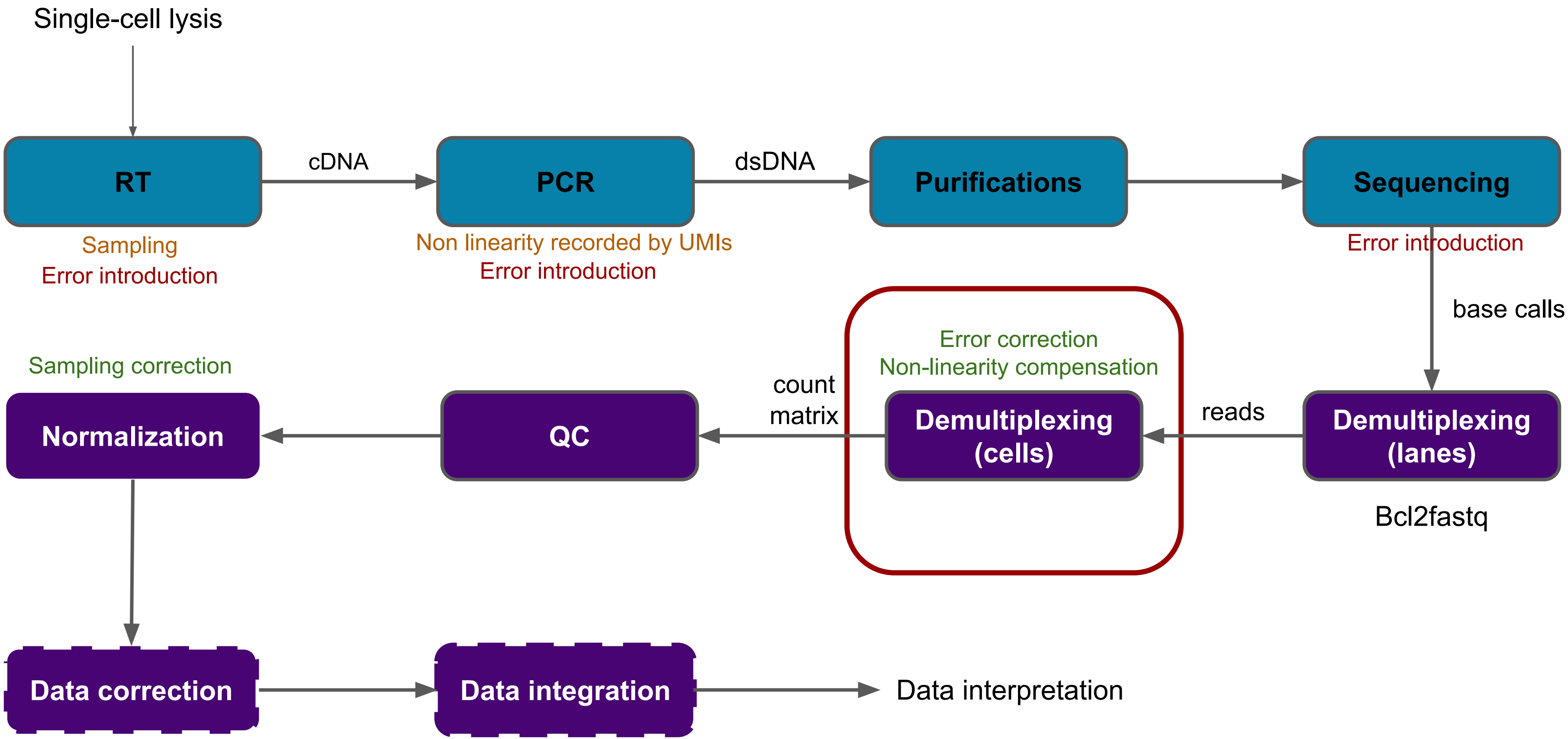
Demultiplexing
from FASTQ files to count matrix

05/11/20 - Gaël Blivet

Different demultiplexing tools with different strategies

- Which one should I use to process unusual -different from 10X- read structure?
- Do these different strategies have an impact on the next stages of the analysis?
- Can it be wise to go back to multiplexing to integrate different analyzes?


Single-cell lysis

RT

cDNA

PCR

dsDNA

Purifications

Sequencing

Sampling

Error introduction

Non linearity recorded by UMIs

Error introduction

Error introduction

base calls

Error correction

Non-linearity compensation

Sampling correction

Normalization

QC

count matrix

Demultiplexing (cells)

reads

Demultiplexing (lanes)

Bcl2fastq

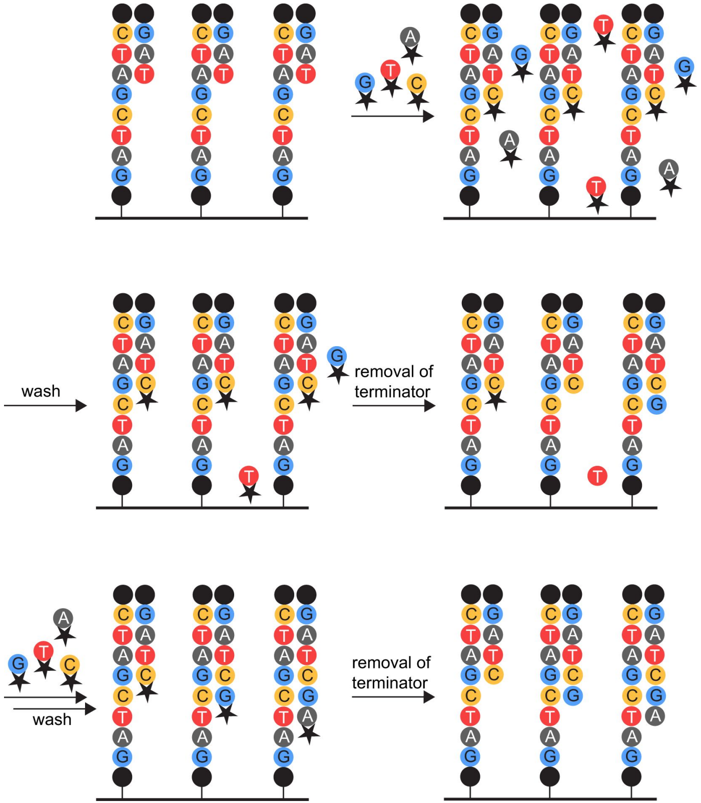
Data correction

Data integration

Data interpretation

What error rate are we talking about?

Step	RT	PCR	Sequencing
Enzyme	Superscript III (MMLV-RT)	Phusion, Q5	?
Error rate / base	~10−5	~10−6 to ~10−7	~10-3 + phasing
References	*Invitrogen; Orton RJ, BMC Genomics. 2015; Potter J, Focus. 2003*	NEB	*Marinier 2015, BMC Bioinformatics; Orton RJ. BMC Genomics. 2015, Tilo Buschmann 2016*



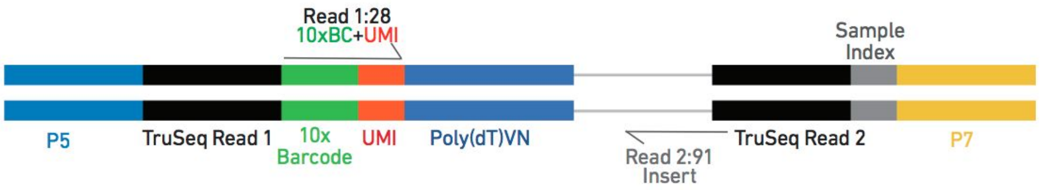
Origin of phasing effects

Pfeiffer, F., Gröber, C., Blank, M. et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep 8, 10950 (2018).

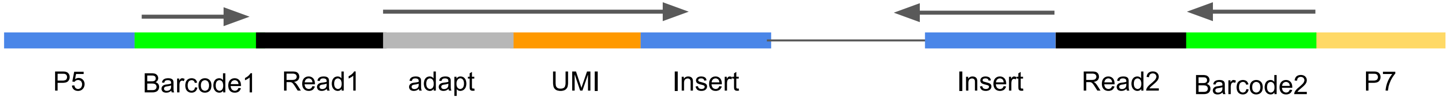
Protocol specific read structure with similar info

- Cell barcode (BC)
- UMI (Unique Molecular identifier)
- Insert (gene)

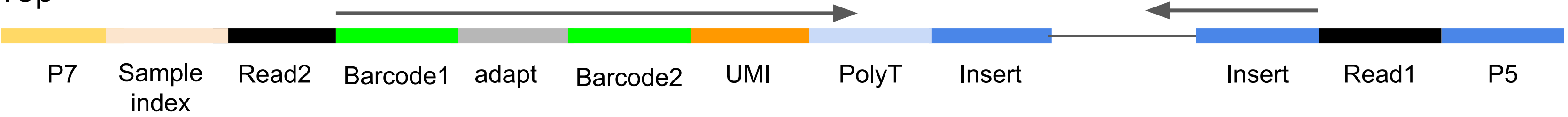
10X



Smart-seq3

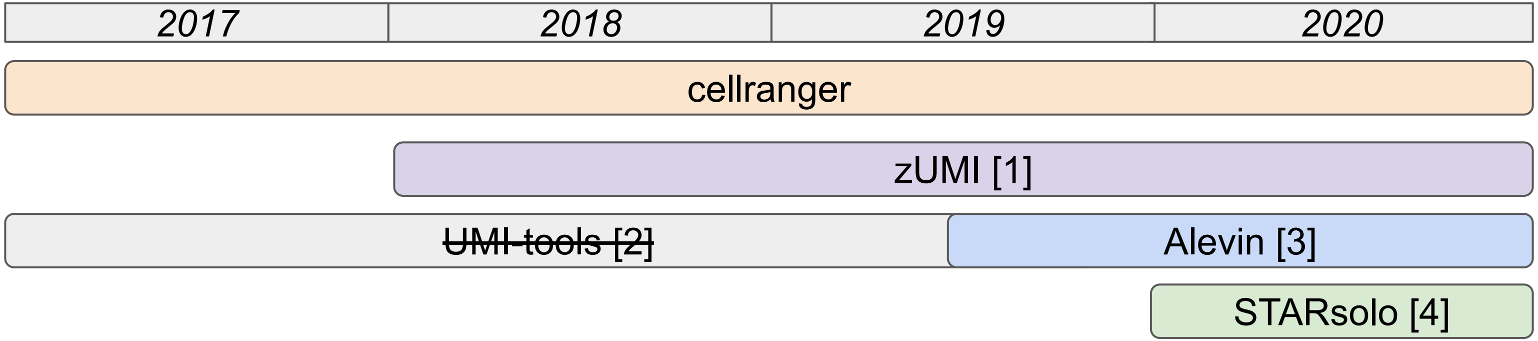


inDrop



Different demultiplexing tools...

- Cellranger is:
  - Not flexible: can’t analyse something else than fastq generated by 10x
  - Slow
  - Resource consuming (memory, space storage?)
  - Black box
- (non-exhaustive) alternatives

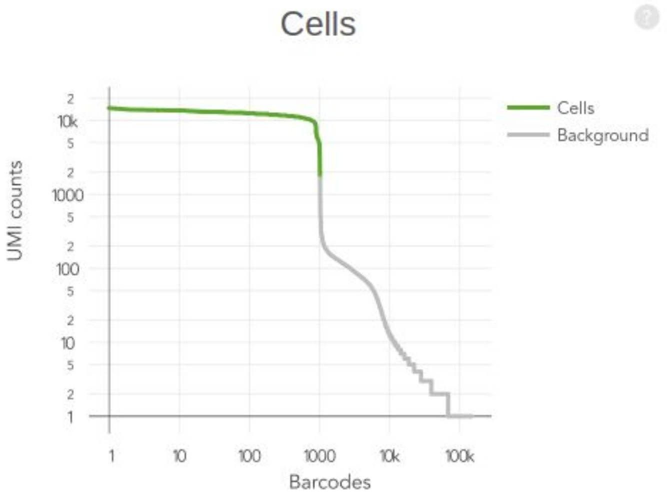


...with different strategies

	Flexible str.	**BC correction**	**UMI correction**	**Mapping**	**Speed est.**
Cellranger	10X compatible	Hamming dist. based	No?	STAR	16h (4 CPUs)
alevin	A bit, more to come	1 error (sub or indel)	Graph-based	Pseudo-alignment	20’ (4 CPUs)
STARsolo	Yes	Hamming dist. based	Various modes	STAR	35’ (3 CPUs)
zUMI	Yes	No?	Hamming dist based	STAR	Fast
LBCpipe	Yes	Seqlev or Hamming dist.	Shift correction	Rsubread	22h (4 CPUs)

Barcode extraction

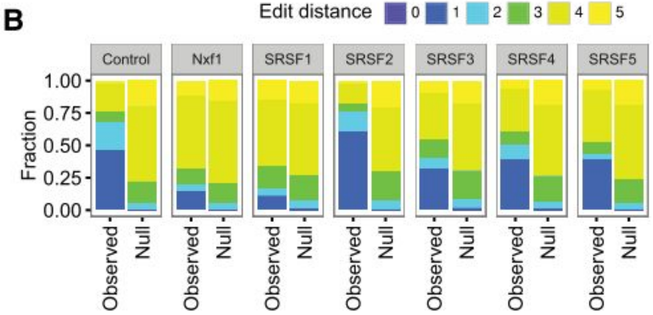
- Pattern flexibility
- Error correction strategies
  - Error type: sub vs indels, shifts
  - Metric: Hamming, Levenshtein, Seqlev, Alevin specific...
  - [LBCpipe] Shift correction
- Barcode database filtering
  - “Knee plot” approach
  - [alevin] second pass with trained classifier



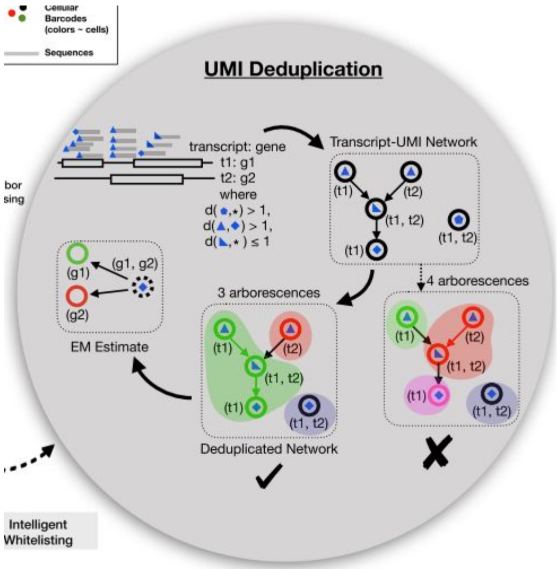
UMI deduplication

- Pattern flexibility
- UMI deduplication strategies
  - No deduplication
  - Hamming distance
  - Graph-based approach
- [alevin] Multimapping strategy (transcript-level)
- [LBCpipe] shift correction

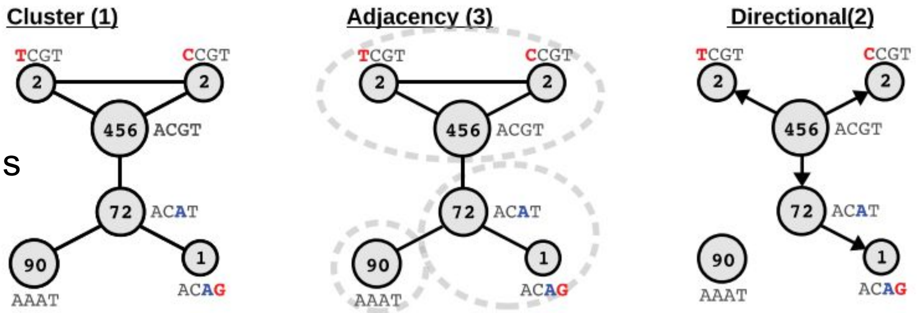
UMI-tools



Alevin



UMI-tools



Insert mapping

- Bulk mapping considerations + 3’ bias
- [alevin] Pseudo-alignment + decoys
- [cellranger] Biotype discard
- Gene/transcript level => impact on UMI correction
- Resources: computing time, memory footprint

Litterature comparison

- Comparison of alevin, cellranger and EmptyDrops: https://lazappi.github.io/phd-thesis/5-analysis.html#pre-processing
- Alevin compares to cellranger in its paper [3]:
  - Better for genes with a lower “sequence uniqueness”
  - Cellranger discards multimappers? (not true anymore since cellranger3?)

Different demultiplexing tools with different strategies

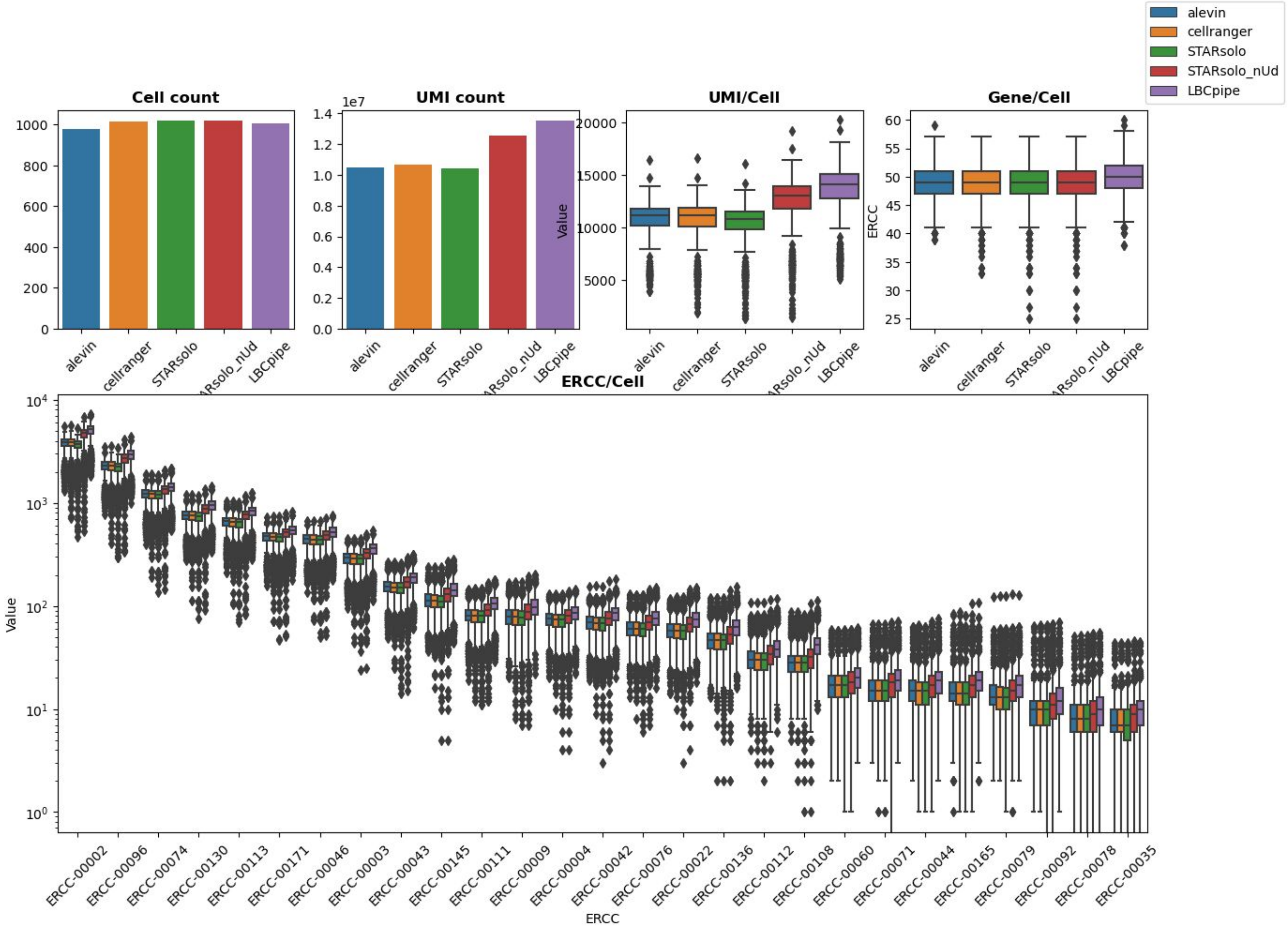
- Which one should I use to process unusual -different from 10X- read structure?
- Do these different strategies have an impact on the next stages of the analysis?
- Can it be wise to go back to multiplexing to integrate different analyzes?

Considered strategy

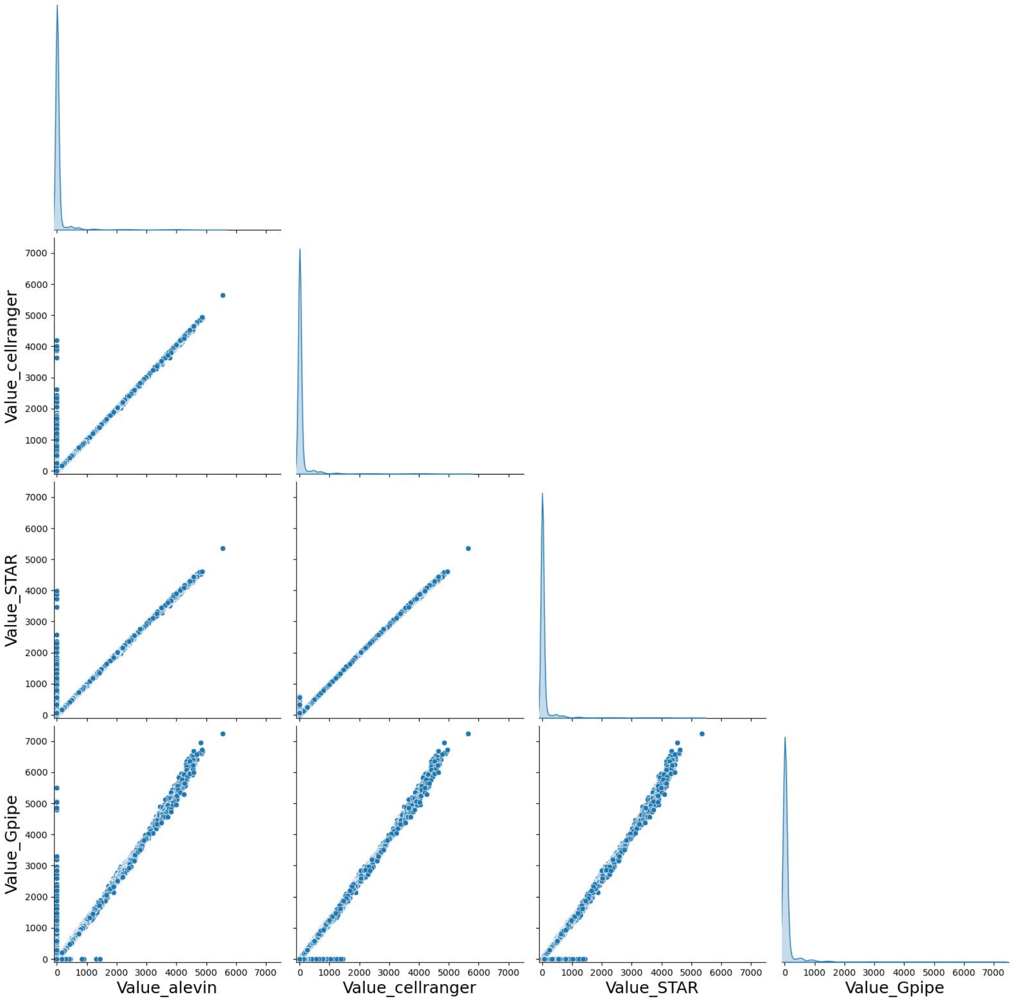
1. Benchmarking on a spike-ins dataset
    - No reference genome issues
    - No biological variations (cell degradation, cell types/states, etc)
    - Only technical variations
    - Dropout exploration
2. Benchmarking on a classical single-cell dataset
    - Reference genome management
3. Benchmarking on a multi-organism single-cell dataset
    - Quality check through secondary analysis
4. Datasets integration
    - From matrices vs from FASTQ
    - Quality check through secondary analysis

Dataset 1

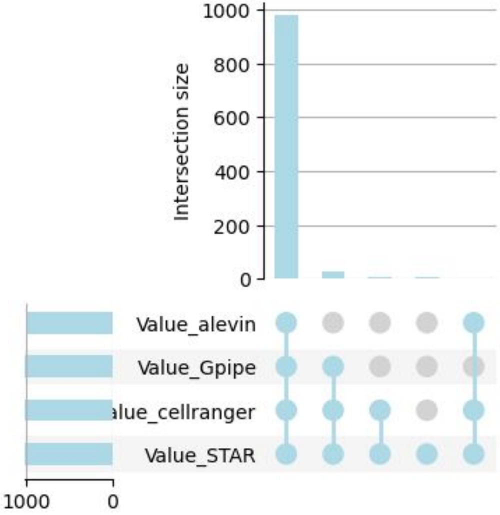
- 10X ERCC 1k cells
  - Massively parallel digital transcriptional profiling of single cells, Zheng *et al.* 2017, Nature communications
  - Listed in 10X datasets
  - 10X v2 chemistry
  - 14bp BC + 10bp UMI
  - 1k droplets
  - 92 ERCCs (spike-ins) in various expected quantities
- No reference genome compatibility issues


alevin
cellranger
STARsolo
STARsolo_nUd
LBCpipe
Cell count
1000
800
600
400
200
0
alevin
cellranger
STARsolo
STARsolo_nUd
LBCpipe
UMI count
1e7
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
alevin
cellranger
STARsolo
STARsolo_nUd
LBCpipe
UMI/Cell
Value
20000
15000
10000
5000
alevin
cellranger
STARsolo
STARsolo_nUd
LBCpipe
Gene/Cell
ERCC
60
55
50
45
40
35
30
25
alevin
cellranger
STARsolo
STARsolo_nUd
LBCpipe
ERCC/Cell
Value
10^4
10^3
10^2
10^1
10^0
ERCC-00002
ERCC-00096
ERCC-00074
ERCC-00130
ERCC-00113
ERCC-00171
ERCC-00046
ERCC-00003
ERCC-00043
ERCC-00145
ERCC-00111
ERCC-00009
ERCC-00004
ERCC-00042
ERCC-00076
ERCC-00022
ERCC-00136
ERCC-00112
ERCC-00108
ERCC-00060
ERCC-00071
ERCC-00044
ERCC-00165
ERCC-00079
ERCC-00092
ERCC-00078
ERCC-00035
ERCC

Counts correlation

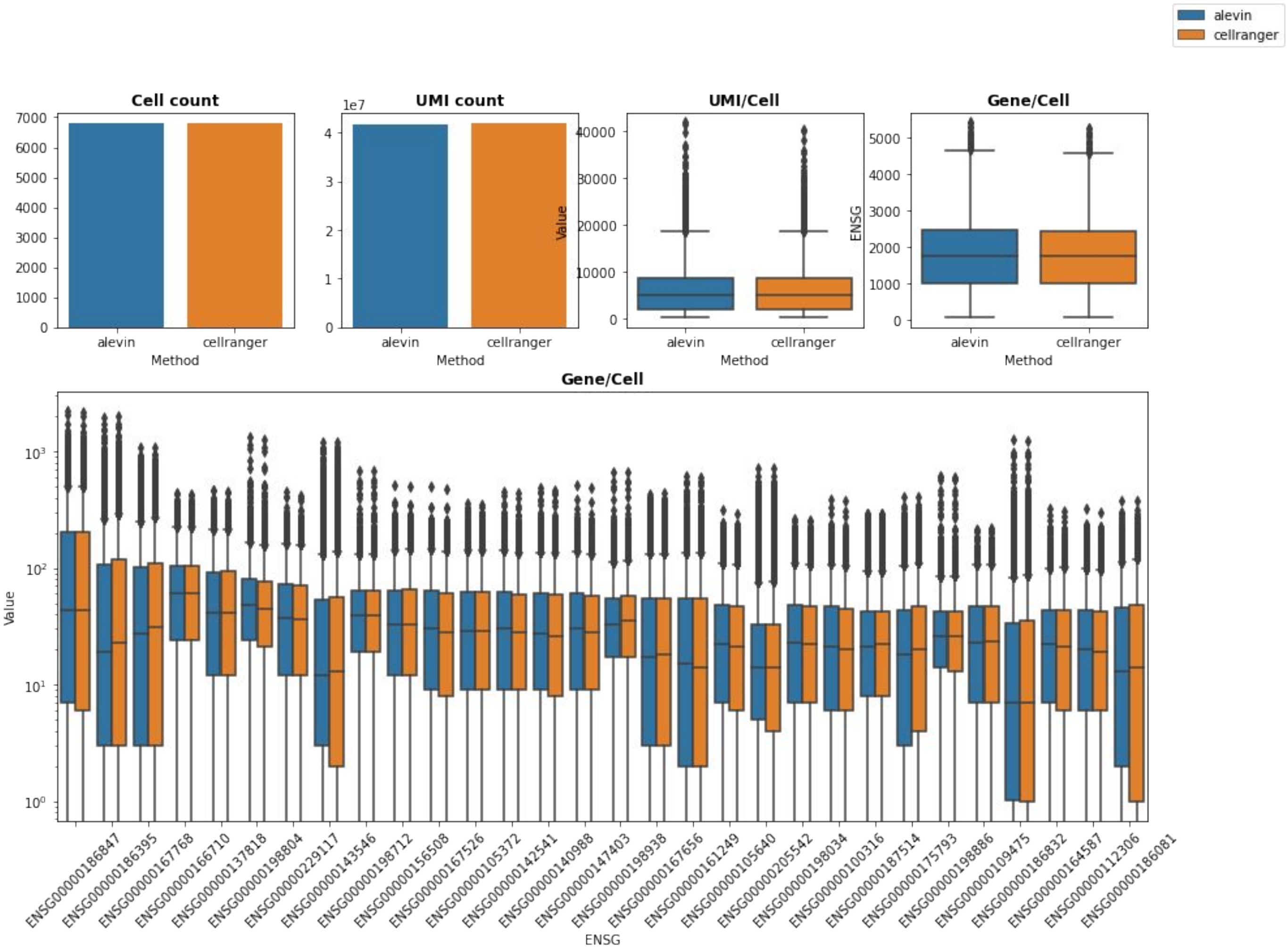


Intersection of detected cells



Dataset 2

- Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma, Ji *et al.*, Cell 2020
  - One “Normal” skin sample, ~320M reads
  - Chromium v2 chemistry
  - 16 bp BC + 10bp UMI
- Reference genome issues
  - Built from Ensembl 101
  - Considering gene IDs (not names)


alevin
cellranger
Cell count
7000
6000
5000
4000
3000
2000
1000
0
alevin
cellranger
Method
UMI count
1e7
4
3
2
1
0
alevin
cellranger
Method
UMI/Cell
Value
40000
30000
20000
10000
0
alevin
cellranger
Method
Gene/Cell
ENSG
5000
4000
3000
2000
1000
0
alevin
cellranger
Method
Gene/Cell
Value
10^3
10^2
10^1
10^0
ENSG00000186847
ENSG00000186395
ENSG00000167768
ENSG00000166710
ENSG00000137818
ENSG00000198804
ENSG00000229117
ENSG00000143546
ENSG00000198712
ENSG00000156508
ENSG00000167526
ENSG00000105372
ENSG00000142541
ENSG00000140988
ENSG00000147403
ENSG00000198938
ENSG00000167656
ENSG00000161249
ENSG00000105640
ENSG00000205542
ENSG00000198034
ENSG00000100316
ENSG00000187514
ENSG00000175793
ENSG00000198886
ENSG00000109475
ENSG00000186832
ENSG00000164587
ENSG00000112306
ENSG00000186081
ENSG

First impressions

- Cellranger can be replaced without results quality loss
  - to go faster
  - to handle unusual read structure
  - STARsolo seems to be the best candidates for these 2 features
- Differences between common demultiplexing tools seem to be marginal though different strategies are applied
- Need to investigate relevancy of shift correction and UMI correction

Perspectives

- Study the impact of these different strategies on secondary analysis
  - On dataset2
  - On a reference dataset from “Benchmarking single-cell RNA-sequencing protocols for cell atlas projects”, Mereu *et al.*, Nature biotechnology 2020
    - ~3k cells
    - A lots of protocols (n=13: 10X, Quartz-seq3, plate-based, etc)
    - Control mix of human, murine and canine cells
- Datasets integration
  - From matrices vs from FASTQ
  - Quality check through secondary analysis

Tools references

- [1] zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs, Parekh *et al.*, GigaScience 2018
- [2] UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy, Smith *et al.*, Genome Research 2017
- [3] Alevin efficiently estimates accurate gene abundances from dscRNA-seq data, Srivastava *et al.*, Genome Biology 2019
- [4] https://github.com/alexdobin/STAR/blob/master/docs/STARsolo.md, Dobin, 2020