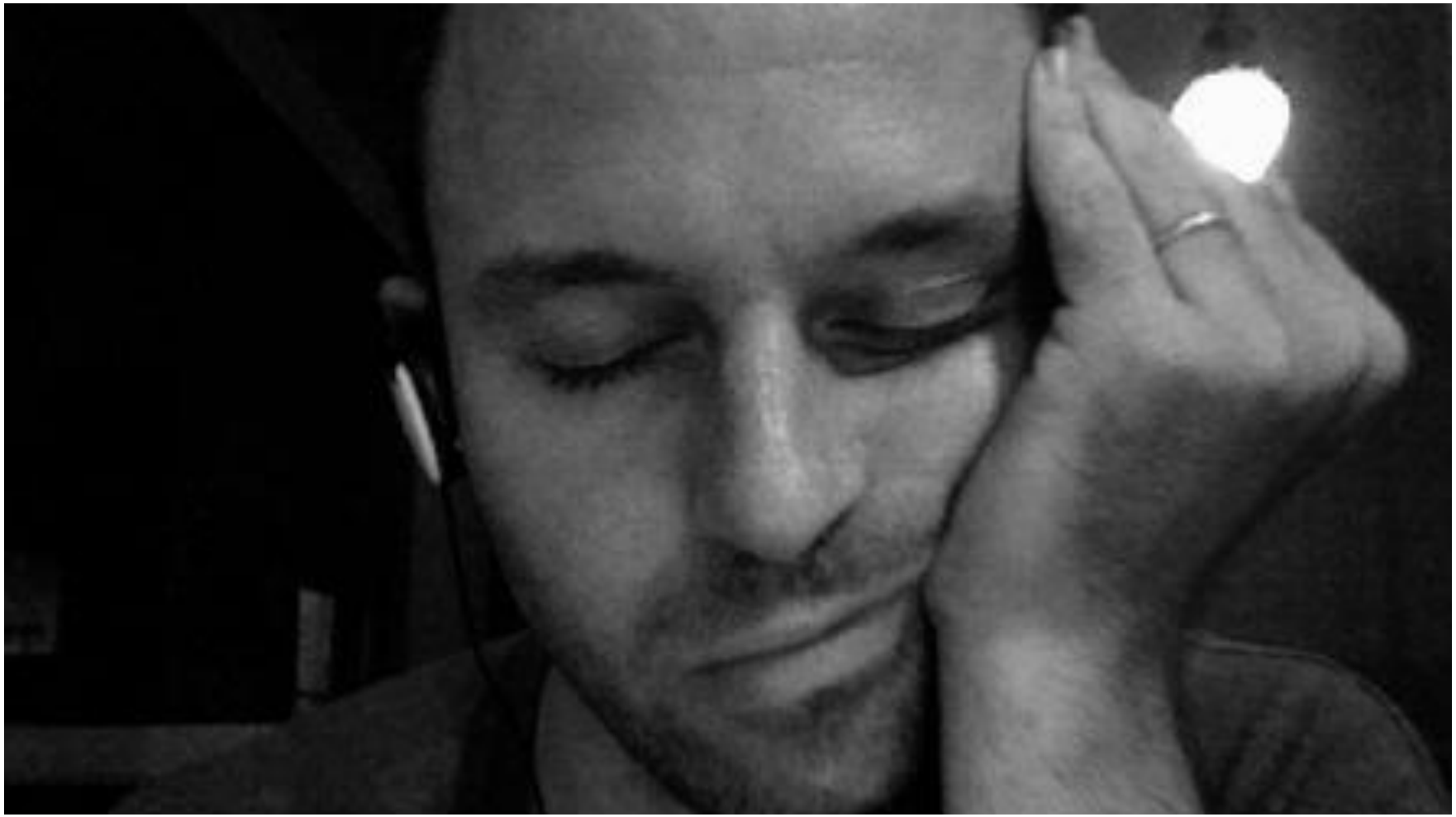
scRNA-seq experimental design and raw data processing

Jenni Westoby

How long can the average human concentrate on someone speaking for?



Lance Shields

What is single-cell RNA-seq?

- A new technology, first published by Tang et al. 2009.
- Became popular in 2014 when new protocols and lower sequencing costs made it more widely accessible.
- Allows RNA molecules from a single cell to be sequenced.
- Typically carried out on a population of cells.

Single Cell RNA Sequencing Workflow

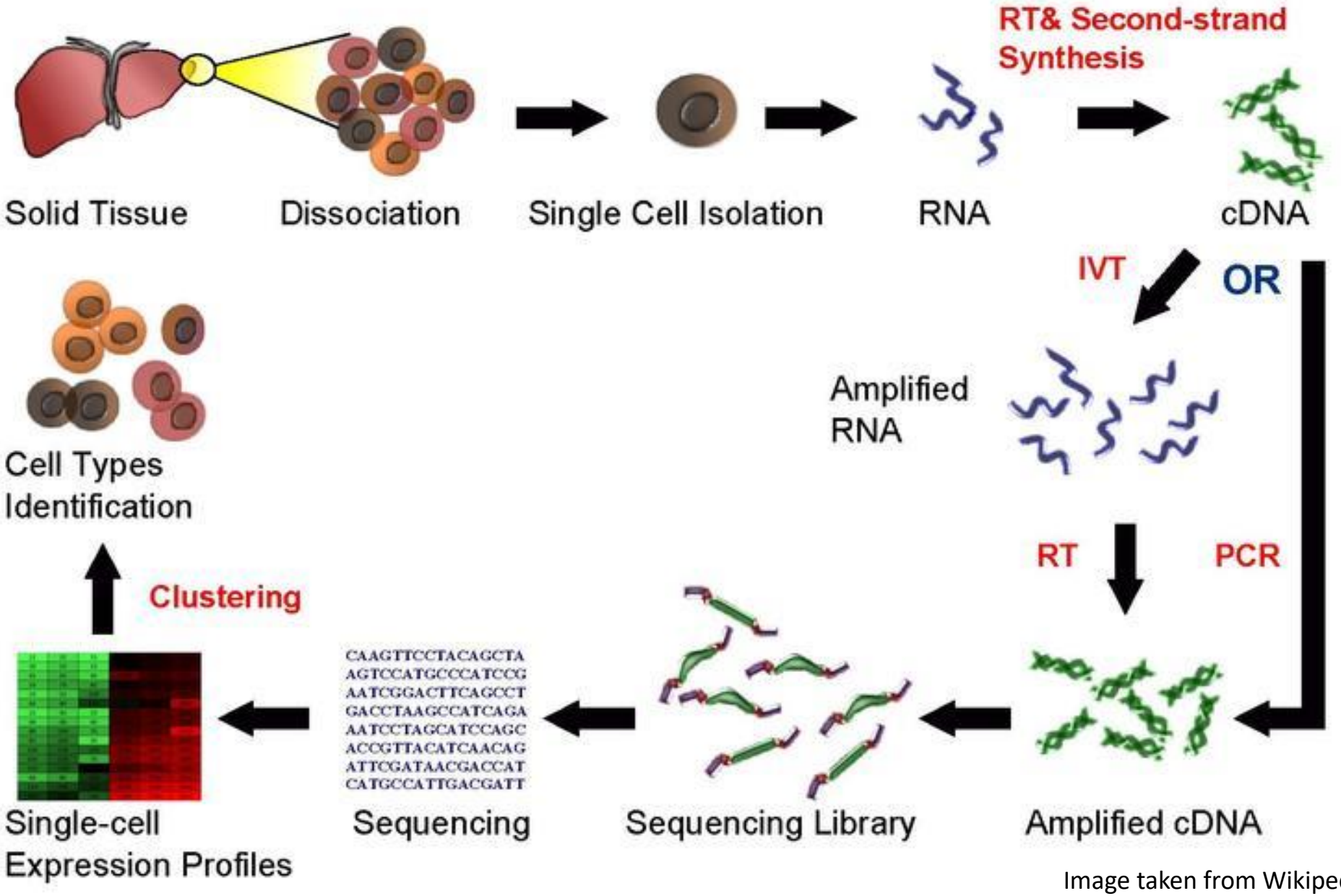


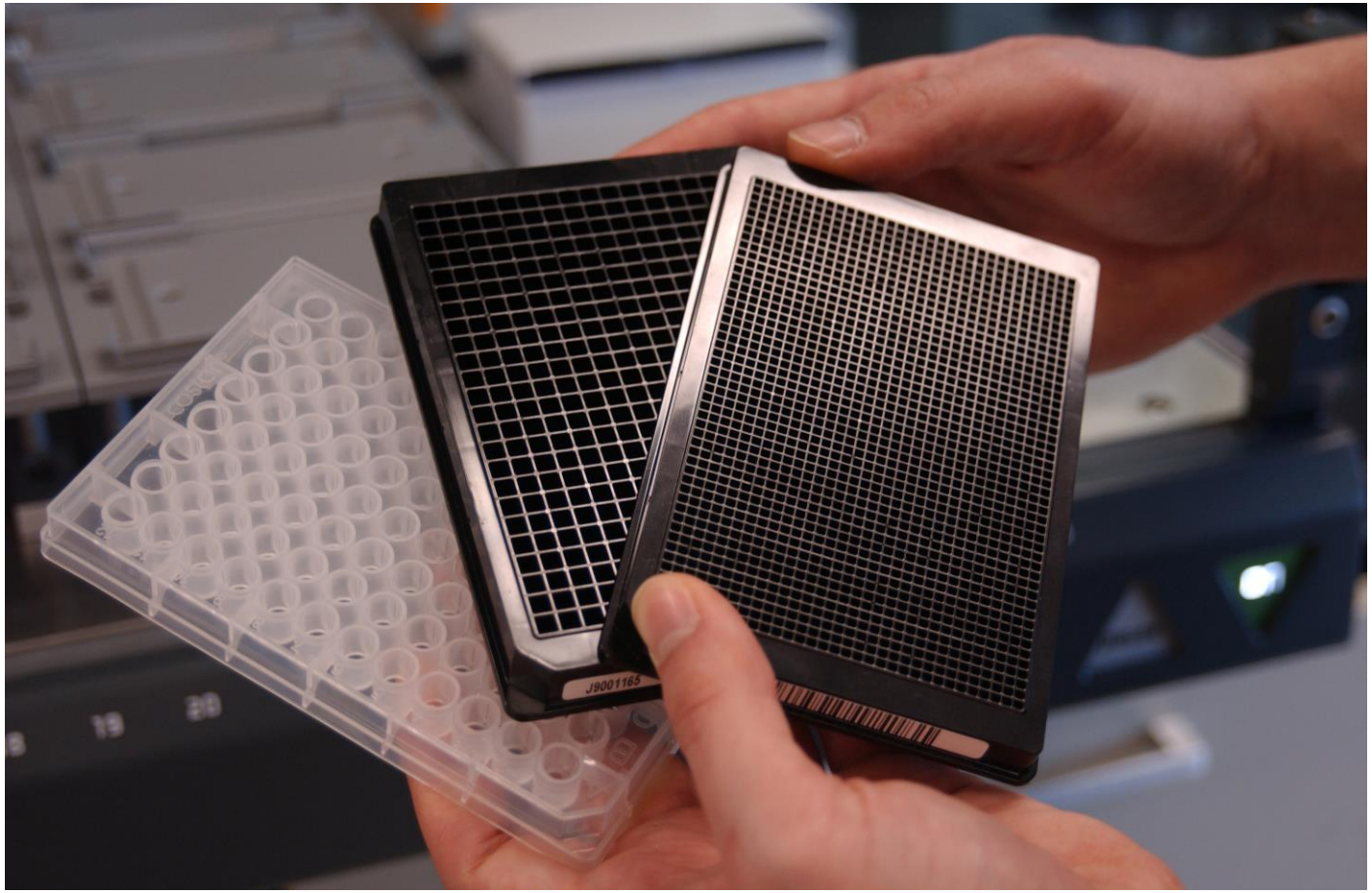
Image taken from Wikipedia by user Yijyechern

Currently available library preparation methods for scRNA-seq

- CEL-seq (Hashimshony et al. 2012)
- CEL-seq2 (Hashimshony et al. 2016)
- Drop-seq (Macosko et al. 2015)
- InDrop-seq (Klein et al. 2015)
- MARS-seq (Jaitin et al. 2014)
- SCRB-seq (Soumillon et al. 2014)
- Seq-well (Gierahn et al. 2017)
- Smart-seq (Picelli et al. 2014)
- Smart-seq2 (Picelli et al. 2014)
- SMARTer
- STRT-seq (Islam et al. 2013)

How are the cells captured?
and
Is the protocol full length or tag based?

Microwell based methods



By S.D. Hamilton (http://www.labautopedia.org/mw/index.php/User:S.d.hamilton) - http://www.labautopedia.org/mw/index.php/Image:Microplates.jpg, GFDL 1.2, https://commons.wikimedia.org/w/index.php?curid=6085654

Microwell based methods

- Can be combined with FACs to collect rare cell types
- Examples of library preparation methods which use microwell based methods include CEL-Seq, MARS-seq, SCRB-seq, Smart-seq, SMARTer, STRT-seq

Microfluidic based methods

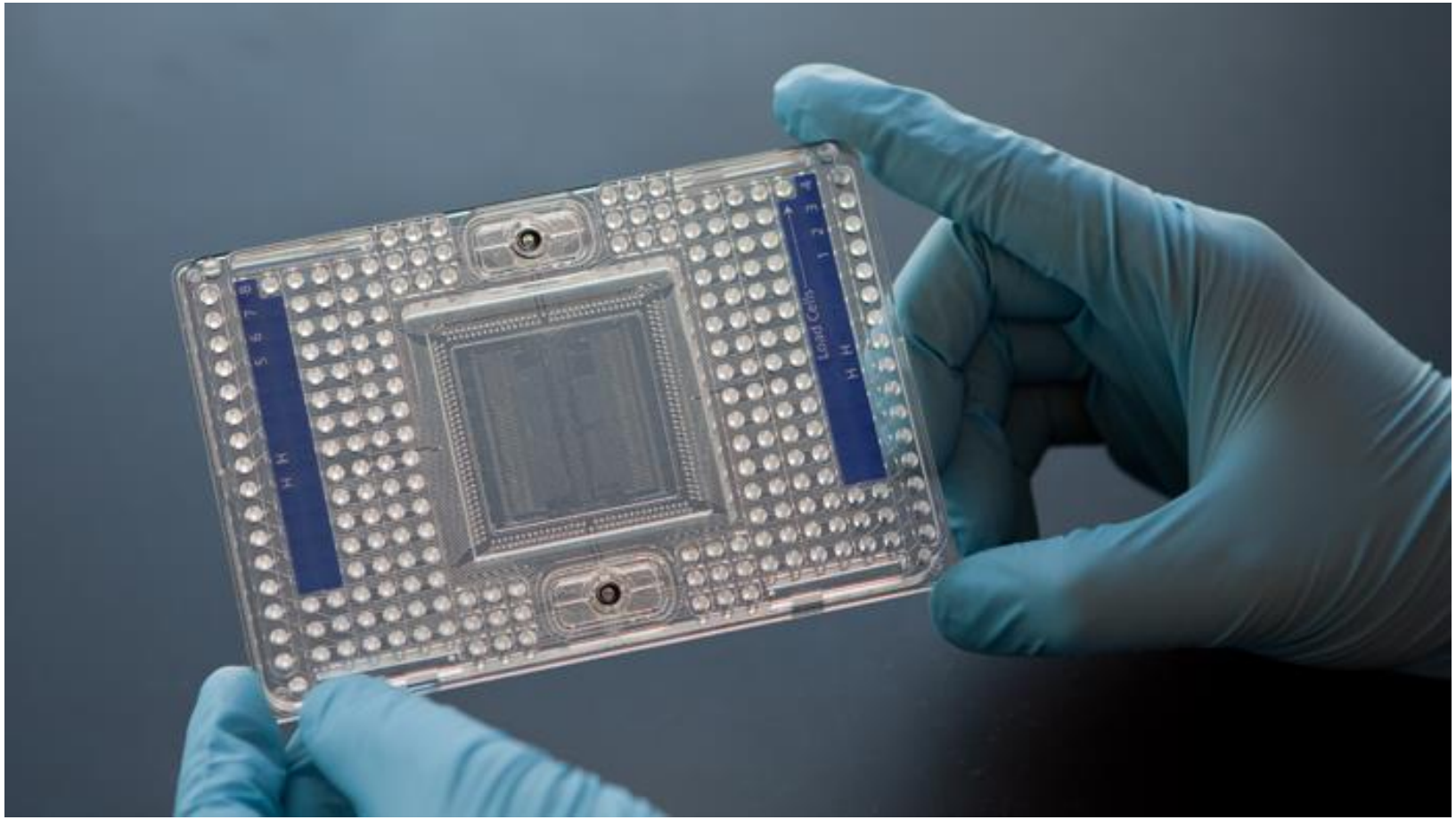


Image taken from Fluidigm

Microfluidic based methods

- Higher throughput than microwell based methods
- Only about 10% of cells are captured, making it unsuitable for analysing rare cell types
- Fluidigm C1 and SMART-seq2 are examples of microfluidic library preparation protocols

Bead based methods

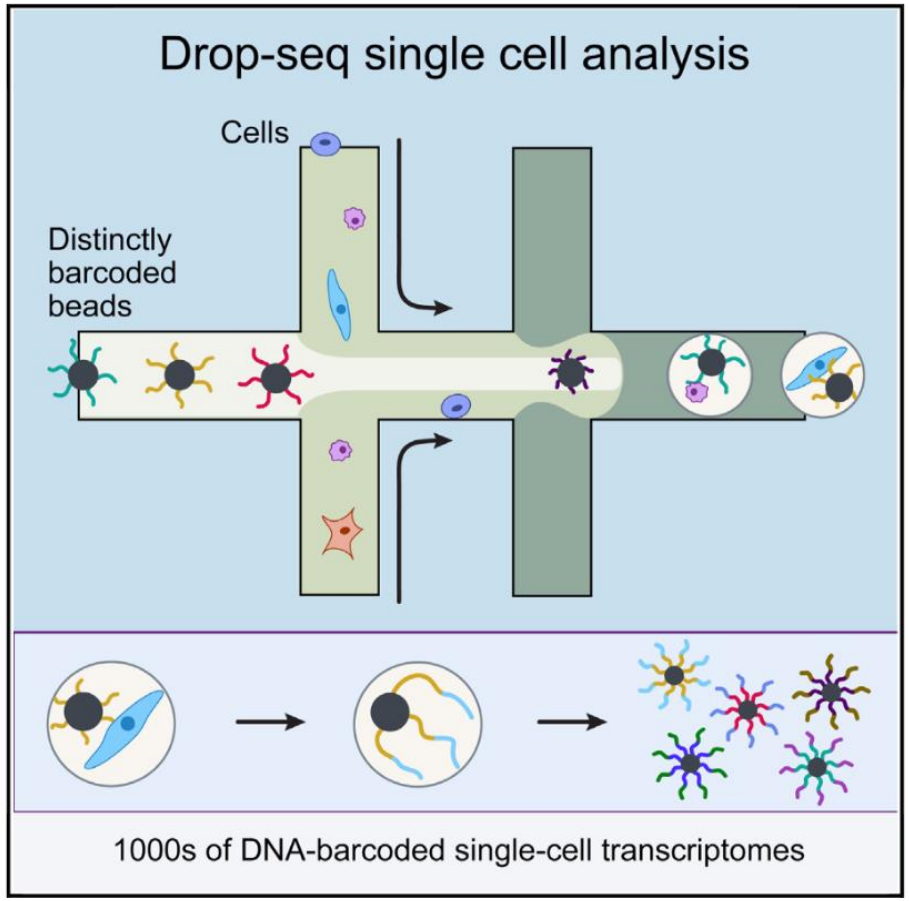


Image taken from Macosko et al.

Bead based methods

- Very high throughput and low cost per cell
- The number of reads sequenced per cell may be very low
- Drop-seq, InDrop and Seq-well are examples of bead based methods

Full length methods: SMART-seq and SMART-seq2

- Most scRNA-seq library preparation protocols show coverage bias, in which more reads are captured from one end of the cDNA than the other
- SMART-seq and SMART-seq2 have reduced coverage bias relative to other scRNA-seq methods (although the bias is not completely eliminated).
- Full length methods are more suitable for studying alternative splicing.

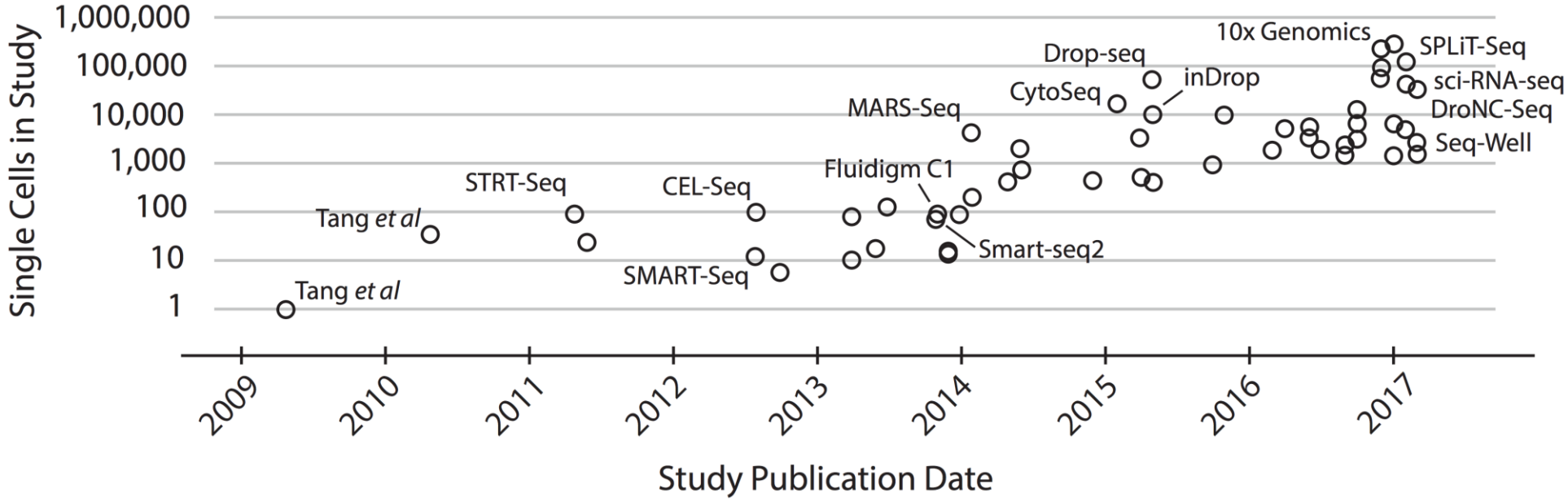
Tag based methods

- Tag based protocols only sequence one end of each cDNA, meaning they exhibit coverage bias.
- An advantage of tag based protocols is that Unique Molecular Identifiers (UMIs) can be added to them.
- UMIs are short sequences which are added to cDNAs at the end of reverse transcription, before PCR amplification. By counting the number of occurrences of UMIs, a high degree of PCR amplification bias can be removed.

Practical considerations when choosing a library preparation method

- How many cells do you want to sequence?

How many cells do you want to sequence?



Svensson, Valentine, Kedar Nath Natarajan, Lam-Ha Ly, Ricardo J Miragaia, Charlotte Labalette, Iain C Macaulay, Ana Cvejic, and Sarah A Teichmann. 2017. "Power Analys
Cell RNA-Sequencing Experiments." *Nat Meth* 14 (4). Springer Nature: 381–87. doi:10.1038/nmeth.4220.

Practical considerations when choosing a library preparation method

- How many cells do you want to sequence?
- Are you interested in rare cells?
- Are you trying to find out about the cellular composition of the tissue?
- Are you interested in alternative splicing?
- Is the accuracy of gene/isoform quantification important to you?
- What is your budget?
- Do you have other experimental considerations?

Experimental Design Interactive Exercise

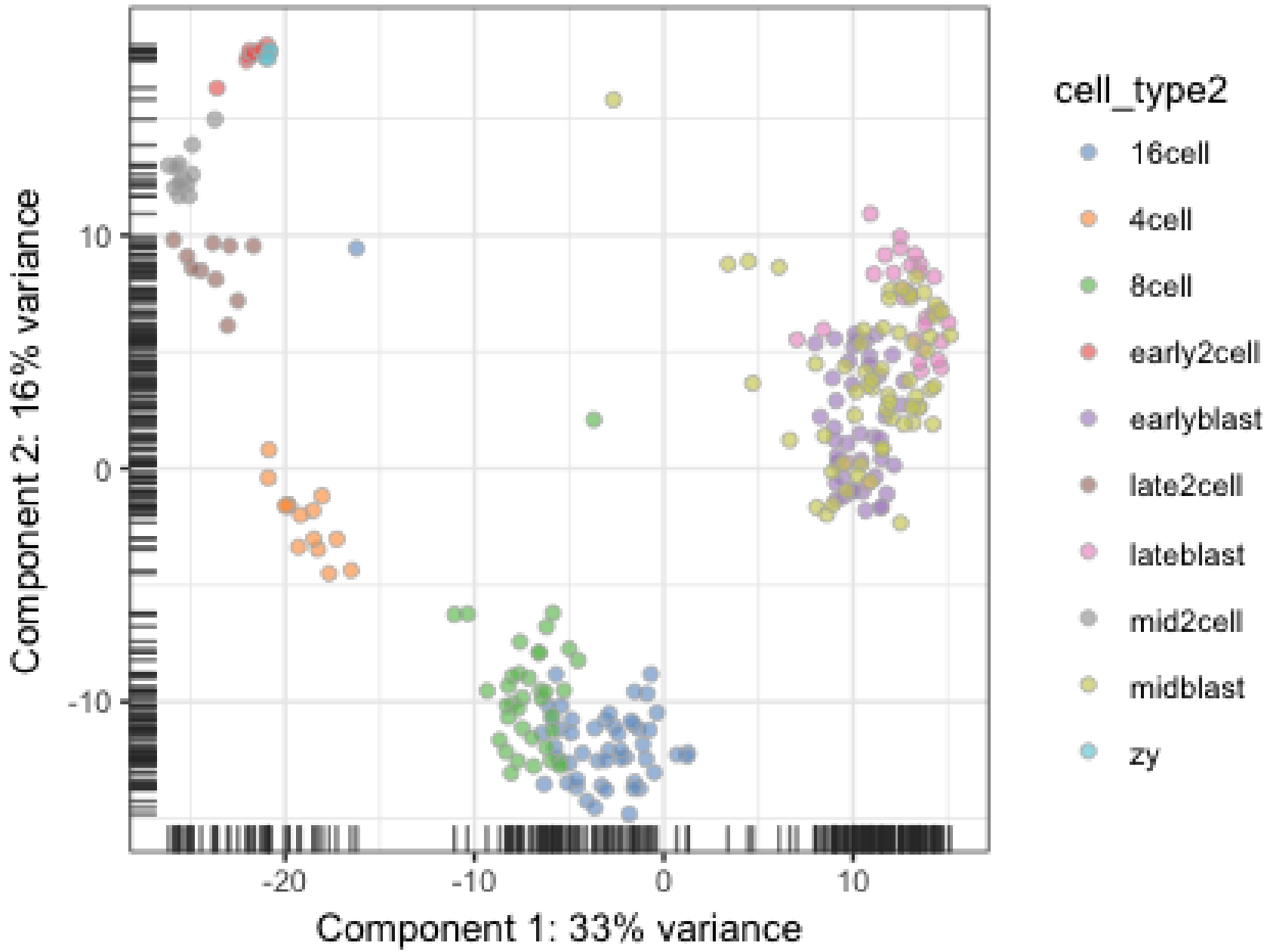
- What library preparation protocol(s) would be suitable if….
  - You wanted to characterise the transcriptomes of the different cell types in the brain?
  - You were interested in the transcriptional profile of a rare cell type in the liver?
  - You wanted to study alternative splicing?
  - You wanted to use machine learning to characterise the relationship between alternative splicing and DNA methylation?

Any questions?

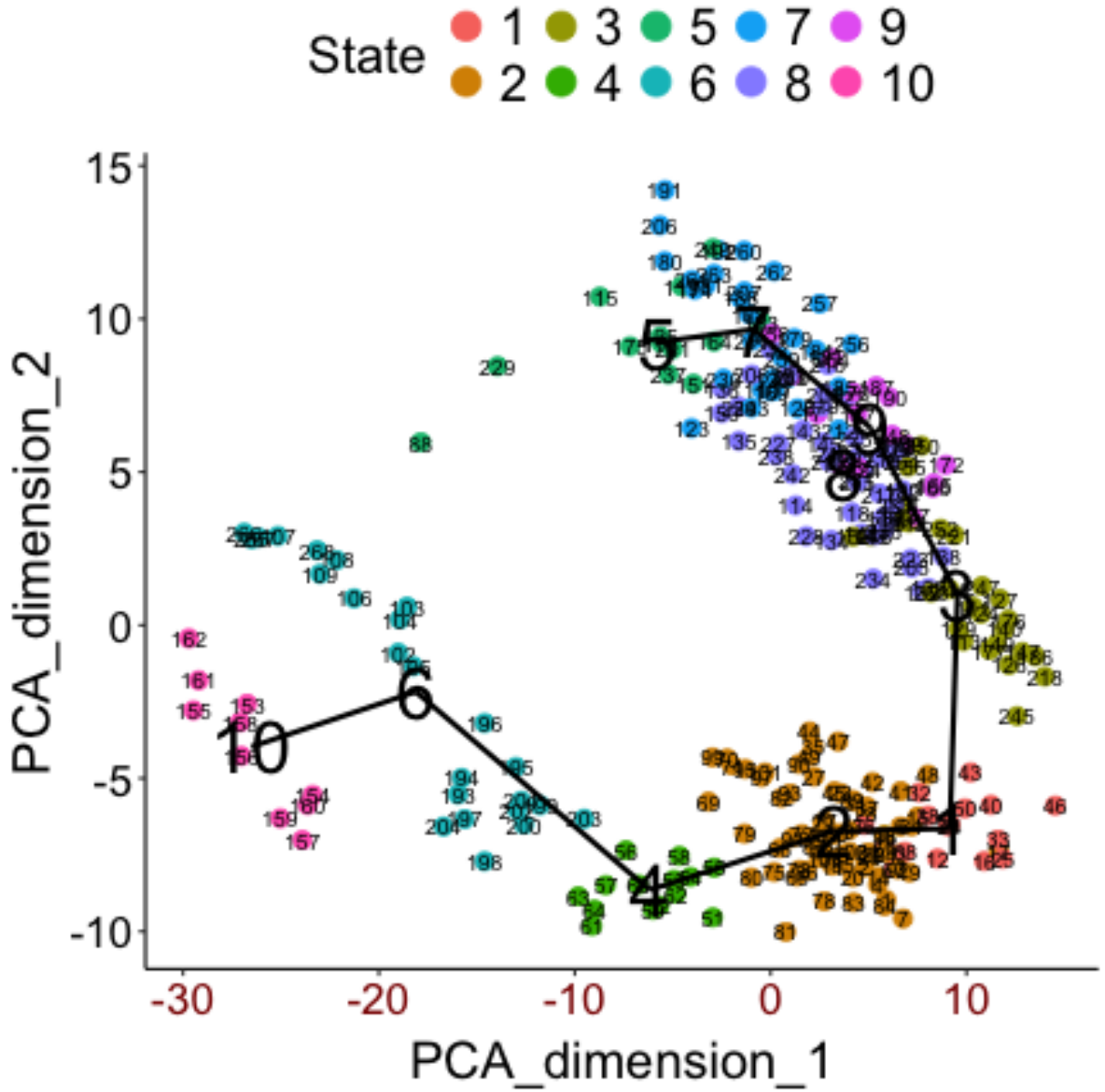
Interactive Exercise: What questions can we answer with scRNA-seq?

- What types of analysis can we do?
- Why single-cell RNA-seq rather than bulk RNA-seq?
- Could technological improvements increase the number of questions we can answer?

Clustering analysis



Pseudotime analysis



Other applications of single-cell RNA-seq

- Gene regulatory network analysis
- Alternative splicing
- Transcriptional noise

Experimental Design

Wet lab considerations

- How will you isolate your single cells?
- If you wanted to sequence rare cells or a particular cell type, how will you isolate them?
- Timing considerations – do you want to study a particular time point in development?
- Choice of model organism?
- Biological replicates
- Costs – of reagents, animals, staff costs etc.

Library preparation protocol considerations

- How many cells do you want to sequence vs how many reads do you want to sequence per cell?
- Costs – of machines, actual sequencing process, reagents
- How complicated is it to set up?
- UMIs or a protocol with less coverage bias?
- Will your choice of library preparation protocol allow you to perform the bioinformatics analysis you want to perform?

Minimising confounding due to batch effects

- Batch effects are technical artefacts which are added to the samples during handling.
- Batch effects can occur:
  - Because different samples were prepared in different labs
  - Because different samples were prepared in the same lab but at different times
  - Because cells from the same sample were sequenced on different plates (plate effects)
- Careful experimental design is required to ensure batch effects do not completely confound our results.

Interactive Exercise: Identify the confounded experiments

1. Cells are simultaneously collected from 2 biological conditions and all prepared on the same plate.
2. Cells are simultaneously collected from 2 biological conditions. Cells from the first condition are prepared on plate 1 and cells from the second condition are prepared on plate 2.
3. Lab 1 collects cells from one biological condition and lab 2 collects cells from another biological condition. The cells from both labs are prepared on the same plate.
4. Cells from biological condition 1 are collected on Monday and cells from biological condition 2 are collected on Tuesday. The cells from condition 1 and 2 are prepared on the same plate.
5. Cells are simultaneously collected from 2 biological conditions. The cells are split between 2 plates so that each plate contains 50% cells from condition 1 and 50% cells from condition 2.

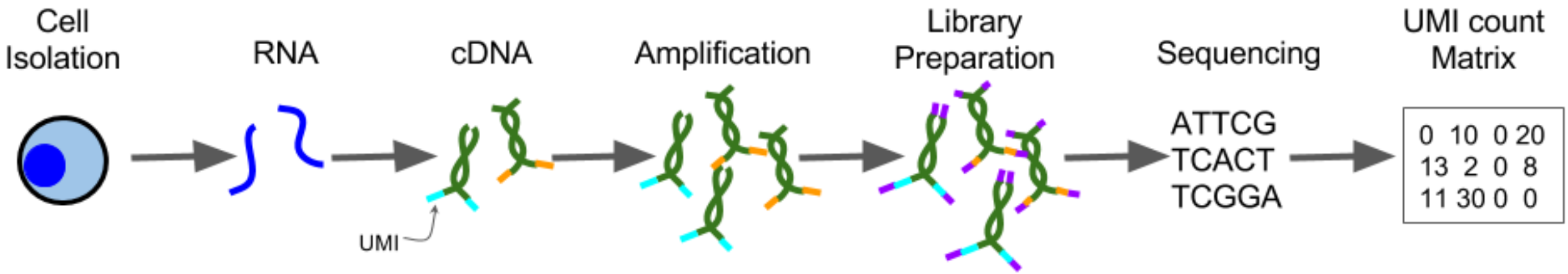
Technical considerations: PCR amplification bias

- PCR amplification bias is an issue for both bulk and single-cell RNA-seq, but is more problematic for single-cell RNA-seq due to the lower amount of starting material.
- During the PCR amplification process, some transcripts become over represented in the final library compared to their true abundance.
- This problem is worsened when a high number of PCR cycles are used to generate the sequencing library, for example in single-cell RNA-seq due to the low amount of starting material.

UMIs

- Unique Molecular Identifiers are short (4-10bp) random barcodes added to transcripts during reverse-transcription.
- They enable sequencing reads to be assigned to individual transcript molecules and thus the removal of amplification noise and biases from single-cell RNA-seq data.
- When sequencing UMI containing data, techniques are used to specifically sequence only the end of the transcript containing the UMI (usually the 3’ end).

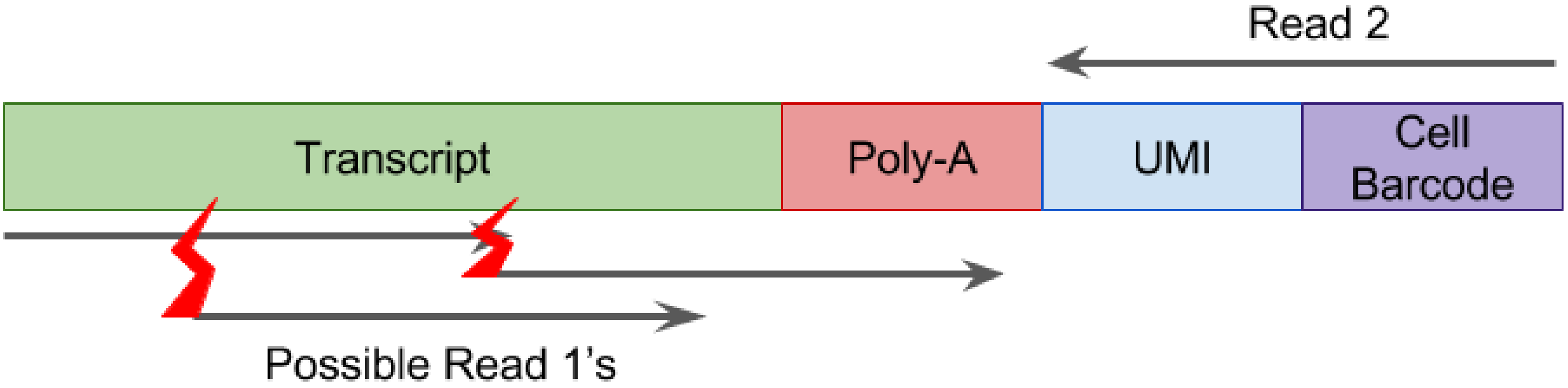
UMIs


Cell Isolation
RNA
cDNA
Amplification
Library Preparation
Sequencing
UMI count Matrix
UMI
ATTCG
TCACT
TCGGA
0 10 0 20
13 2 0 8
11 30 0 0

UMIs

- Since the number of unique barcodes (4^N, where N is the length of UMI) is much smaller than the total number of molecules per cell (~10^6), each barcode will typically be assigned to multiple transcripts.
- Hence, to identify unique molecules both barcode and mapping location (transcript) must be used.
- UMI-sequencing typically consists of paired-end reads where one read from each pair captures the cell and UMI barcodes while the other read consists of exonic sequence from the transcript

UMIs



Coverage bias

- A disadvantage of using UMIs is that UMI based protocols typically exhibit 3’ coverage bias.
- This reduces our ability to distinguish between reads originating from alternative splice isoforms which only differ at their 5’ ends.
- This may not be a particular problem if we are interested in gene level expression analysis.
- However if we want to analyse alternative splicing, a protocol with less coverage bias should be used (ie. SMART-seq or SMART-seq2).

Dropouts

- Dropouts are events in which reads mapping to a gene are detected in some cells but not others.

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5
Gene 1	100.73	99.99	99.57	75.76	105.49
Gene 2	0.00	0.00	0.00	0.00	0.00
Gene 3	20.73	0.00	29.57	0.00	25.49

Dropouts

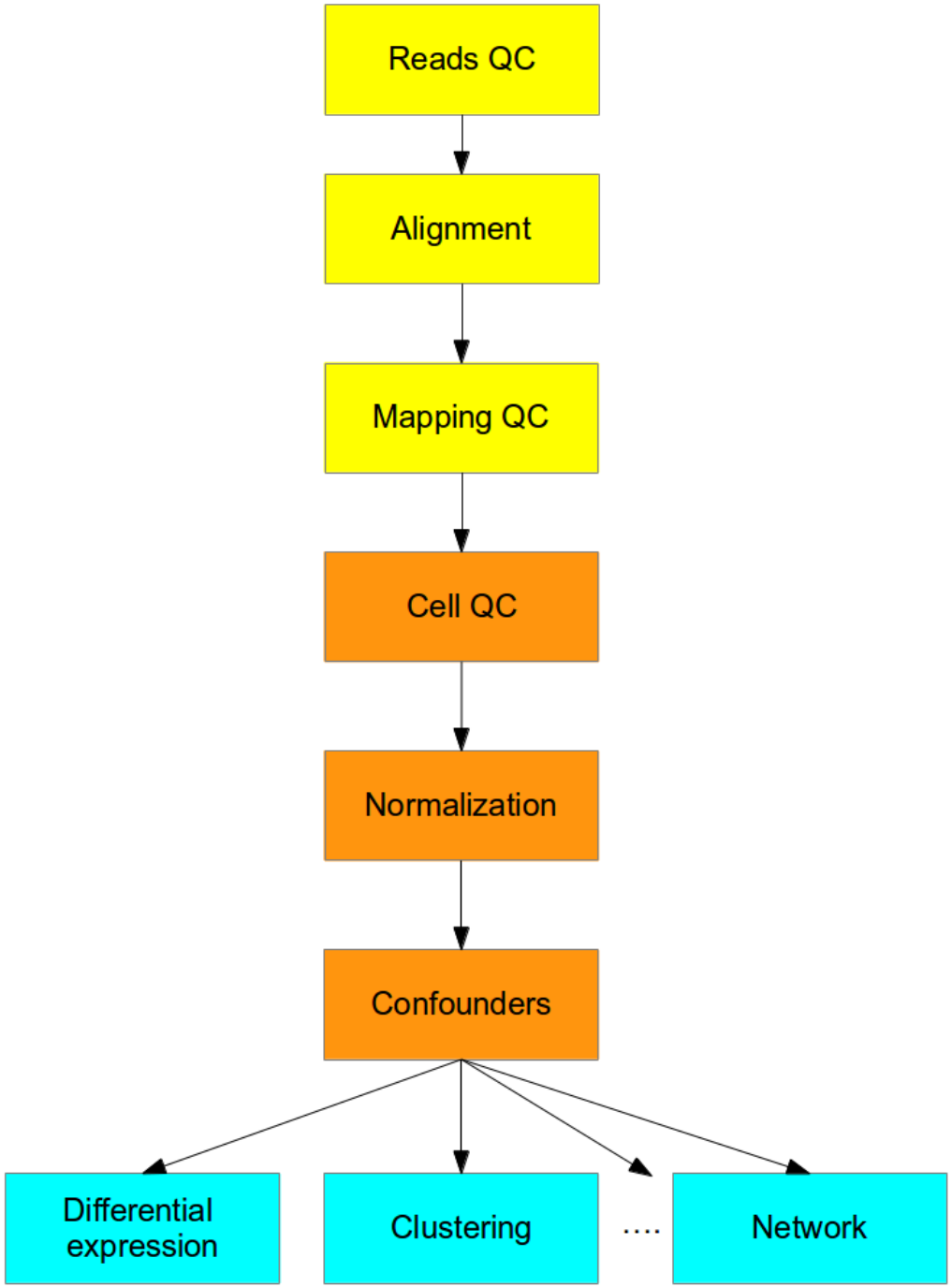
- Dropouts are events in which reads mapping to a gene are detected in some cells but not others.
- Dropouts can occur for biological reasons eg. Differential gene expression between different cell types, transcriptional bursting, spatial gene regulation.
- Due to the low amount of starting material in single-cell RNA-seq, a high number of technical dropouts occur due to failure to capture reads from expressed transcripts.
- Dropout rates have been shown to differ between different library preparation protocols.

Biological and technical causes of inter-cellular variability

- Cells may be at different stages of the cell cycle at the time of capture.
- Cells may be of different sizes, affecting the total amount of RNA extracted from each cell.
- Cells can undergo stress or become physically damaged during cell capture.
- RNA capture and PCR amplification efficiency may not be uniform between cells. This can be tested using ERCC spike-ins as a control.

Interactive Exercise: What are the data processing steps between sequencing and clustering analysis?

Flowchart of data processing steps

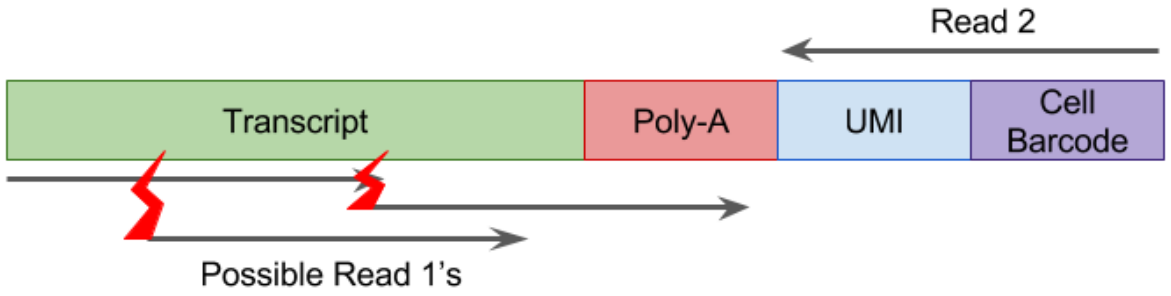


Step 1: Obtain your data

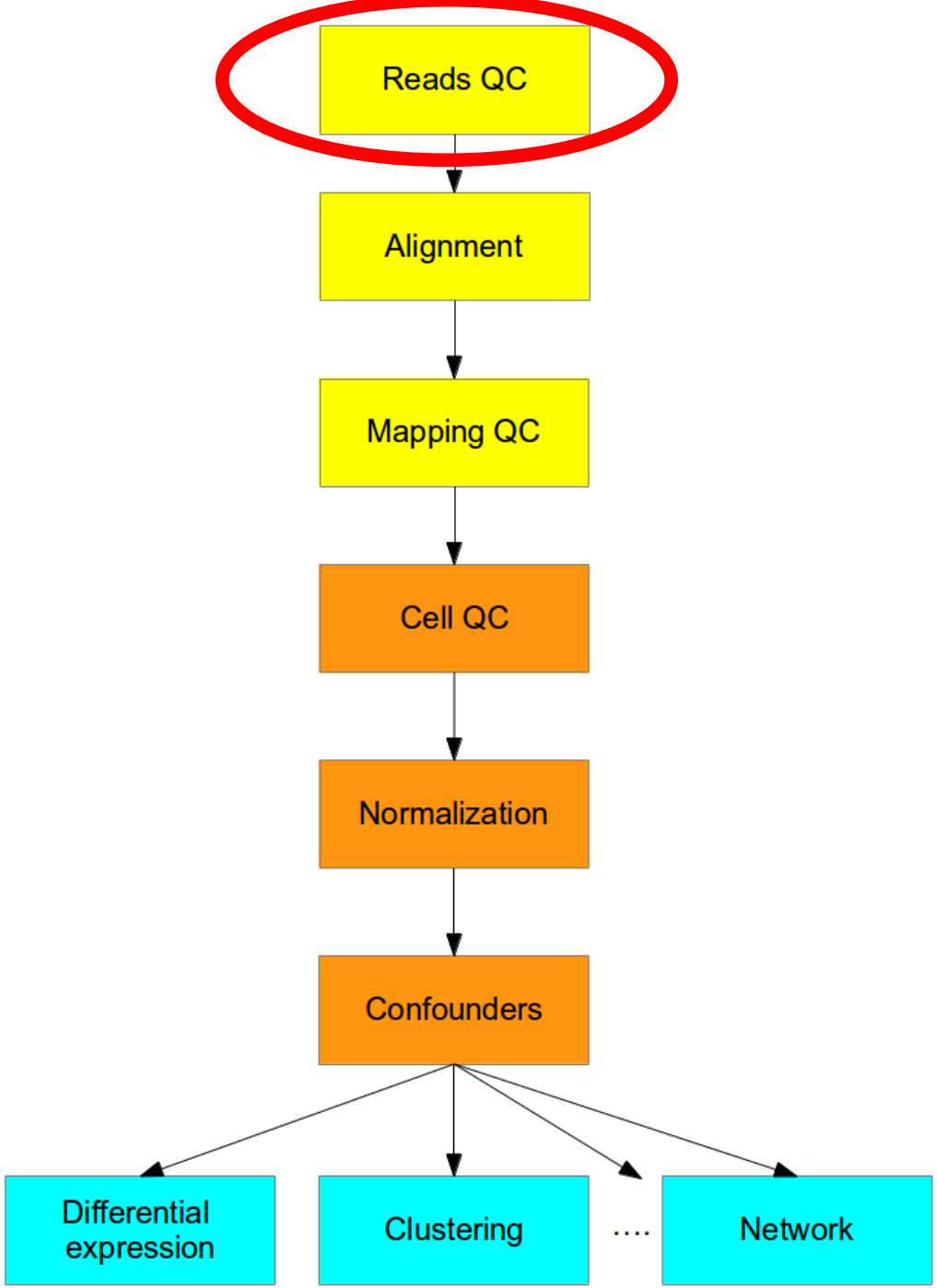
- Previously published data?
  - Download your data from Array Express or GEO
- Unpublished data produced by your/ your collaborators lab?
  - You may need to demultiplex your data (although Illumina may have done this for you)

Demultiplexing

- Because multiple cells are sequenced on the same sequencing lane, initially the sequence from multiple cells will be present in the same FASTQ file.
- Ideally, we would prefer to have one FASTQ file per cell.
- To achieve this, we can use cell barcodes to extract the reads from each cell in our sample from the multiplexed FASTQ file and generate one FASTQ file per cell.



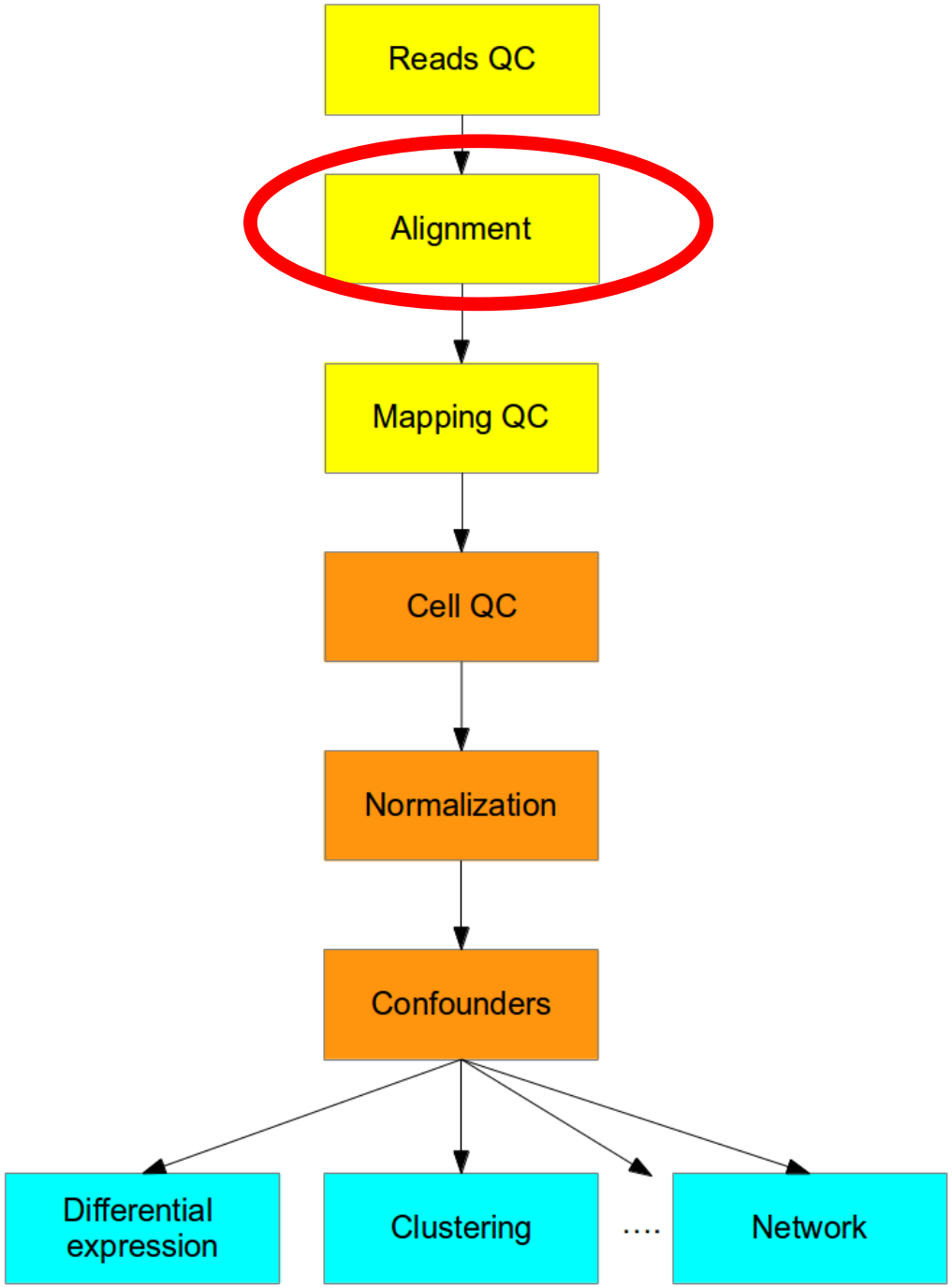
Flowchart of data processing steps



Step 2: Reads QC

- Once you have obtained your sequencing data, you should evaluate the quality of your reads.
- This can be done with standard tools also used in bulk RNA-seq.
- In the practical this afternoon, we will use a tool called FastQC to evaluate the reads quality.
- A common step to improve reads quality is trimming. We will use a tool called cutadapt to trim sequencing adapters in the practical this afternoon.

Flowchart of data processing steps

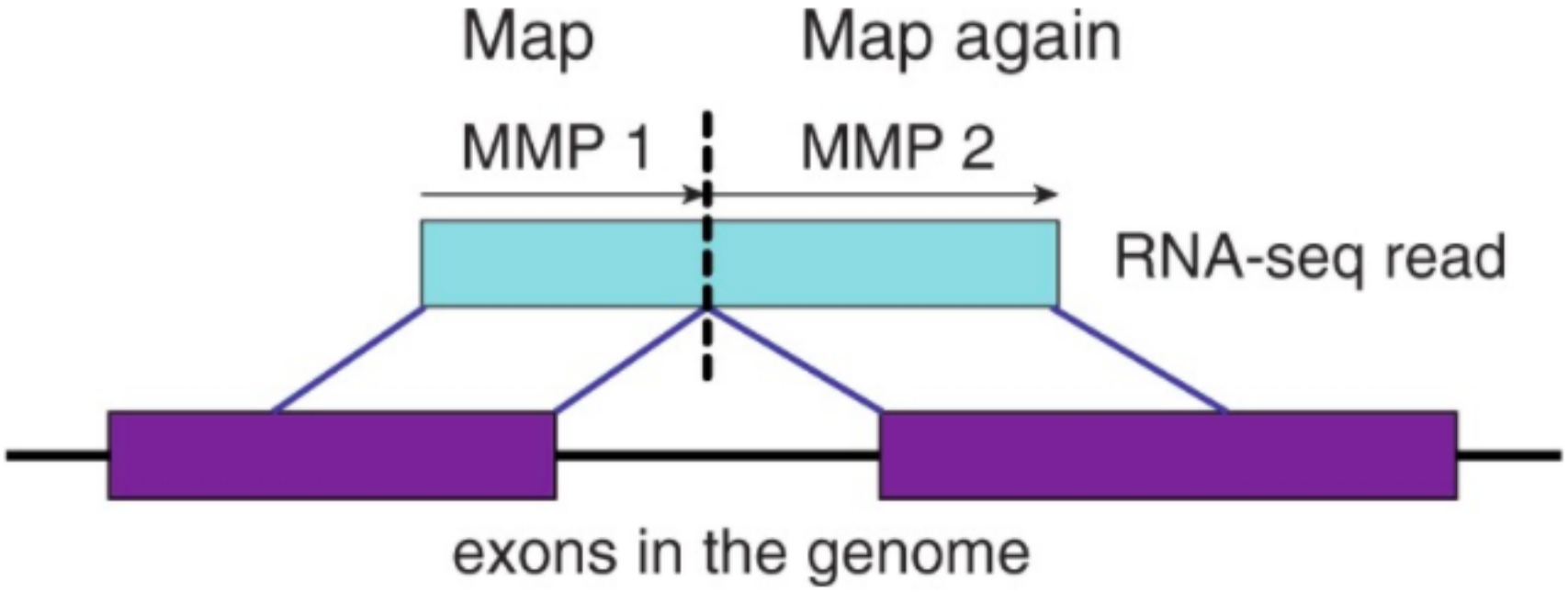


Step 3: Alignment

- Alignment describes the process of mapping reads to a reference genome or transcriptome.
- Once you have filtered your reads based on quality, you will need to perform an alignment step.
- In the practical tomorrow morning, we will consider two aligners – STAR and Kallisto

STAR

- STAR is a traditional aligner which works by trying to find the longest possible sequence which matches one or more sequences in the reference genome.



Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. and Gingeras, T.R. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1), pp. 15–21.

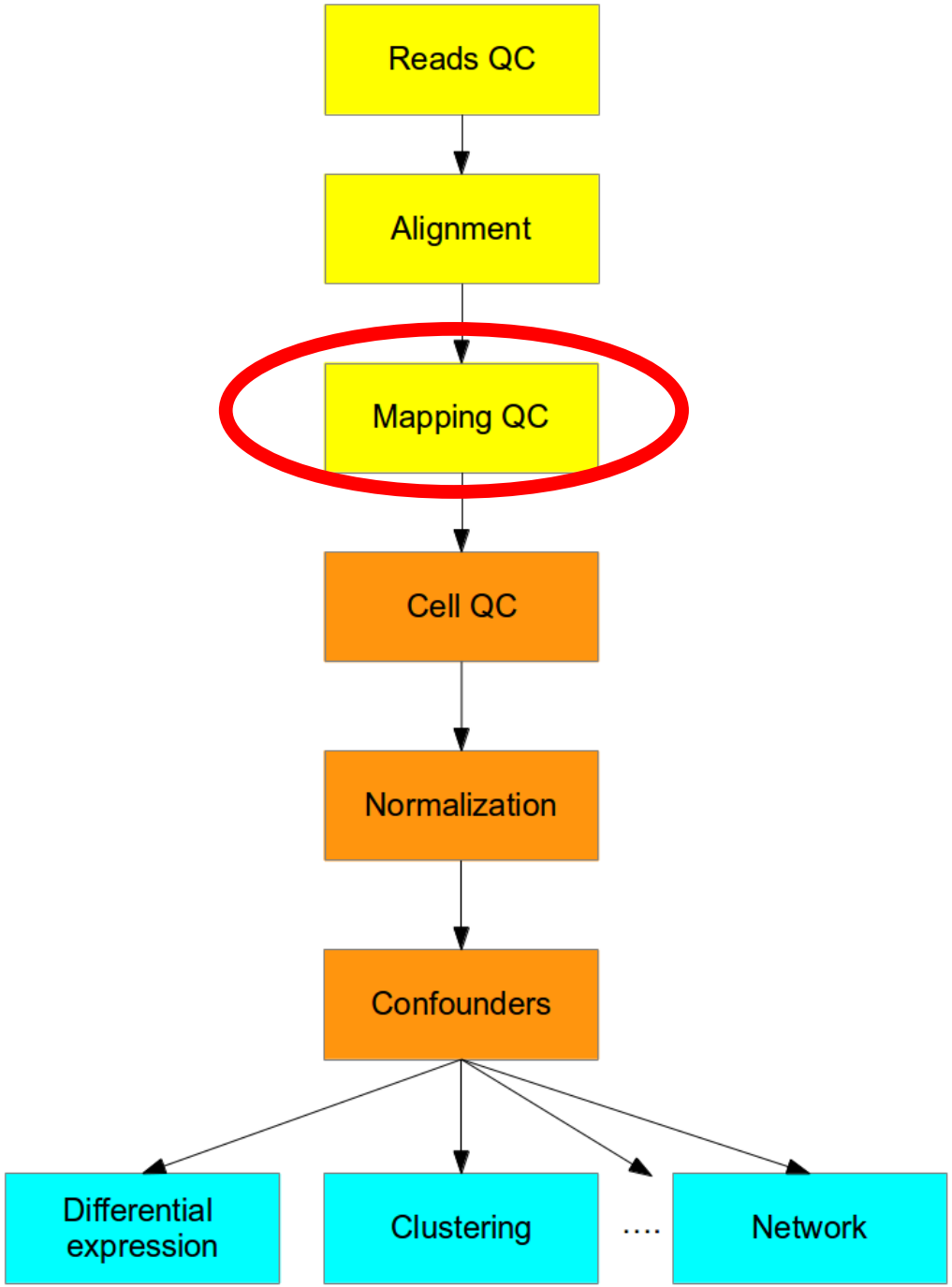
STAR

- STAR is a traditional aligner which works by trying to find the longest possible sequence which matches one or more sequences in the reference genome.
- STAR is a splice aware aligner, making it suitable if you are interested in studying alternative splicing.
- A disadvantage of STAR is that it requires a lot of RAM. If you can not meet the RAM requirements for STAR, HISAT2 is a good alternative.
- STAR works with a reference genome. If a reference is not available for your model organism, you will need to take an assembly based approach instead.

Kallisto

- Unlike STAR, Kallisto is a pseudo-aligner
- Aligners map reads to a reference genome/transcriptome
- Pseudo-aligners break up reads into chunks of sequence called k-mers, which they then map to a reference transcriptome.
- Pseudo-aligners are generally much faster than traditional aligners, making them attractive for single-cell studies.

Flowchart of data processing steps

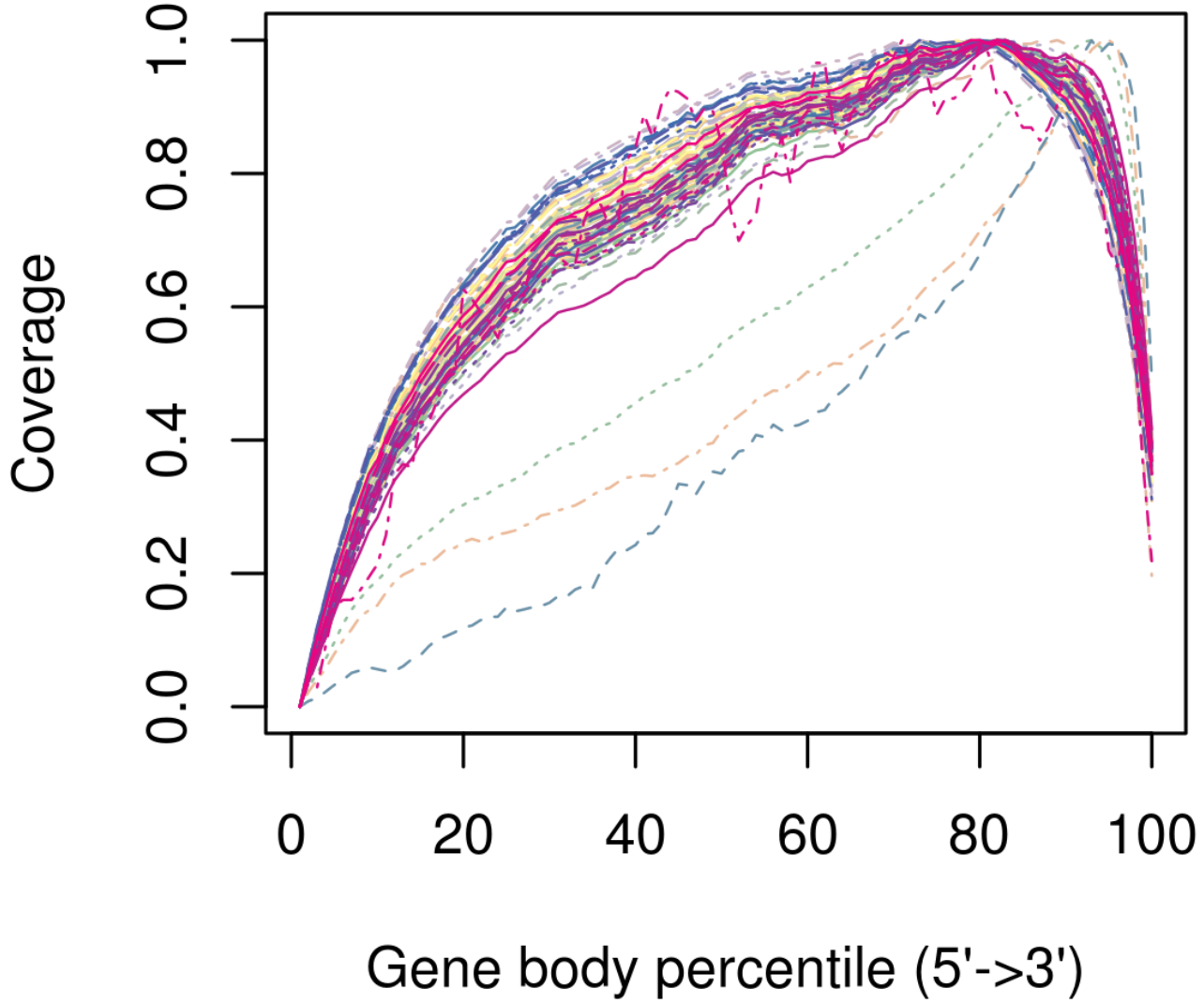


Step 4: Mapping QC

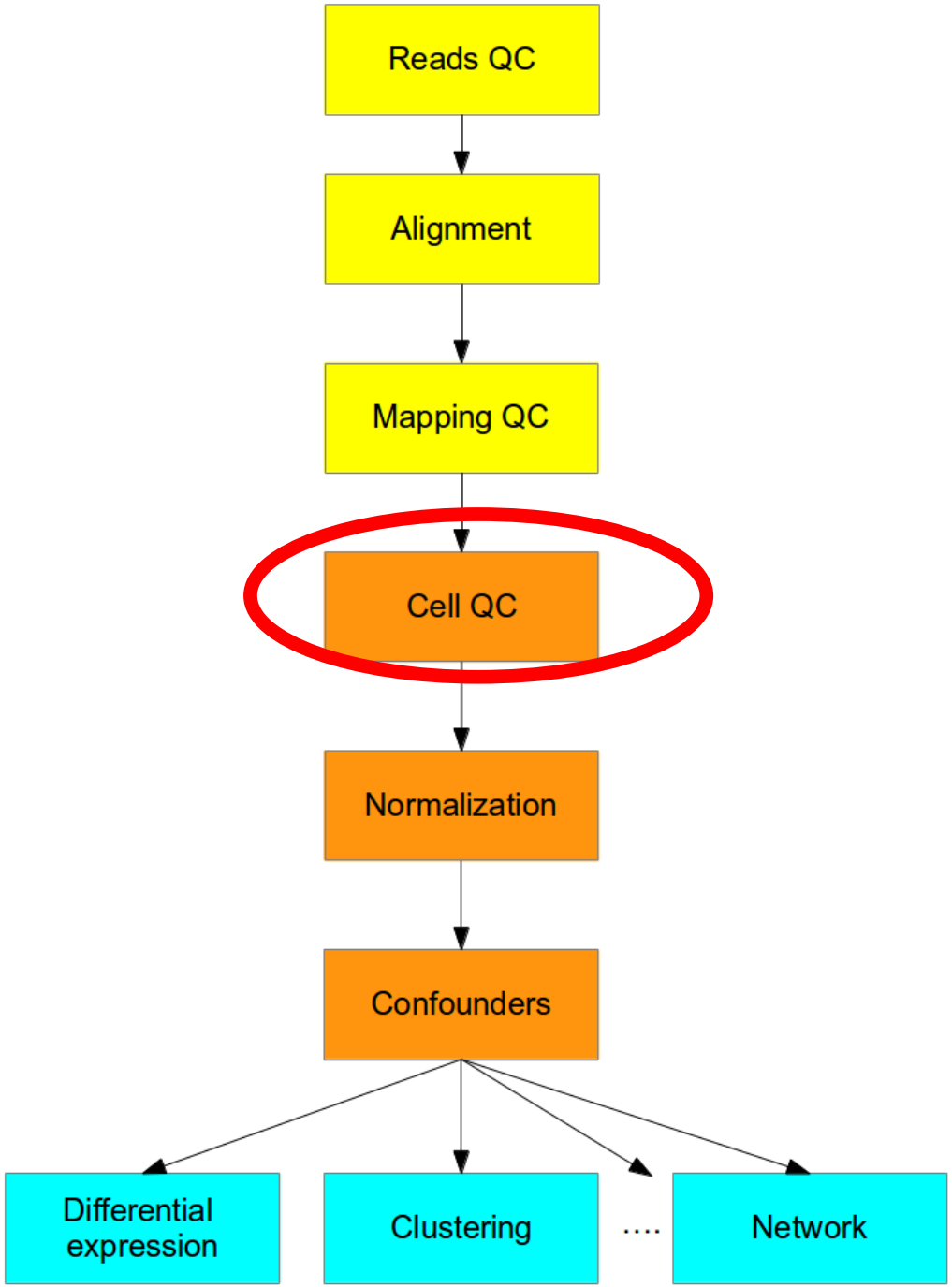
- In this step, we consider the quality of our alignments
  - Amount of reads mapping to rRNA/tRNAs
  - Proportion of uniquely mapping reads
  - Reads mapping across splice junctions
  - Read depth along the transcripts.

- Tools used in bulk RNA-seq mapping QC such as RSeQC are applicable to single-cell RNA-seq data

Step 4: Mapping QC



Flowchart of data processing steps



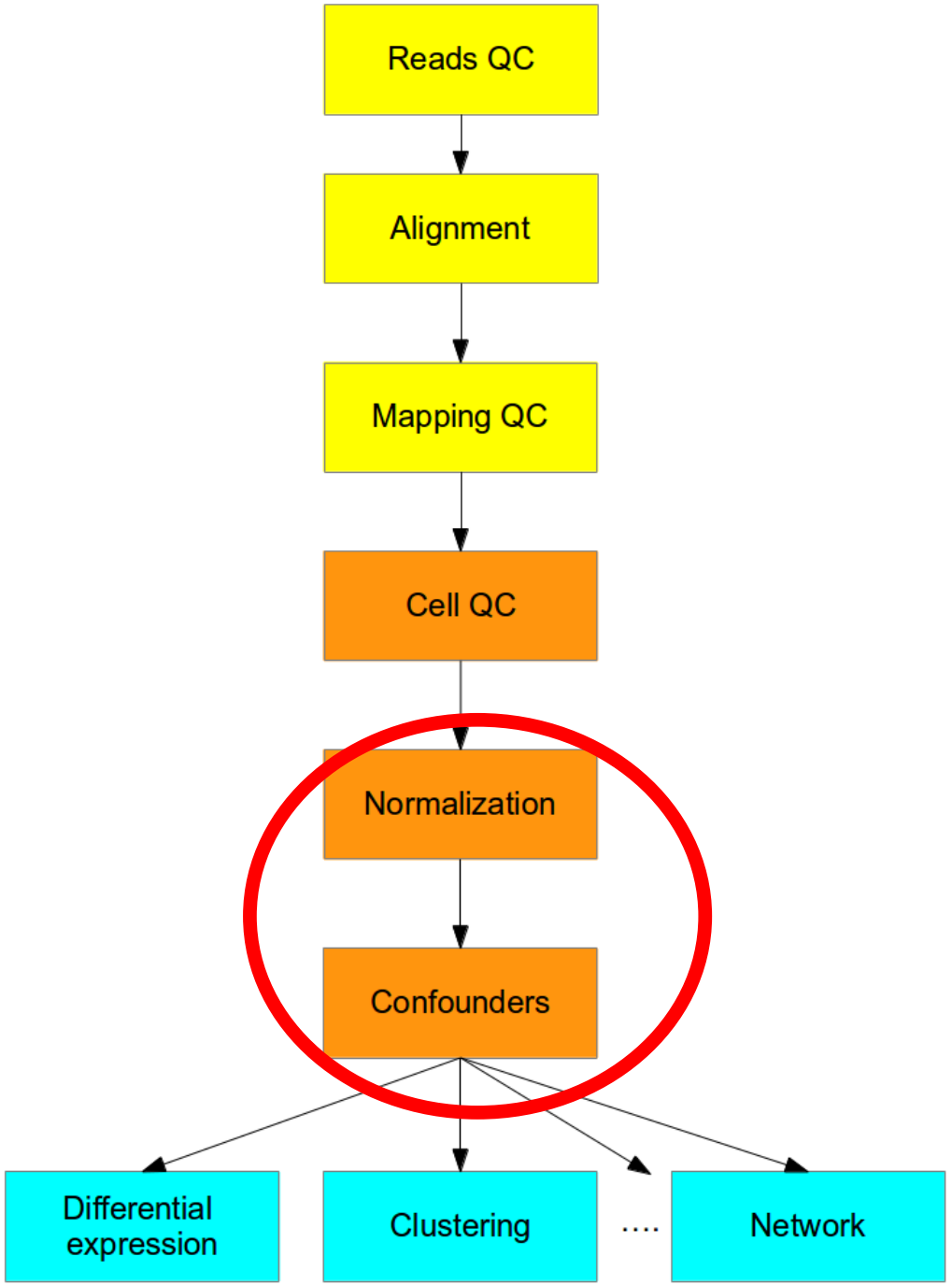
Step 5: Cell QC

- In this step, low quality cells (eg. Doublets, cells damaged during sequencing) should be removed.
- Indicators of low quality include:
  - Cells with a low number of reads
  - Cells with a low number of detected genes
  - Cells with a high percentage of mitochondrial reads
  - Cells with an unusual ratio of reads from ERCC spike-ins vs reads from endogenous genes
- We will do a lab on this topic on Wednesday.

ERCC spike-ins

- A pool of 96 synthetic RNAs of differing length.
- If a known amount of spike-ins is added to each cell, information from spike-ins can be used to estimate variables such as:
  - Variation in cell size.
  - Capture efficiency.
  - Technical noise.

Flowchart of data processing steps



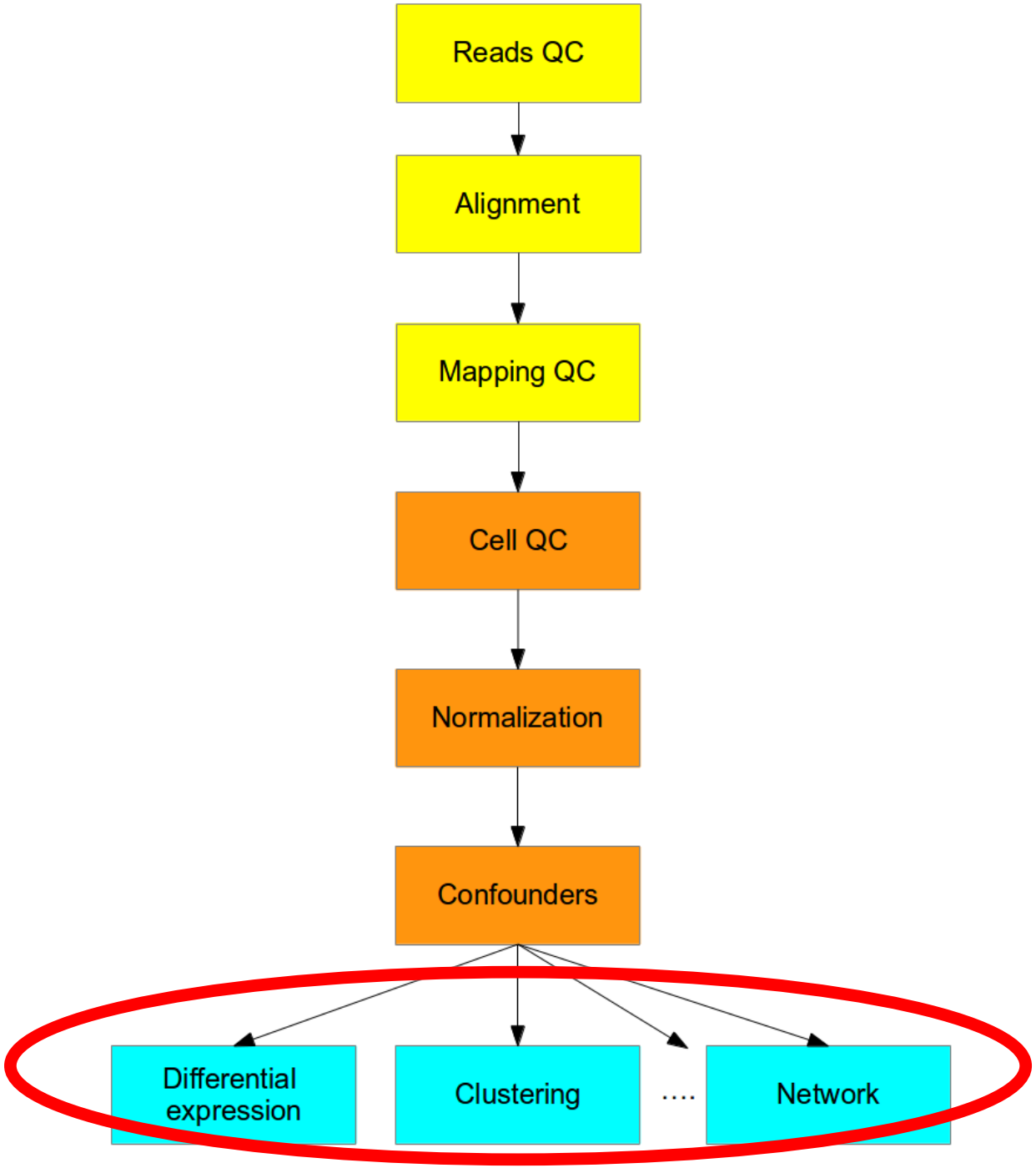
Step 6: Normalisation & Confounders

- Even the best designed single-cell RNA-seq experiment will have technical confounders.
- These confounders should be normalised out as much as possible prior to analysis.
- Examples of technical confounders include:
  - Variation in library sizes between cells
  - Number of detected genes per cell
  - Batch effects
  - Cell cycle effects
  - ….and many more

Step 6: Normalisation & Confounders

- Fortunately, tools exist to characterise and normalise for many potential confounders.
- We will use scater to characterise the some of the common confounders in single-cell RNA-seq

Flowchart of data processing steps



Optional Interactive Exercise

- Use the information presented in this session to make an outline for how you will design scRNA-seq experiments for your research and/or make a flowchart of how you will analyse your results.
- Feel free to discuss your ideas with other participants and/or with myself and Tallulah.