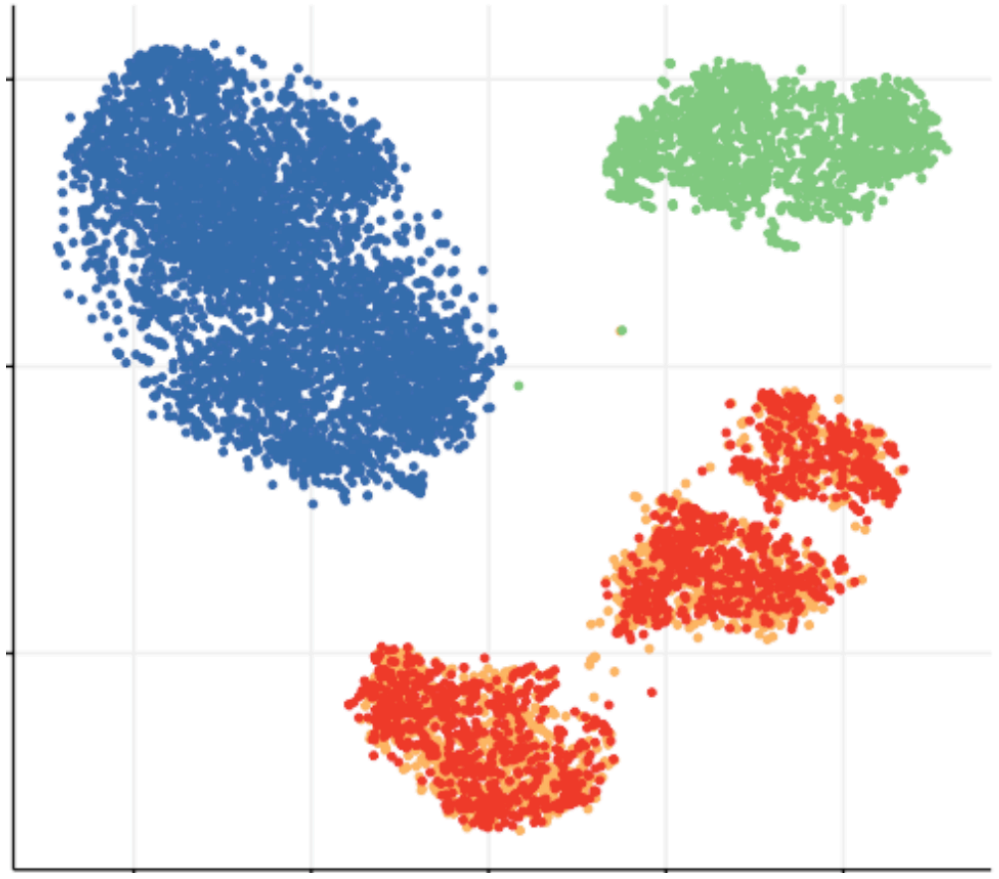
Introduction to single cell RNA sequencing

CRUK Bioinformatics Summer School 2018

Mike Morgan
Comp Bio Postdoc
Marioni Group



wellcome sanger institute

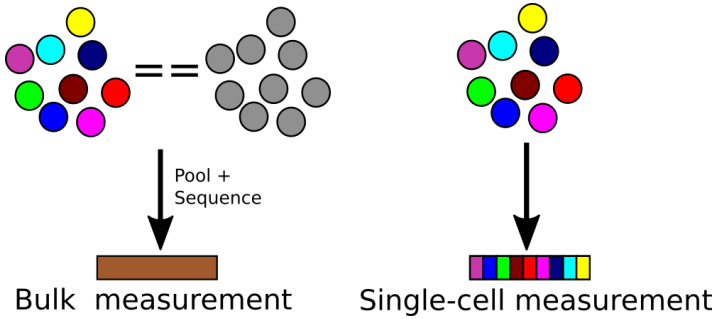
CANCER RESEARCH UK

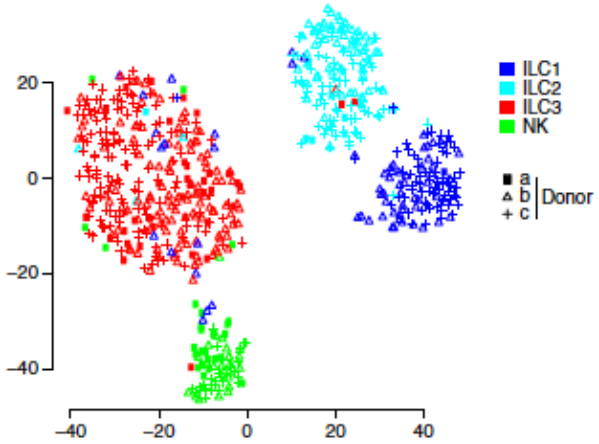
Why study single cells?

Unravel tissue heterogeneity:
- Novel and rare cell types
- Unknown cellular states

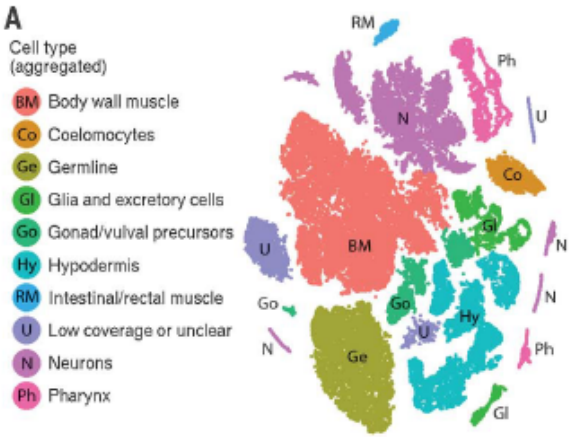
Transcriptional dynamics

Can also measure single-cell:
- Chromatin accessibility
- Mutation & CNV (scDNA-seq)
- Methylation

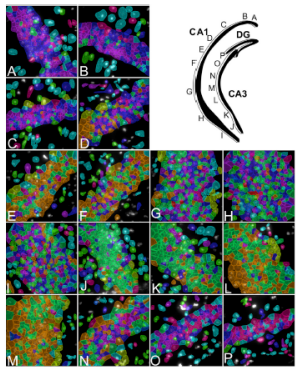




Innate-lymphoid cells
Bjorklund *et al.*, Nature Immunology (2016)



Whole *C. elegans* larva
Cao *et al.*, Science (2017)



Mouse hippocampus
Shah *et al.*, Neuron (2017)

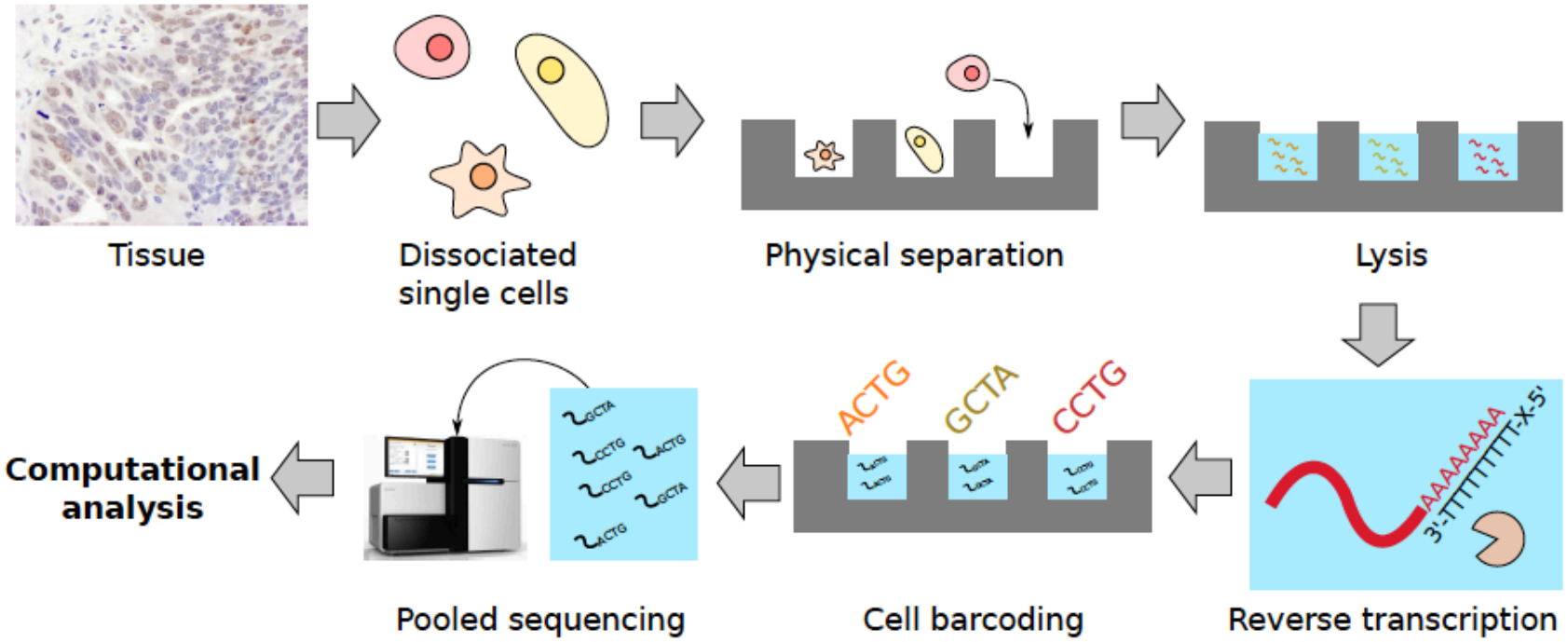
How can we study single cells?

Technology	Measurements (P)	Cells (N)	Throughput	Pro	Con
Flow cytometry	1-15	1k-100k	big N, small P	Technically easy	Limited targets
Mass cytometry	20-50	1k-100k	big N, medium P	>P than flow	Limited targets
RNA FISH	1	~100	small N, small P	Spatial resolution	Technically hard, low throughput
Multiplex FISH	~100	100's	medium N, medium P	Spatial resolution	Technically and analytically hard
SS2 scRNA-seq	~20,000	100-1000	medium N, big P	High throughput	Sparse, low input material
Droplet scRNA-seq	~20,000	100-1M	big N, big P	High throughput	*Very* sparse, low input material

NB – every method has it's pros and cons. There is no all-encompassing single cell methodology.

It depends on your biological question!

A typical scRNA-seq experiment

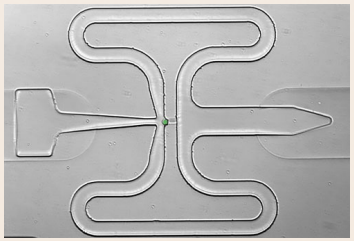


- Dissociation can be easy (blood) or hard (collagenous tissue)
- Separation and RT differ by protocol

Image courtesy of Aaron Lun

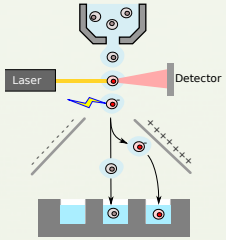
Physical separation defines main scRNA-seq protocols

Microfluidic device



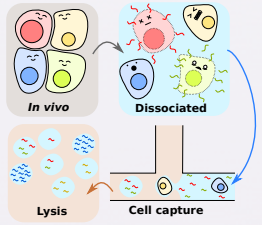
- 96 or 800 well format
- Physically check presence of cells
- High capture efficiency
- Doublet issues
- Expensive
- Full-length cDNA (SMART-seq{2})
- Spike-in control RNA
- **High gene coverage**

Plate-based



- 96 or 384 well format
- Sort specific population(s) of cells
- High capture efficiency
- Experimental design considerations
- Full-length cDNA (SMART-seq(2) or end-tagging; UMIs)
- Spike-in control RNA
- **High gene coverage**

Droplet-based



- 100-1000's of cells
- Doublet issues
- Variable capture efficiency
- Low per-cell cost
- 3' end tag; UMIs
- No spike-in control RNA
- **High cell coverage**

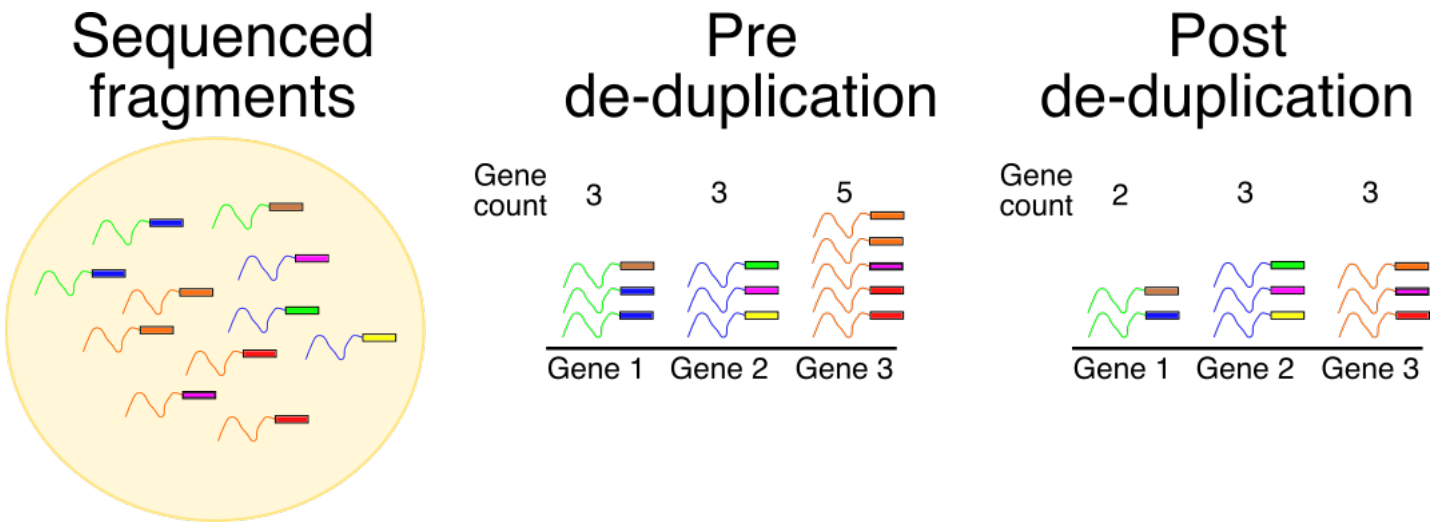
wellcome sanger institute | CANCER RESEARCH UK

Images courtesy of Aaron Lun

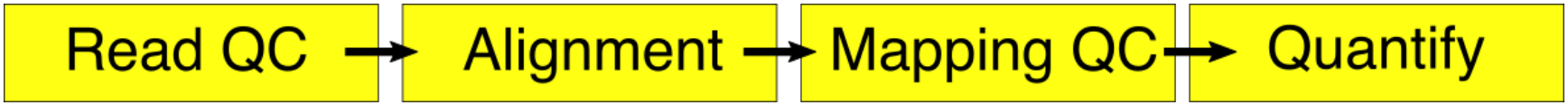
What are UMIs?

Unique molecular identifiers give (almost) exact molecule counts in sequencing experiments.

They reduce the amplification noise by allowing (almost) complete de-duplication of sequenced fragments.



A typical SMART-seq workflow

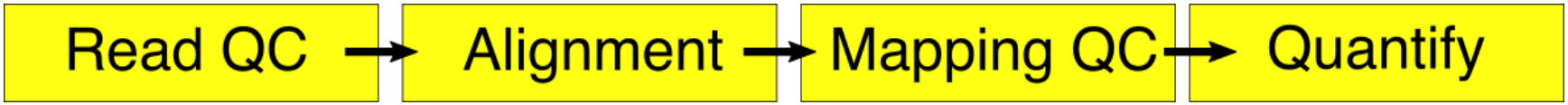


The same tools used for bulk RNA-seq, e.g. FastQC, Star, PicardTools
(Deduplication is essential)

Typically 1 library per cell, potentially many 100's of FASTQ files
Need to be able to handle many files in parallel – e.g. high performance computing cluster.

Pipelining tools exist (beyond the scope of this tutorial – see resources).

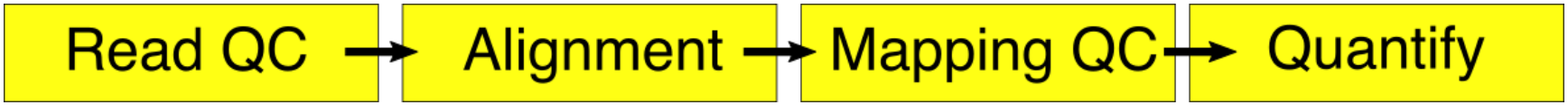
A typical SMART-seq workflow



The same tools used for bulk RNA-seq, e.g. FastQC, Star, PicardTools
(Deduplication is essential)

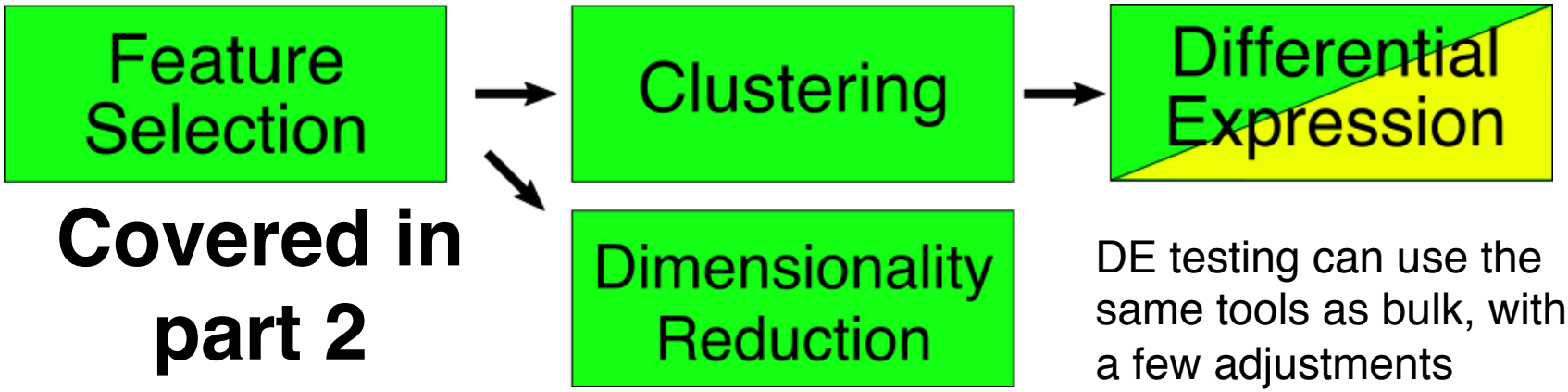
Single-cell specific tools (generally performed in R; Practical 1)

A typical SMART-seq workflow



The same tools used for bulk RNA-seq, e.g. FastQC, Star, PicardTools
(Deduplication is essential)

Single-cell specific tools (generally performed in R; Practical 1)



Covered in part 2

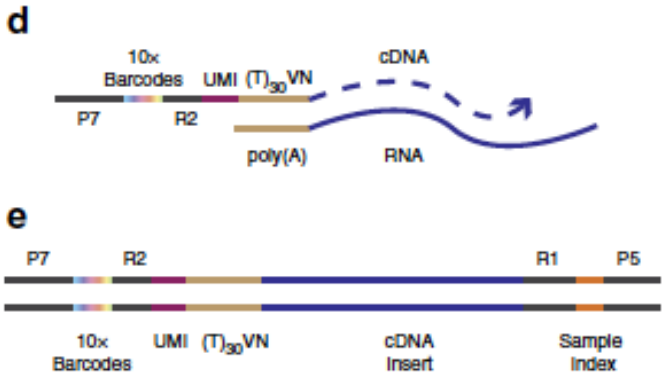
DE testing can use the same tools as bulk, with a few adjustments

A typical droplet workflow

Droplet-based methods create a new problem, and solution:

- Many 100's-1000's cells == 1000's small FASTQ files
- Prohibitively expensive to sequence 20,000 cells to 1M reads

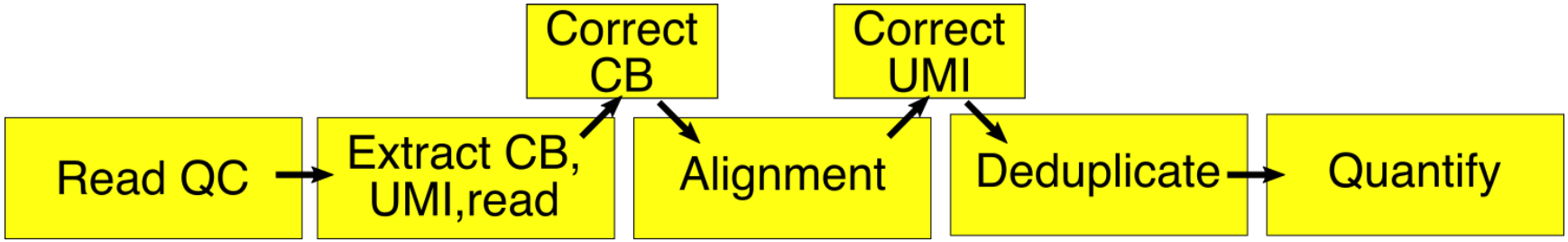
Solution: multiplex cells using barcodes



10X Genomics Chromium v1 chemistry design
Zheng *et al.,* Nature Comms (2017)

A single 10X Genomics Chromium library generates 3 FASTQ files: R1, R2, Index

A typical droplet workflow

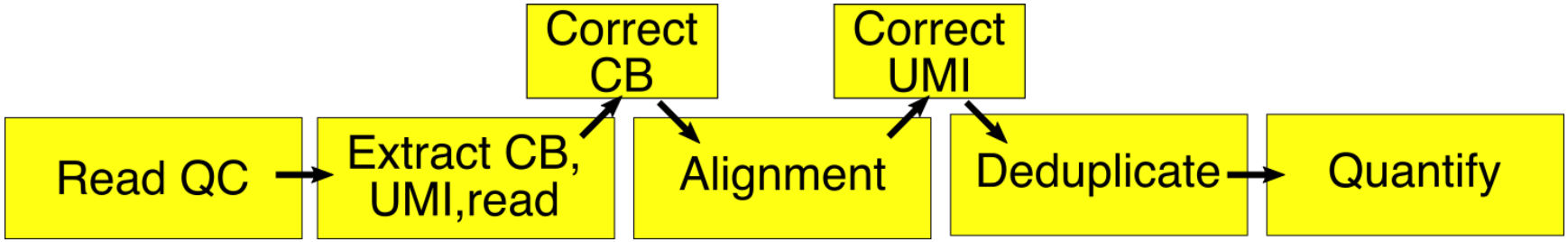


Generally run in a single pipeline, e.g. Cellranger (10X specific), DropSeq (Macosko *et al.*) or custom (not recommended if just starting).

Sequencing errors in cell barcodes and UMIs are a source of technical noise – must be dealt with

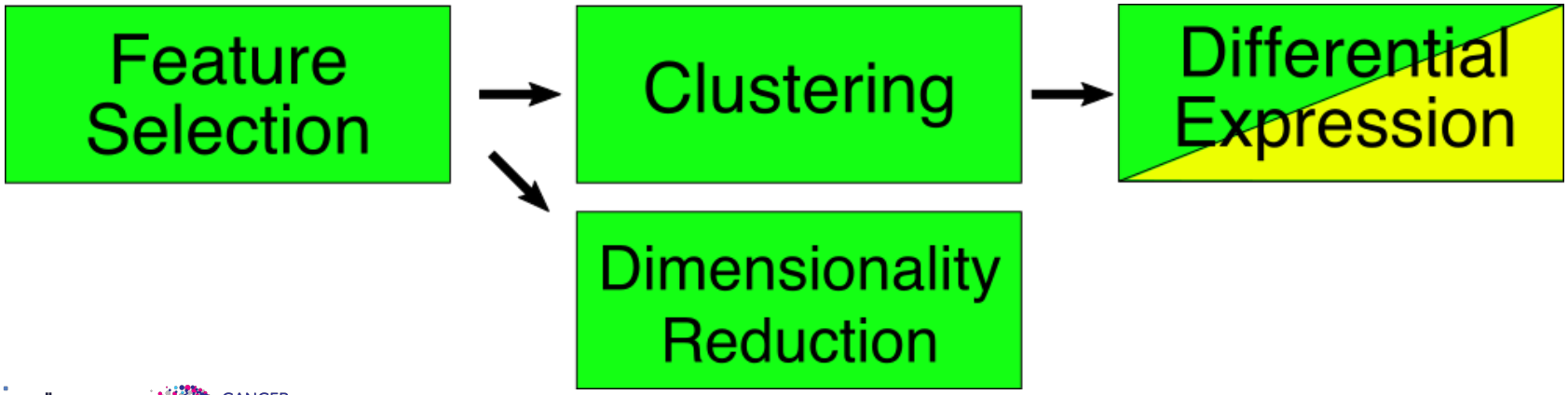
Recent development: Rob Patro & co have a new end-to-end (i.e. FASTQ to counts matrix) lightweight pipeline:
https://salmon.readthedocs.io/en/latest/alevin.html

A typical droplet workflow



Generally run in a single pipeline, e.g. Cellranger (10X specific), DropSeq (Macosko *et al.*) or custom (not recommended if just starting).

Single-cell specific tools (generally performed in R; Practical 1)



Dealing with single cells

Regardless of technology, our goal is to derive/extract real biology from technically noisy data.

Single cell analysis workflow

Starting with a counts matrix:

Cell	A	B	C	D
Gene X	10	20	30	40
Gene Y	15	30	45	60
Gene Z	20	40	60	80
	…	…	…	…

- Quality control
- Normalization
- Batch correction [if required]
- Dimensionality reduction and visualisation (part 2)
- Clustering (part 2)
- Differential expression testing (same as bulk RNA seq… mostly)

Quality control on cells

- Low sequencing depth
- Low numbers of expressed genes (i.e. any non-zero count)
- High spike-in (if present) or mitochondial content

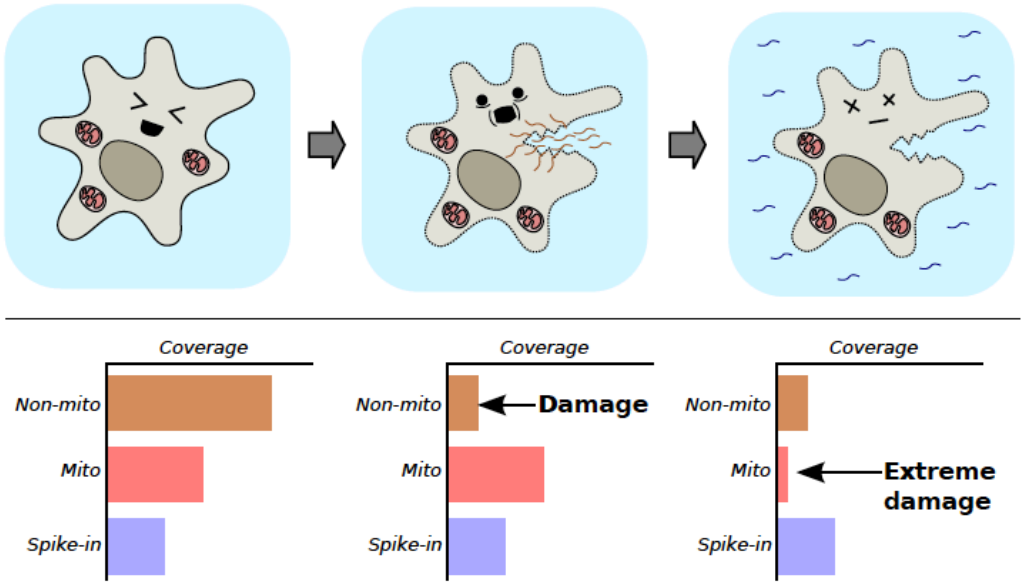


Image courtesy of Aaron Lun

Normalization

- The aim is bring all cells onto the same *distribution* to remove biases between them
- We want to preserve biological variability, not introduce new technical variation
- Primary source of bias is sequencing depth – scale down counts accordingly
- Need a method that is robust to sparsity and composition bias
  - TMM & DESeq size factors are not!

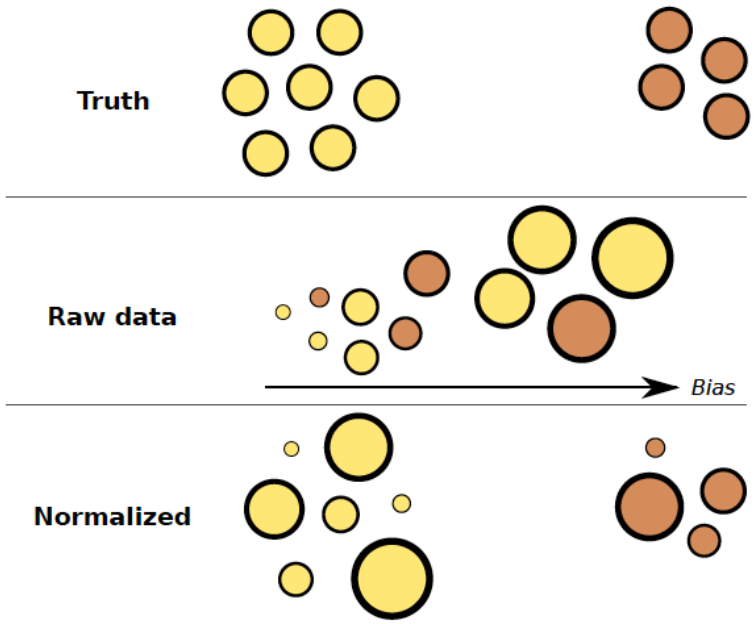
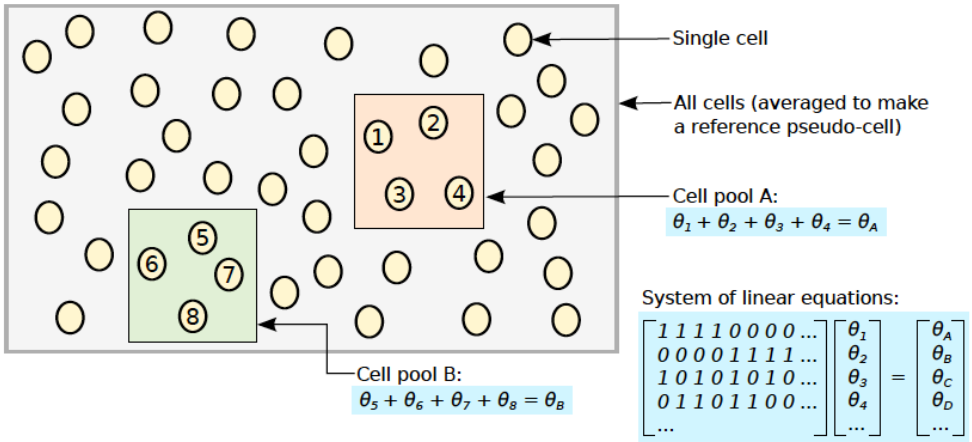


Image courtesy of Aaron Lun

Normalization by deconvolution

- Estimate cell-specific size factors.
- Handles sparsity and is robust to DE.



Lun *et al.,* Genome Biology (2016)

1. Cluster cells together
2. Pool cells to increase counts, reduce 0's
3. Robust estimate of each pool size factor
4. Wash & repeat for multiple pools
5. Solve the linear system of equations to obtain *per-cell* size factors

Image courtesy of Aaron Lun

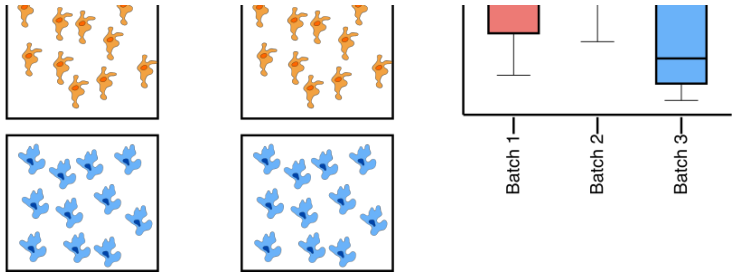
Confounders and batch correction

- A segue into proper experimental design
- Some batch effects cannot be avoided
- Some can, make sure you know which is which

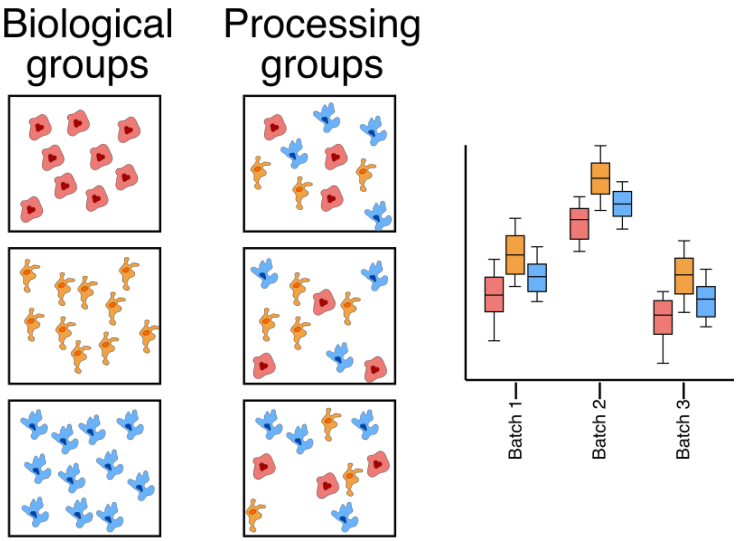
Confounded design

Biological groups | Processing groups

Please don't design your experiment like this!!!



Not confounded design



Adapted from Hicks *et al.*, bioRxiv (2015)

What if I still have batch effects?

Good experimental design doesn't remove batch effects, it prevents them from biasing your results (hopefully)

If you still have batch effects then they can be dealt with (if necessary) <- important for clustering and visualization

Simple batch correction

If you have a single cell type and multiple conditions:

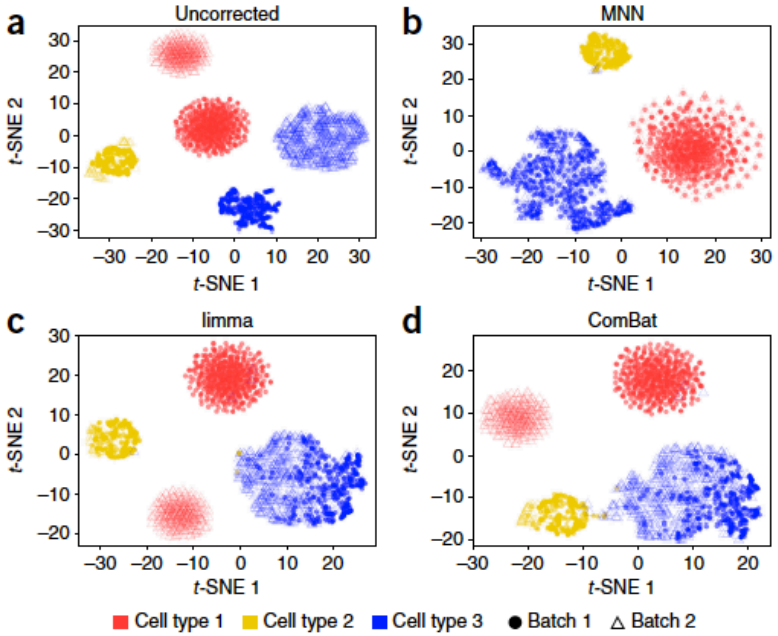
Use a linear model to regress gene expression on batch

More complex batch correction

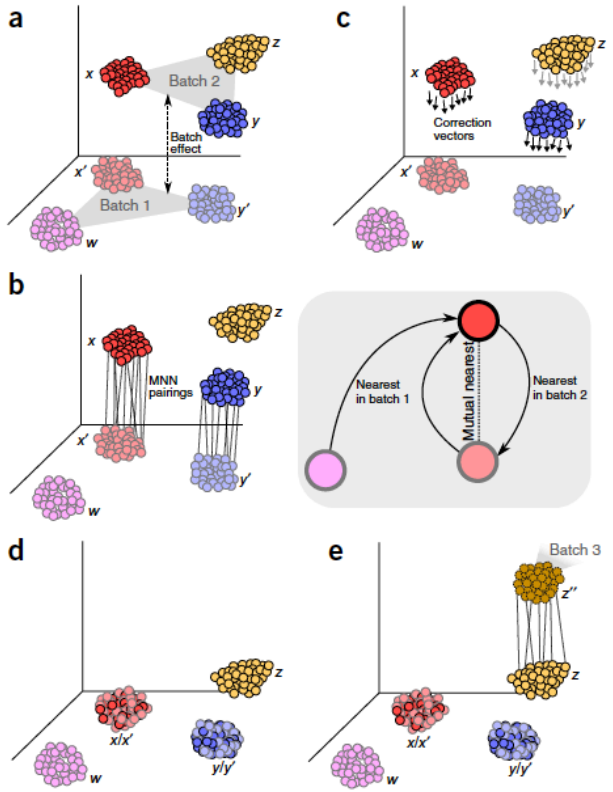
Linear models (and bulk batch correction methods) can't handle composition differences between batches.

Need a method that handles multiple batches, i.e. > 2, and corrects expression values properly

Match cells between batches that share the same biological subspace, remove the orthogonal components (mnnCorrect).



Haghverdi *et al.*, Nature Biotech (2018)



Resources

Single Cell Resources:

Single cell course (Hemberg Lab; Wellcome Sanger Institute):
http://hemberg-lab.github.io/scRNA.seq.course/index.html

Aaron Lun's single cell workflow (very detailed):
https://www.bioconductor.org/packages/release/workflows/html/simpleSingleCell.html

Cellranger pipeline:
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

Resources

Workflow Resources:

Snakemake (Python):
http://snakemake.readthedocs.io/en/stable/#

Nextflow (Java/agnostic):
https://www.nextflow.io

Ruffus (Python):
http://www.ruffus.org.uk

make (bash):
https://www.tutorialspoint.com/unix_commands/make.htm

Recommended reading

Study design

Hicks *et al.,* bioRxiv (2015):

https://www.biorxiv.org/content/biorxiv/early/2015/08/25/025528.full.pdf

Batch correction:

Haghverdi *et al*, Nature Biotech (2018):

https://www.nature.com/articles/nbt.4091

Butler *et al*., Nature Biotech (2018):

https://www.nature.com/articles/nbt.4096