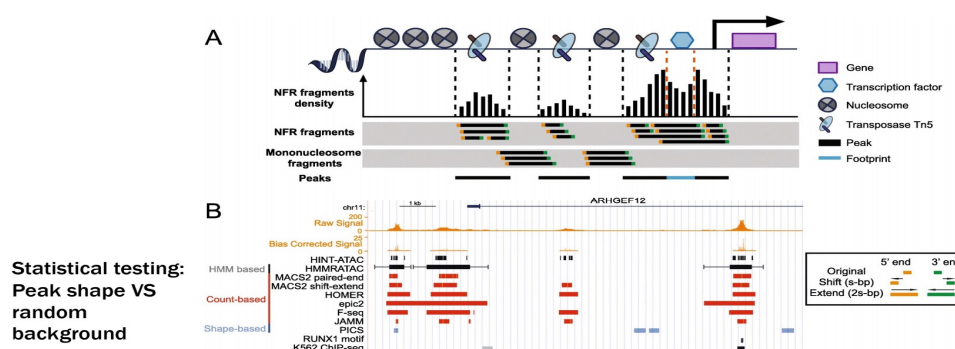## Single-Cell RNA-seq

Epigenetics

- Gene expression regulation: structure and environment
- Data analytics pipeline

Single-cell RNA-seq

- Why single cell?
- What is single-cell RNA-seq?
- How to get the single-cell RNA-seq data?

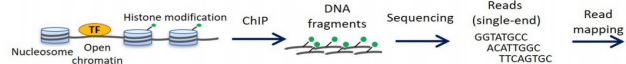Single-cell RNA-seq data analytics

Epigenetics – Peak Calling

Statistical testing:
Peak shape VS random background

Peak Calling Output – BED file

Browser Extensible Data (BED) format

- Chromosome
- Start
- End
- Label
- ...

```
track name="ItemRGBDemo" description="Item RGB demonstration" visibility=2 itemRgb="On"
chr7    127471196    127472363    Pos1    0    +    127471196    127472363    255,0,0
chr7    127472363    127473530    Pos2    0    +    127472363    127473530    255,0,0
chr7    127473530    127474697    Pos3    0    +    127473530    127474697    255,0,0
chr7    127474697    127475864    Pos4    0    +    127474697    127475864    255,0,0
chr7    127475864    127477031    Neg1    0    −    127475864    127477031    0,0,255
chr7    127477031    127478198    Neg2    0    −    127477031    127478198    0,0,255
chr7    127478198    127479365    Neg3    0    −    127478198    127479365    0,0,255
chr7    127479365    127480532    Pos5    0    +    127479365    127480532    255,0,0
chr7    127480532    127481699    Neg4    0    −    127480532    127481699    0,0,255
```
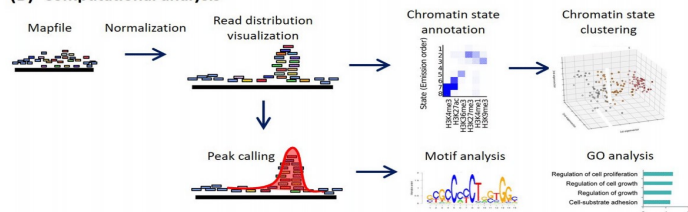
The Overall Data Analytics Pipeline for Epigenetics

Histone Marks and Chromatin Accessibility
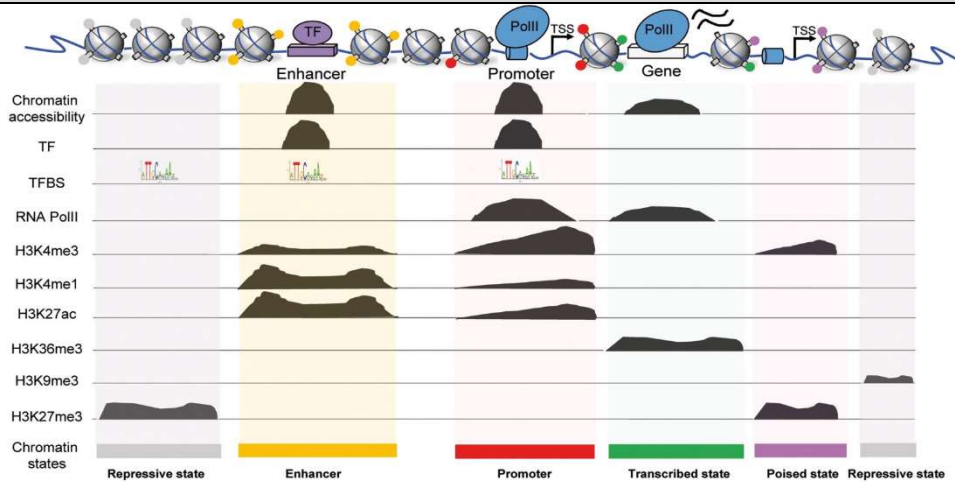
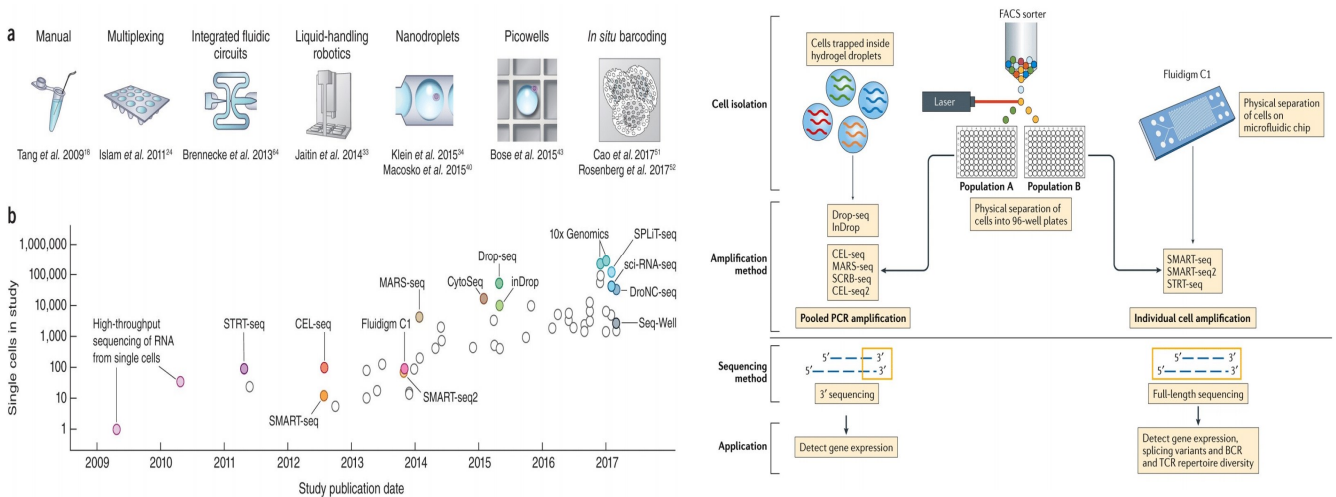## Why Single-Cell Analysis?
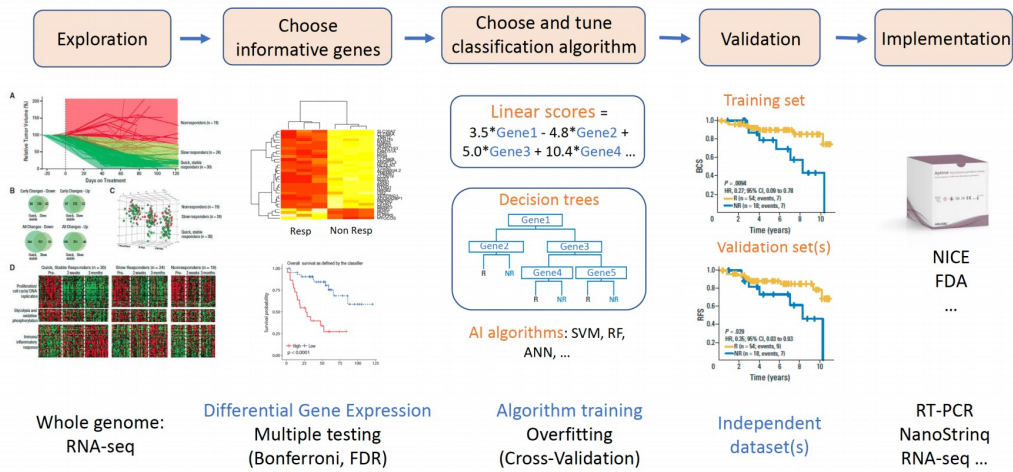
- **Tumour Micro-Environment:**
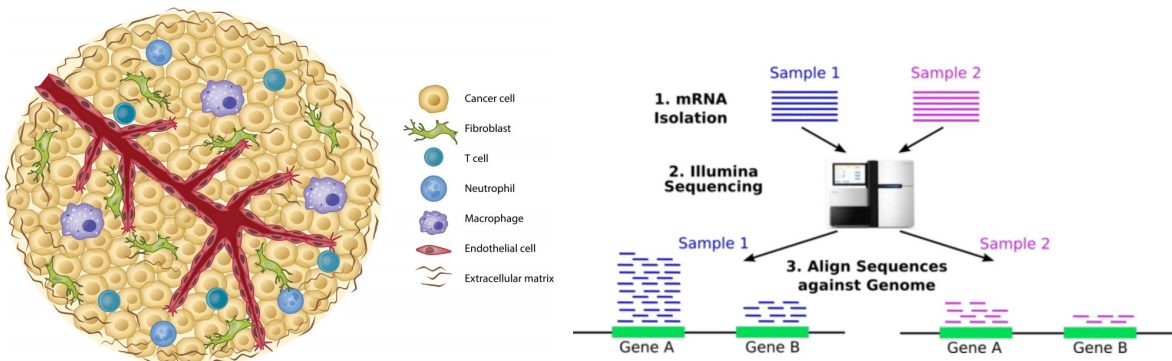
What is Single-Cell Analysis?

Single cell sequencing examines the sequence information from individual cells with optimized next-generation sequencing (NGS) technologies, providing a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment.

What can Single-Cell Sequencing do?

- Understanding heterogeneous tissues
- Identification and analysis of rare cell types
- Changes in cellular composition
- Dissection of temporal changes
- Example of applications:
    - Differentiation Paths
    - Cancer Heterogeneity
    - Neural Cell Classification
   - Embryonic Development
    - Drug Treatment Response

How to do Single-Cell Sequencing?

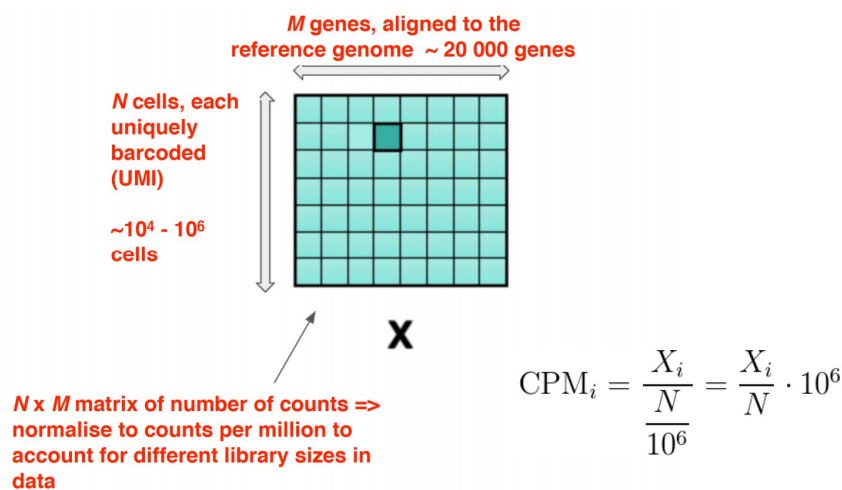| Paplexi et al. 2017	FACS ✖	CyTOF ✖	qPCR ✖	Plate-based protocols (STRT-seq, SMART-seq, SMART-seq2) ⬇	Fluidigm C1 ✖	Pooled approaches (CEL-seq, MARS-seq, SCRB-seq, CEL-seq2) ⬇	Massively parallel approaches (Drop-seq, InDrop) ⬇
Cell capture method	Laser	Mass cytometry	Micropipettes	FACS	Microfluidics	FACS	Microdroplets
Number of cells per experiment	Millions	Millions	300–1,000	50–500 ⭐	48–96 ✖	500–2,000 ⭐	5,000–10,000 ⭐
Cost	$0.05 per cell	$35 per cell	$1 per cell	$3–6 per well	$35 per cell	$3–6 per well	$0.05 per cell
Sensitivity	Up to 17 markers ✖	Up to 40 markers ✖	10–30 genes per cell ✖	7,000–10,000 genes per cell for cell lines; 2,000–6,000 genes per cell for primary cells	6,000–9,00 genes per cell for cell lines; 1,000–5,000 genes per cell for primary cells	7,000–10,000 genes per cell for cell lines; 1,000–6,000 genes per cell for primary cells	5,000 genes per cell for cell lines; 1,000–3,000 genes per cell for primary cells
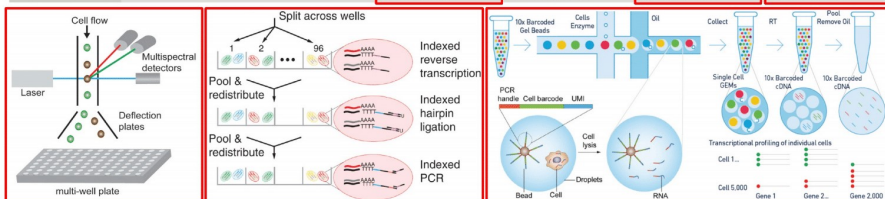
Single-Cell Data Analytics
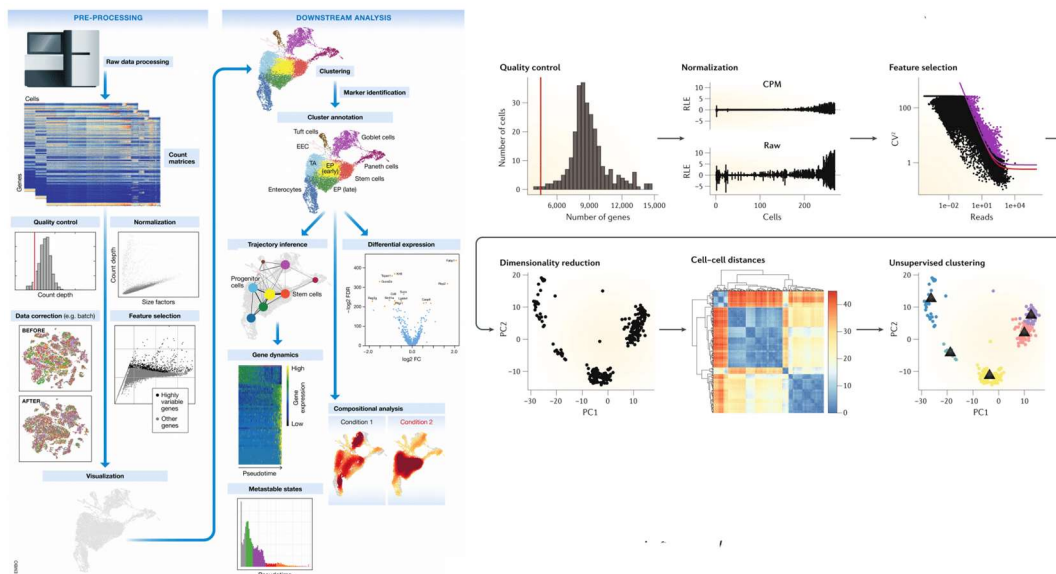

Challenges in Single-Cell Data Analytics

- Noise
- Doublet
- Dropout
- Batch Effect

The Gene Expression Matrix

M genes, aligned to the reference genome ~ 20 000 genes

N cells, each uniquely barcoded (UMI)

~10^4 - 10^6 cells

N x M matrix of number of counts => normalise to counts per million to account for different library sizes in data

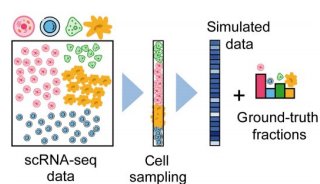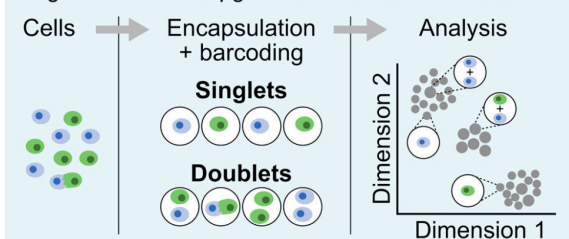$$\mathrm{CPM}_i = \frac{X_i}{\frac{N}{10^6}} = \frac{X_i}{N} \cdot 10^6$$

Noise in the Gene Expression Matrix

- The number of genes expressed in the count matrix
- The total counts per cell
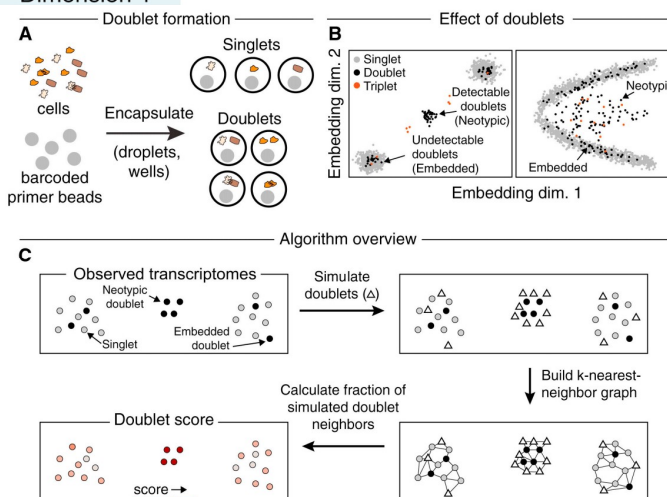- The percentage of counts in mitochondrial genes
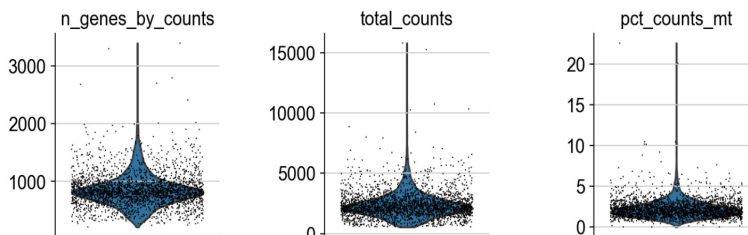
## Doublets

Single-cell RNA-seq generates doublet artifacts

Identify doublet data points using the simulated doublet data points (similarity)

Dropout

- A dropout event occurs when a transcript is expressed in a cell but is entirely undetected in its mRNA profile
- Dropout events occurs due to low amounts of mRNA in individual cells
- Trade-off: given the same budget, more cells, more dropouts

- Missing values
  - Advanced statistical/ML methods
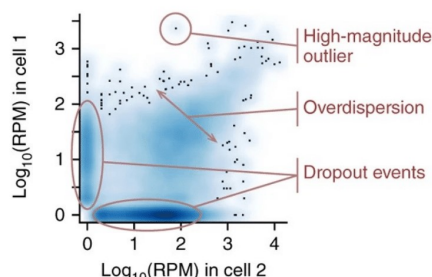
Batch Effect

In molecular biology, a batch effect occurs when non-biological factors in an experiment cause changes in the data produced by the experiment.

Some Approaches to Batch Correction:

- Normalization
- Alignment
- Removing genes correlated with batch
- Regression of residuals with technical covariates
  - Statistical methods
- Latent space representations
  - Machine learning methods