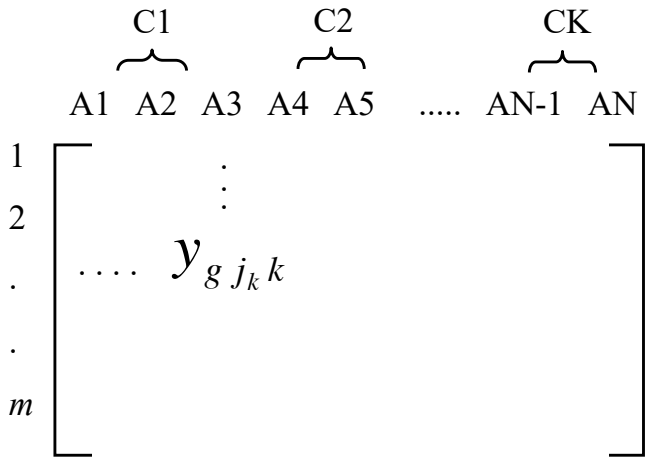
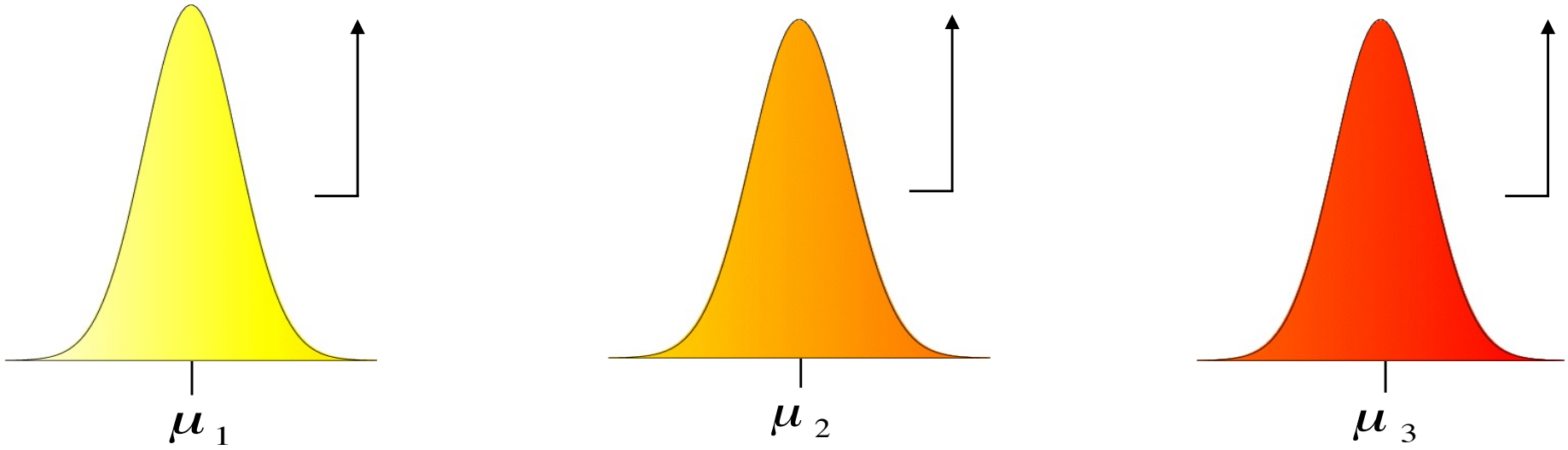
Data Structure

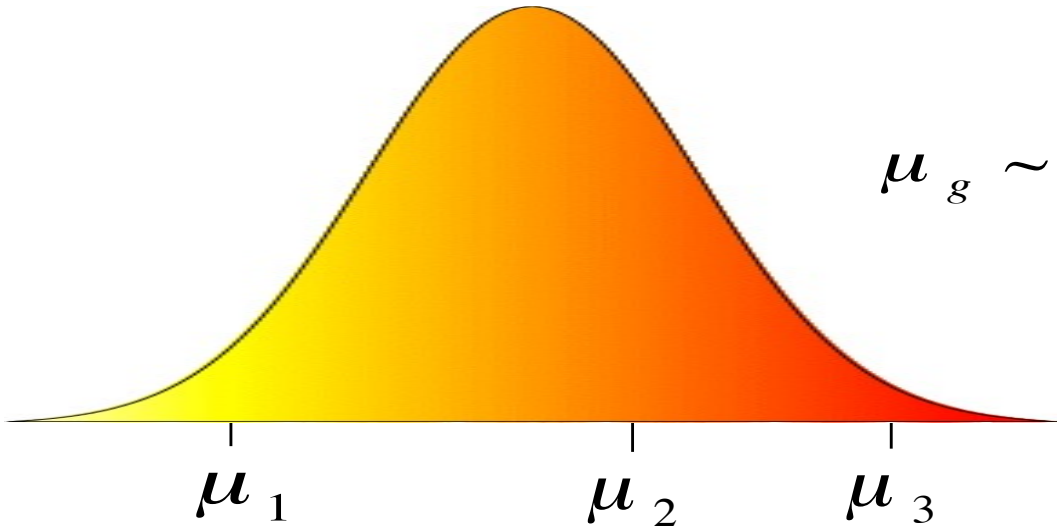


- $g = 1,2, \ldots, m$ genes (transcripts)
- $k = 1,2,..., K$ conditions
- $j = 1,2,... j_k$ replicates

Hierarchical Model for Expression Data

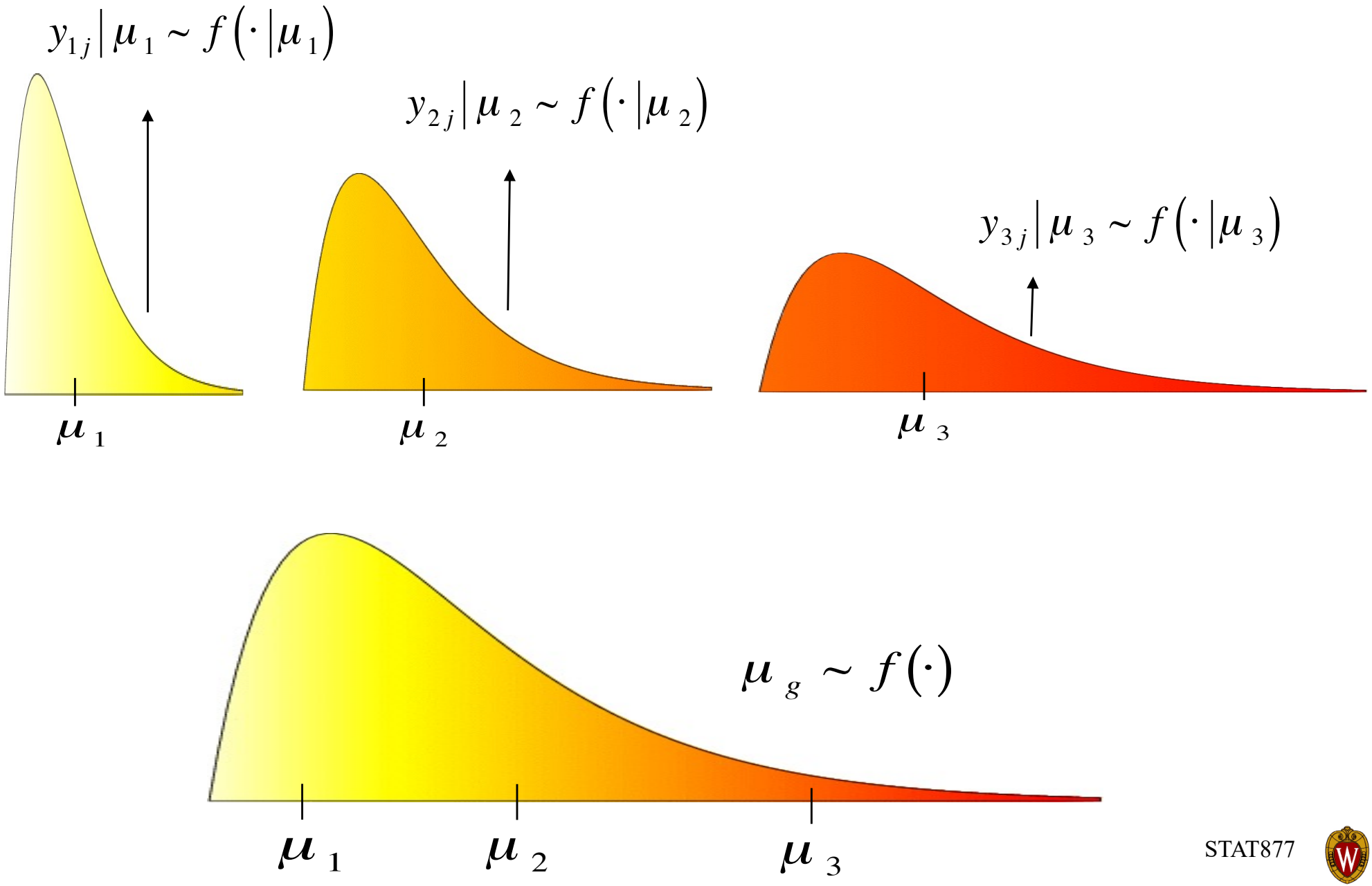
$$l(y_{1j}) | \mu_1 \sim f(\cdot | \mu_1) \qquad l(y_{2j}) | \mu_2 \sim f(\cdot | \mu_2) \qquad l(y_{3j}) | \mu_3 \sim f(\cdot | \mu_3)$$





$$\mu_g \sim f(\cdot)$$

Hierarchical Model for Expression Data (One condition)



Hierarchical Model for Expression Data (Two conditions)

- Let $y = [y_{c1}, y_{c2}]$ denote data (one gene) in conditions C1 and C2 (not denoting gene or replicates here).
- Two patterns of expression:

$$\text{P0 (EE)}: \quad \mu_{c1} = \mu_{c2}$$

$$\text{P1 (DE)}: \quad \mu_{c1} \neq \mu_{c2}$$

- For P0, $y \sim \int f(y|\mu) f(\mu) d\mu \equiv f_0(y)$
- For P1, $y \sim \int f(y|\mu_{c1}, \mu_{c2}) f(\mu_{c1}, \mu_{c2}) d\mu_{c1} d\mu_{c2}$

$$\equiv \underbrace{\int f(y_{c1}|\mu_{c1}) f(\mu_{c1}) d\mu_{c1}}_{f_0(y_{c1})} \underbrace{\int f(y_{c2}|\mu_{c2}) f(\mu_{c2}) d\mu_{c2}}_{f_0(y_{c2})} \equiv f_1(y)$$

Hierarchical Mixture Model for Expression Data

- Two conditions:

$$y \sim p_0 f_0(y) + p_1 f_1(y) \quad => \quad p(P1|y) = \frac{p_1 f(y|P1)}{p_0 f(y|P0) + p_1 f(y|P1)}$$

- Multiple conditions:

$$y \sim \sum_{k=1}^{K} p_k f_k(y) \quad => \quad p(Pk'|y) = \frac{p_{k'} f(y|Pk')}{\sum_{k \neq k'} p_k f(y|Pk)}$$

- Parameter estimates via EM (ask if you want derivation)

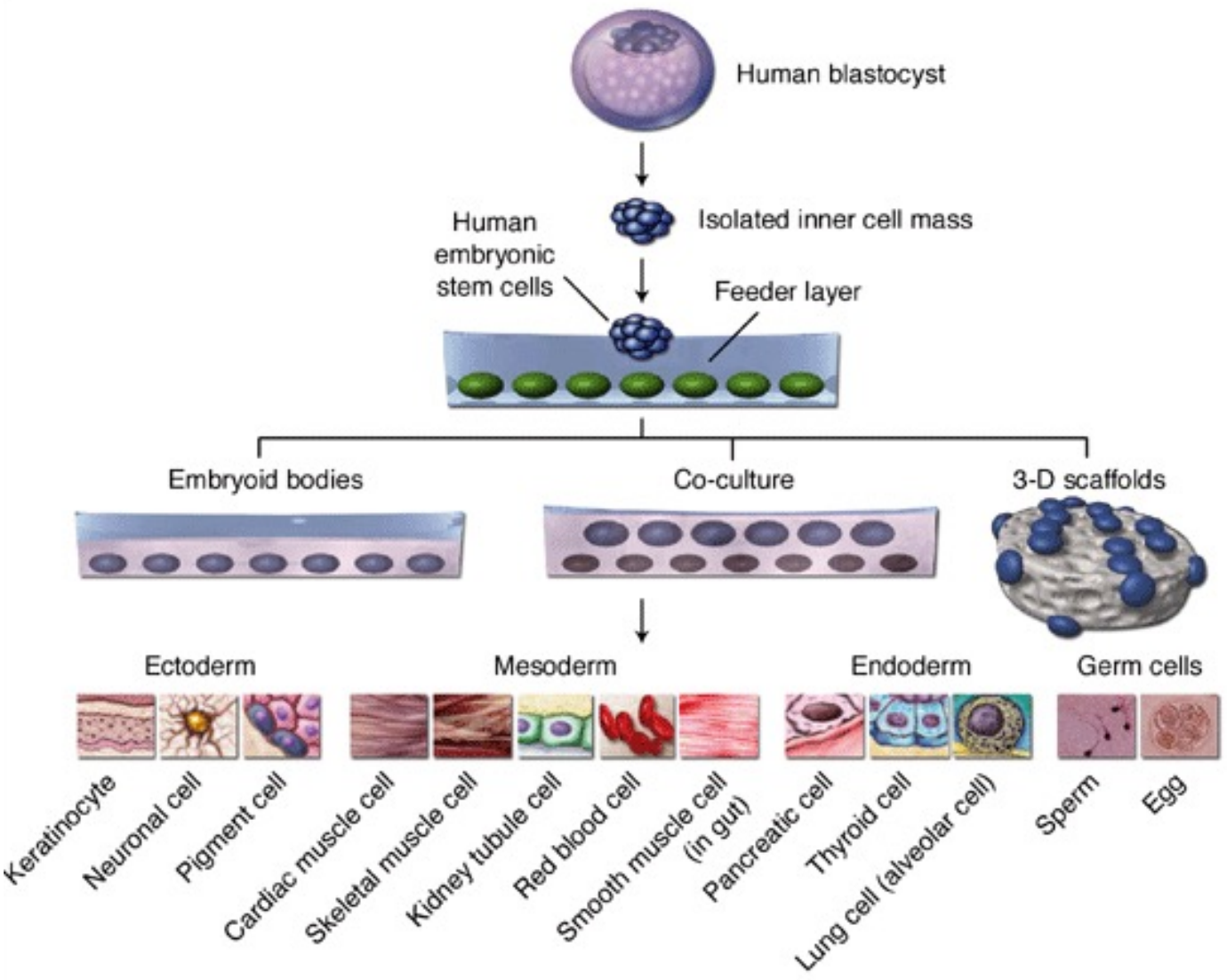
- Posterior probabilities can be used to determine FDR based threshold (MNewton lectures).

Comments on EBarrays

- Hierarchical model is used to identify patterns of expression. The model accounts for the measurement error process and for natural fluctuations in absolute expression levels.
- Posterior probabilities of expression patterns are calculated for every transcript.
- Multiple conditions are handled in the same way as two conditions (no extra work required!).
- Threshold can be adjusted to target a specific FDR (Mnewton).
- In Bioconductor, with families GG, LNN, and LNNMV.

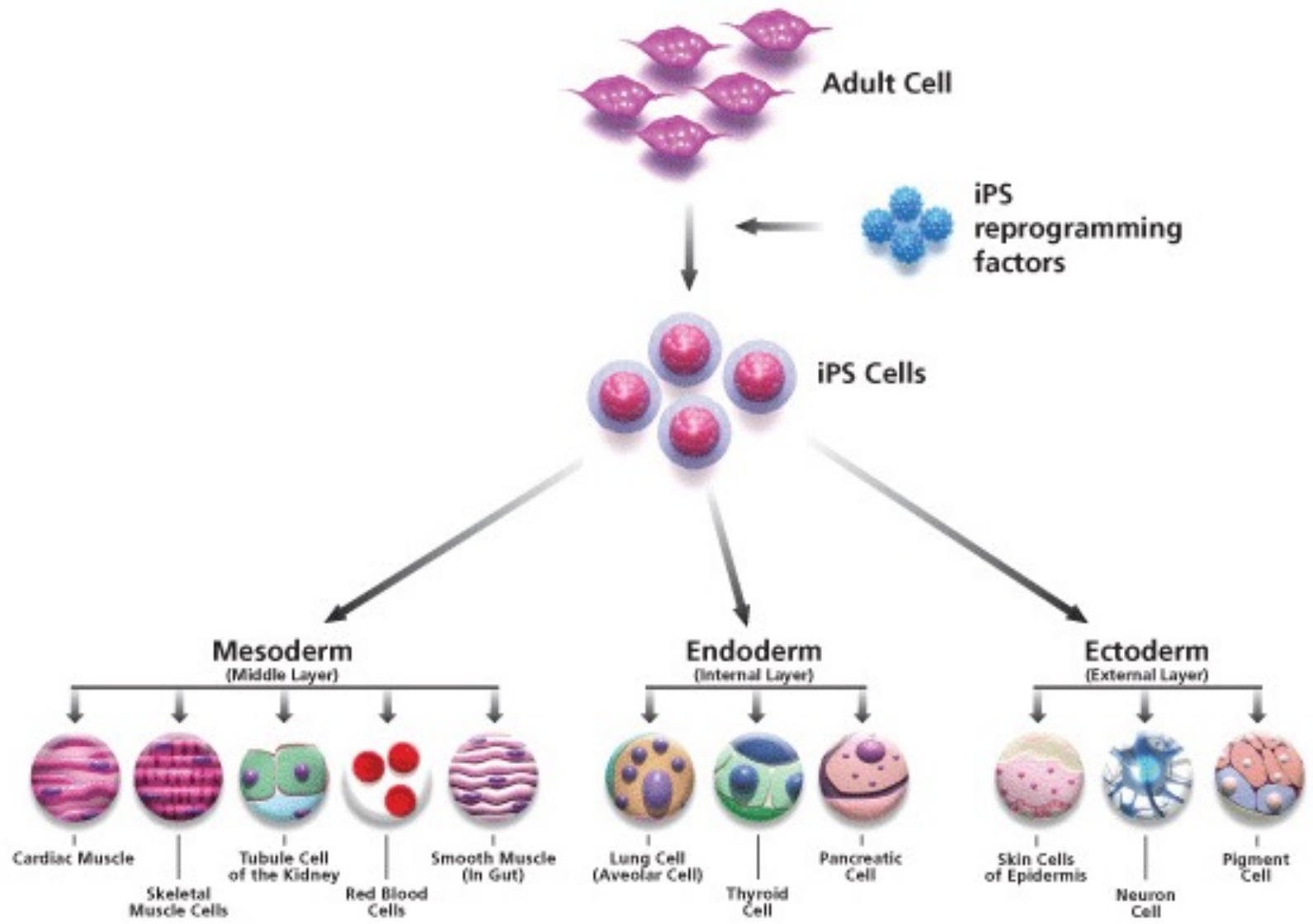
Moving on to RNA-seq data analysis

The Thomson lab isolated the first hESC in 1998



Thomson *et al.*, Science, 1998

The Thomson lab developed iPSC lines in 2007



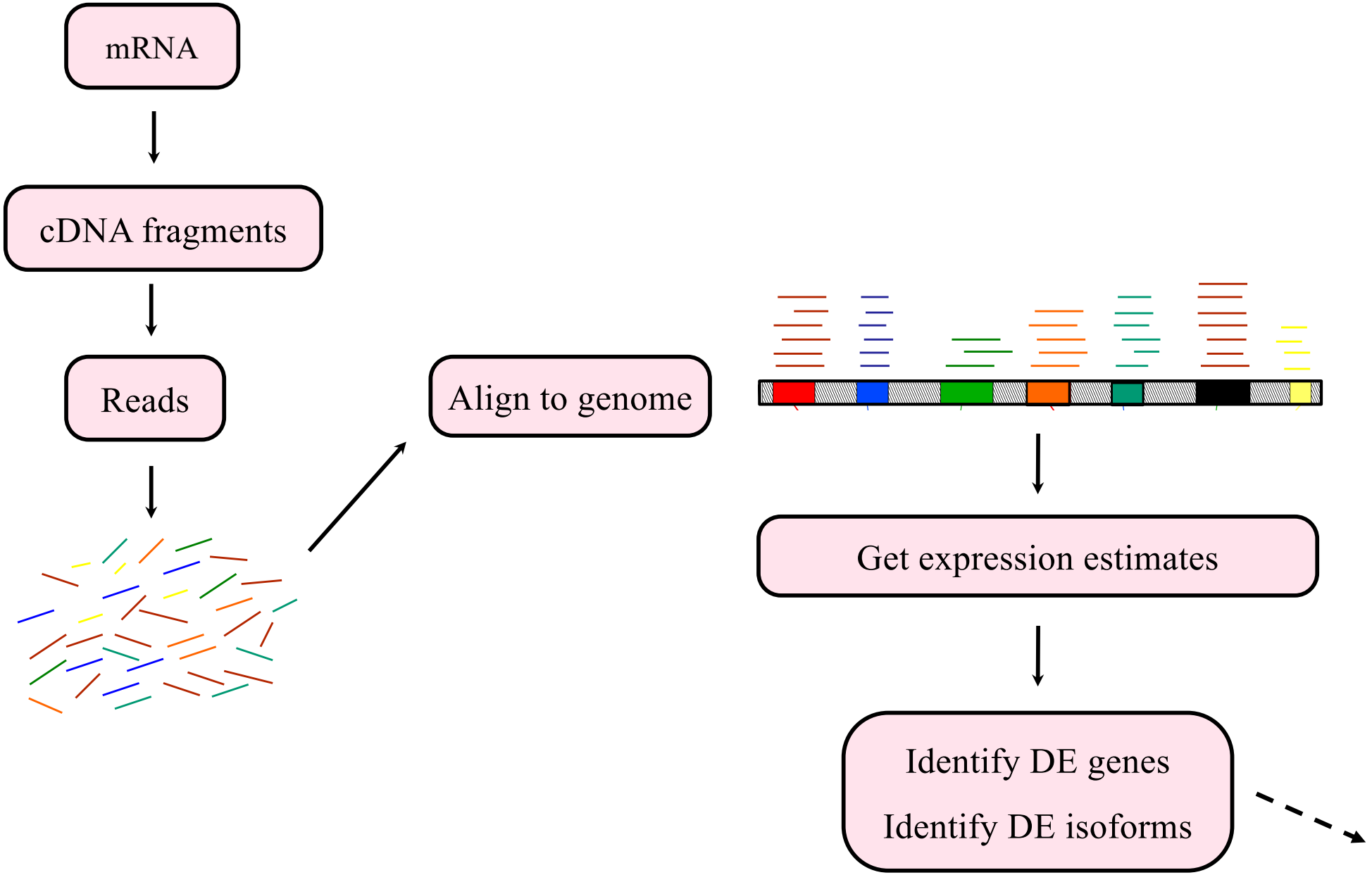
Yu *et al*., Science, 2007

Characterize differences between hESCs and iPSCs

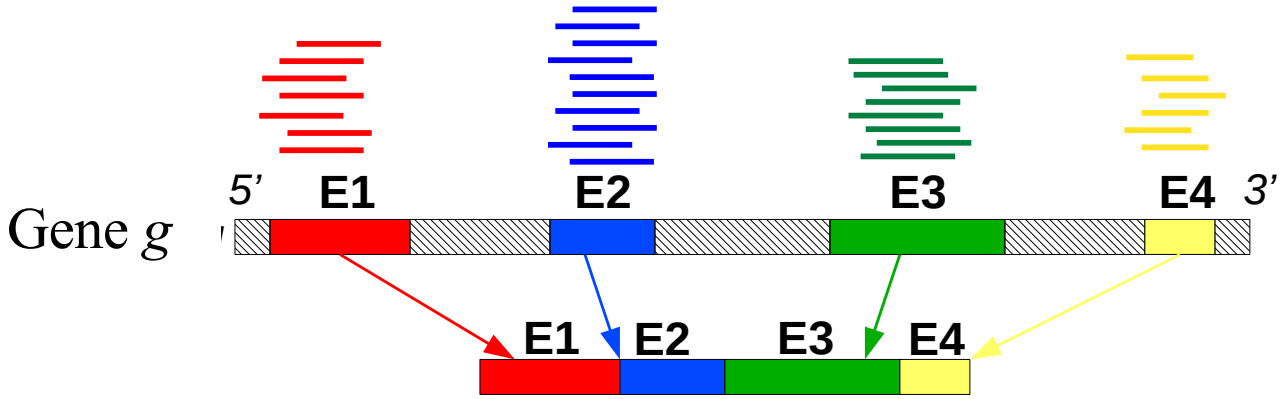
- Mutations
- Methylation
- Protein
  - .
  - .
  - .
- Gene expression
- Isoform expression
  - 4 hESCs vs. 4 iPSCs
  - Illumina Genome Analyzer II ->Bowtie -> RSEM
  - Repeated in triplicate

Can we accurately identify differentially expressed isoforms in an RNA-seq experiment ?

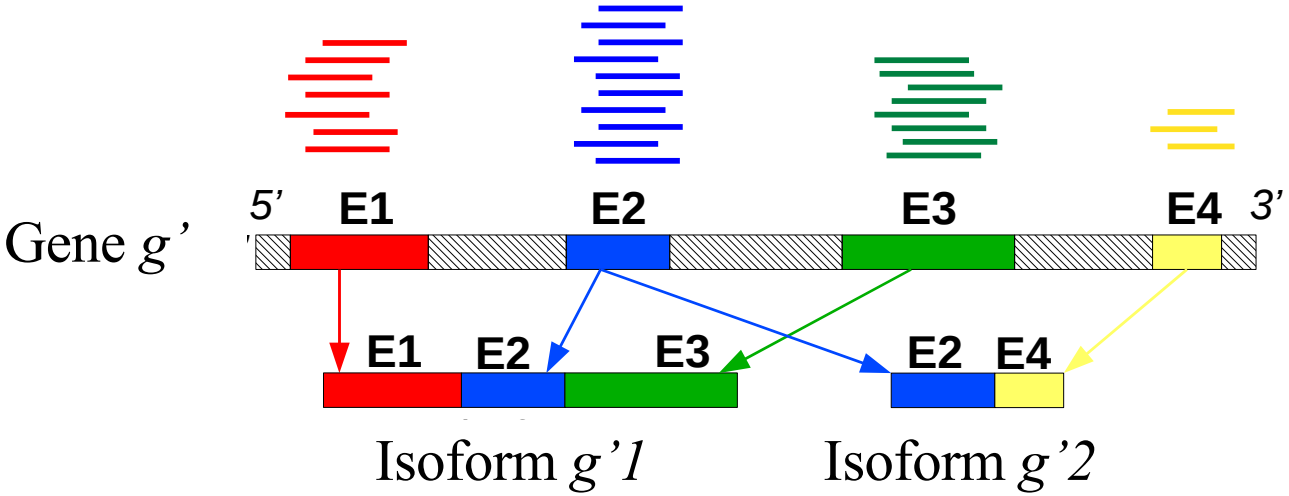
RNA-Seq: Data Collection



Estimating isoform expression

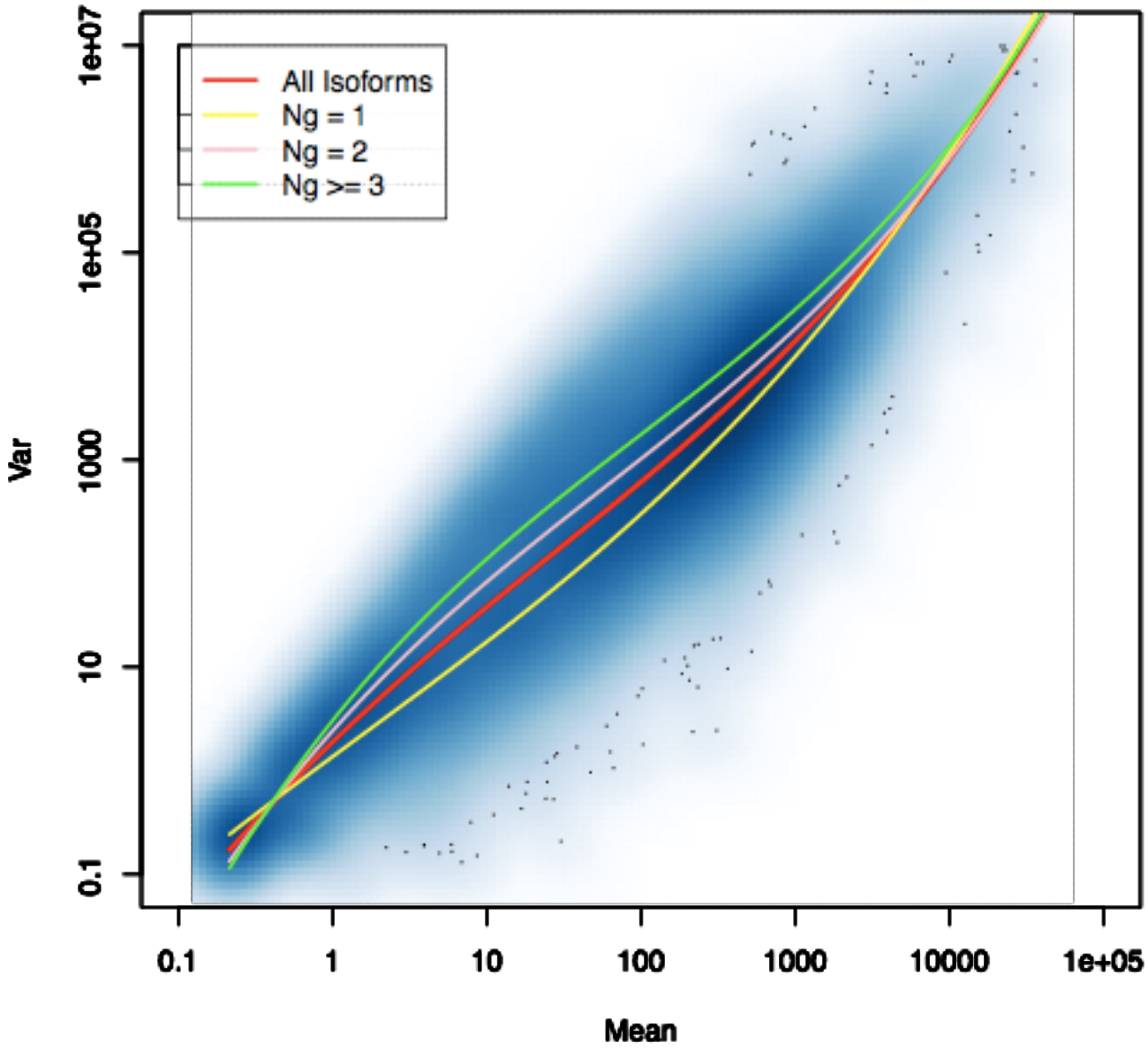


$N_g = 1$

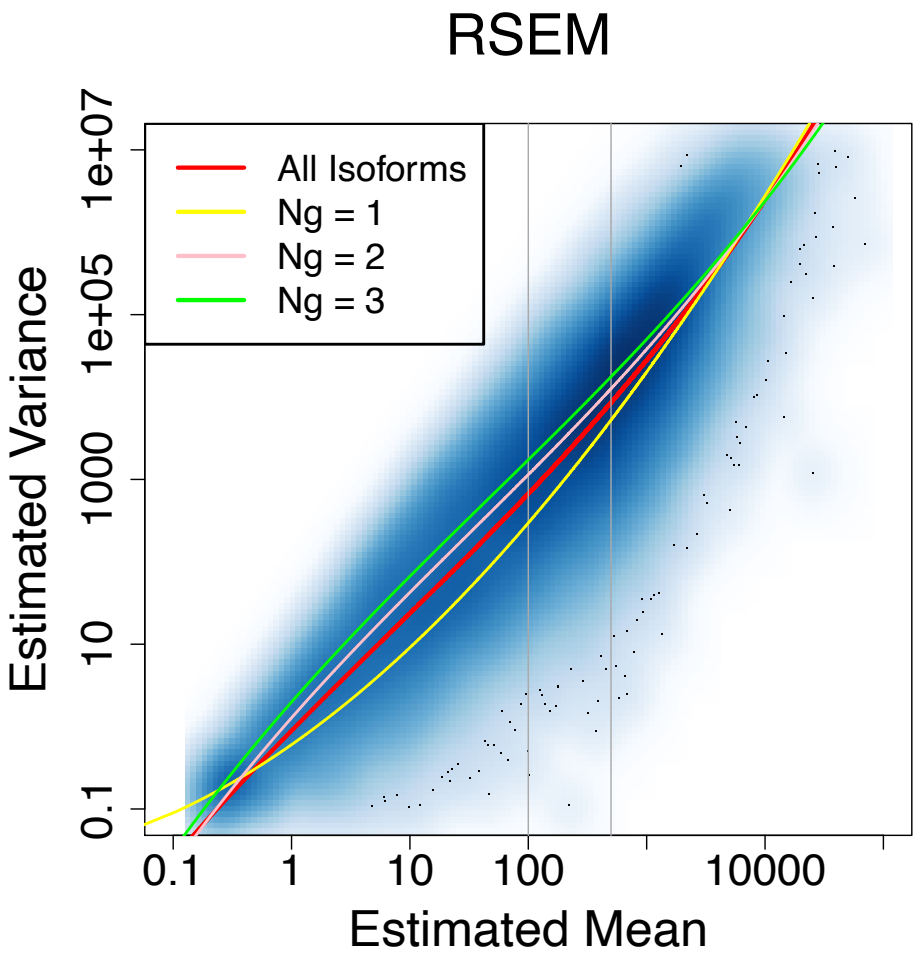


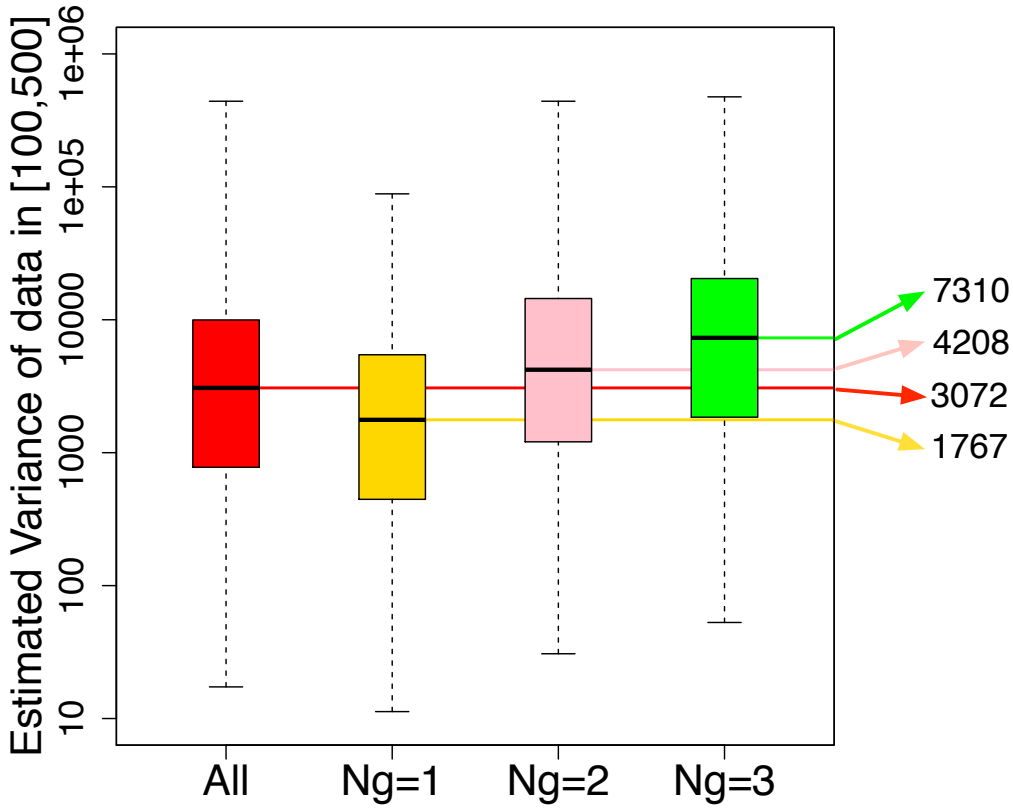
$N_{g'} = 2$

Mean-var relationship changes with N_g

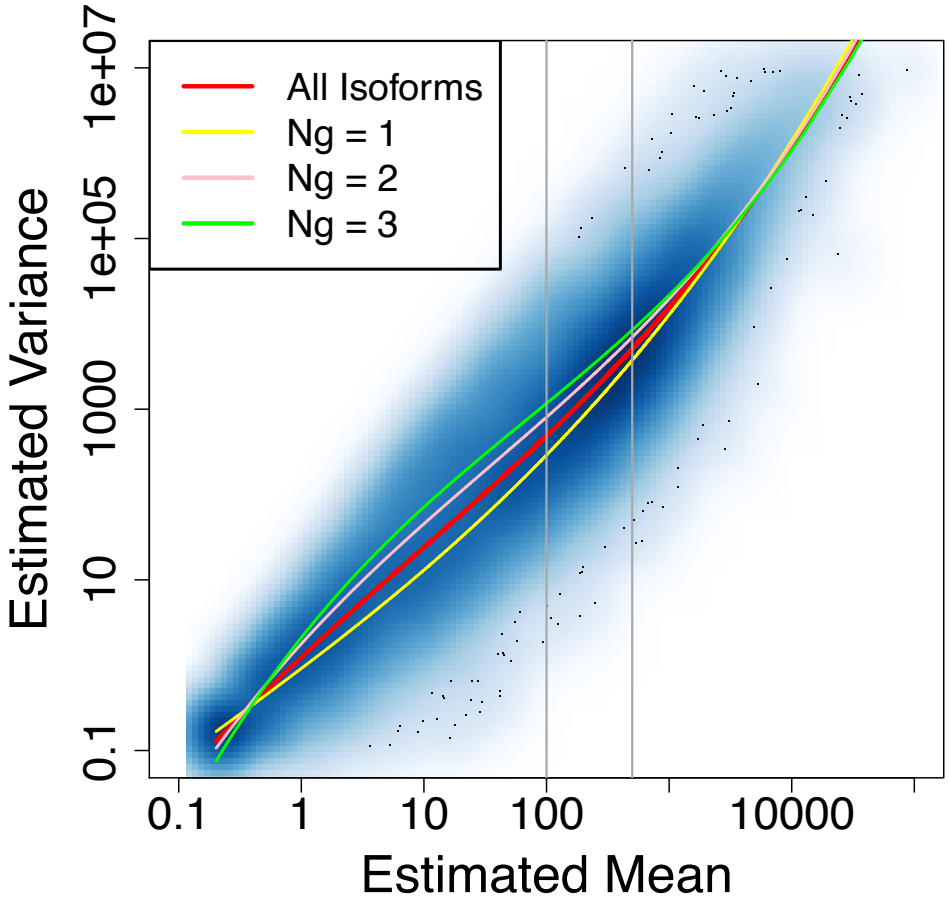


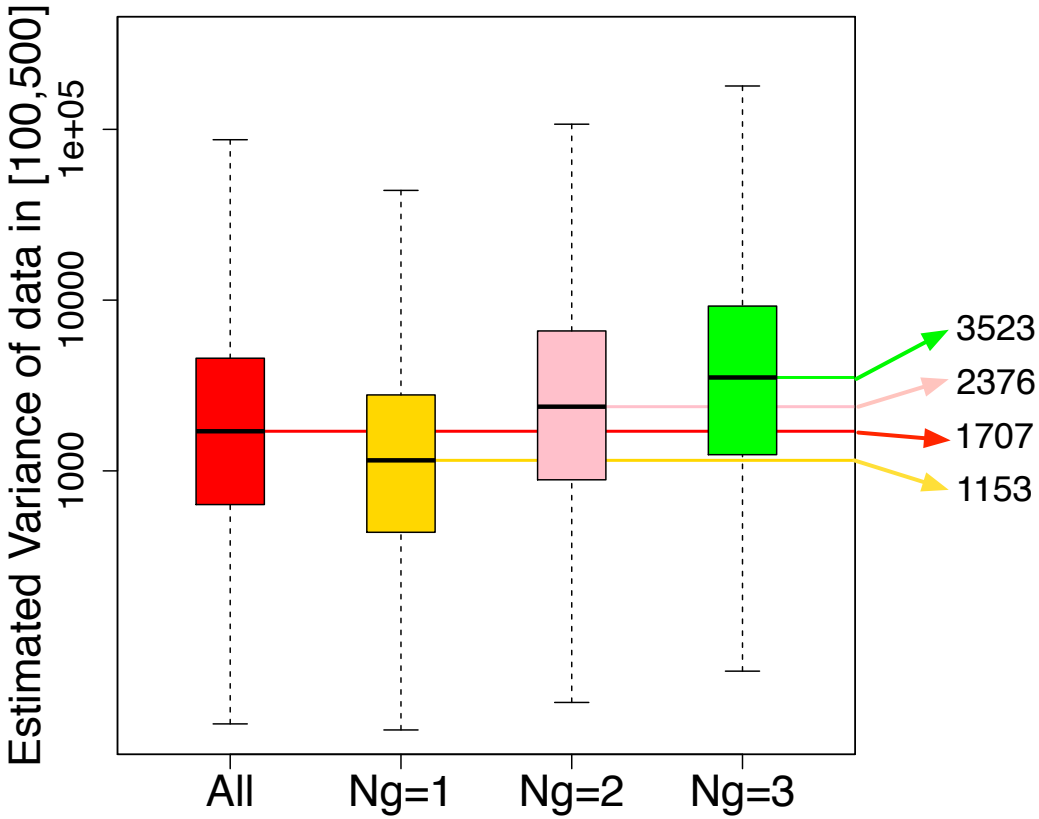
RSEM processed Thomson lab data



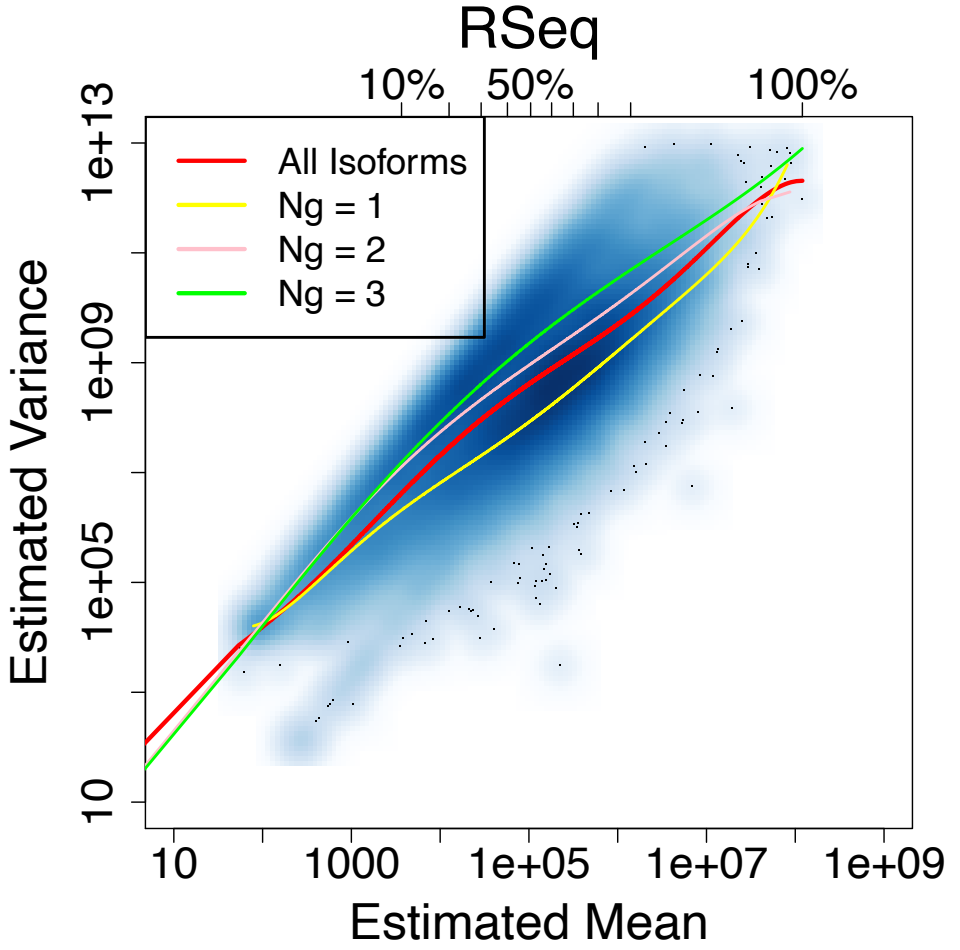


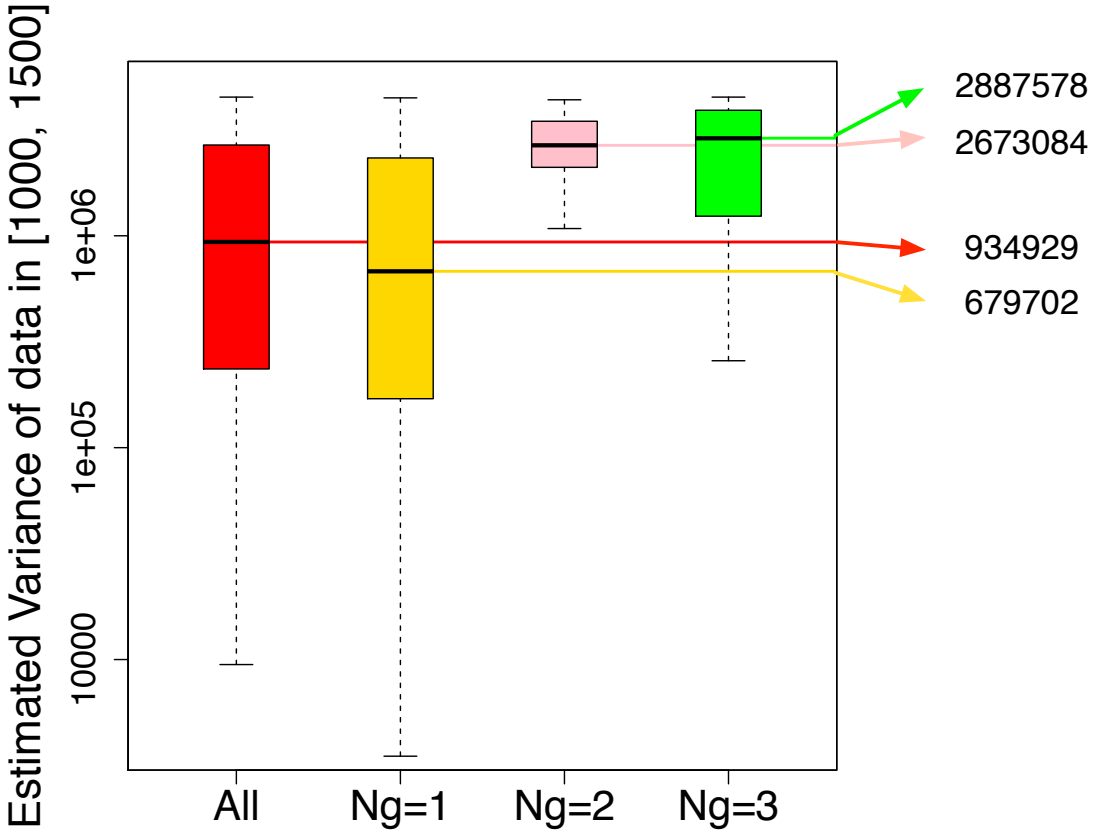
RSEM processed Gould lab data



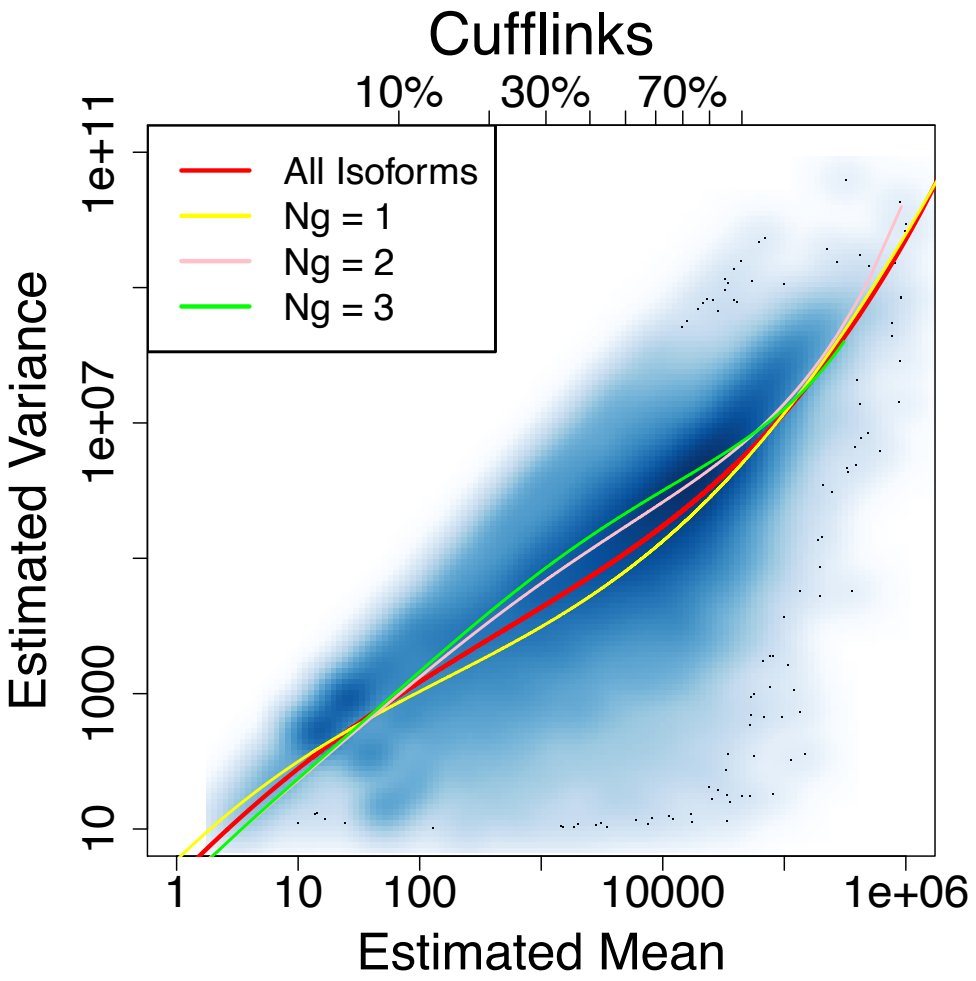


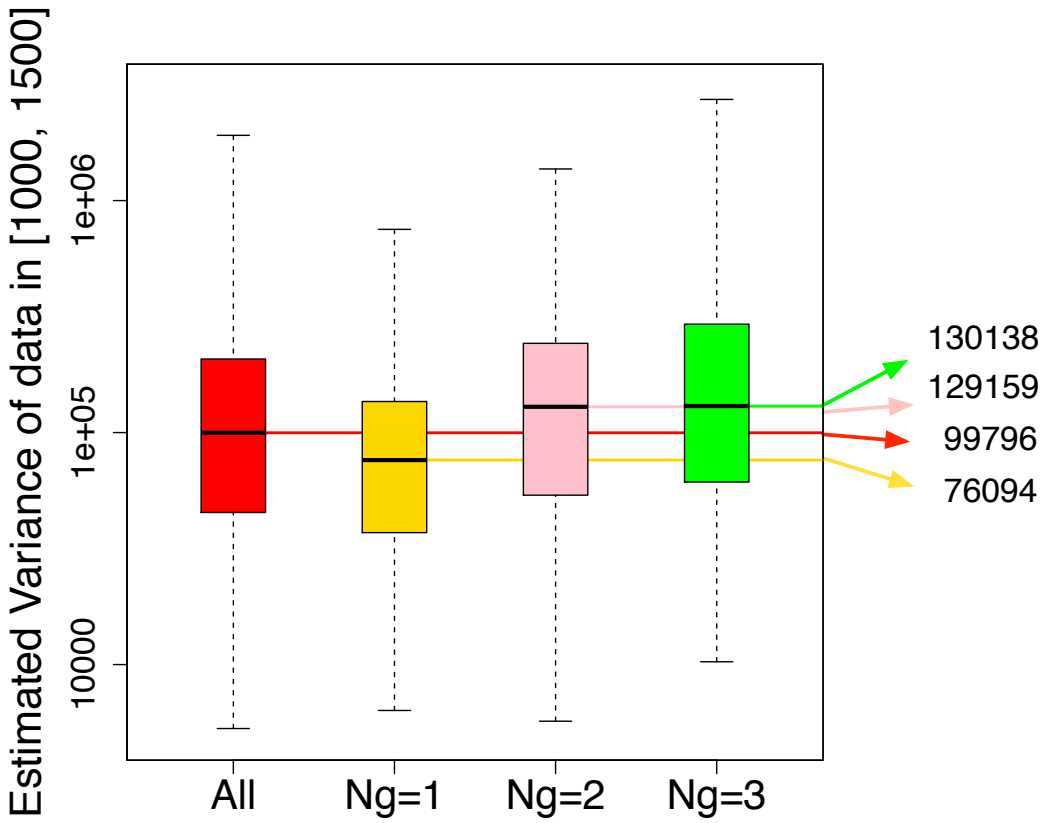
RSeq processed MAQC brain data





Cufflinks processed Wold lab data





RNA-Seq: Methods for identifying DE genes

Each method assumes:

$$X_{gs} \sim NB(\mu_{gs}, \sigma^2_{gs});$$
$$E(X_{gs}) = \mu_{gs};$$
$$Var(X_{gs}) = \sigma^2_{gs}.$$

Condition C; Sample s; Gene g;
Counts X_{gs}; Normalization Factor l_s.

- edgeR: Robinson *et al*. 2007 (still widely used)

$$Var(X_{gs}) = \mu_{gs}(1 + \mu_{gs}\phi_g); \quad \mu_{gs} = l_s \mu_{g0C};$$
Test $H_0 : \mu_{g0C1} = \mu_{g0C2}$

- DESeq: Simon Anders and Wolfgang Huber 2010 (still widely used)

$$Var(X_{gs}) = \mu_{gs} + l_s^2 v_{gC}$$
$$\mu_{gs} = l_s \mu_{g0C}; \quad v_{gC} = v_g(\mu_{g0C})$$
Test $H_0 : \mu_{g0C1} = \mu_{g0C2}$

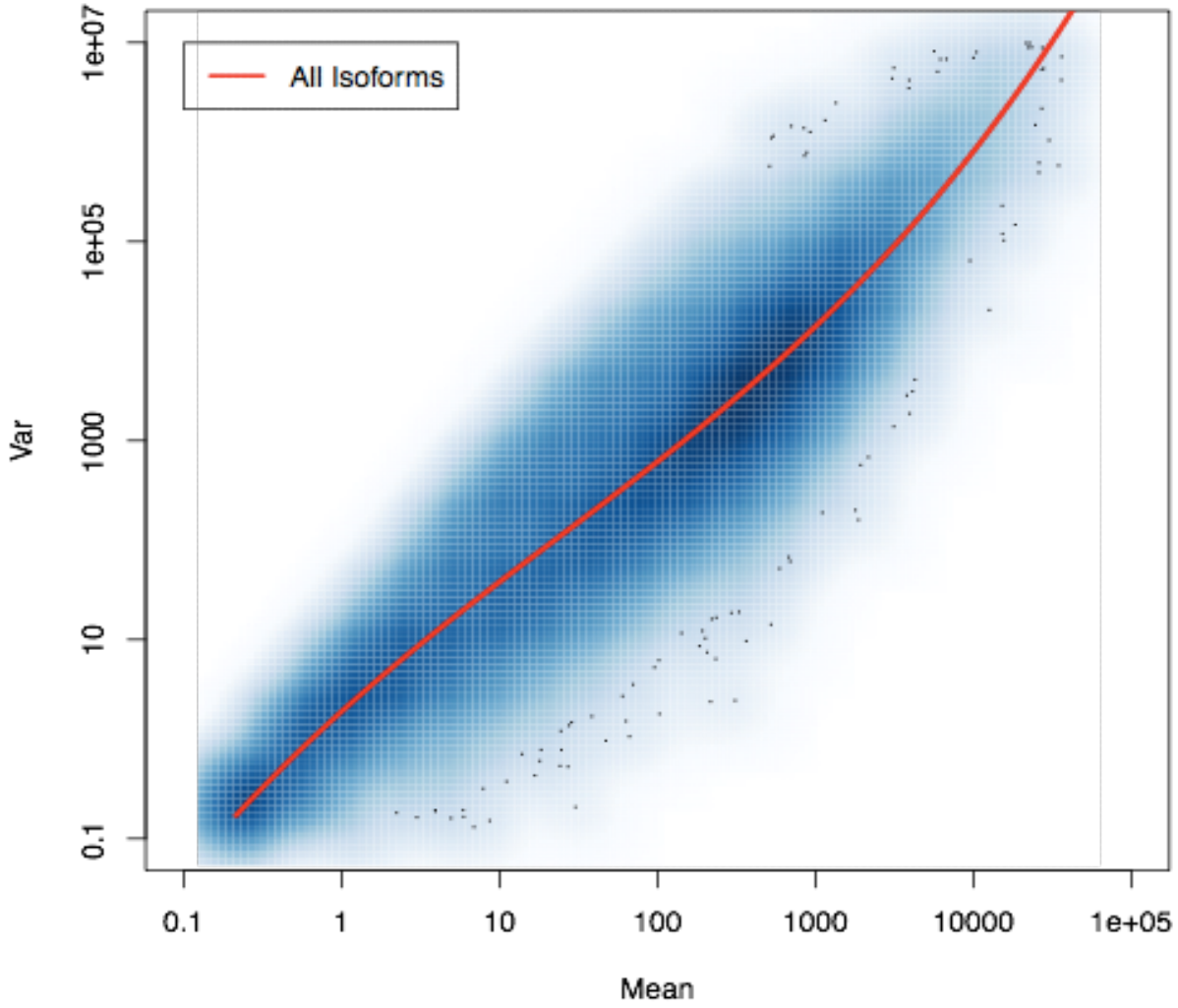
- BaySeq: Thomas J. Hardcastle *et al*. 2010

$$Var(X_{gs}) = \mu_{gs}(1 + \mu_{gs}\phi_g)$$

  – derive an empirical prior from the data

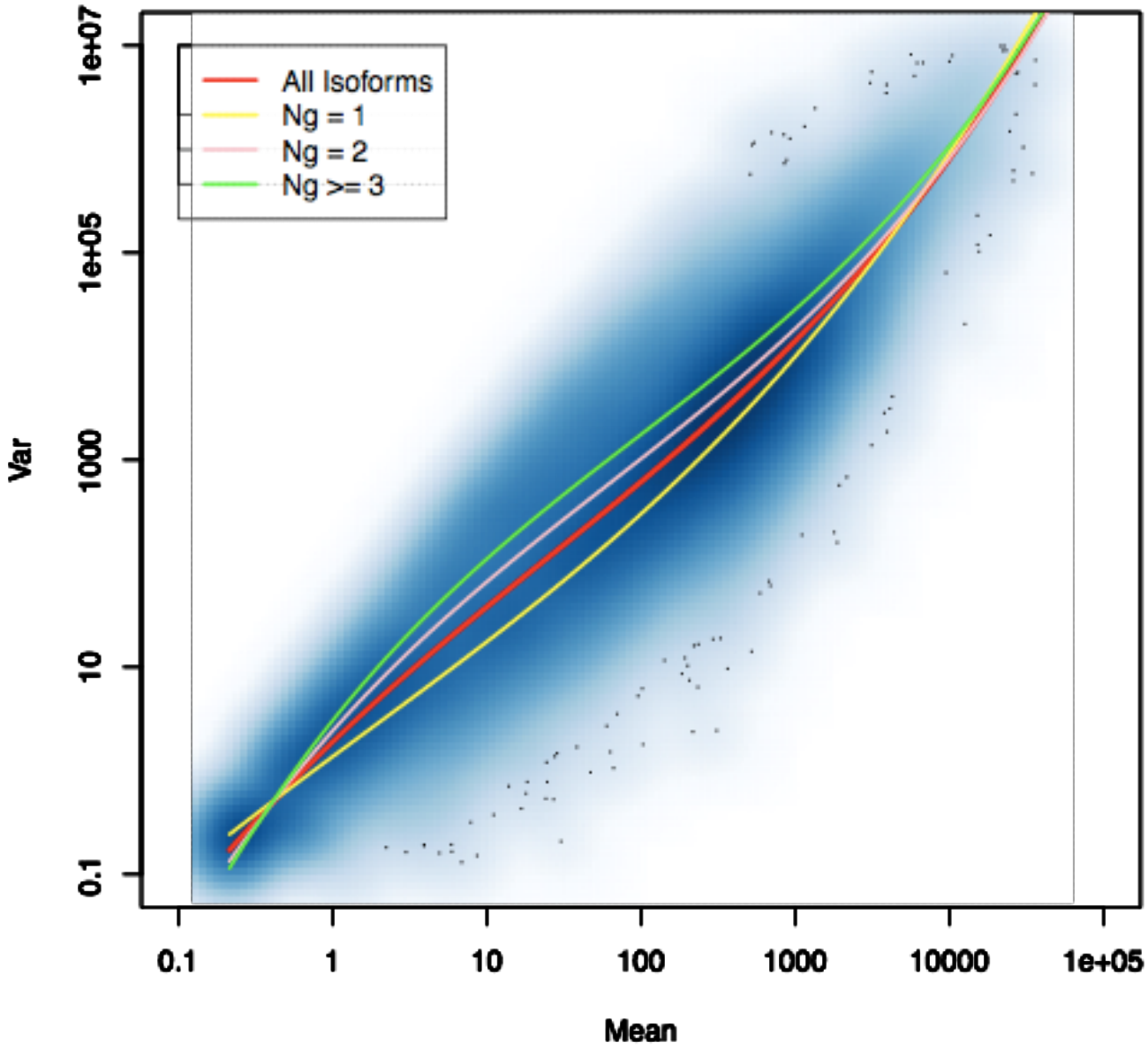
RNA-Seq: Methods for identifying DE genes

Most methods assume $X_{gs} \sim NB(r_g, q_g)$



Mean-variance relationship changes with isoform complexity

Mean-var relationship changes with N_g



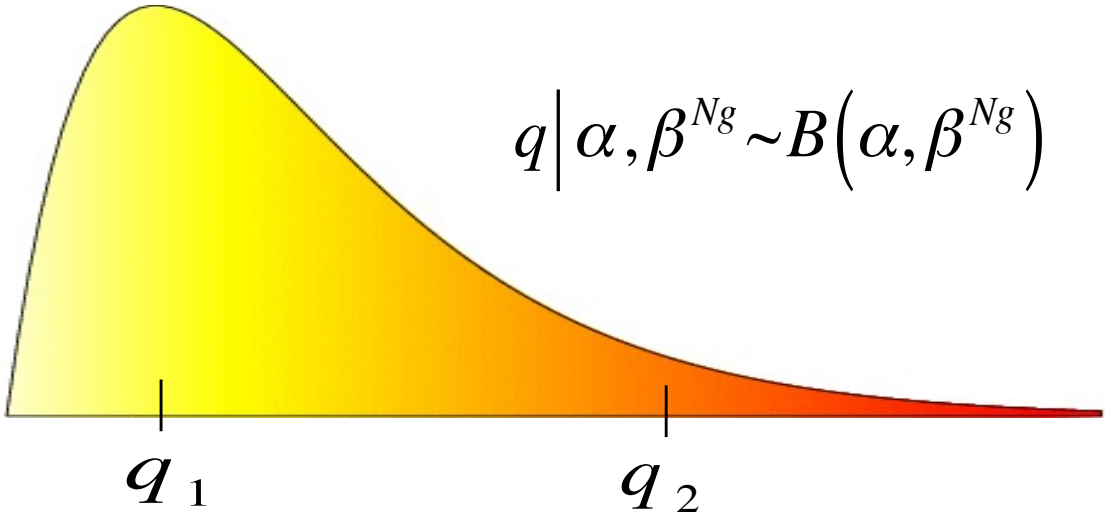
Need to account for the fact that expression estimates for complex isoforms (those from $N_g>1$ genes) have increased variability

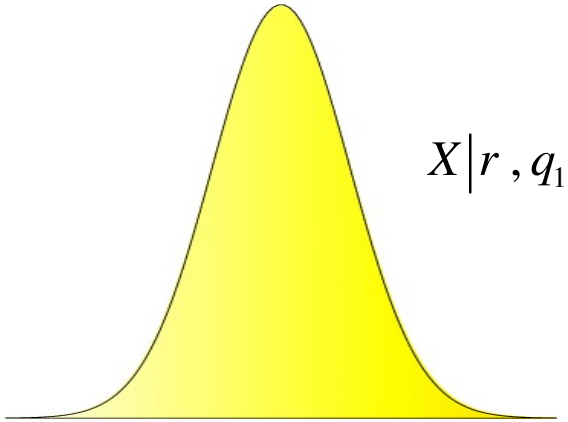
EBSeq: An Empirical Bayes Method for Identifying Differentially Expressed Genes and Isoforms in an RNA-seq experiment

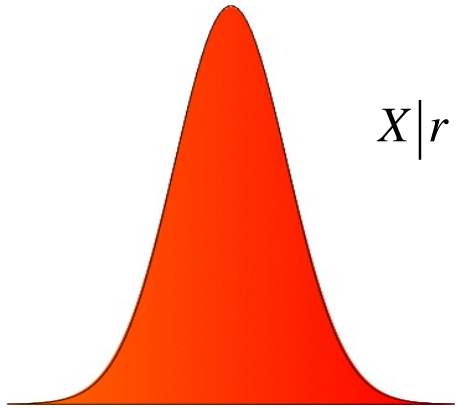
Leng *et al*., *Bioinformatics* 2013

EBSeq: An empirical Bayes NB-Beta Model

$$X|r,q \sim NB(r,q) \text{ and } q|\alpha,\beta^{Ng} \sim B(\alpha,\beta^{Ng})$$







EBSeq

X : Expression of isoform i in gene g and sample s

p_0, p_1 : The prior probability of being EE, DE

Prior depends on Ng

$$X \mid r, q^C \sim NB(r, q^C) \equiv NB\left(\mu = \frac{r(1-q^C)}{q^C}, \sigma^2_{gi,s} = \frac{r\,(1-q^C)}{(q^C)^2}\right); \; q^C \mid \alpha, \beta^{Ng} \sim Beta(\alpha, \beta^{Ng})$$

The isoform is EE if $q^{C1} = q^{C2}$ and DE if $q^{C1} \neq q^{C2}$

Then $X \sim p_0 f_0(X) + p_1 f_1(X)$ where

$$BNB\left(\mu = \frac{r\,\beta^{Ng}}{\alpha - 1}\right)$$

EE: $$f_0(X) = \int \prod_{X_{gi,s} \in X_{gi}} P(X \mid r, q) P(q \mid \alpha, \beta^{Ng}) dq$$

$$f_0\left(X^{C1}\right) f_0\left(X^{C2}\right)$$

DE: $$f_1(X) = \int \prod_{X_{gi,s} \in X^{C1}_{gi}} P(X \mid r, q) P(q \mid \alpha, \beta^{Ng}) dq \int \prod_{X_{gi,s} \in X^{C2}_{gi}} P(X \mid r, q) P(q \mid \alpha, \beta^{Ng}) dq$$

Of primary interest is $P(DE \mid X_{gi}) = \dfrac{p_1 f_1(X)}{p_0 f_0(X) + p_1 f_1(X)}$;

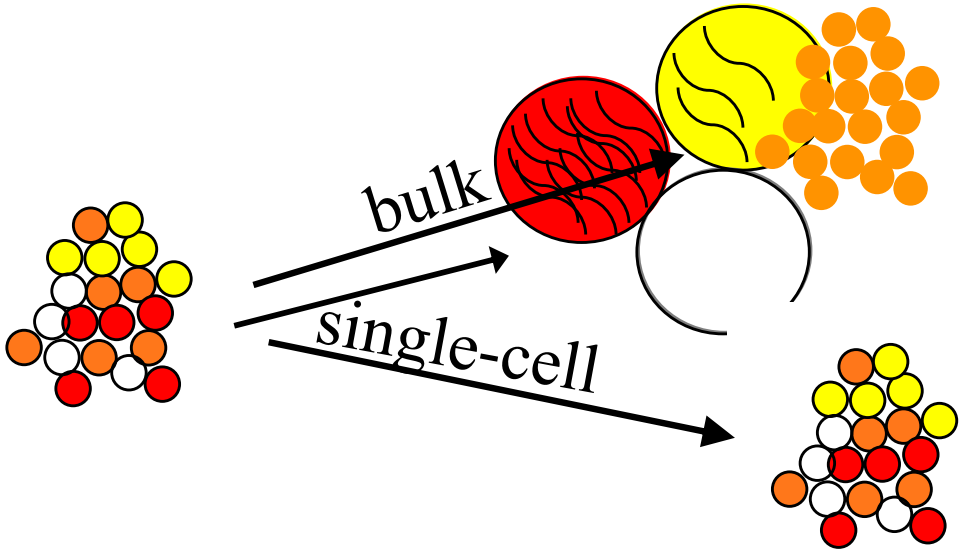
$$P(EE \mid X_{gi}) = \frac{p_0 f_0(X)}{p_0 f_0(X) + p_1 f_1(X)}$$

Summary of EBSeq

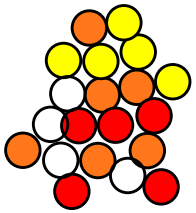
- Methods for identifying DE genes in an RNA-seq experiment do not work well for isoform inference as they do not accommodate uncertainty in isoform expression estimation.
- EBSeq identifies both DE isoforms and genes, accommodates uncertainty, and is robust to outliers.
- EBSeq can be used with more than two conditions, and to quantify EE.
- The approach is in BioConductor, Galaxy, and a GUI. Details are in Leng *et al*., *Bioinformatics*, 2013.
- EBSeq-HMM is in *Bioinformatics*, 2015 (for time course data).

scRNA-seq plus projects

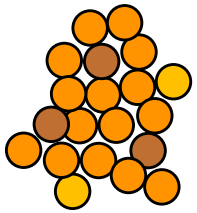
Single-cell vs. bulk RNA-seq



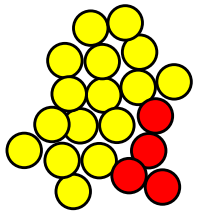
Heterogeneous



Homogeneous

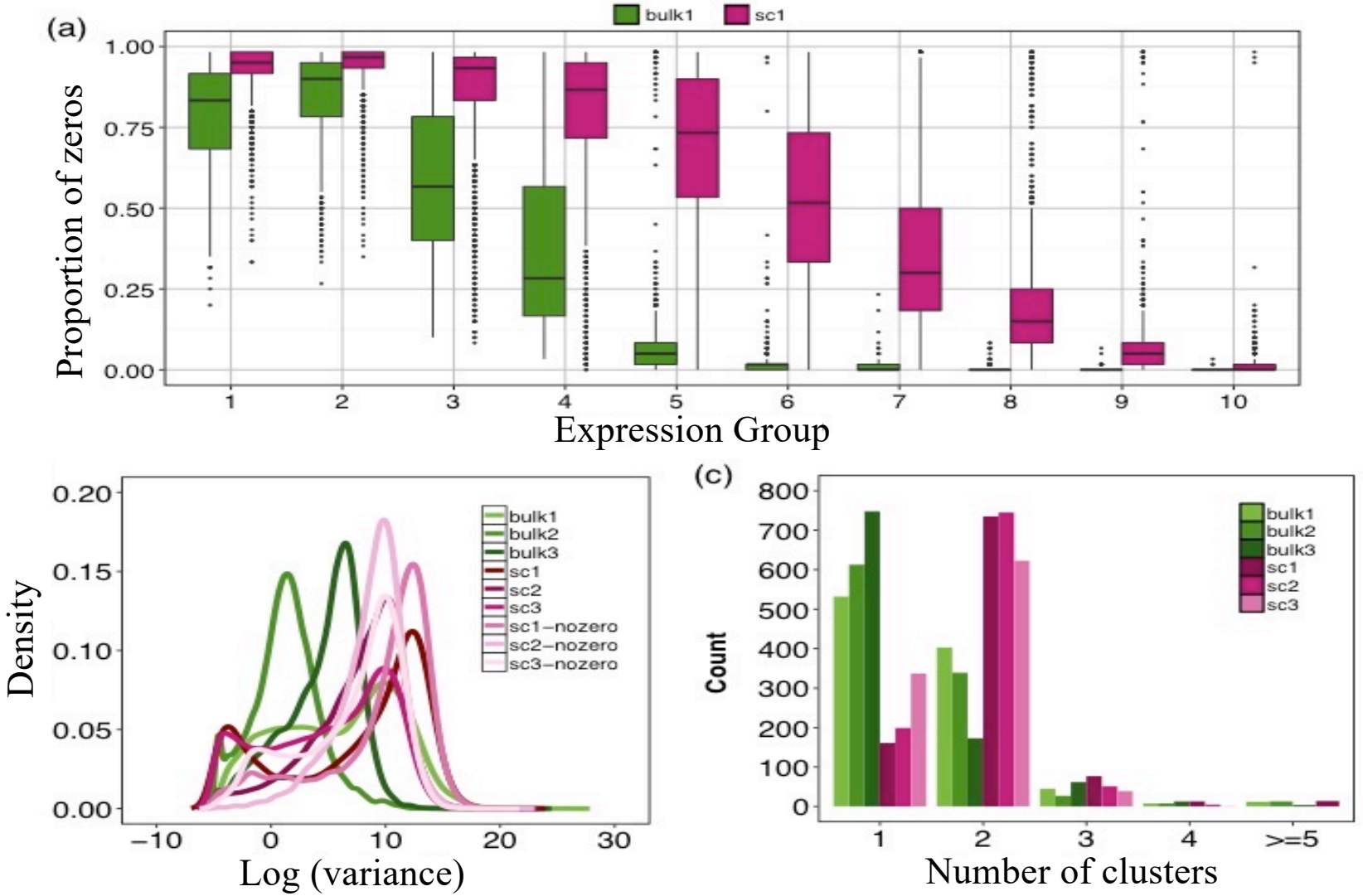


Sub-population



Features of single-cell RNA-seq data

- Abundance of zeros, increased variability, complex distributions



Bacher and Kendziorski, *Genome Biology*, 2016.

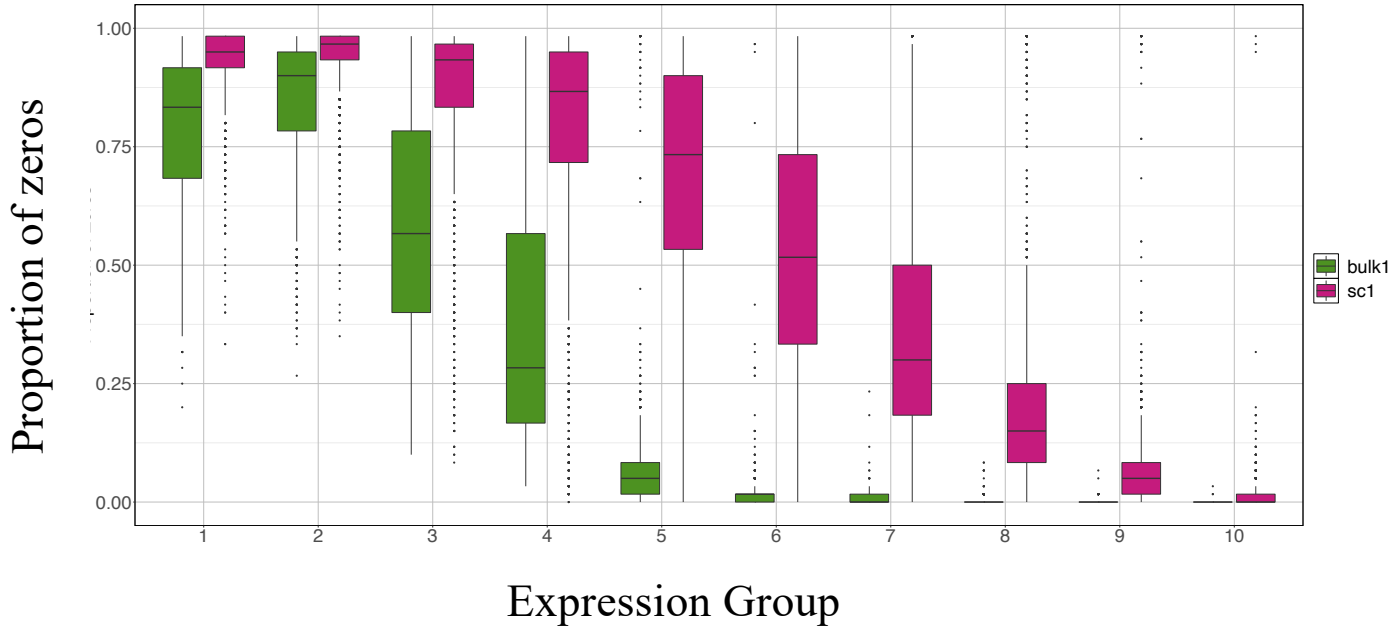
Challenges in scRNA-seq

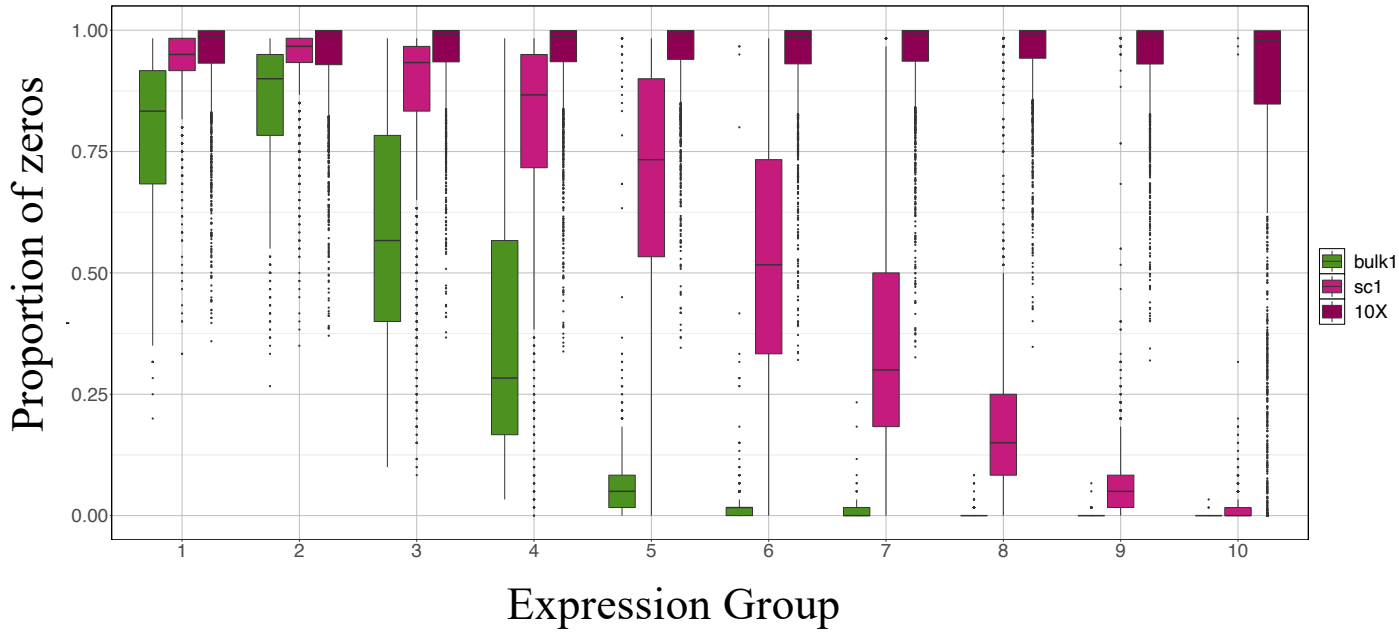
- Normalization
- Technical vs. biological zeros
- Clustering; Identifying sub-populations
- De-noising
    - Adjusting for technical variability
    - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions)
- Pseudotime reordering
- Network reconstruction
- Developmental timing

Challenges in scRNA-seq

- Normalization → Bacher, Chu *et al.*, *Nature Methods*, 2017
- Technical vs. biological zeros
- Clustering; Identifying sub-populations
- De-noising
  - Adjusting for technical variability → Leng *et al*. *Bioinformatics*, 2017
  - Adjusting for biological variability (oscillatory genes) → Leng, Chu *et al.*, *Nature Methods*, 2016
- Identifying and characterizing differences in gene-specific expression distributions (aka. identifying differential distributions) → Korthauer *et al.*, *Genome Biology*, 2016
- Pseudotime reordering → Chu, Leng *et al.*, *Genome Biology*, 2016
- Network reconstruction
- Developmental timing → Brown *et al.*, *PLoS Comp Bio*, 2021

Features of single-cell RNA-seq data





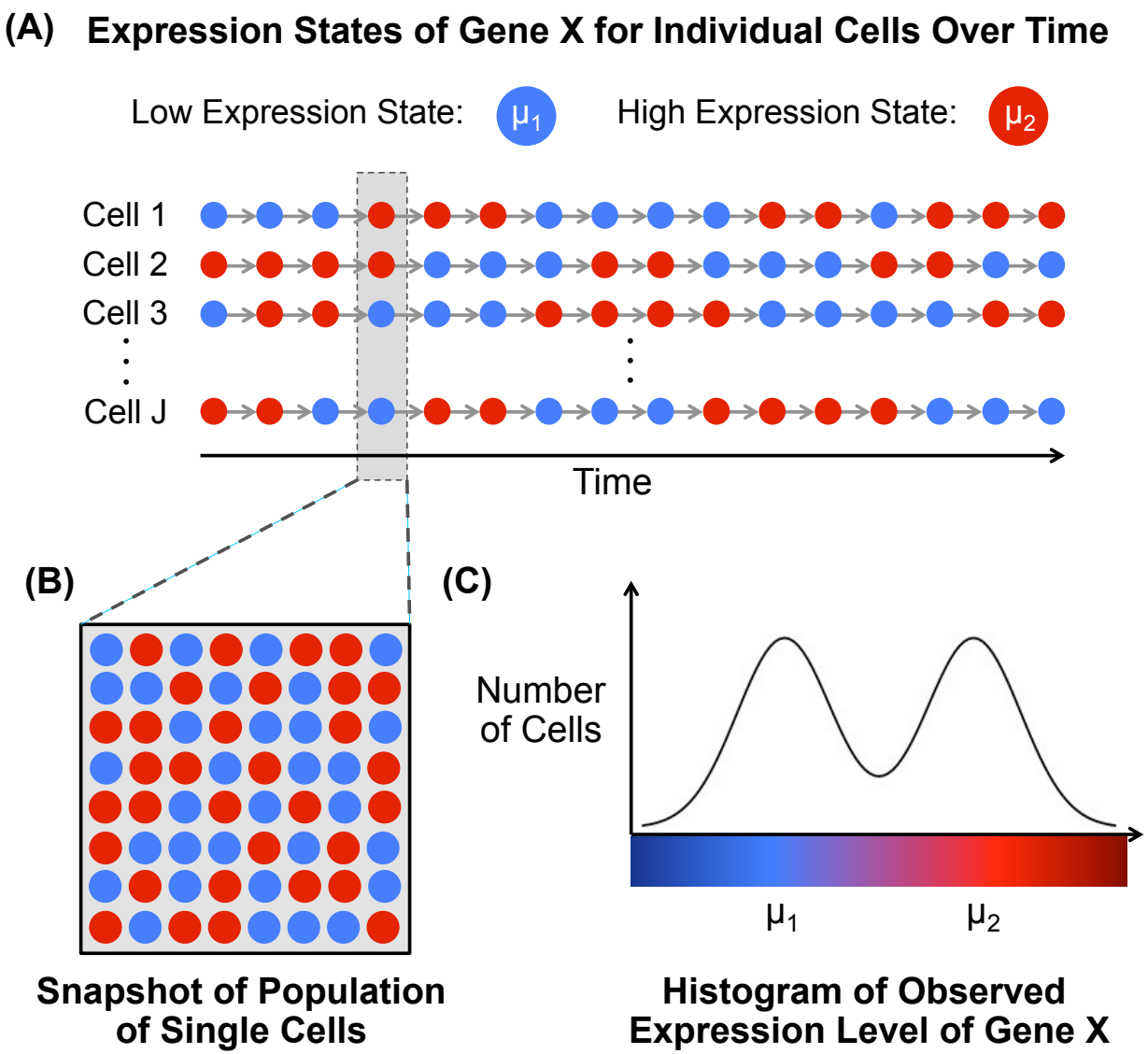
Challenges in scRNA-seq generated using 10X genomics

- Protocol improvement
- Distinguishing real cells from background barcodes
- Normalization
- Clustering; Identifying sub-populations
- De-noising
  - Adjusting for technical variability
  - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differential distributions
- Pseudotime reordering
- Data integration and visualization

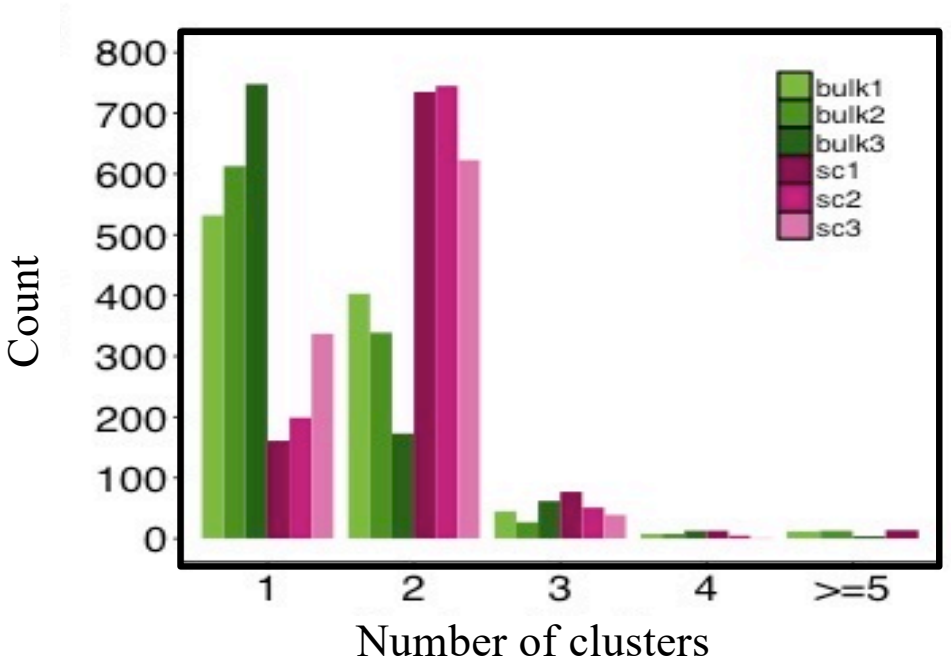
Challenges in scRNA-seq generated using 10X genomics

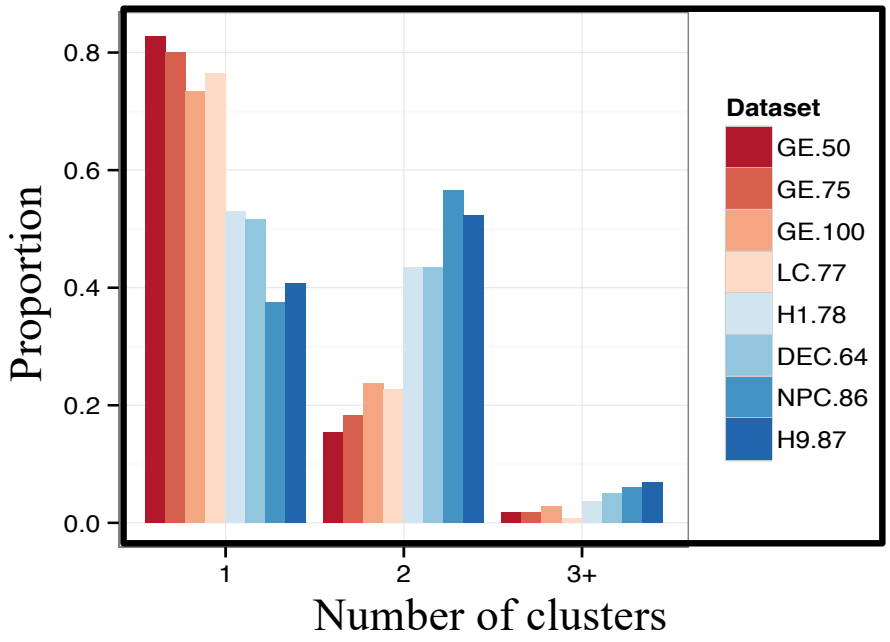
- Protocol improvement → Bacher *et al., Nucleic Acids Research, 2021*
- Distinguishing real cells from background barcodes → CB2: Ni, *et al.*, *Genome Biology*, 2020
- Normalization → Dino: Jared Brown *et al., Bioinformatics 2021*
- Clustering; Identifying sub-populations
- De-noising
  - Adjusting for technical variability
  - Adjusting for biological variability (oscillatory genes)
- Identifying and characterizing differential distributions → scDDboost: Ma *et al.*, *Annals of Applied Statistics,* 2021
- Pseudotime reordering
- Data integration and visualization → CHARTS: Bernstein *et al.*, *BMC Bioinformatics*, 2021

Gene-specific multi-modality



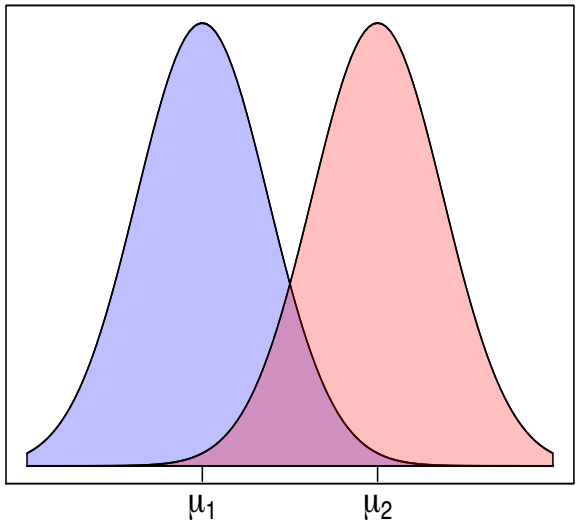
Many genes show multi-modal expression distributions



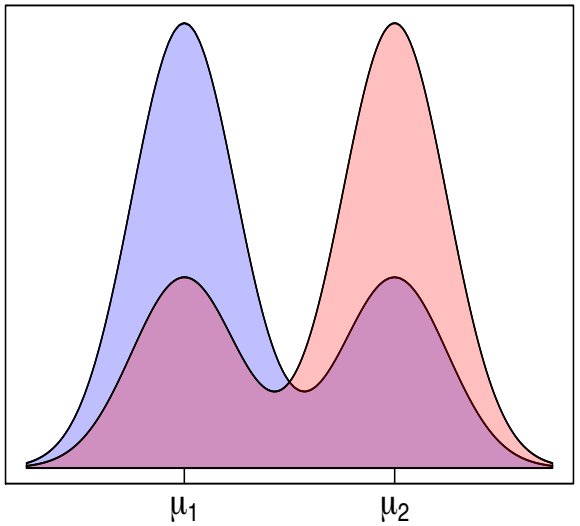


Opportunity to identify differences beyond traditional DE

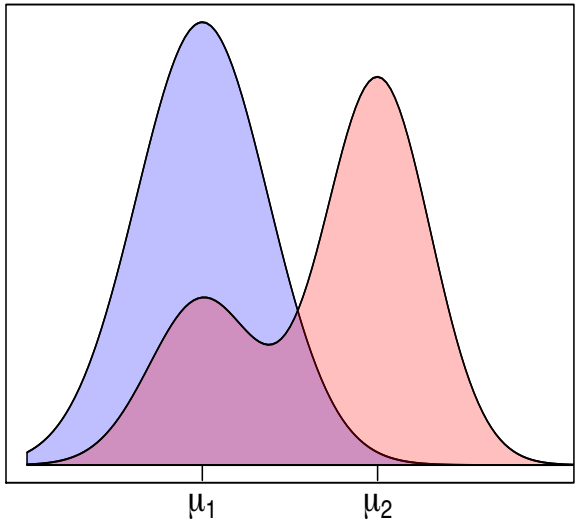
Differential expression (DE)



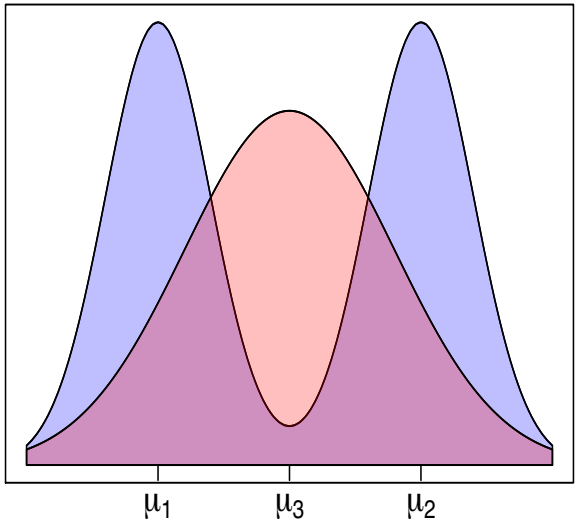
Differential proportions (DP)



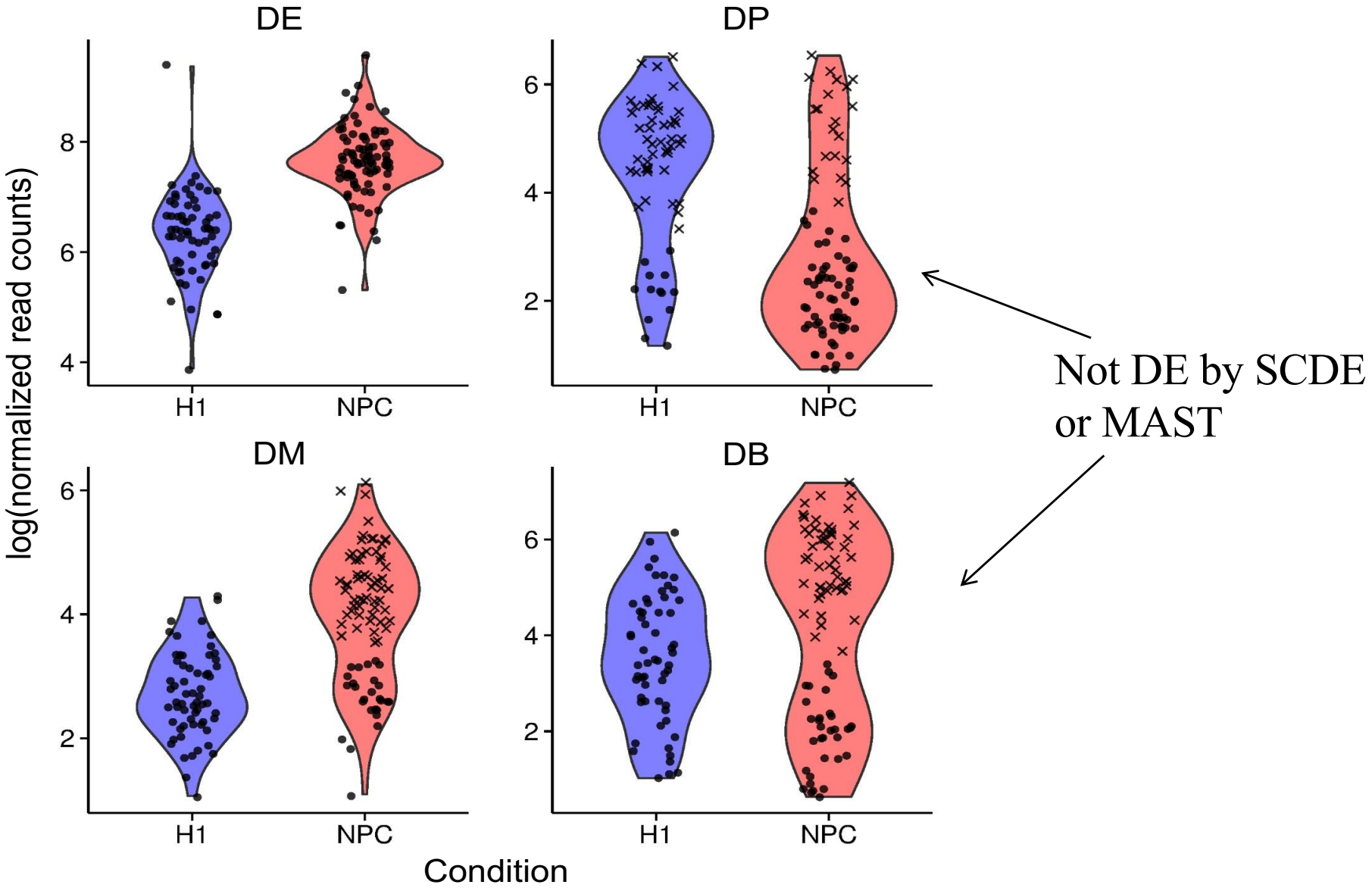
Differential modes (DM)



Both DM and DE*



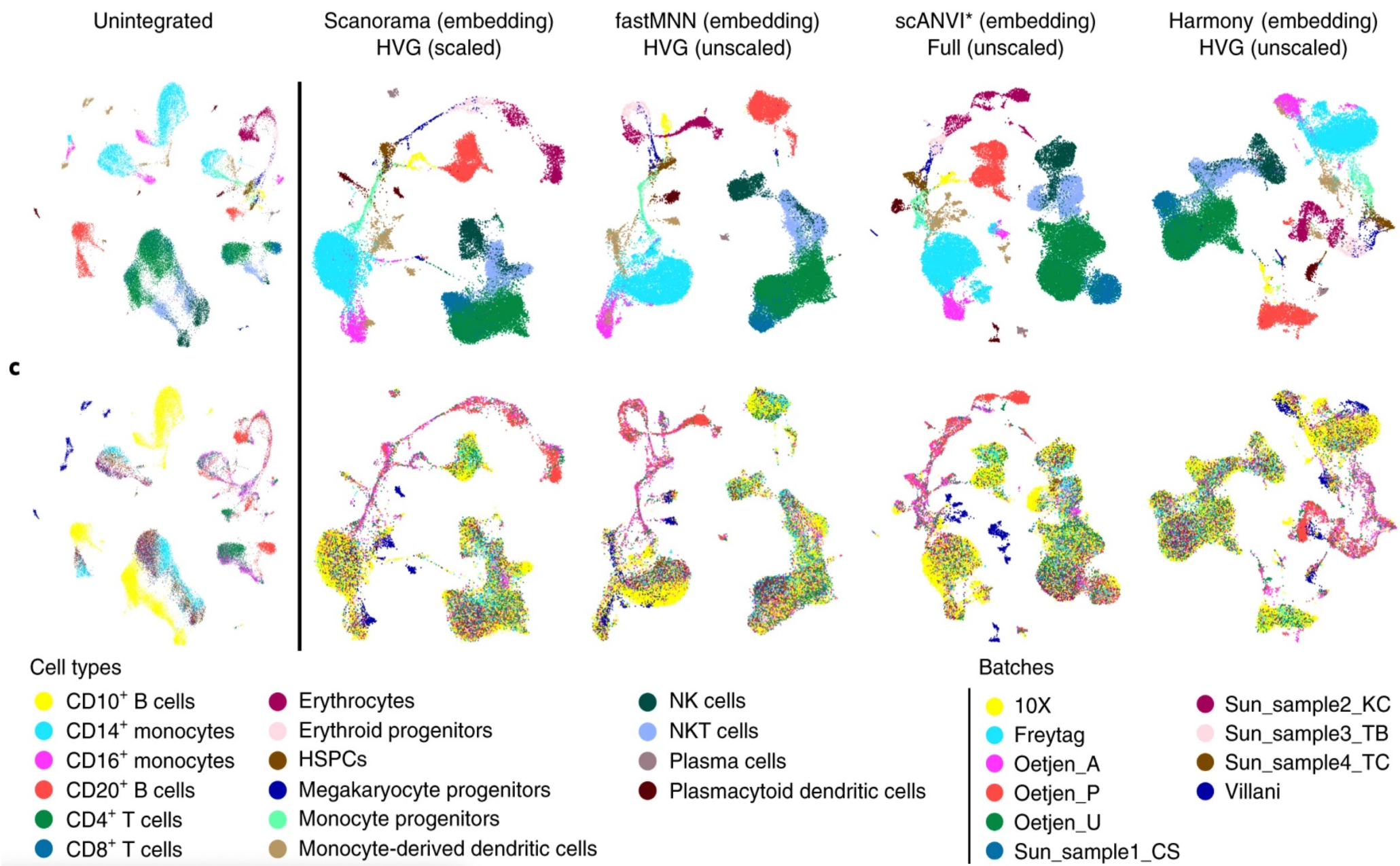
Genes identified in H1 vs. NPC comparison



Single-cell RNA-seq DE project

- Using single-cell expression data with at least four sub-populations (and, optionally, simulated data), use the methods in the Seurat pipeline to identify DE genes across multiple sub-populations. How are multiple comparisons handled? Data and code are at the course website.
- The code easily applies to the provided data, but it is not quite "copy-and-paste". You should, however, be able to format the provided data to fit this problem, and it allows some flexibility in how to run the analysis. Questions? Ask CK and/or ZN.
- You'll have to read documentation in the Seurat pipeline to see what methods they use for DE analysis, and how they handle multiple comparisons. Is it ideal? Hint: no.
- What methods will you evaluate, and how? Is there a better way to handle multiple comparisons across methods?

Single-cell RNA-seq data integration project (Figure 2 from Leucken et al., *Nature Methods*, 2022)

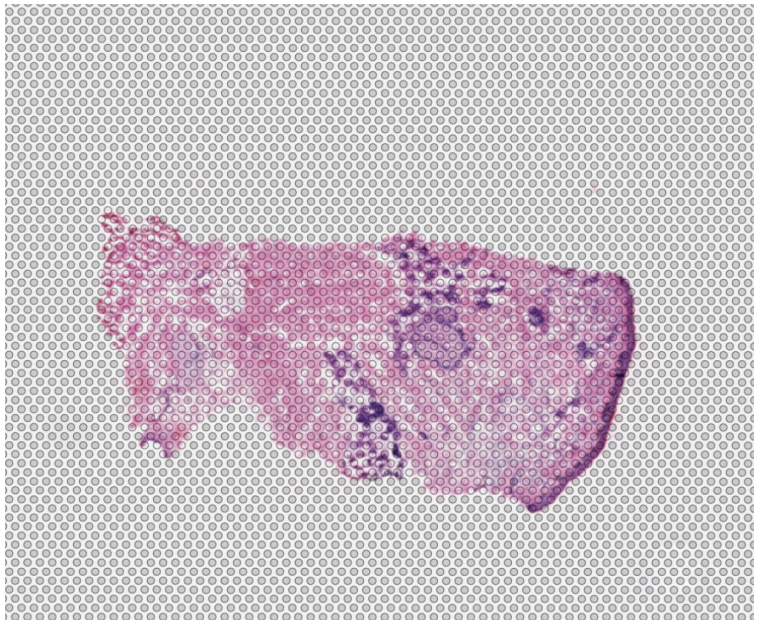


Single-cell RNA-seq data integration project

- Luecken *et al*., *Nature Methods*, 2021 benchmarked single-cell data integration methods. For the sake of comparison, they chose to keep the normalization method the same for all the methods. Moreover, they use default parameters in their analyses. This project will focus on finding the best result from each of the methods.
- Specifically, choose 4 or more samples and integrate them using 3 or more methods. Compare the results. The authors use results from simulated data as well as visual inspection of real data in their evaluations. It is recommended that you do the same and/or justify additional techniques for evaluating each method's performance.

Data from a spatial transcriptomics experiment

H&E Image



UMI counts
S spots

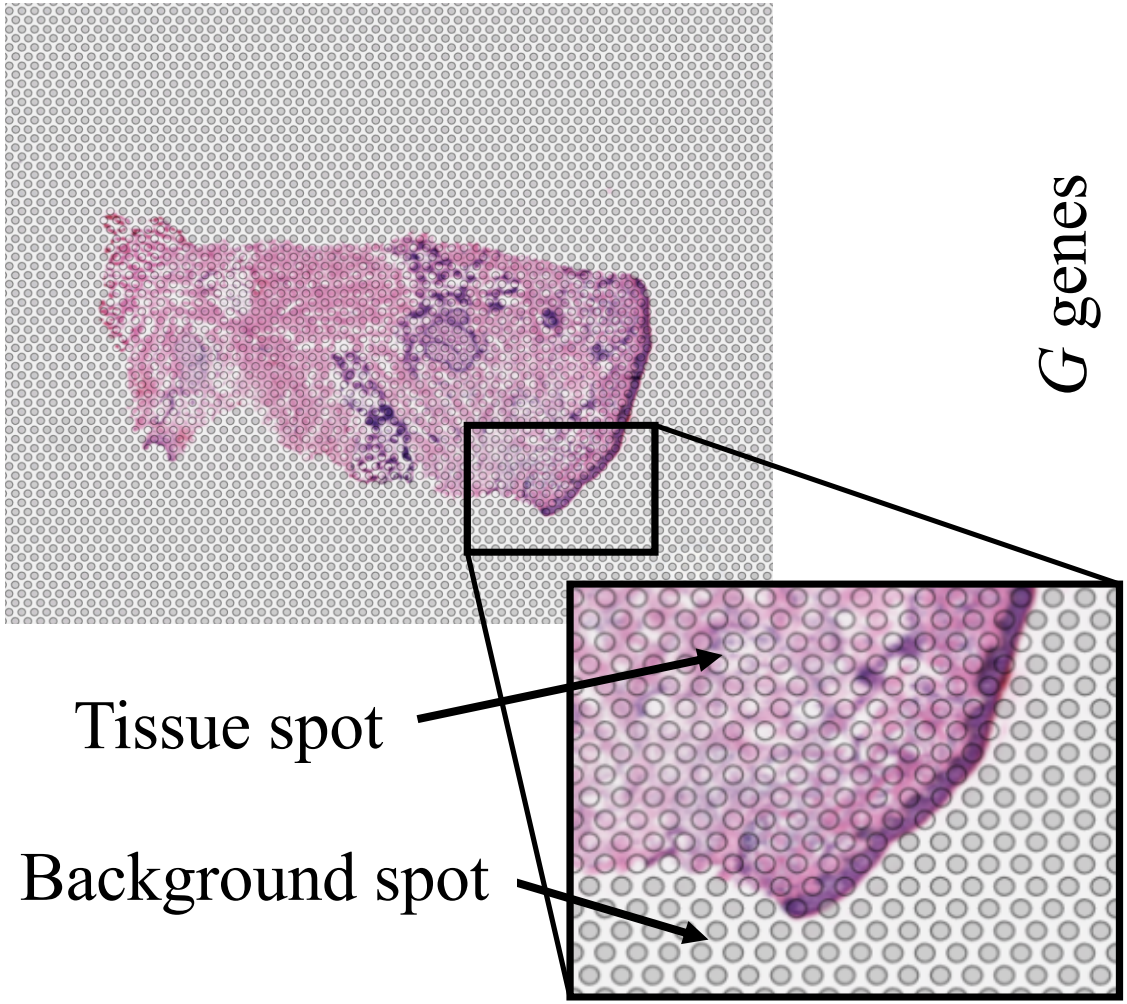
G genes

$$\begin{bmatrix} X_{gj} \end{bmatrix}$$

X_{gj} : UMI count of gene g at spot j

Data from a spatial transcriptomics experiment

H&E Image



UMI counts
S spots

tissue bg

G genes

$$\begin{bmatrix} X_{gj} & \vdots & \end{bmatrix}$$

X_{gj} : UMI count of gene g at spot j (should be 0 for background spots)

Method to watch

Published: 03 January 2018

Method to Watch: Spatial Transcriptomics

Tal Nawy

Nature Methods **15**, 30 (2018) | Cite this article

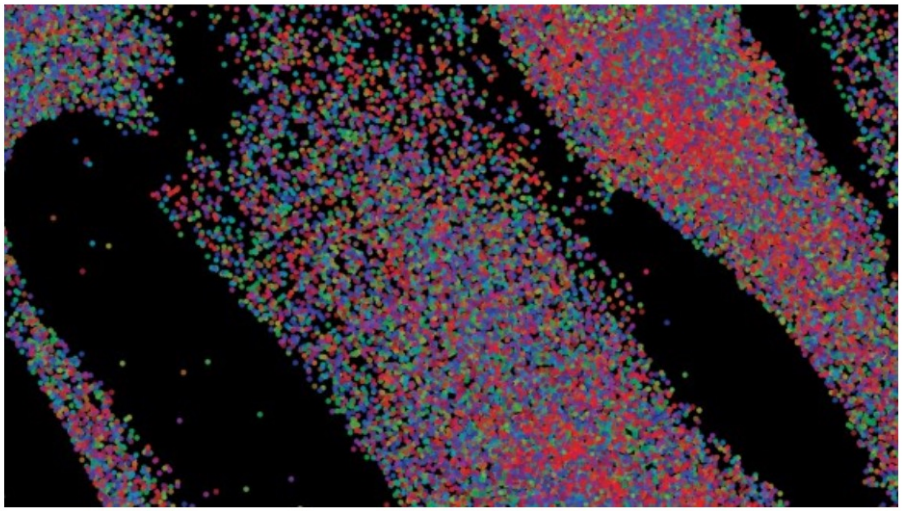
"We anticipate that improvements in data generation and analysis will bring spatial transcriptomics into wider practice and will be transformative for biology."

Applications of spatial transcriptomics

FOCUS | 06 JANUARY 2021

Method of the Year 2020: spatially resolved transcriptomics

Spatially resolved transcriptomics is our Method of the Year 2020, for its ability to provide valuable insights into the biology of cells and tissues while retaining information about spatial context.



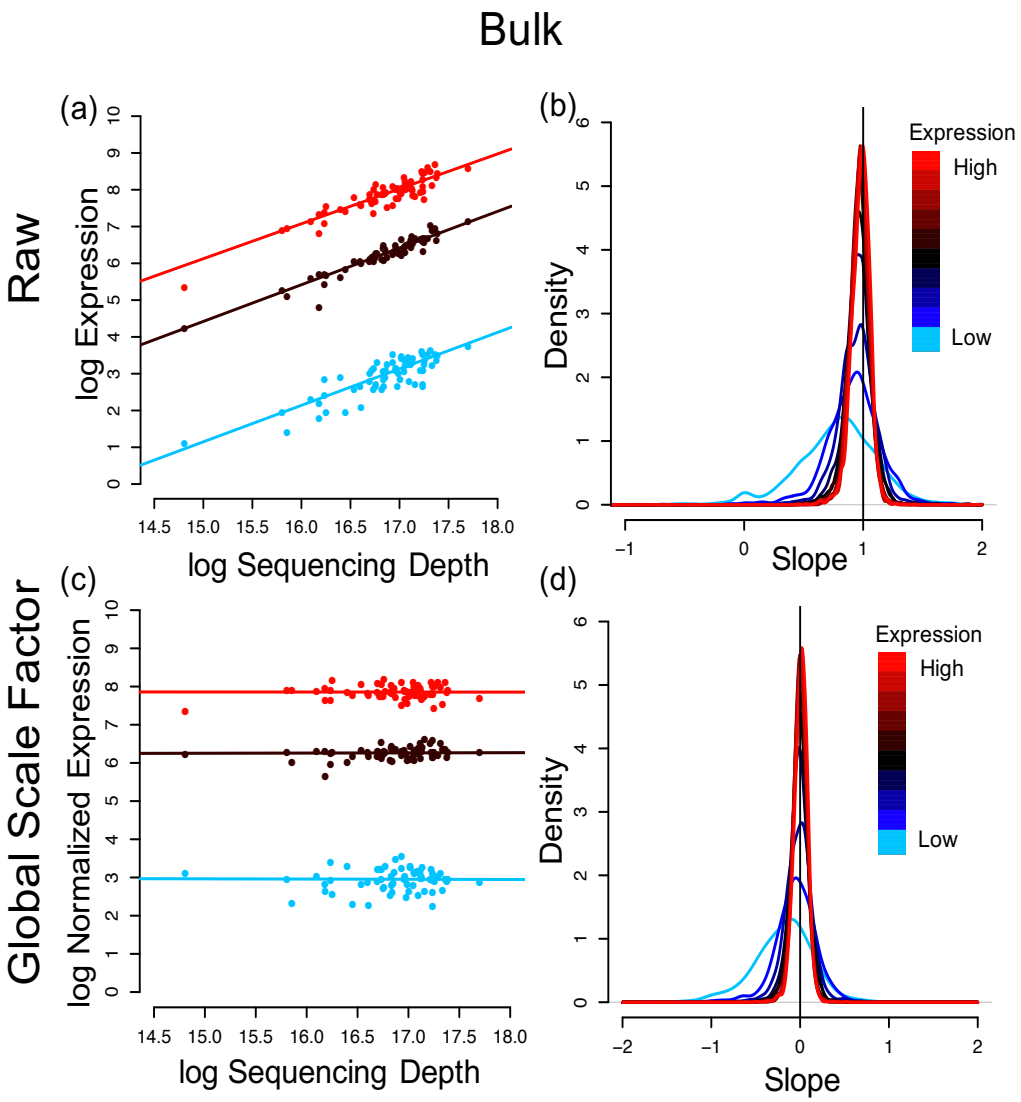
Method of the Year 2020: spatially resolved transcriptomics

Spatially resolved transcriptomics methods are changing the way we understand complex tissues.

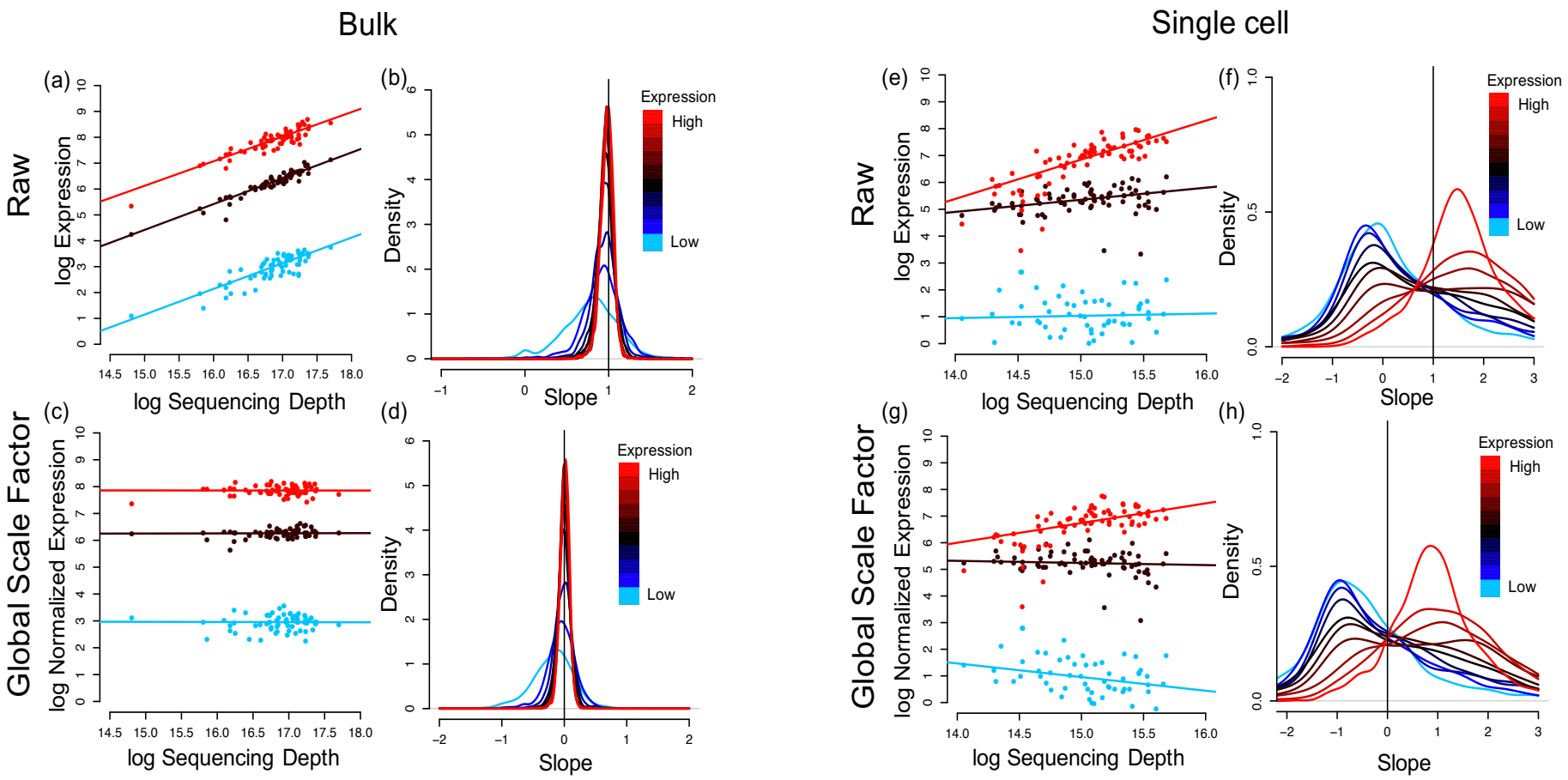
Normalization for single-cell RNA-seq data

- Goal: correct for technical artifacts and/or gene-specific features
  - Sequencing depth
  - Length, GC content
  - Amplification and other technical biases
- Most single-cell normalization methods calculate global scale factors (one or a few) as in bulk RNA-seq
  - One scale factor is calculated per sample and applied to all genes in that sample.
- No methods specifically for spatial transcriptomics data

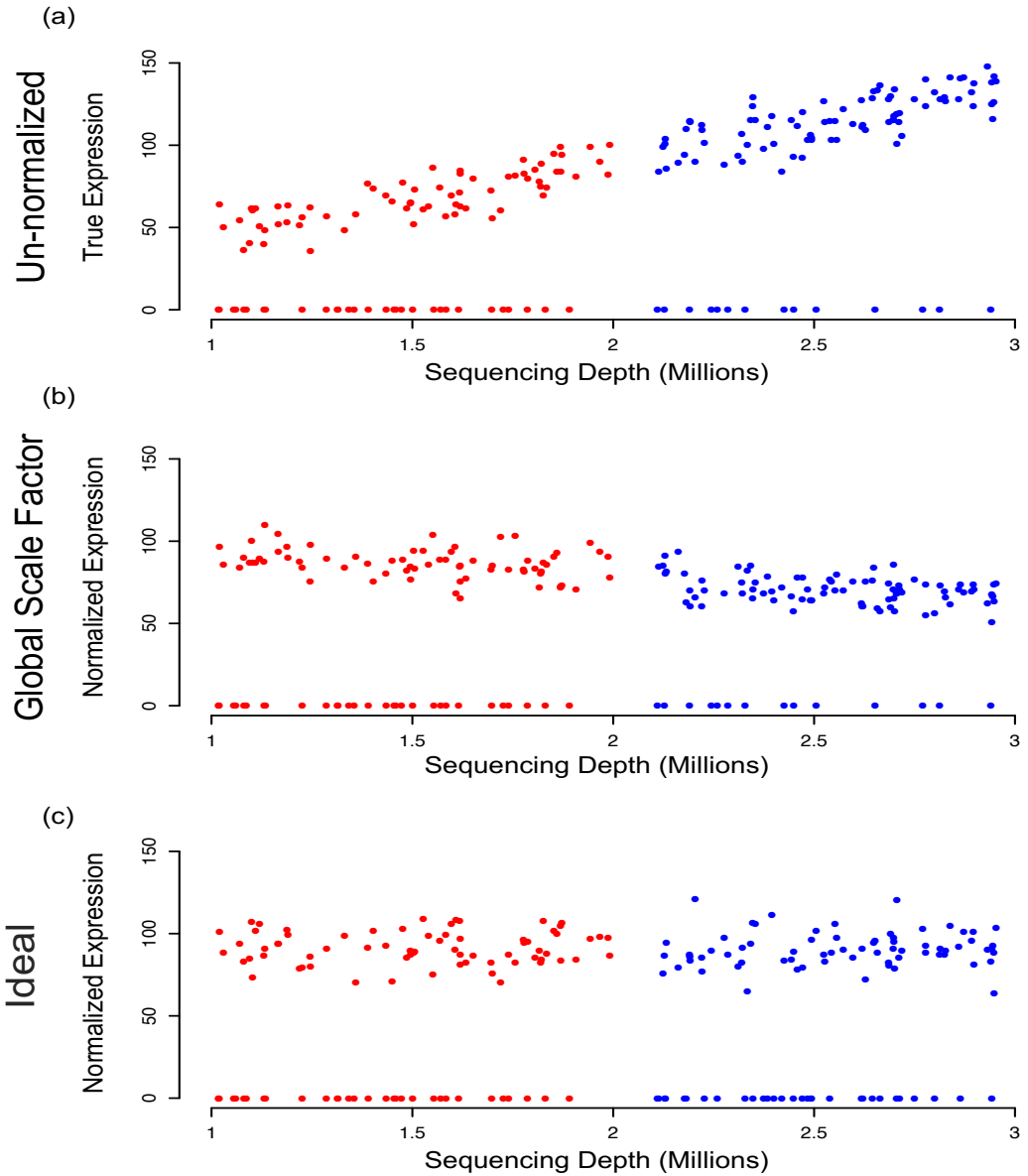
Bulk: Global scale-factor normalization for sequencing depth



Expression vs. depth varies with expression in scRNA-seq

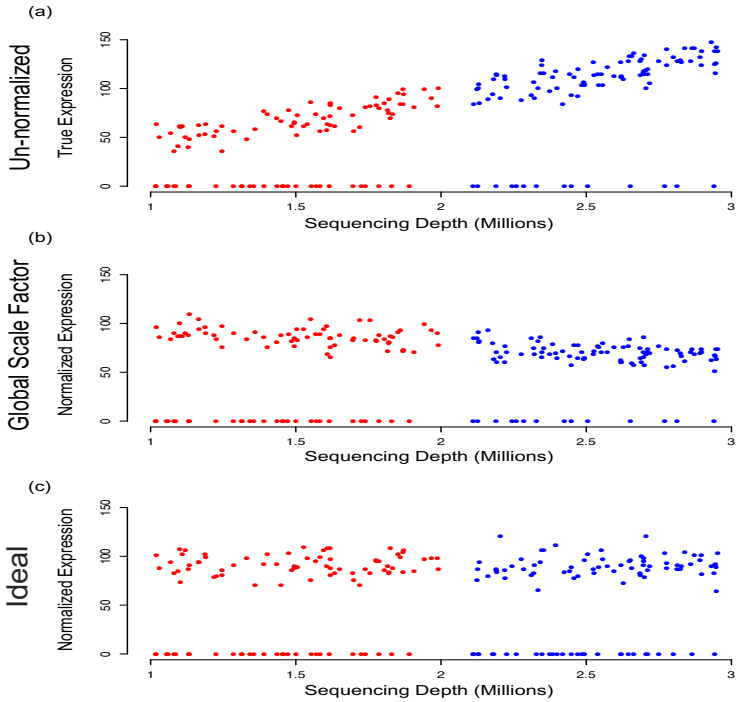


Implications for DE analysis



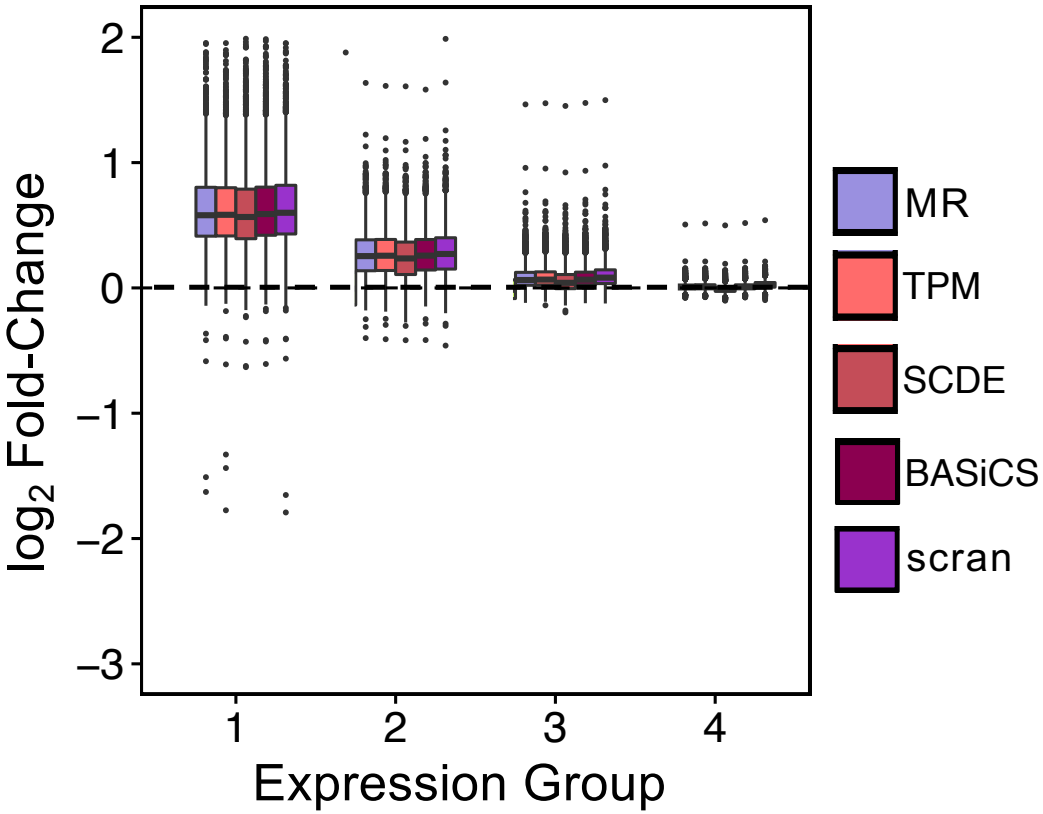
FC= H1-1/H1-4

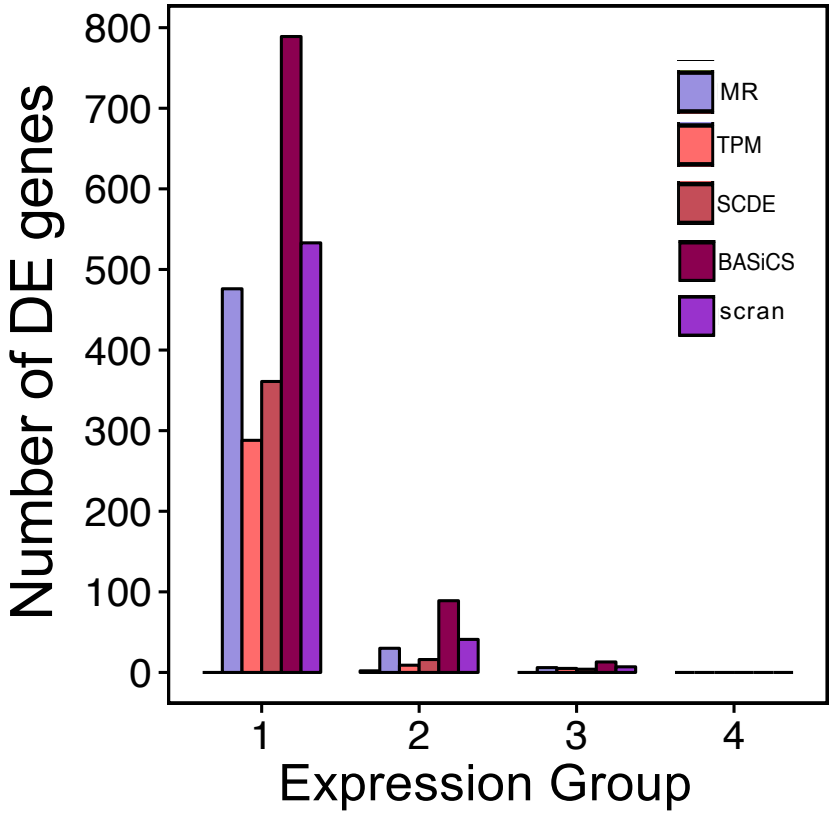
- H1-1: ~100 H1 cells profiles at ~1 million reads per cell
- H1-4: Same H1 cells profiled at ~4 million reads per cell
- Prior to normalization, H1-1/H1-4 should be about ¼
- Post normalization, H1-1/H1-4 should be about 1
- If over-normalization is going on, H1-1/H1-4 will be greater than 1.



FC= H1-1/H1-4

- H1-1: ~100 H1 cells profiles at ~1 million reads per cell
- H1-4: Same H1 cells profiled at ~4 million reads per cell





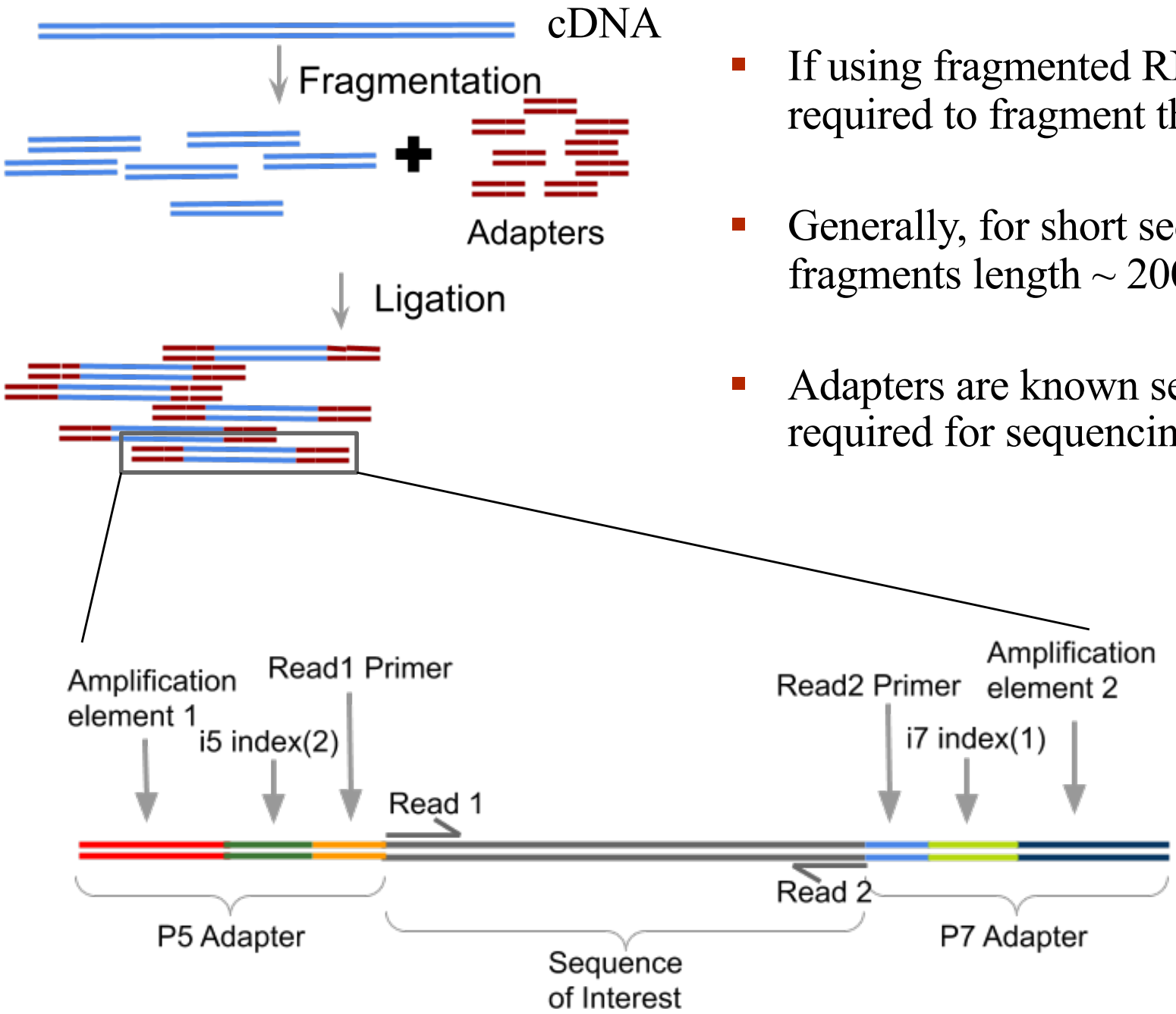
Normalization of spatial transcriptomics data

- This project will evaluate different normalization methods (TPM, scran, scTransform, and Dino) on spatial transcriptomics data. Initial evaluations will be conducted to assess the impact of various normalization methods on known tissue types (e.g. known regions in brain from the SpatialLIBD dataset).

- The authors of the SpatialLIBD paper built an R package to provide easy access to their data. Details at our website.

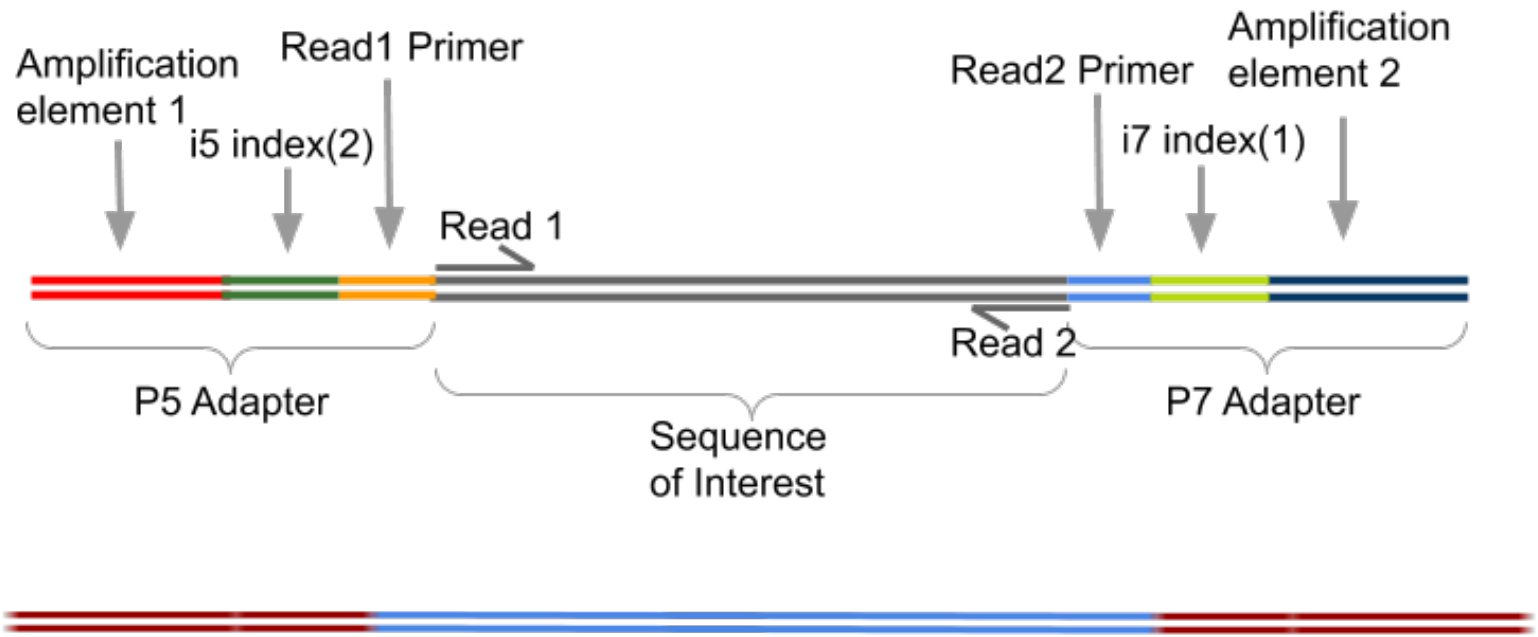
QC

Library Preparation



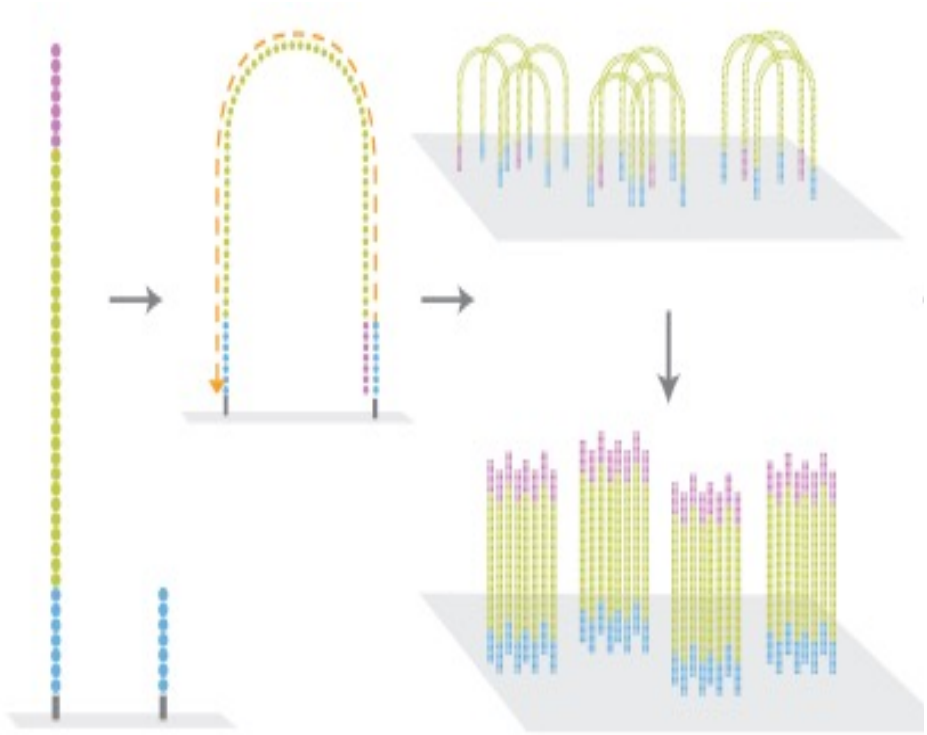
- If using fragmented RNA, it is not required to fragment the cDNAs.
- Generally, for short sequencing, fragments length ~ 200-300bp
- Adapters are known sequences required for sequencing on a flowcell

Library Preparation

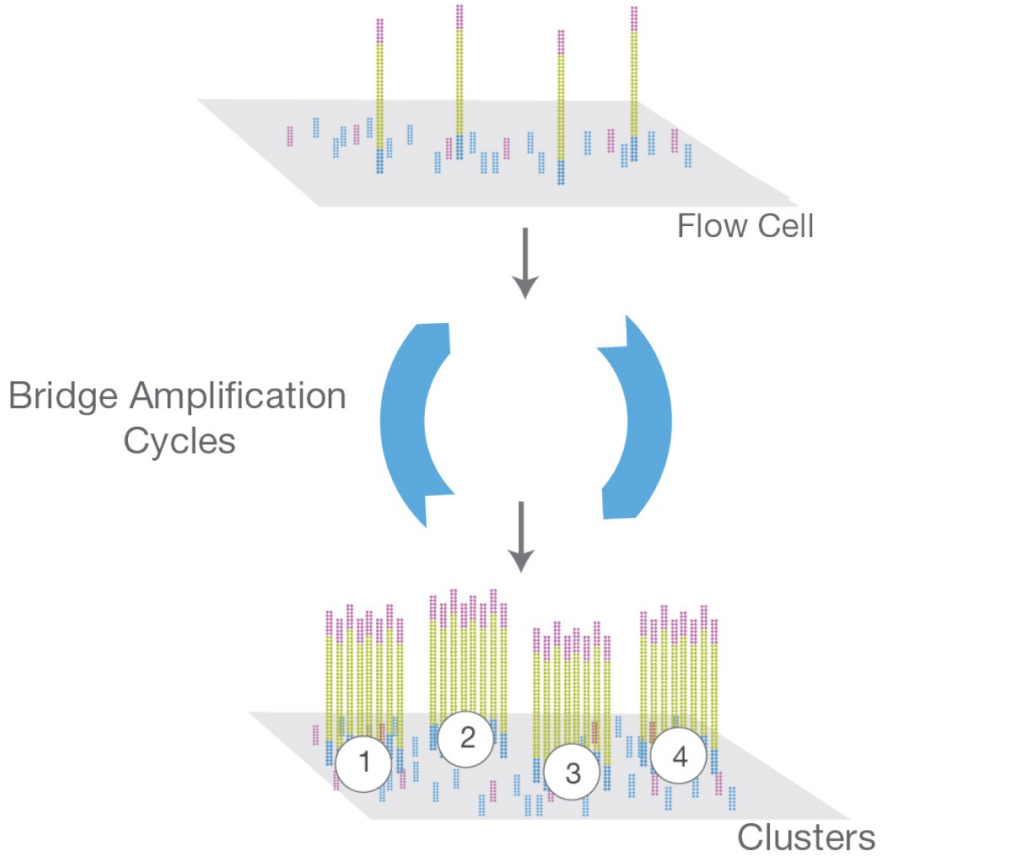


- Illumina sequencing uses sequencing by synthesis which requires adapters to facilitate sequencing on the flow cell. The adapters are short sequences that are known; there are relatively few used in sequencing.
- You can read all about Illumina adapters here: https://teichlab.github.io/scg_lib_structs/data/illumina-adapter-sequences-1000000002694-14.pdf

Cluster Amplification (outside class)



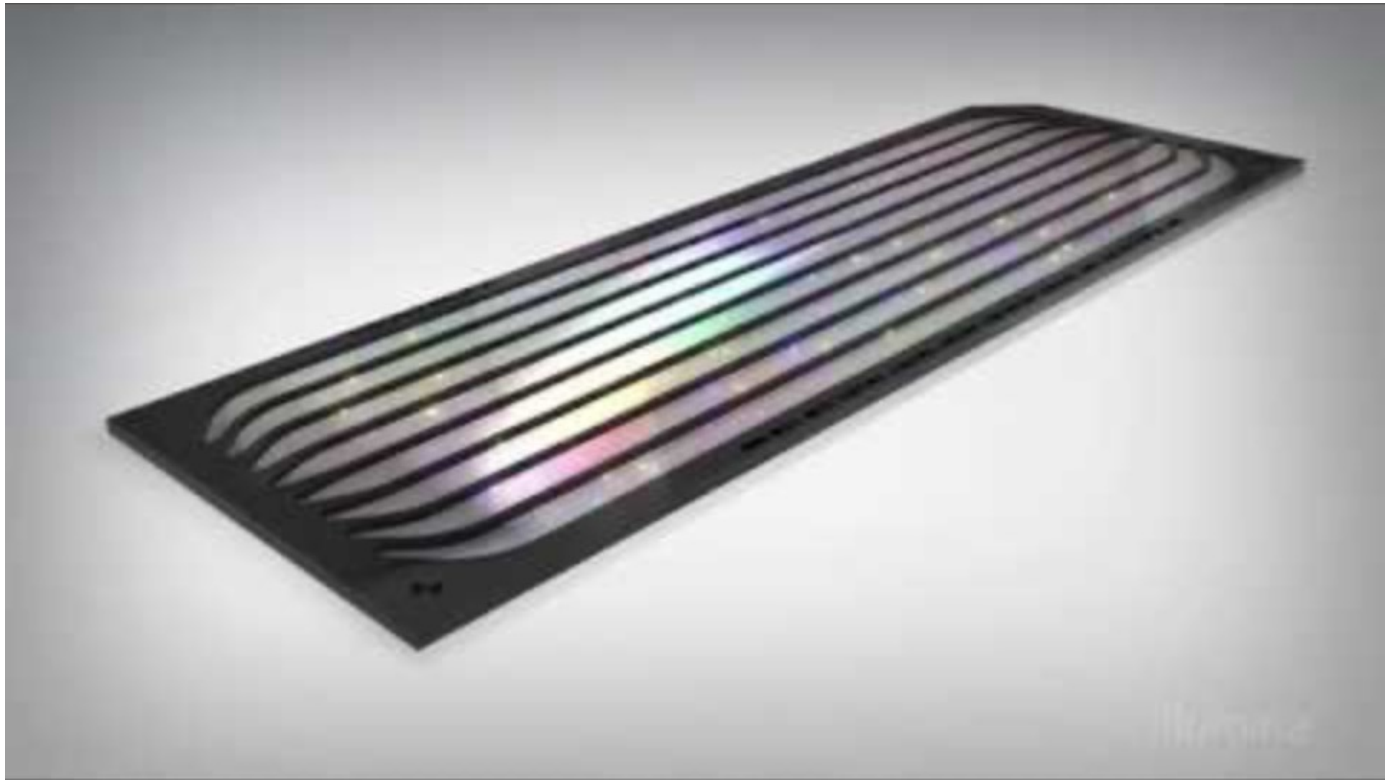
Fragments are replicated and clusters are formed using bridge amplification.



Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification

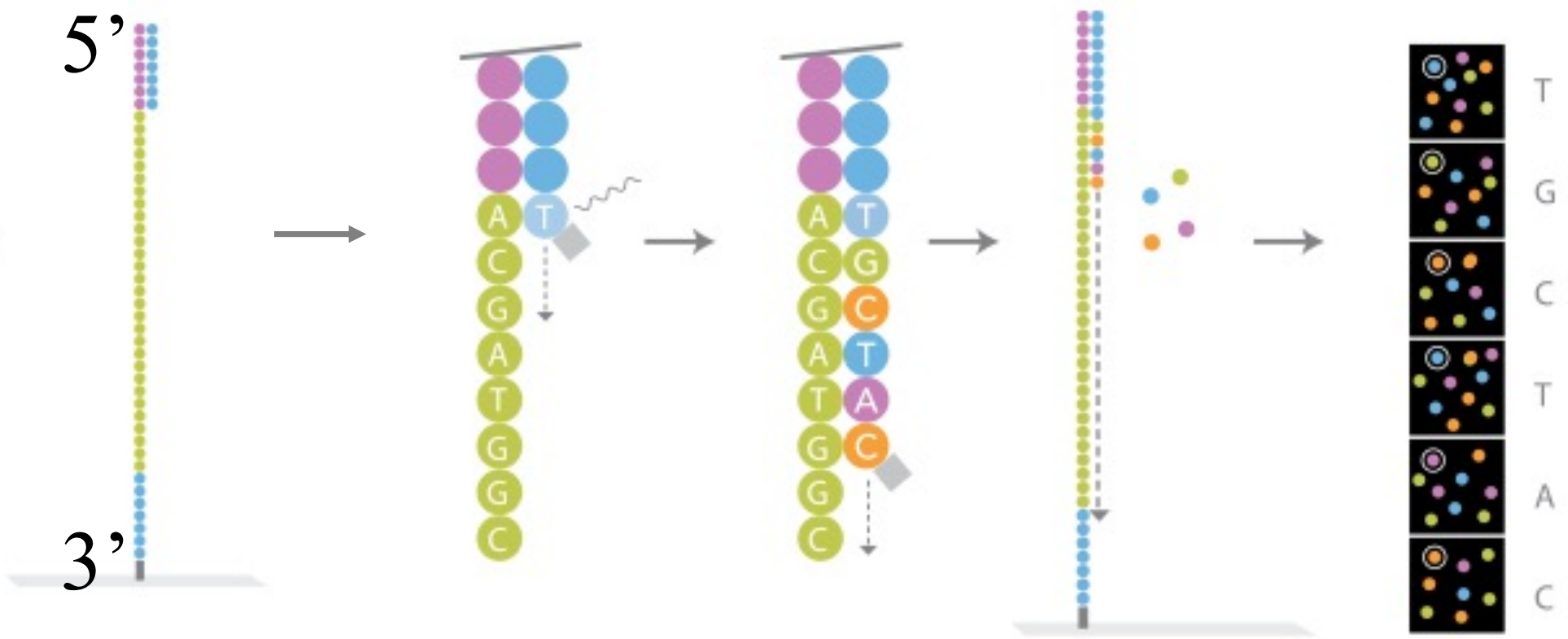
Source : Illumina

Flowcell (outside class)



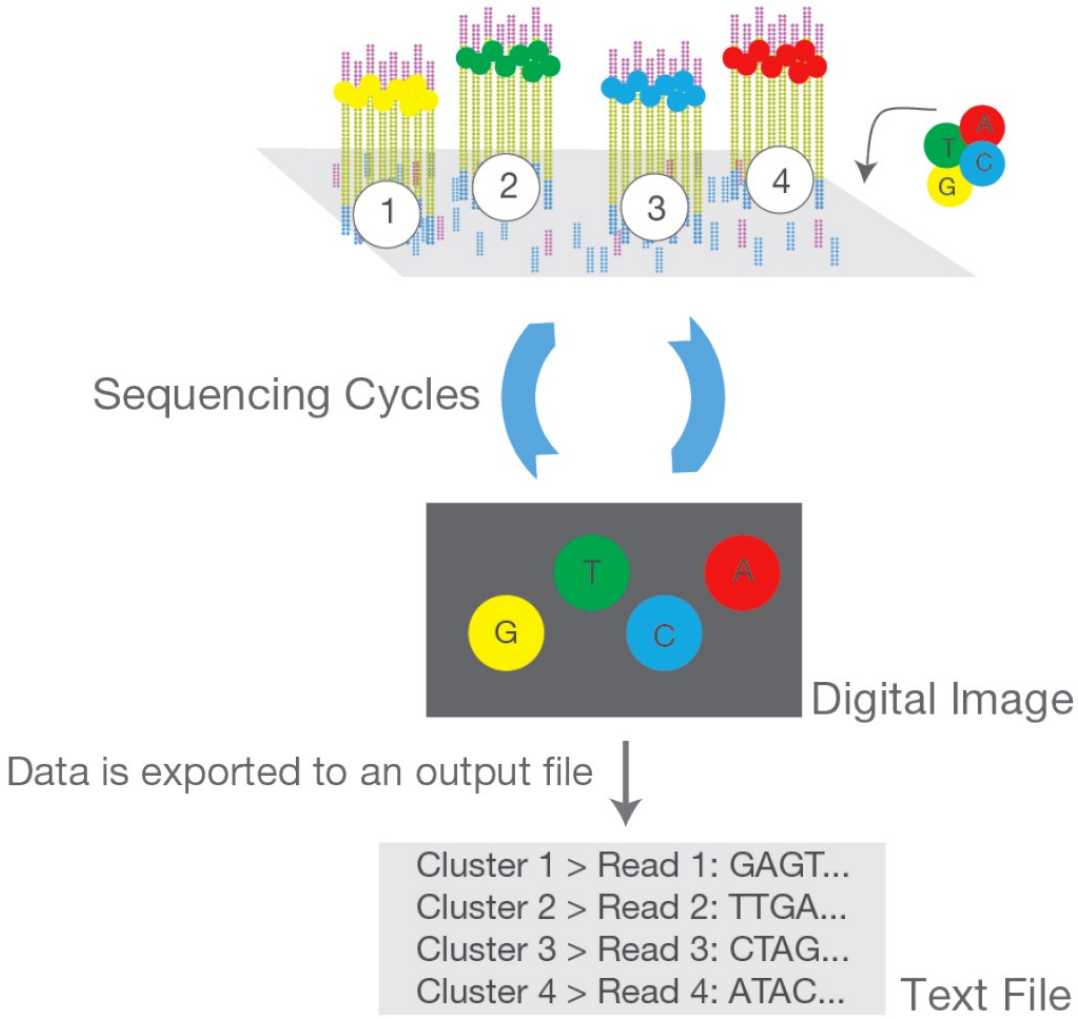
https://youtu.be/pfZp5Vgsbw0?t=5

Sequencing by Synthesis (outside class)



- In each sequencing cycle one fluorescently tagged nucleotide attached to the strand and the base is detected using emitted wavelength and signal strength.
- Fragment read length is same as the number of cycles.

Sequencing by Synthesis (outside class)



Emission recorded from each cluster.

https://www.youtube.com/watch?v=fCd6B5HRaZ8&feature=youtu.be&t=133

Source : Illumina

FASTQ

```
@A00589:158:HK2TMDRXX:1:2101:1705:1000
GGCTAATATTAGAAAATGGTTAAGTCCTAAATAACTTCAAGTGTGGTTATATAATGGACACTGTCAAT G TTCA
+
, F, FFF, F: , F: FF, FFFFFFFFFFFF: FFFFFFFFF: FF: FFF: , , : F: F, , FFFFF, FFFFF: FFF F  , ,FF
```

- Line 1: Begins with '@' followed by a **sequence identifier** and other optional information.
- Line 2: **Sequence** read by sequencer
- Line 3: Begins with '+' and optionally followed by sequence identifier or **other stuff**.
- Line 4: **Quality score** for the corresponding base in line 2

Single-End Read Example – Line 1

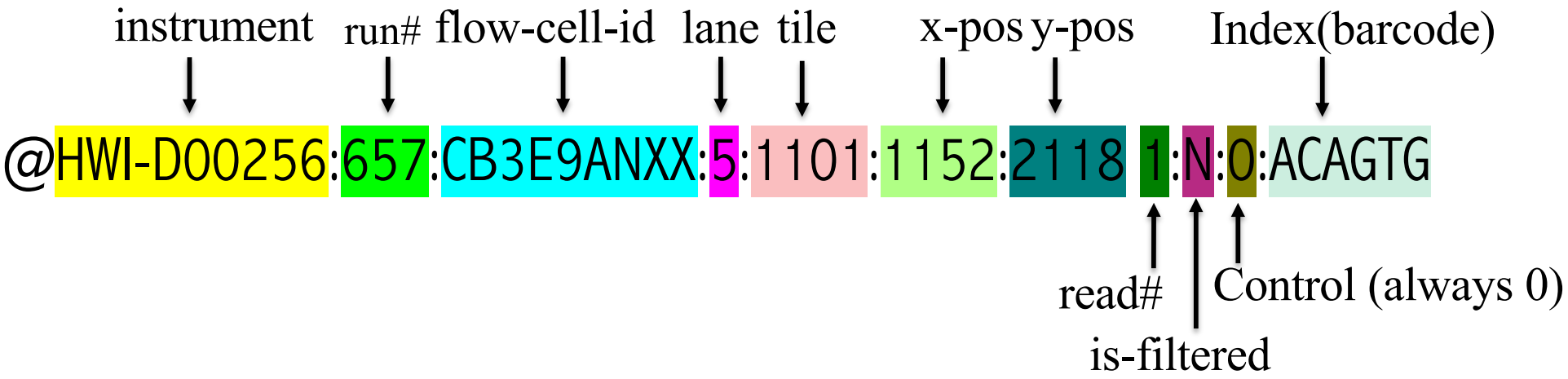
Read1

```
@HWI-D00256:657:CB3E9ANXX:5:1101:1152:2118 1:N:0:ACAGTG

CTGGGAACTGCGCCAGGAGAGCAGGGTCCTGACCCGGGCCTTCAGGAGGTGAGGCCANCTGGTGGGCAGGA
GGCTGTGGTAGAGGCAGCTCAGTTCTAGGA

+

BBBBBFFFFFFFFFFFFFFFFFFFFBFFFFFFFFFFFFFFFFBFFFF<<FFFFF#</FFFFFFFFBF
FFFFFFFFFFFFFFFFF/BFFFFFFFFB
```



Ref : https://help.basespace.illumina.com/articles/descriptive/fastq-files/

Paired-End Read Example - Line 1

Read1

SampleName_S1_L001_R1_001.fastq.gz

```
@A00589:158:HK2TMDRXX:1:2101:1705:1000 1:N:0:AAGATTGGAT+AGCGGGATTT
NGGTGTTTCAGCTATGAATTGTGCTCAT
+
#::F,FFF:FFFFFFFFFFFFF:FFFFF
```

Read #

Read2

SampleName_S1_L001_R2_001.fastq.gz

```
@A00589:158:HK2TMDRXX:1:2101:1705:1000 2:N:0:AAGATTGGAT+AGCGGGATTT
GGCTAATATTAGAAAATGGTTAAGTCCTAAATAACTTCAAGTGTGGTTATATAATGGACACTGTCAATGTTCATA
ACTTAAACCTGGGTAC
+
',F,FFF,F:,F:FF,FFFFFFFFFFFFF:FFFFFFFFF:FF:FFF:,,:F:F,,FFFFF,FFFFF:FFFF,,FFFFFFFFFFF:FFFF
F:F'
```


Quality scores – Line 4

The sequencing quality score is also known as the phred score of a given base, Q, is defined by

$$Q = -10 \log_{10} p$$

Where p is the estimated probability of the base call being wrong.

Higher Q scores correspond to higher accuracy.

Examples:

1. Q=20 the probability of an incorrect base call is 1 in 100 (Q20; ?)
2. Q=30 the probability of an incorrect base call is 1 in 1000 (Q30; 5)
3. Q=37 the probability of an incorrect base call is 1 in 10^3.7 (Q37, F)

Each quality score is encoded with a single ASCII character in FASTQ file

Quality Score Encoding: https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm

FASTQ

```
@A00589:158:HK2TMDRXX:1:2101:1705:1000
GGCTAATATTAGAAAATGGTTAAGTCCTAAATAACTTCAAGTGTGGTTATATAATGGACACTGTCAAT G TTCA
+
, F, FFF, F: , F: FF, FFFFFFFFFFFF: FFFFFFFF: FF: FFF: , , : F: F, , FFFFF, FFFFF: FFF F , ,FF
```

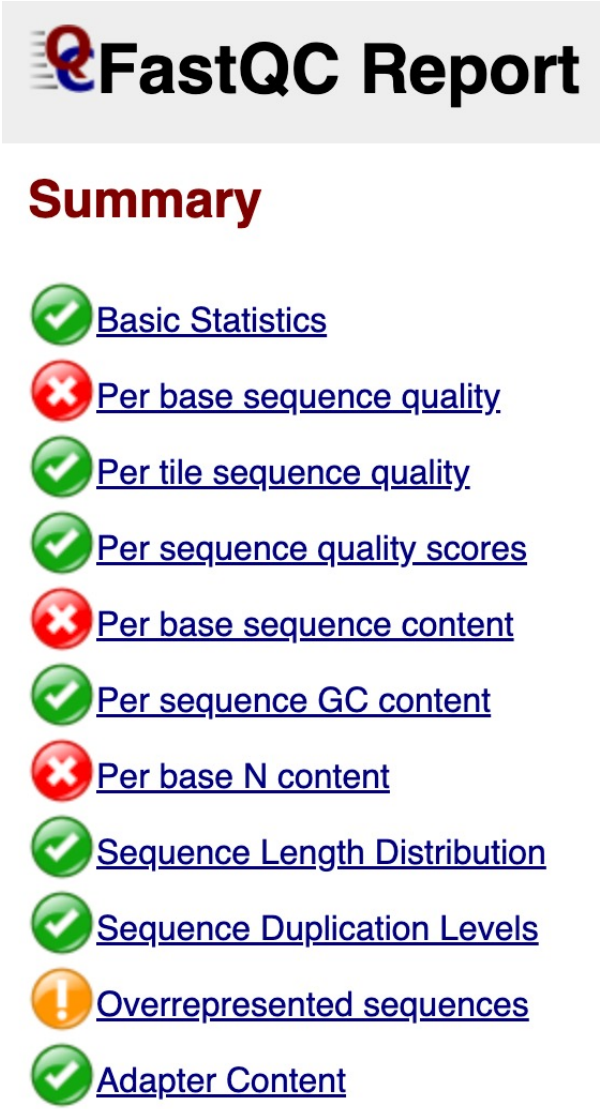
Base = G

Score = F = 37

- Line 1: Begins with '@' followed by a **sequence identifier** and other optional information.
- Line 2: **Sequence** read by sequencer
- Line 3: Begins with '+' and optionally followed by sequence identifier or **other stuff**.
- Line 4: **Quality score** for the corresponding base in line 2

FastQC

- Simple checks for identifying issues originating during library preparation or sequencing.
- Supports fastq (also Gzip-ed), SAM, BAM, among others.
- Stand alone interactive mode
  - Suitable for small number of fastq files.
- Non-interactive mode
  - Suitable for analysis pipelines
- Be cautious about relying on the flags.



FastQC Ref:
https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

FastQC – Basic Statistics

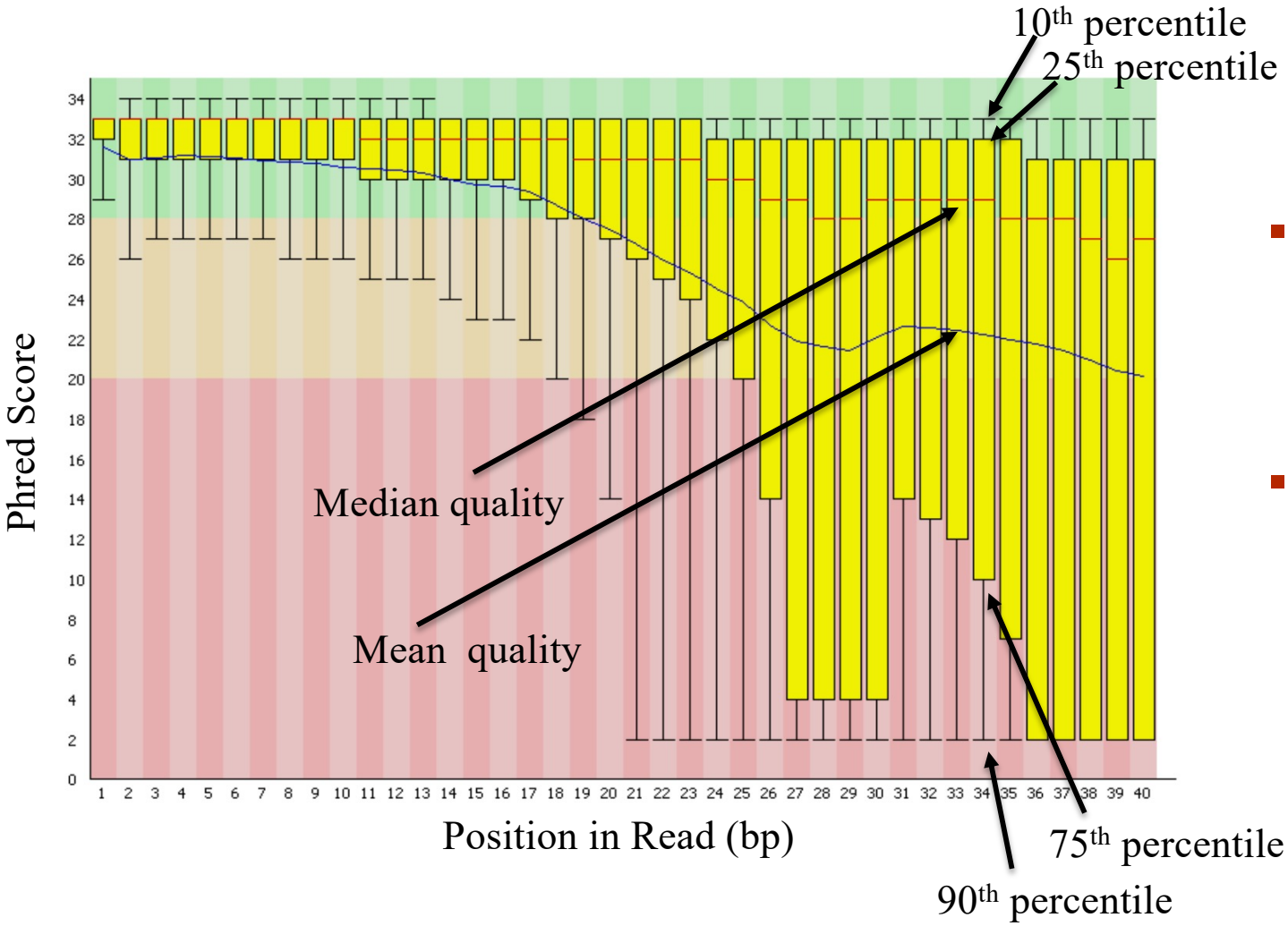
Basic Statistics

Measure	Value
Filename	E67-1_ACAGTG_L005_R1_001.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	33342243
Sequences flagged as poor quality	0
Sequence length	101
%GC	50

All reads are of same length.
Otherwise provides a range

- Basic statistics almost never raises a warning or an error.

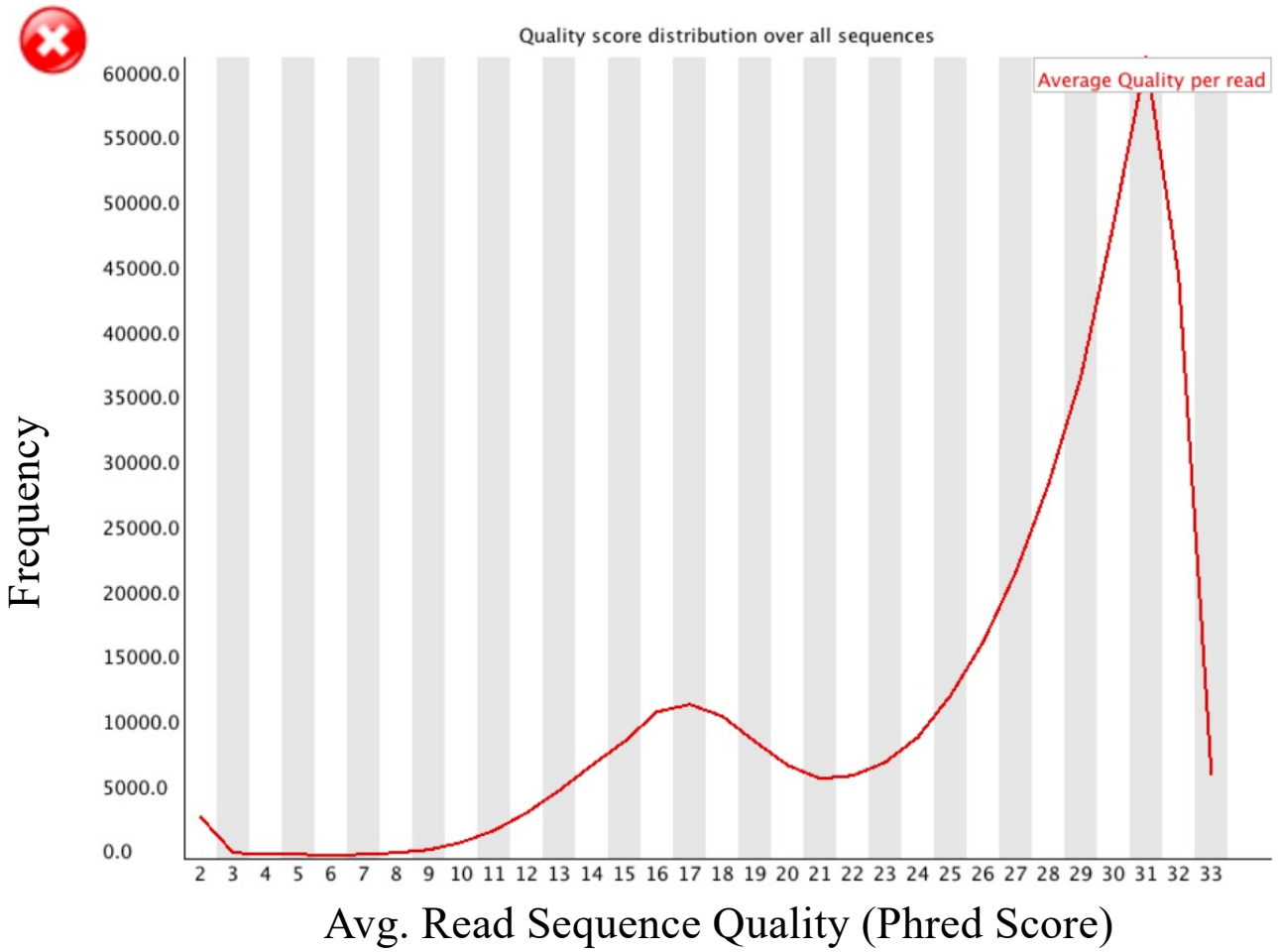
FastQC – Per Base Sequence Quality

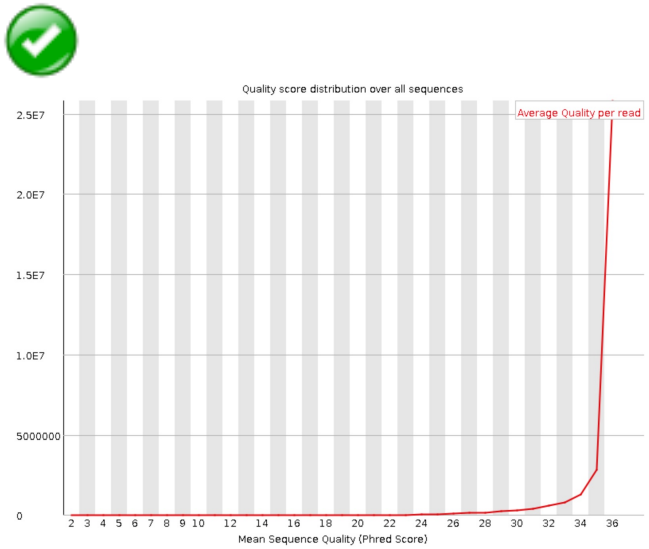


- Warning if lower quartile for any base is less than 10, or if the median for any base is less than 25.
- Failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

- The average quality score typically drops over the length of the read.

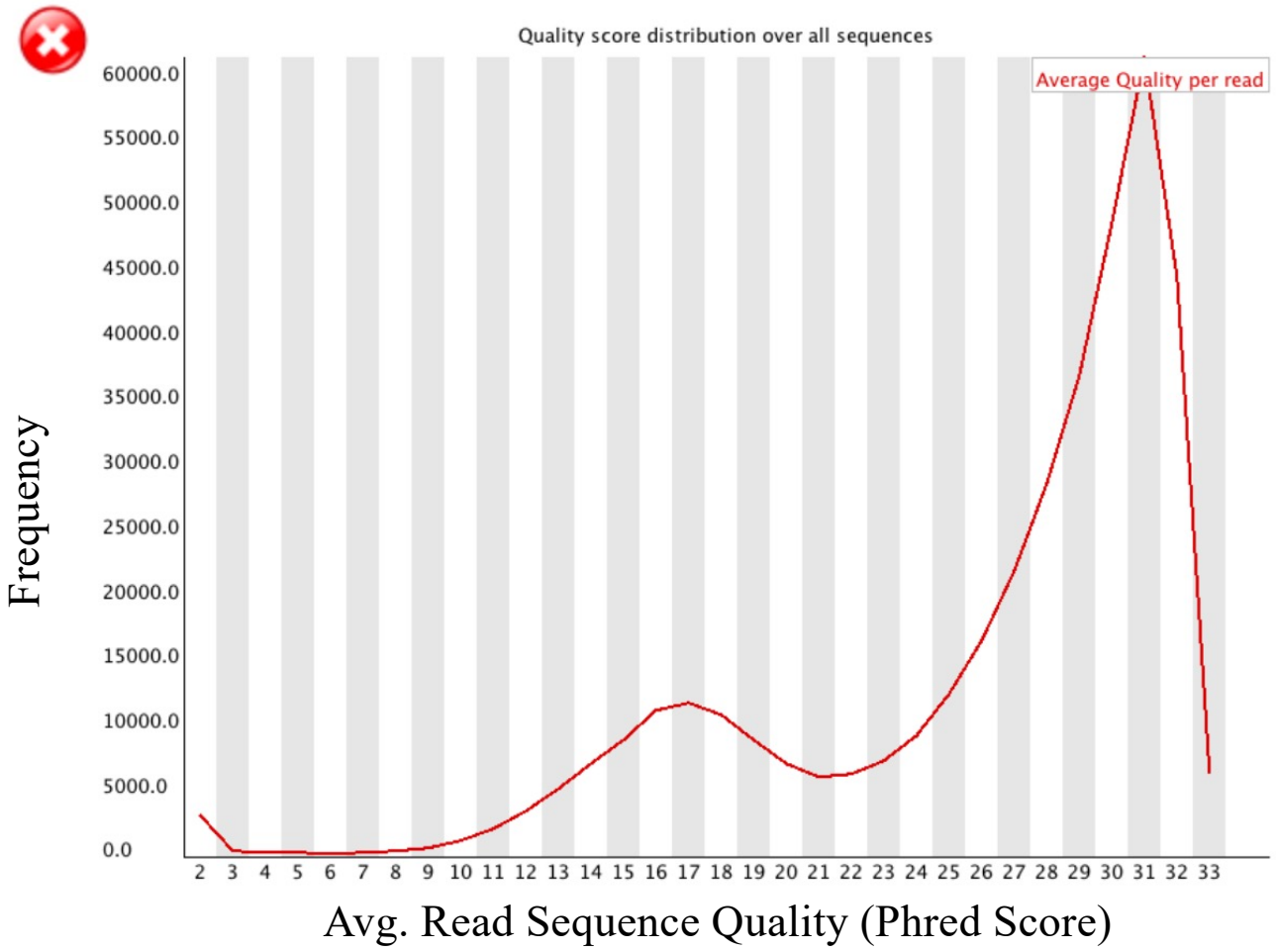
FastQC – Per Sequence Quality Scores





- Determines if a subset of reads have low quality
- Average read quality should be tight in the upper range of the plot.

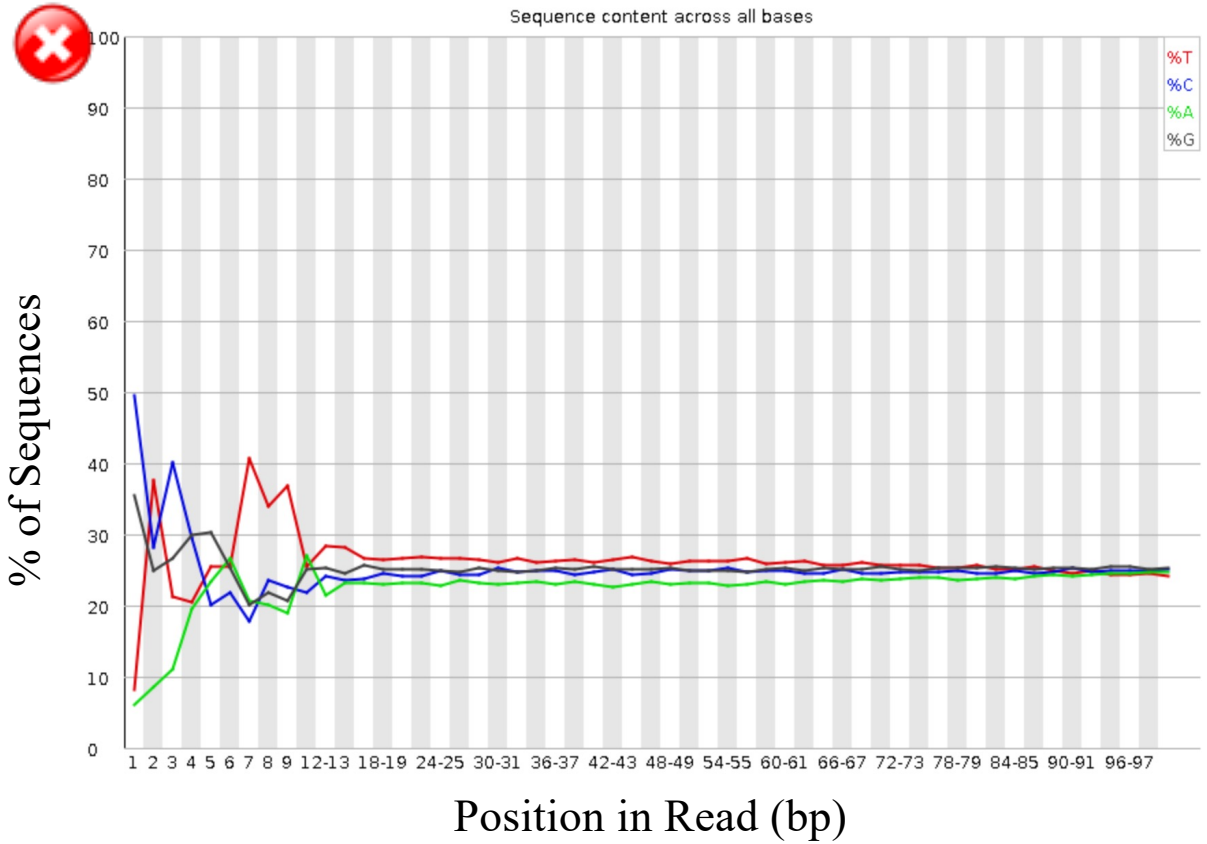
FastQC – Per Sequence Quality Scores

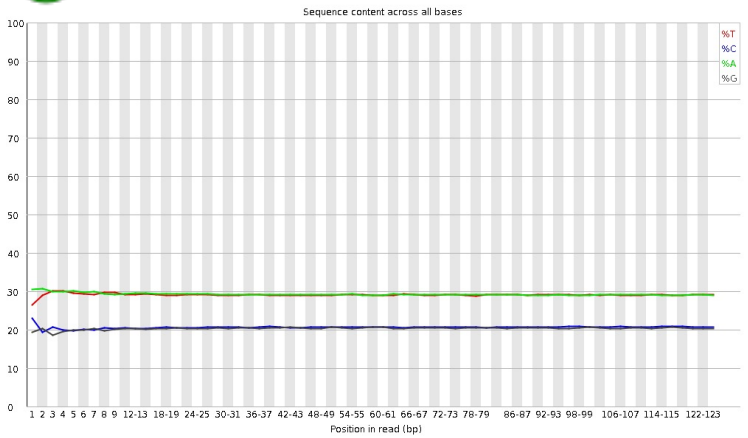


- Warning if the most frequently observed mean quality is below 27.
- Failure if the most frequently observed mean quality is below 20.

- Determines if a subset of reads have low quality
- Average read quality should be tight in the upper range of the plot.

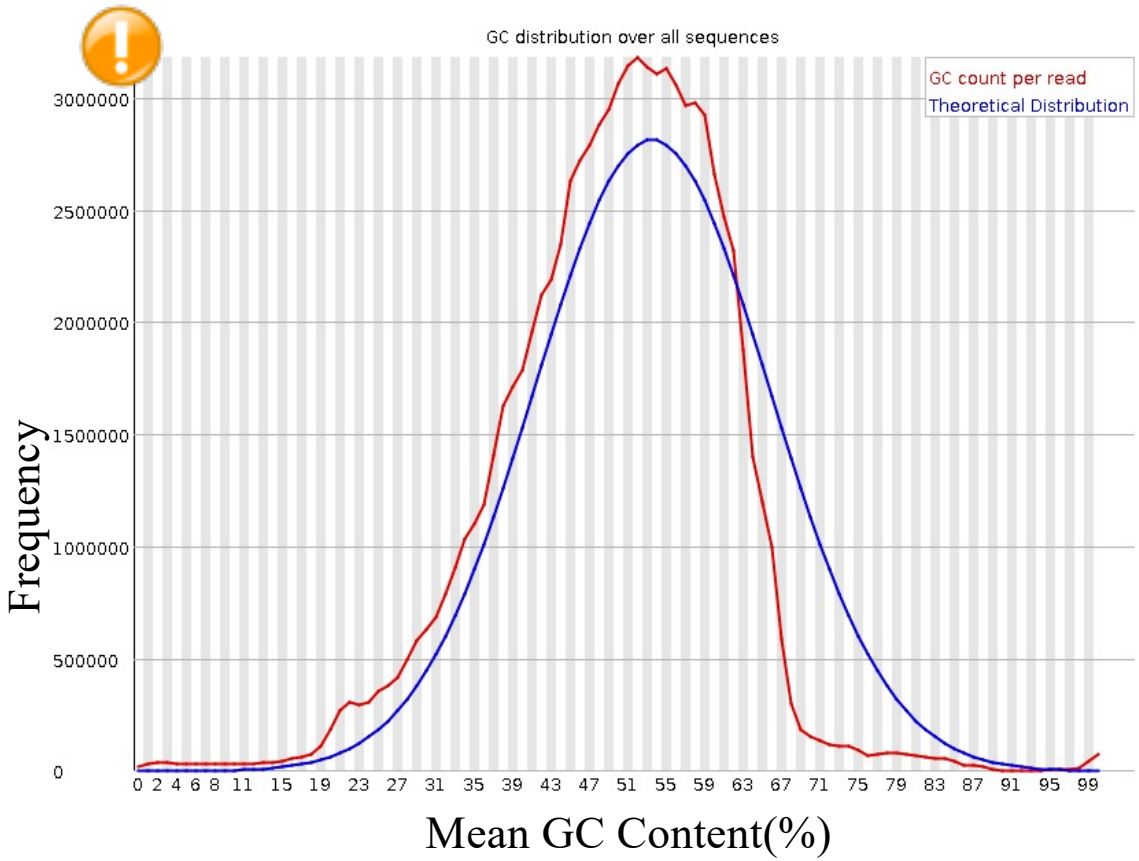
FastQC – Per Base Sequence Content

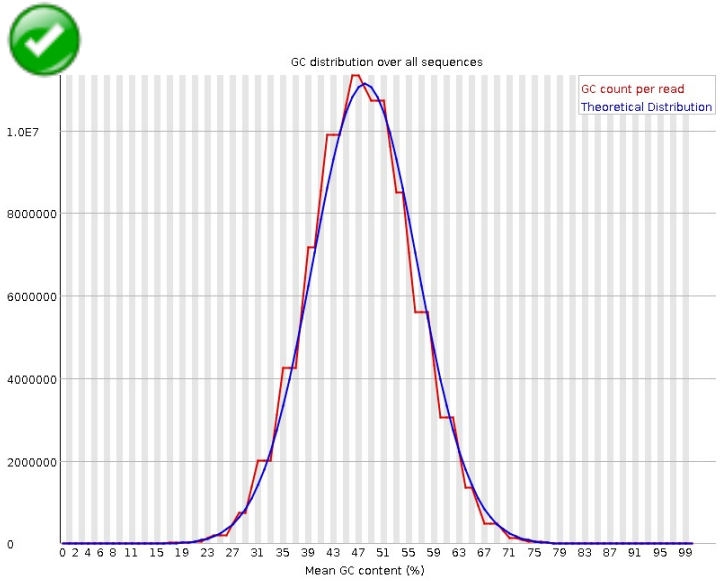




- % of bases called at each position across all reads. In a random library the lines should run parallel with each other with %A=%T, %G = %C.
- This may not be true for most of the RNA-seq library preparation protocols because of priming using random hexamers or selection bias (effects first 10-15 bases, alike above example). That's OK.

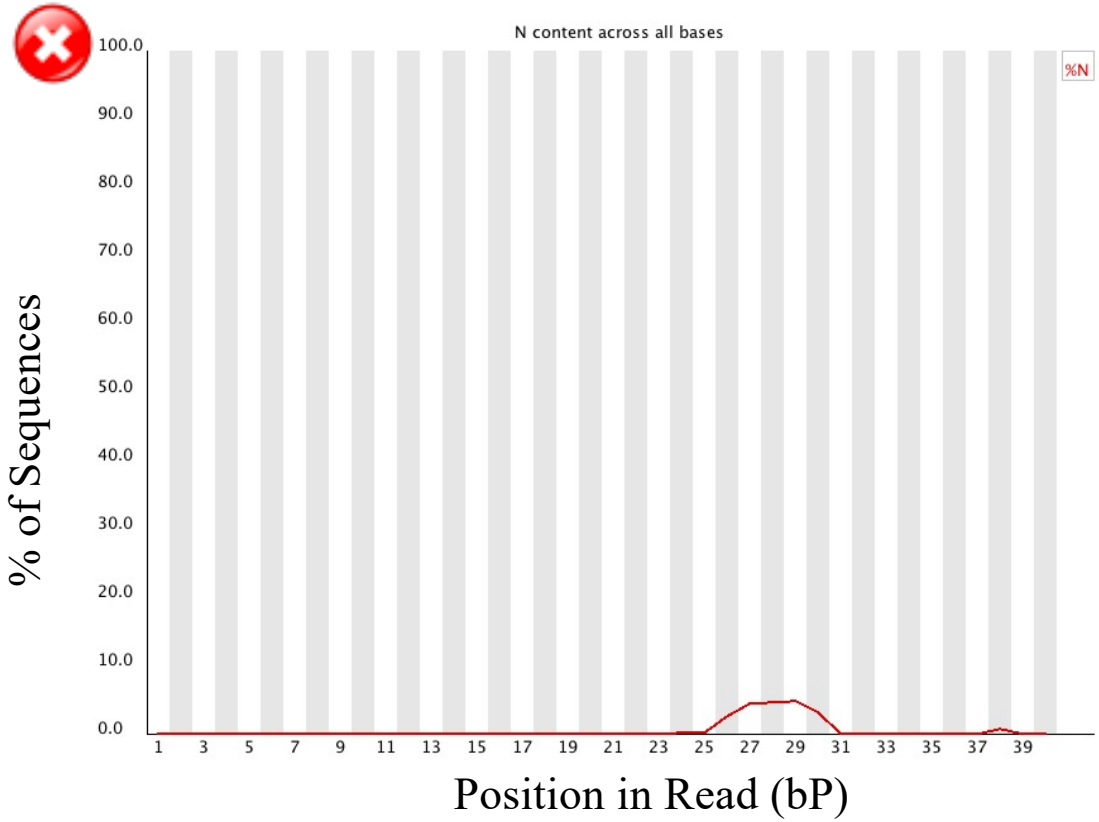
FastQC – Per Sequence GC Content

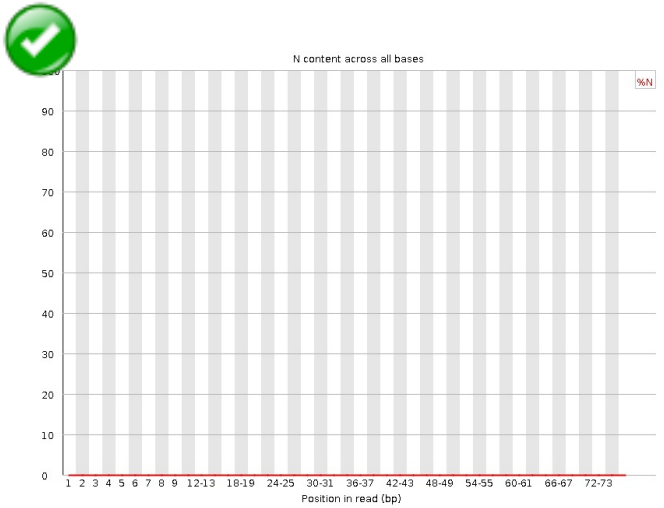




- Measures the GC content across the whole length of each sequence and compares it to a modelled normal distribution of GC content.
- An unusually shaped distribution could indicate a contaminated library.
- Wider peak may represent contamination with a different species.

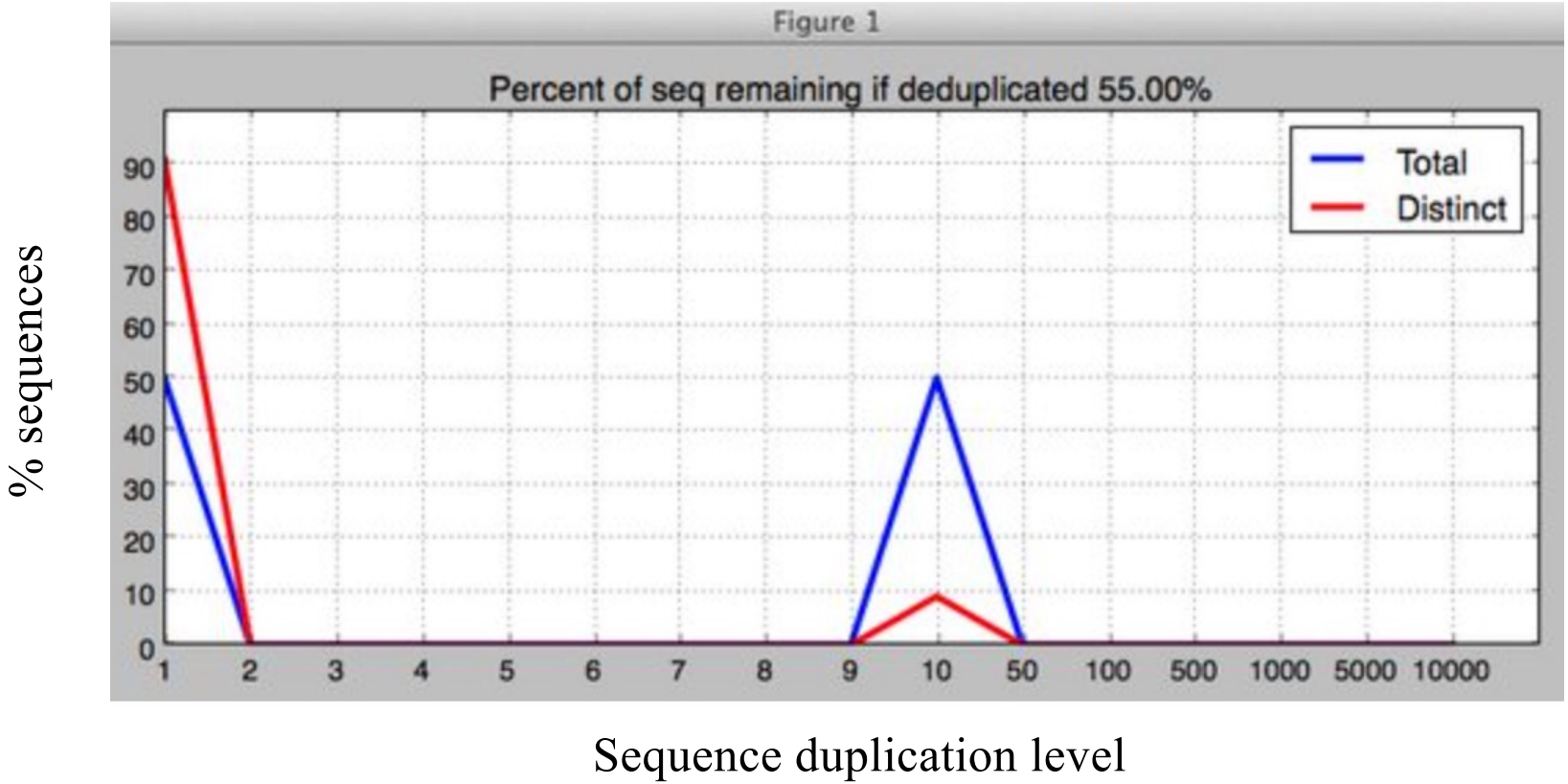
FastQC – Per Base N Content





- If a sequencer is unable to make a base call with sufficient confidence, then it will substitute N instead of a base.
- A noticeable curve indicates a problem occurred during the sequencing run.

QC Figures - Duplication Sequences



- The plot shows the proportion of the library which is made up of sequences in each of the different duplication level bins (10 unique reads; one read present 10 times).
- Warn: If non-unique sequences make up more than 20% of the total
- Fail: If non-unique sequences make up more than 50% of the total

Duplicated sequences (in more detail)

- Consider two scenarios, each with 20 reads in total:
  - Case1: 10 unique reads and 5 reads that are doubled
  - Case2: 10 unique reads and 1 read present 10 times
- Percent remaining reads after de-duplication (reported top of plot)
  - Case1: 15/20=75%
  - Case2: 11/20=55%

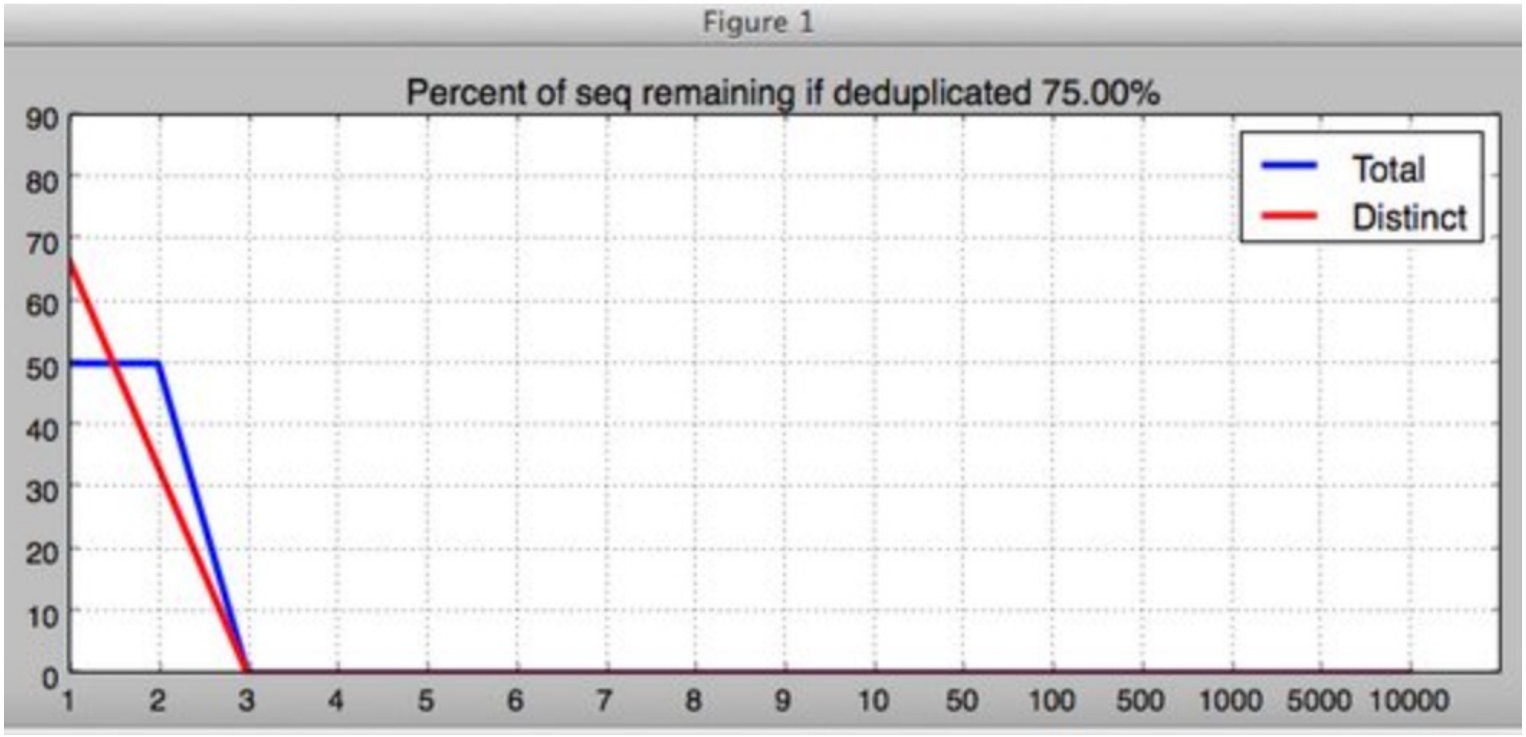
Duplicated sequences (in more detail)

- Consider two scenarios, each with 20 reads in total:
  - Case1: 10 unique reads and 5 reads that are doubled
  - Case2: 10 unique reads and 1 read present 10 times
- Blue line: percentage of all sequences that are duplicated at a given rate
- Percent singles, doubles, …, tens, … from total. This is calculated as the number of reads that are singles over the total, the number of reads that are doubles over the total…
  - Case1: 10/20=50% (singles); 10/20=50% (doubles)
  - Case2: 10/20=50% (singles); 10/20=50% (tens)

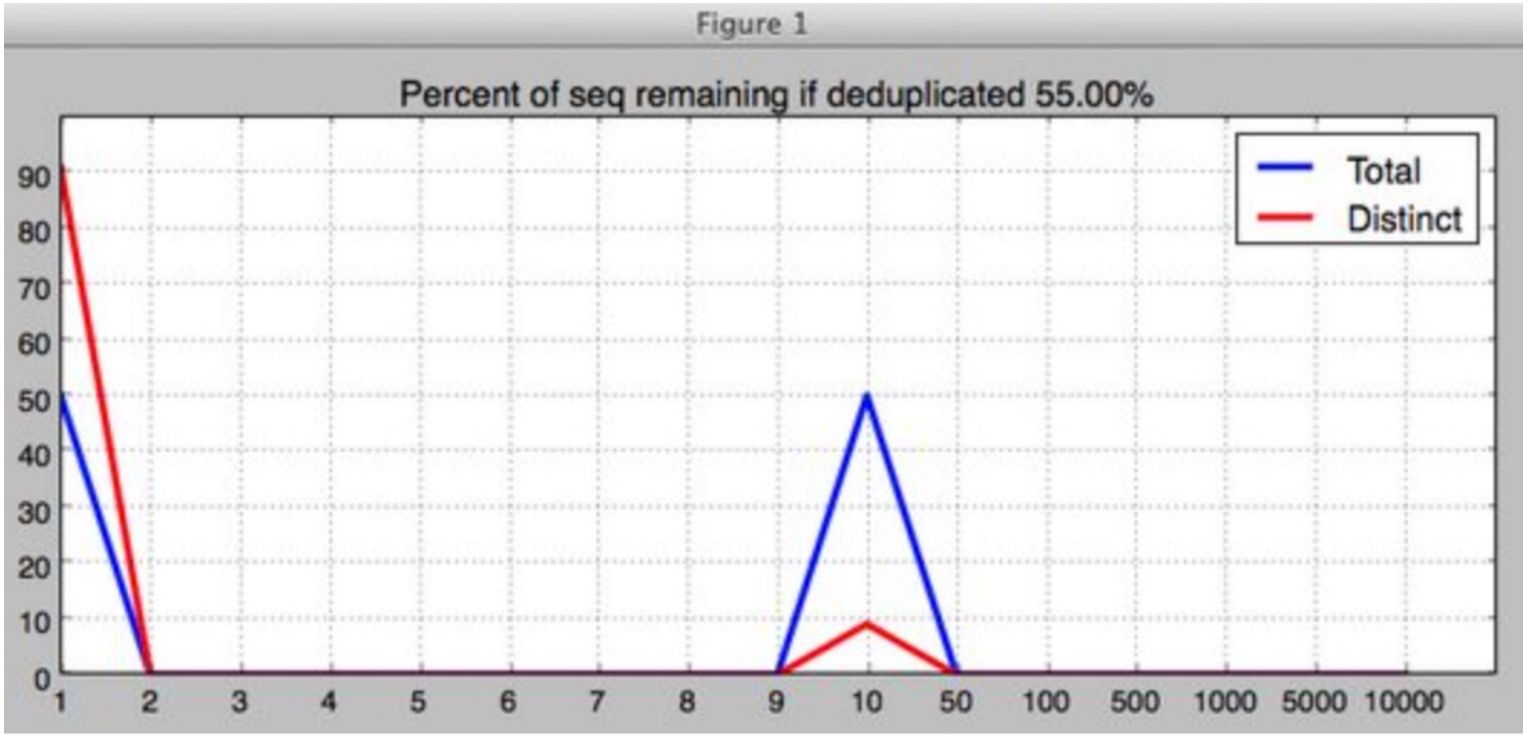
Duplicated sequences (in more detail)

- Consider two scenarios, each with 20 reads in total:
  - Case1: 10 unique reads and 5 reads that are doubled
  - Case2: 10 unique reads and 1 read present 10 times
- Red line: percentage of distinct sequences that are duplicated at a given rate
- Consider the de-duplicated set of reads (set of reads after duplicates are removed). The red line is the percent of de-duplicated singles, doubles, …, tens, … from the total number of de-duplicates.
  - Case1: 10/15=66% (singles); 5/15=33% (doubles)
  - Case2: 10/11=91% (singles); 1/11=9% (tens)

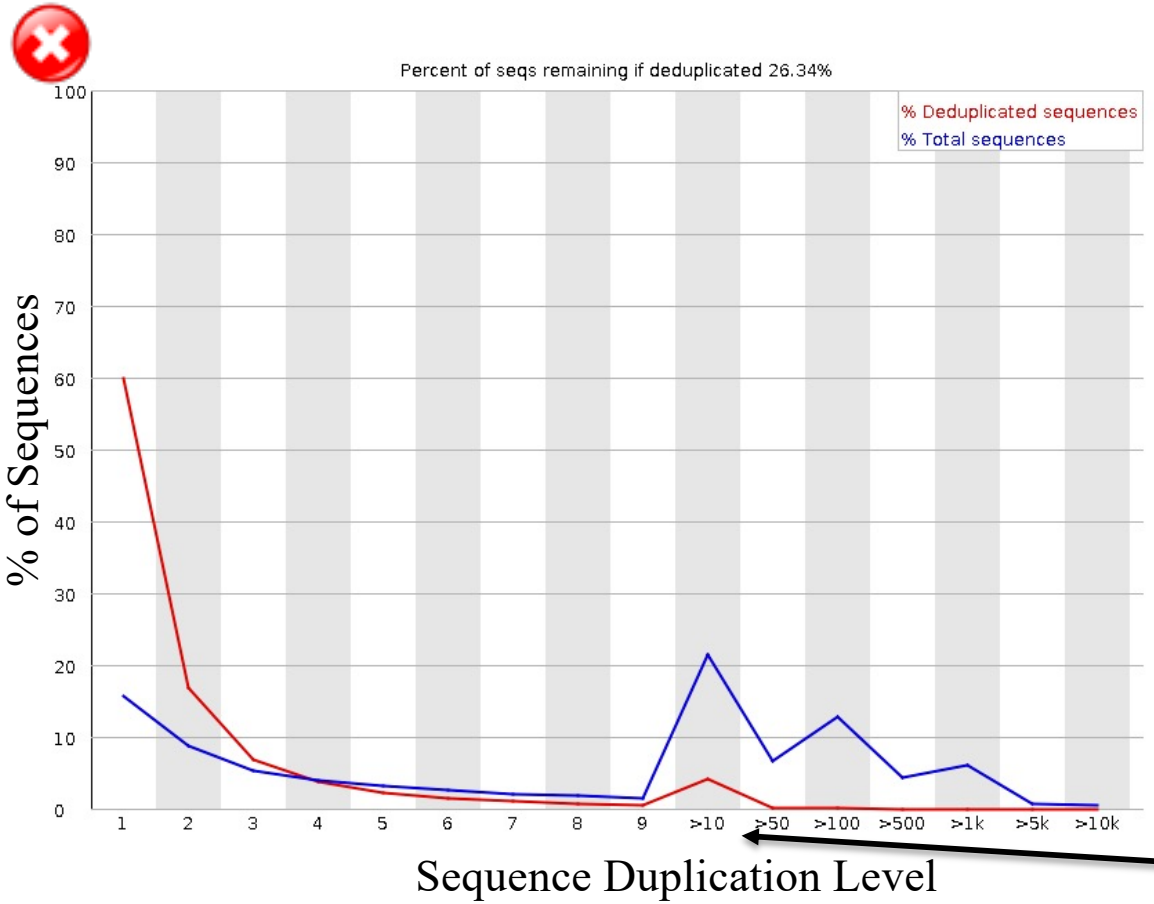
Duplication Sequences: Case 1

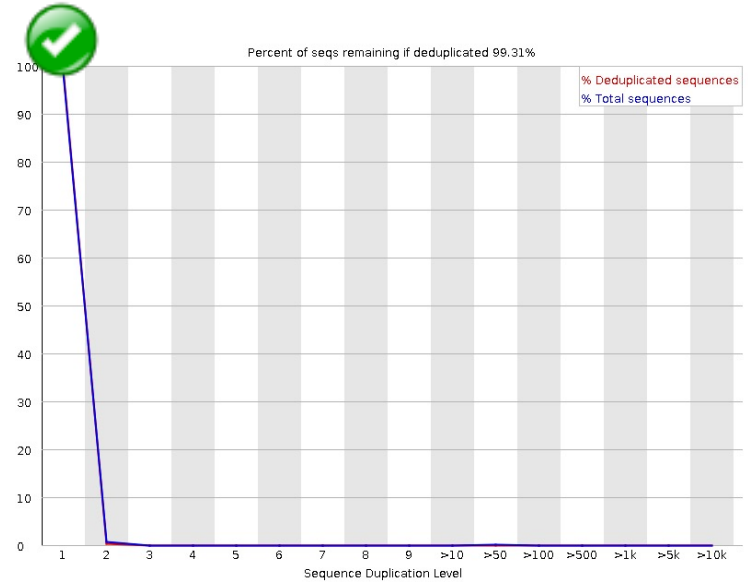


Duplication Sequences: Case 2



FastQC – Sequence Duplication Levels





For counts 10 or more are binned.

- First 200,000 sequences are analysed; this should be a good indicator of duplication levels in the whole file. Each sequence is tracked to the end of the file to give a representative count of the overall duplication level.
- Reads over 75bp are truncated to 50bp for the purposes of this analysis.

Common reasons for duplicate sequences

- Technical duplicates arising from PCR artifacts (what we try to avoid…if there are too many of these, sample quality is poor)
- Biological duplicates where different copies of exactly the same sequence are randomly selected (possible, but large numbers of duplicates are not likely in a diverse library…some are expected in RNA-seq from the most highly expressed transcripts)
- From a sequence level there is no way to distinguish between these two types. However, fastqc reports duplicates and you can blast them.
- You can find more details at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/

Click on analysis modules (3) and then duplicate sequences (8)

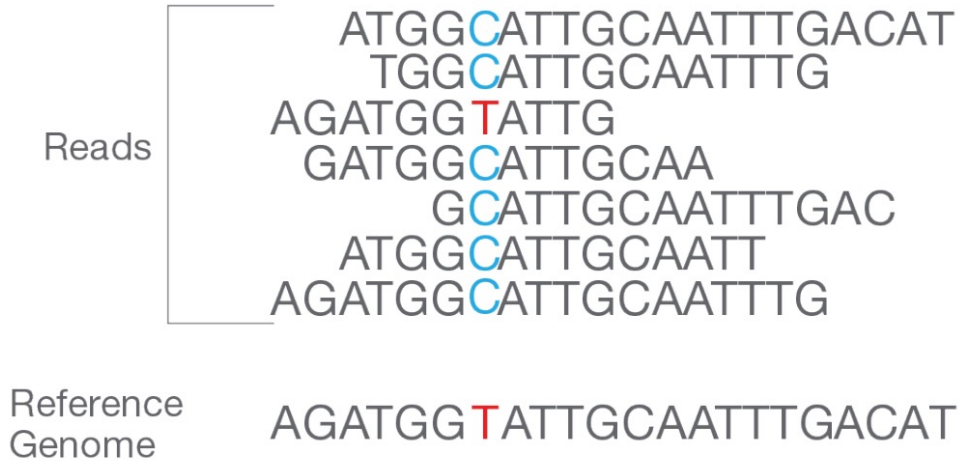
FastQC – Overrepresented Sequences

Sequence	Count	Percentage	Possible Source
ATCGGAAGAGCACACGTCTGAACTCCAGTCACACAGTGATCTCGTATGCC	98162	0.29440730787067926	TruSeq Adapter, Index 5 (100% over 50bp)

Matches the sequence with list of known adapters at FastQC/Configuration/containment_list.txt, else 'no hit'

- Similar to sequence duplication, only first 100,000 reads are tracked. Any reads over 75 bp are trimmed to 50 bp.
- Hits with more than 20 bp and having no more than 1 mismatch are considered. A sequence is considered overrepresented if it accounts for $\geq 0.1\%$ of the total reads.
- Many adapter sequences are very similar to each other so they may be reported here.
- When small RNA libraries are sequenced, it's possible that some sequence may naturally be present in a significant proportion of the library.

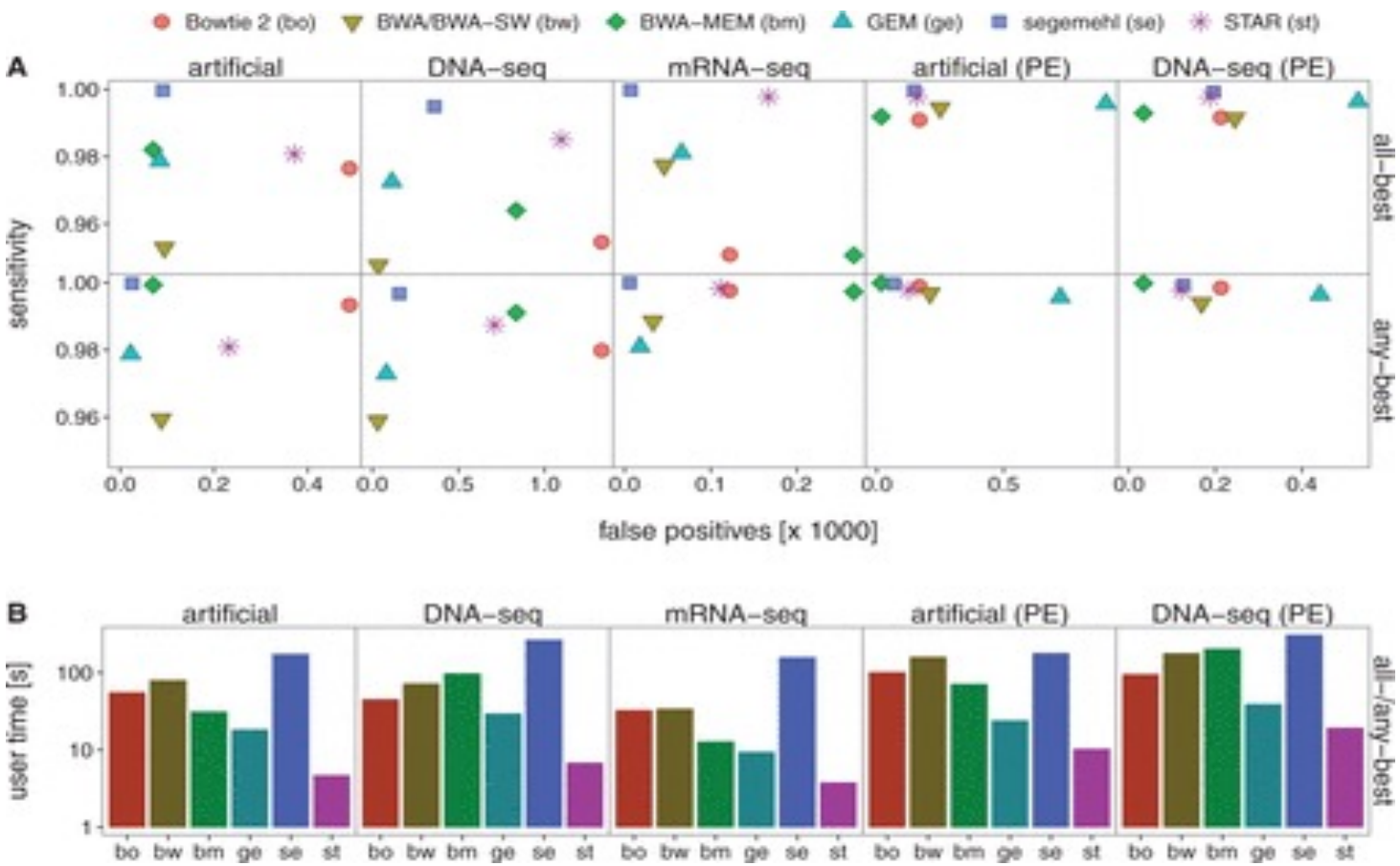
Alignment



- Subsequently the reads are aligned to a reference genome using any of the alignment software.

Aligners

- Bowtie2
- BWA
- BWA-MEM
- GSNAP
- STAR
- GEM
  - .
  - .
  - .



Otto, et al., 2014, https://doi.org/10.1093/bioinformatics/btu146

fastqc

- FastQC was developed by the Bioinformatics group at the Babraham Institute in Cambridge, UK

- Many details can be found at
  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

For Wednesday February 9

Please download and install FastQC on your laptop before coming to class on Wednesday. We will have an in-class lab that day. Details on installation are given at our Canvas course website.