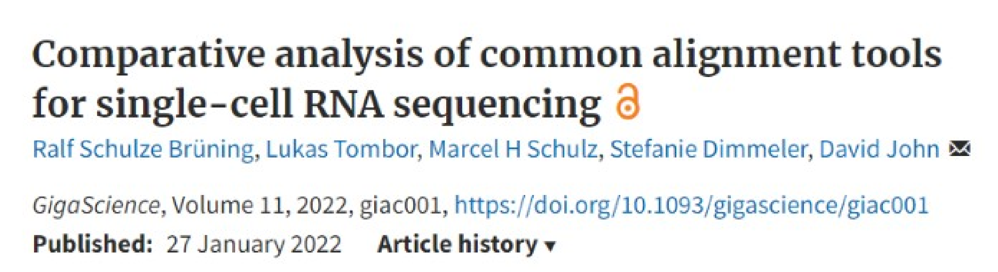
☺

- For modular pipeline construction
- Not proprietary
- Allows analysis of non 10X technologies
- The fastest and less resource consuming (can run on a laptop)
- Easy to include in a pipeline
- Compatible with HPC managers

☹

- Not a turnkey solution
- No secondary analysis
- Gap with cellranger
- No Add sample_name and well range

Conclusion

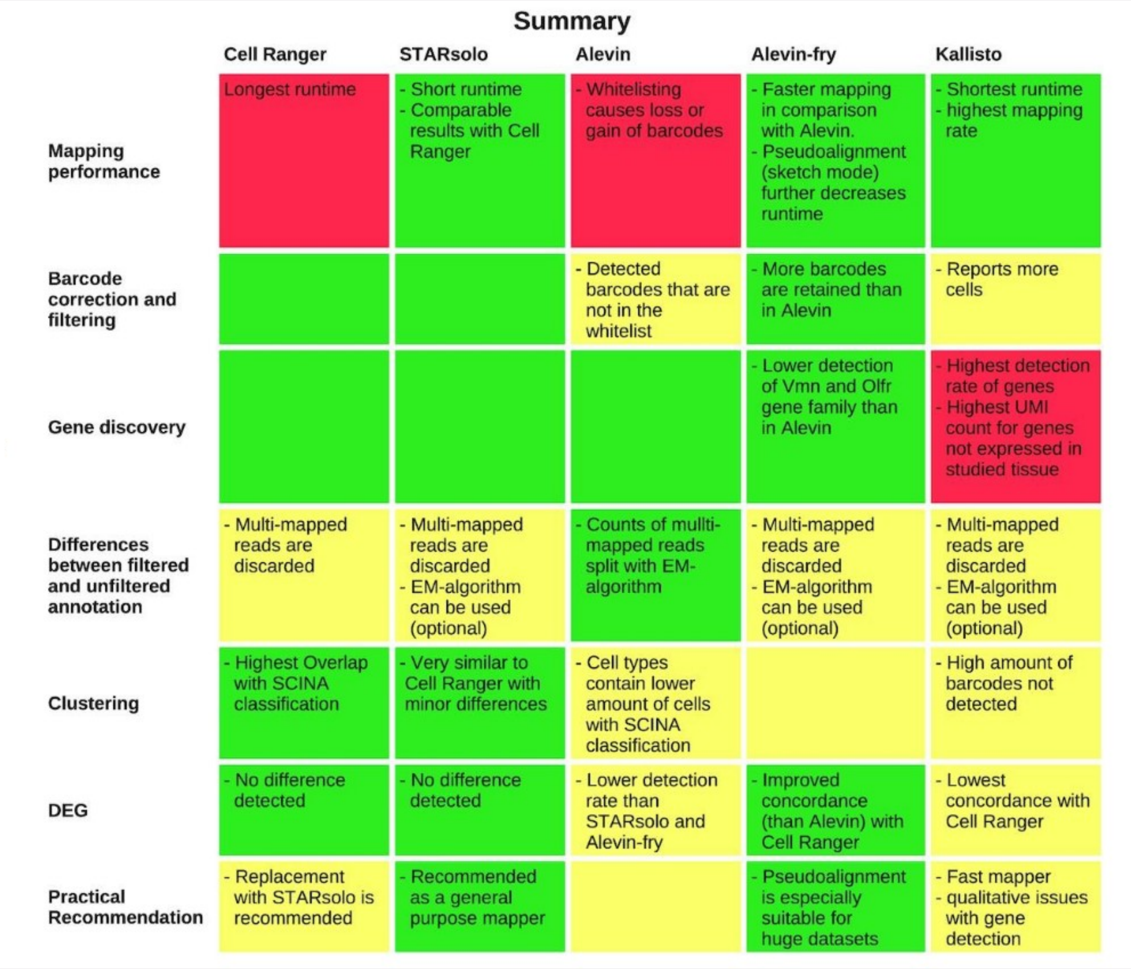


Comparative analysis of common alignment tools for single-cell RNA sequencing

Ralf Schulze Brüning, Lukas Tombor, Marcel H Schulz, Stefanie Dimmeler, David John

GigaScience, Volume 11, 2022, giac001, https://doi.org/10.1093/gigascience/giac001

Published: 27 January 2022 **Article history** ▾



Summary

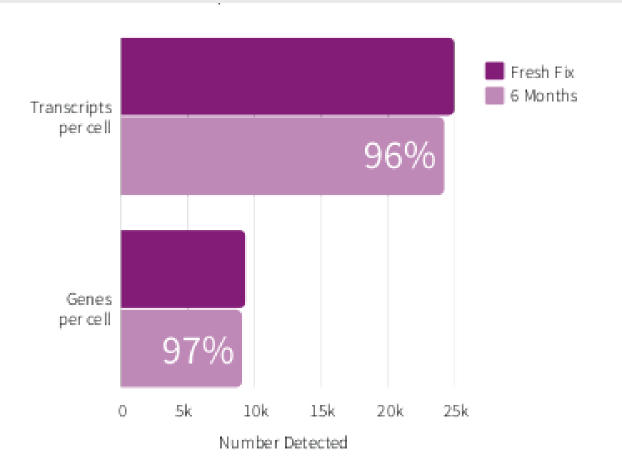
	Cell Ranger	STARsolo	Alevin	Alevin-fry	Kallisto
Mapping performance	🟥 Longest runtime	🟩 - Short runtime - Comparable results with Cell Ranger	🟥 - Whitelisting causes loss or gain of barcodes	🟩 - Faster mapping in comparison with Alevin. - Pseudoalignment (sketch mode) further decreases runtime	🟩 - Shortest runtime - highest mapping rate
Barcode correction and filtering	🟩	🟩	🟨 - Detected barcodes that are not in the whitelist	🟩 - More barcodes are retained than in Alevin	🟨 - Reports more cells
Gene discovery	🟩	🟩	🟩	🟩 - Lower detection of Vmn and Olfr gene family than in Alevin	🟥 - Highest detection rate of genes - Highest UMI count for genes not expressed in studied tissue
Differences between filtered and unfiltered annotation	🟨 - Multi-mapped reads are discarded	🟨 - Multi-mapped reads are discarded - EM-algorithm can be used (optional)	🟩 - Counts of mullti-mapped reads split with EM-algorithm	🟨 - Multi-mapped reads are discarded - EM-algorithm can be used (optional)	🟨 - Multi-mapped reads are discarded - EM-algorithm can be used (optional)
Clustering	🟩 - Highest Overlap with SCINA classification	🟩 - Very similar to Cell Ranger with minor differences	🟨 - Cell types contain lower amount of cells with SCINA classification	🟨	🟨 - High amount of barcodes not detected
DEG	🟩 - No difference detected	🟩 - No difference detected	🟨 - Lower detection rate than STARsolo and Alevin-fry	🟩 - Improved concordance (than Alevin) with Cell Ranger	🟨 - Lowest concordance with Cell Ranger
Practical Recommendation	🟨 - Replacement with STARsolo is recommended	🟩 - Recommended as a general purpose mapper	🟨	🟩 - Pseudoalignment is especially suitable for huge datasets	🟨 - Fast mapper - qualitative issues with gene detection

From Shulze Bruning *et al.*, *GigaScience* 2022

Other technology: Parse Biosciences

Design Flexible Experiments that Scale

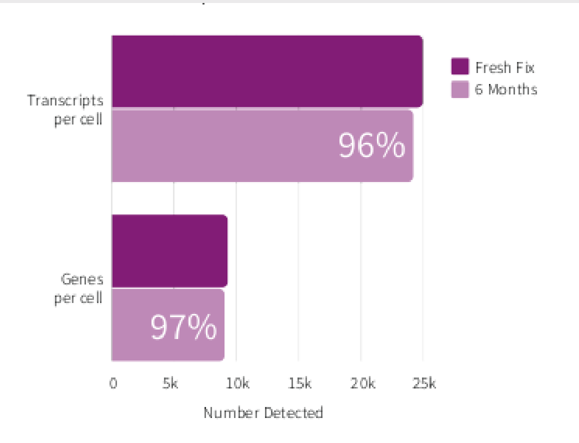
- Multiples samples are fixed and can be sequenced up to 6 months later

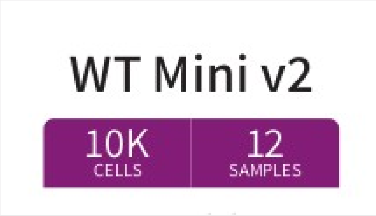


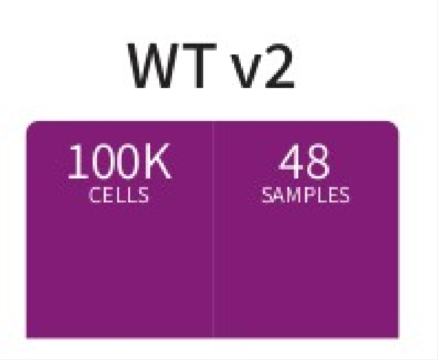
Other technology: Parse Biosciences

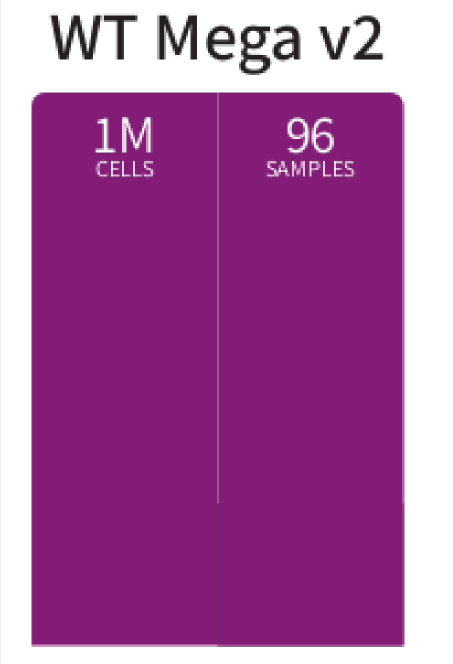
Design Flexible Experiments that Scale

- Multiples samples can be fixed sequenced up to 6 months later
- 3 kits are available







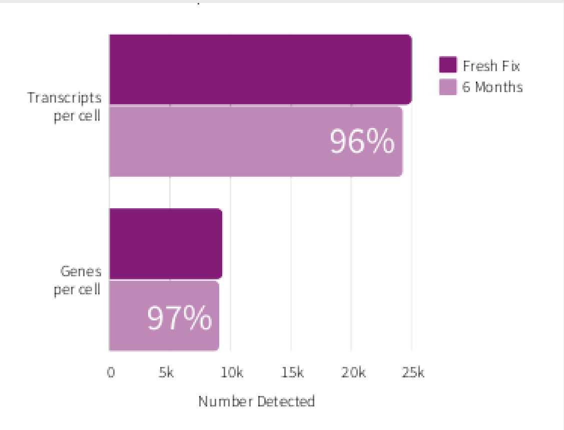


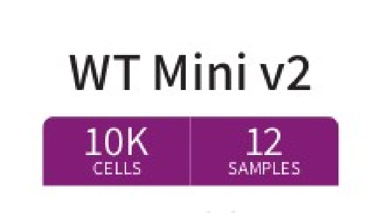
Other technology: Parse Biosciences

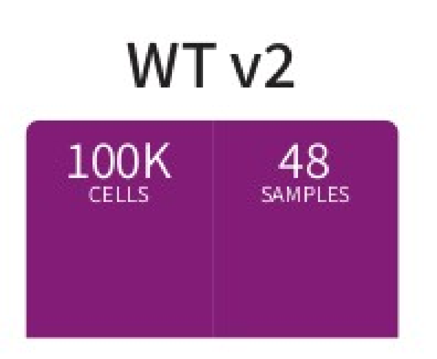
Design Flexible Experiments that Scale

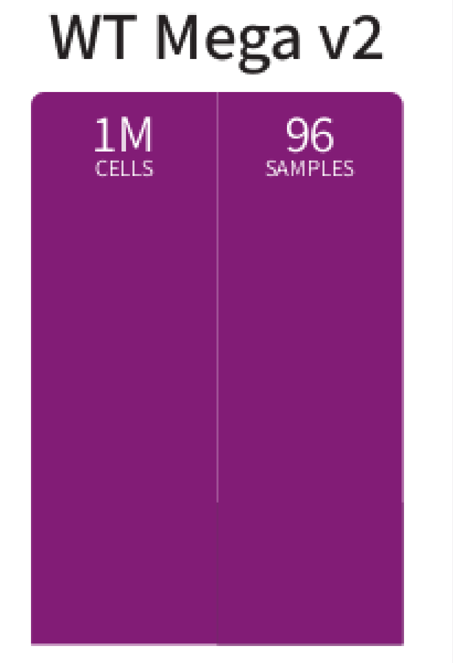
Multiples samples can be fixed sequenced up to 6 months later

- 3 kits are available
- four barcoding steps are required





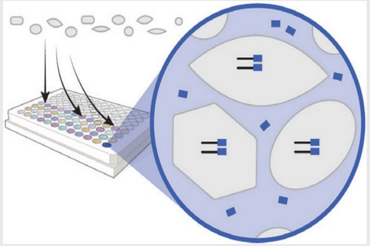




Parse Biosciences Workflow

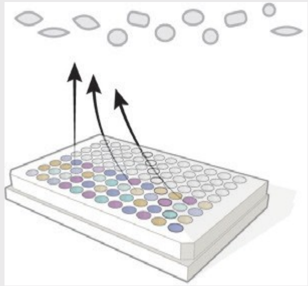
1 Reverse Transcription

Split : Fixed cells/nuclei are distributed into wells, and the first sample-specific barcodes are added by in-cell reverse transcription.



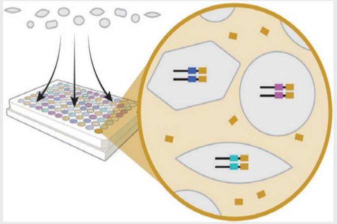
2

Pool : All the cells are pooled together.

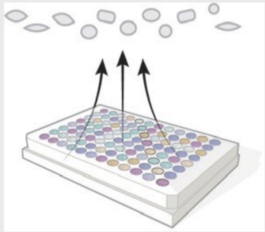


3 Ligation

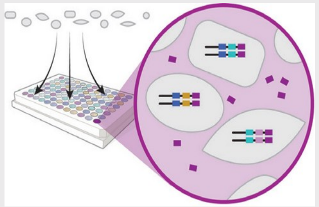
Split : The pooled cells are distributed across a plate, and an in-cell ligation adds the second barcode.



4 Pool : All the cells are pooled together.



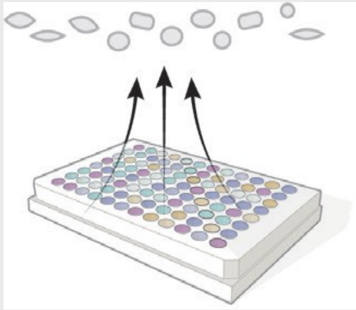
5 Split | The pooled cells are again distributed across a plate, and a third barcode is added via in-cell ligation reaction.



Parse Biosciences Workflow

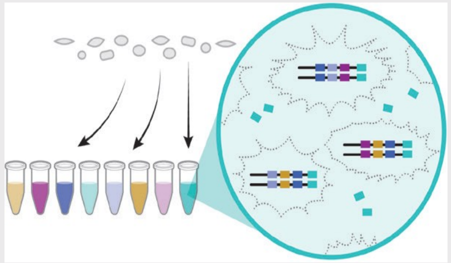
6

Pool : All the cells are pooled together.



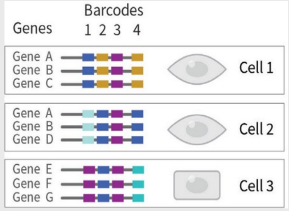
7 Lysis and Library Prep

Split : The pooled cells are distributed across several sub-libraries then lysed. The fourth barcode is added via PCR.



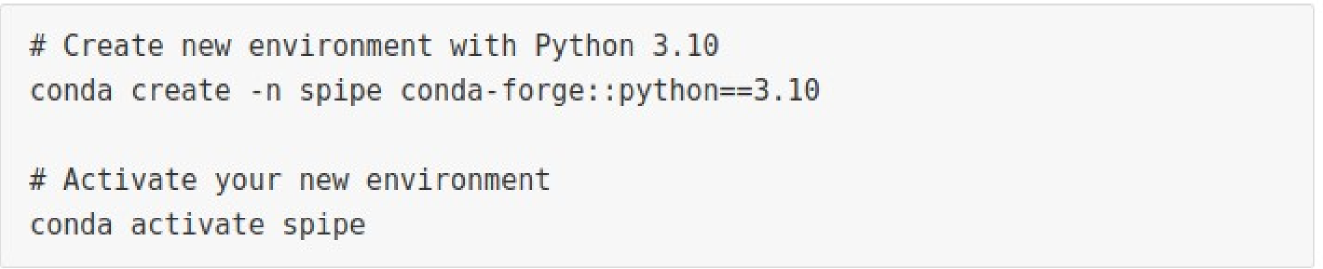
8 Sequencing with Illumina

Each transcript is assigned to a single cell based on a unique combination of barcodes.



9 Data Analysis

ParseBiosciences-Pipeline.1.2.0.zip



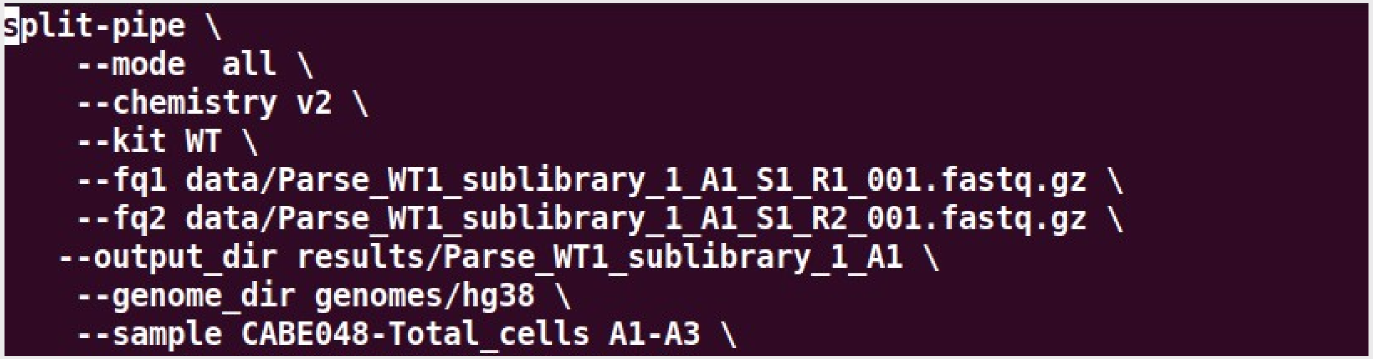
```
# Create new environment with Python 3.10
conda create -n spipe conda-forge::python==3.10

# Activate your new environment
conda activate spipe
```

Building the count matrix

ParseBiosciences-Pipeline

- Reference preparation : use split-pipe --mode mkref
- Aligner: STAR
- single cell gene expression : Parse count
- specify samples name for well ranges
- Running the pipeline for each library
- Combine libraries results



```
split-pipe \
    --mode  all \
    --chemistry v2 \
    --kit WT \
    --fq1 data/Parse_WT1_sublibrary_1_A1_S1_R1_001.fastq.gz \
    --fq2 data/Parse_WT1_sublibrary_1_A1_S1_R2_001.fastq.gz \
   --output_dir results/Parse_WT1_sublibrary_1_A1 \
    --genome_dir genomes/hg38 \
    --sample CABE048-Total_cells A1-A3 \
```

Conclusion

🙂

- Turnkey solution
- Many QC-metrics, results summarized in 1 html.
- Some secondary analysis
- No empty droplet
- Versions for TCR-Seq and BCR-seq
- Batch effect reduced

☹

- Proprietary
- Analyze only Parse product (cannot customize BC and UMI)
- Has its own scheduler: hard to include in another pipeline

SincellTE 2024

Aziza CAIDI

Thank you for your attention!

Thanks to Marine AGLAVE

GUSTAVE ROUSSY
CANCER CAMPUS
GRAND PARIS

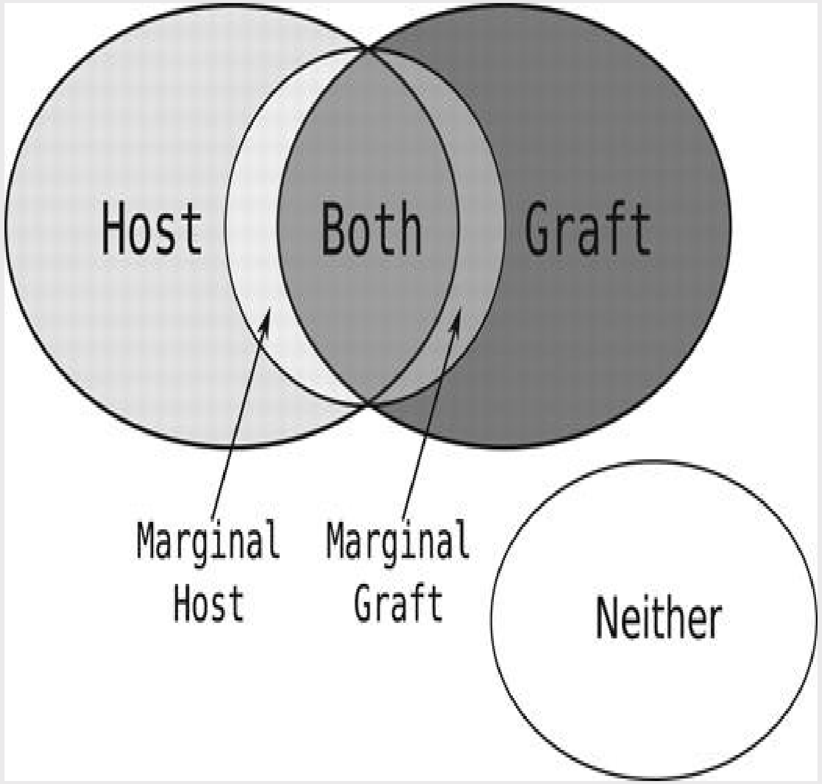
Quality Check

Multiple species: Xenome

- For xenografts or contaminated samples

- 5 fastq files :
  - Graft
  - Host
  - Both
  - Neither
  - Ambiguous

- For single cell, apply to R2 only and sync R1: e.g. seqkit:
  - *seqkit seq* lists the selected read names.
  - *seqkit grep* filters R1 by keeping only reads in this list.
  - *seqkit pair* pairs filtered R1 with R2.

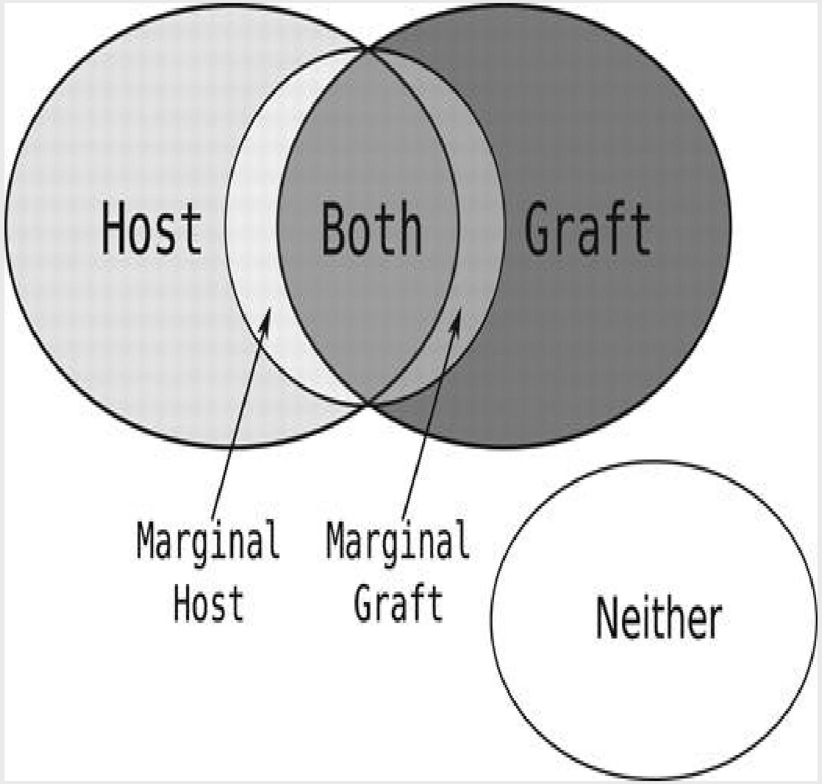


https://github.com/data61/gossamer/blob/master/docs/xenome.md

Multiple species: Xenome

- Xenome version is bugged: patch gossamer
- https://github.com/data61/gossamer

- Alternatives :
  - Xengsort (Zentgraf and Rahmann, S. Mol Biol 2021).
  - XenofilteR (Kluin *et al*, BMC Bioinfo 2018)
  - Bamcmp (Khandelwal *et al.*, MCR 2017).
  - XenoSplit: (https://github.com/goknurginer/XenoSplit Unpublished 2019).



https://github.com/data61/gossamer/blob/master/docs/xenome.md