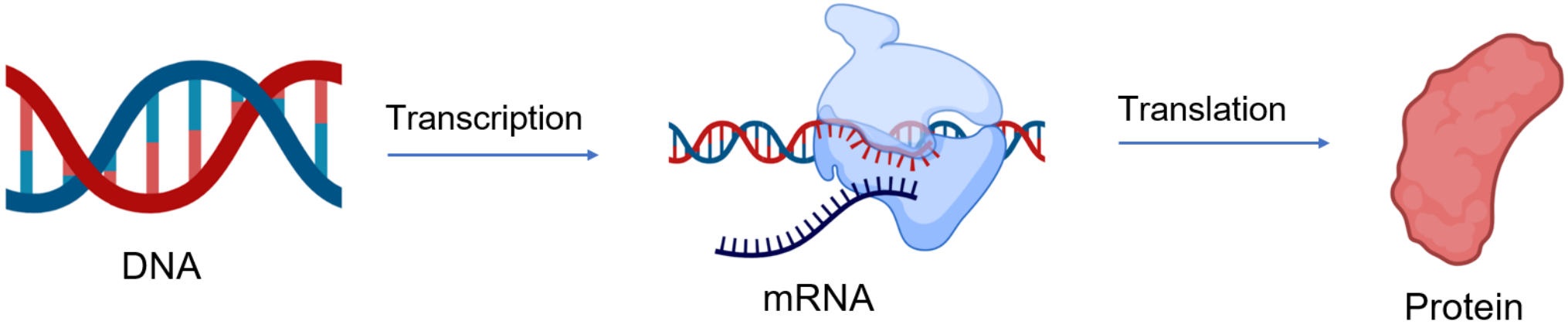
BGGN 239 – Week 2

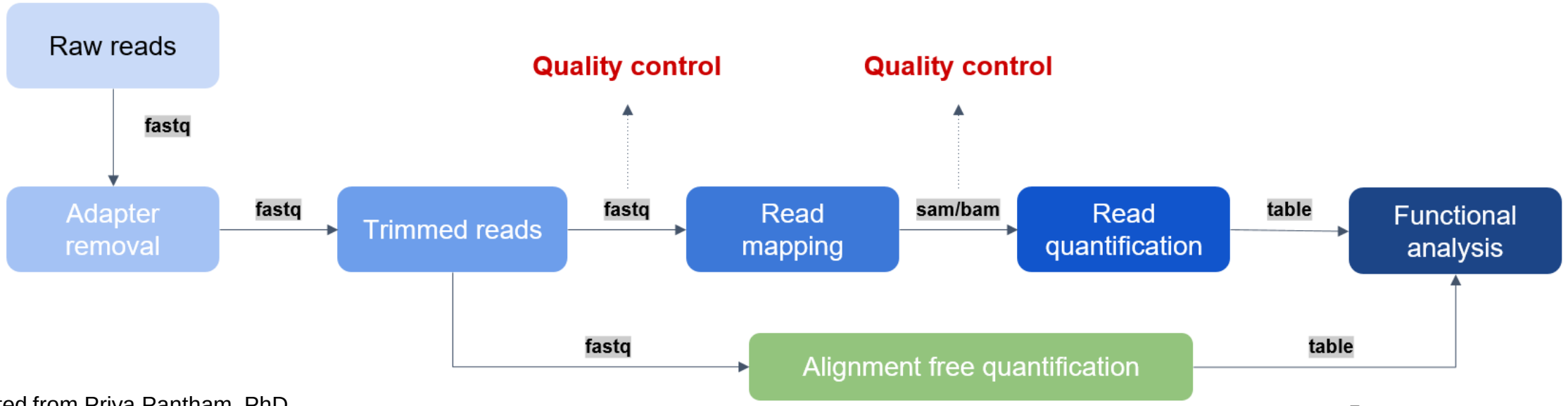
Bulk and single-cell gene expression analysis

Ferhat Ay
ferhatay@lji.org

Associate Professor of Computational Biology, LJI
Department of Pediatrics & BISB PhD Program, UCSD

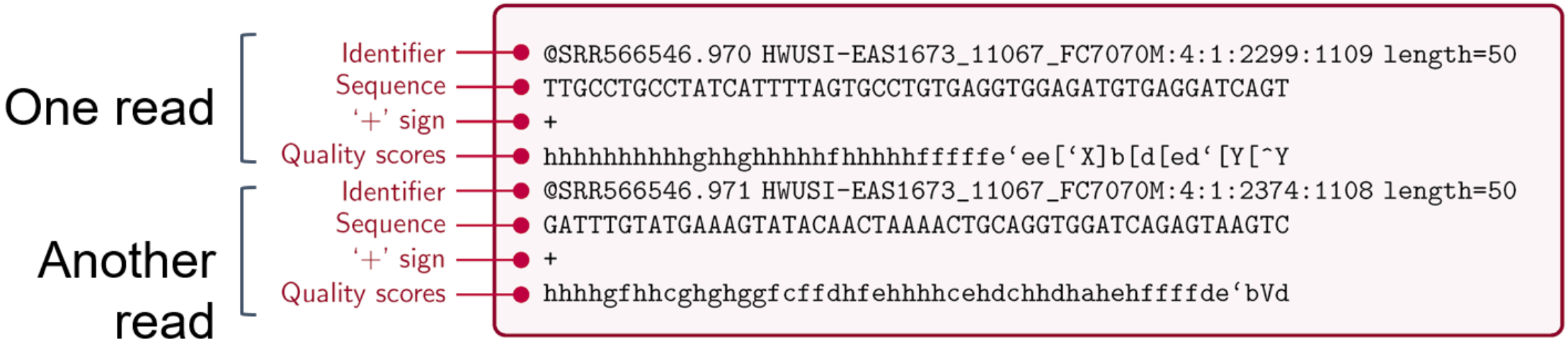
RNA-seq data flow





Slide adapted from Priya Pantham, PhD

Reads and mapping them



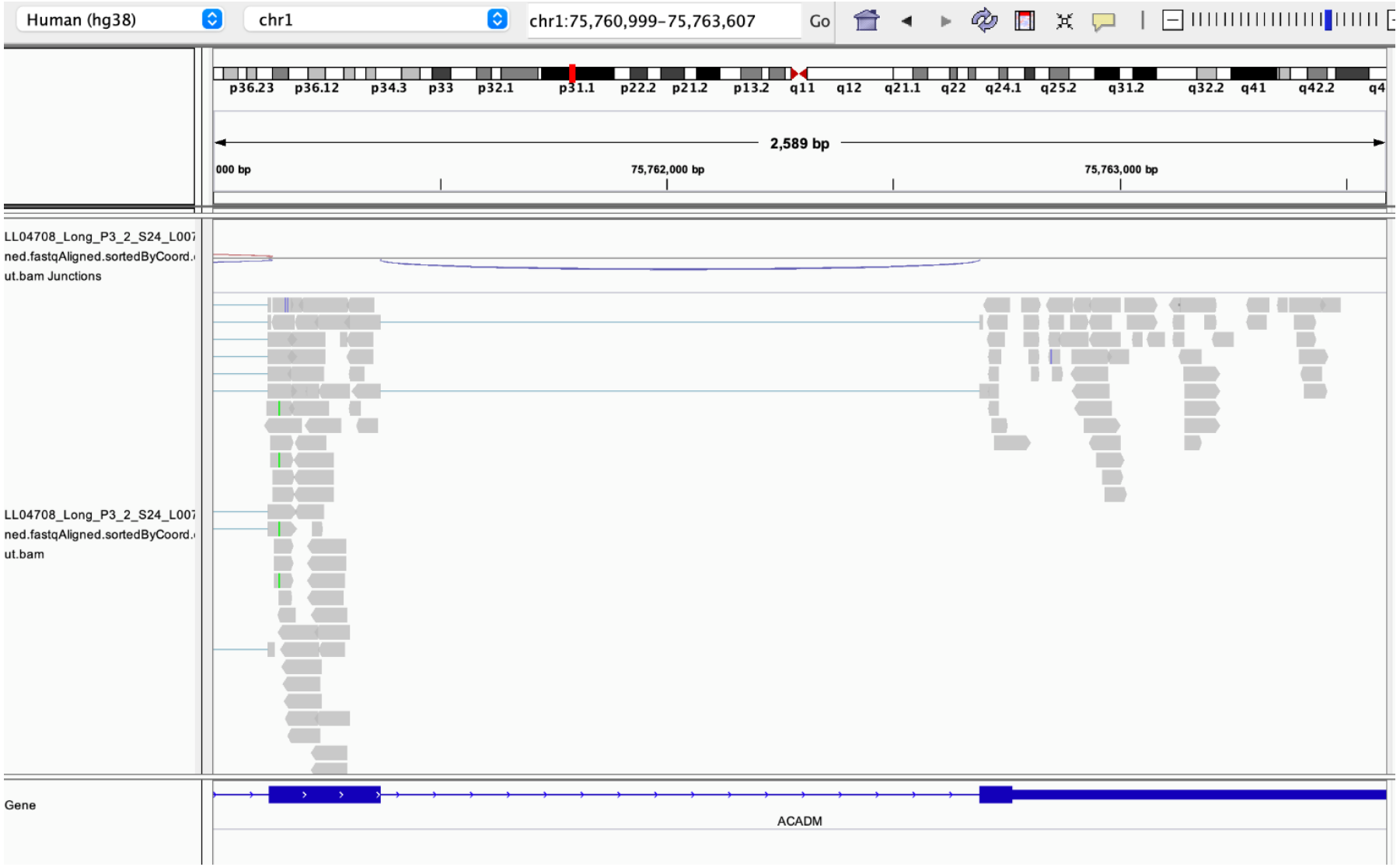
Must match exactly?

Read mapping – popular tools

- STAR, Bowtie, HISAT2, TopHat

Slide adapted from Priya Pantham, PhD

Visualizing mapped reads



Slide adapted from Priya Pantham, PhD

What to do with counts?

Issues:

1. Gene length (does not matter much for across sample comparisons)
2. Read depth
3. GC content, mappability

What to do with counts?

- Reads per million or Counts per million
- Does not account for transcript length
- OK to use for sequencing protocols where reads are generated irrespective of gene length

$$\text{RPM or CPM} = \frac{\text{Number of reads mapped to gene} \times 10^6}{\text{Total number of mapped reads}}$$

What to do with counts?

- RPKM: Reads Per Kilobase of transcript per Million mapped reads
- FPKM*: Fragments Per Kilobase of transcript per Million mapped reads

- FPKM (or RPKM) attempts to normalize for gene size and library depth

*Fragments can mean either individual reads (SE) or paired-end reads that map together (PE)

$$\mathrm{RPKM} = \frac{\text{Number of reads mapped to gene} \times 10^3 \times 10^6}{\text{Total number of mapped reads} \times \text{gene length in bp}}$$

What to do with counts?

- TPM: Transcripts per million (Transcripts Per Kilobase Million)

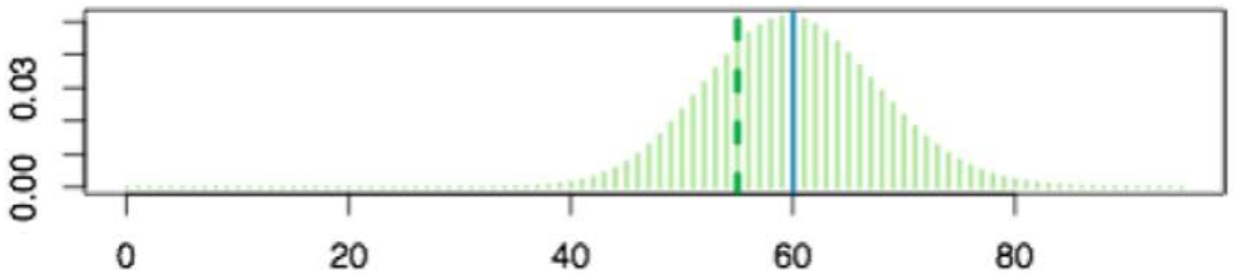
- Another form of normalization for gene length and sequencing depth, but in a slightly different order

$$\mathrm{TPM} = A \times \frac{1}{\sum(A)} \times 10^6$$

$$\text{Where A} = \frac{\text{total reads mapped to gene} \times 10^3}{\text{gene length in bp}}$$

$$\mathrm{TPM} = \frac{RPKM}{\sum(RPKM)} \times 10^6$$

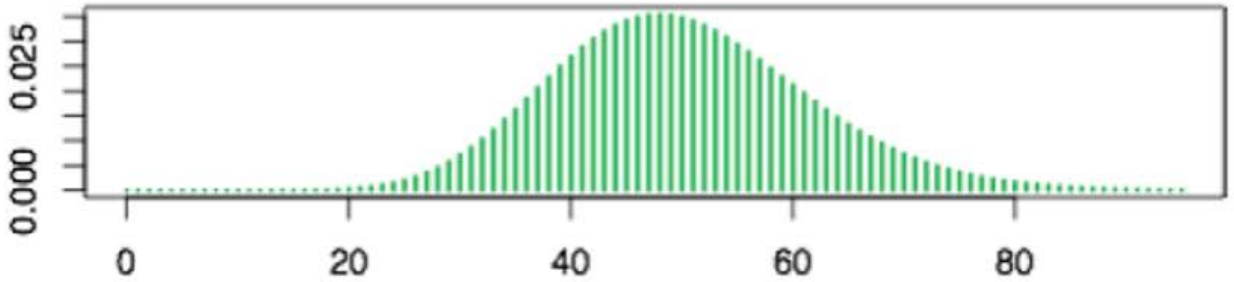
Which distribution better captures count data?



Poisson

- Minimum variance of count data:

$v = \mu$ (Poisson)



Negative binomial

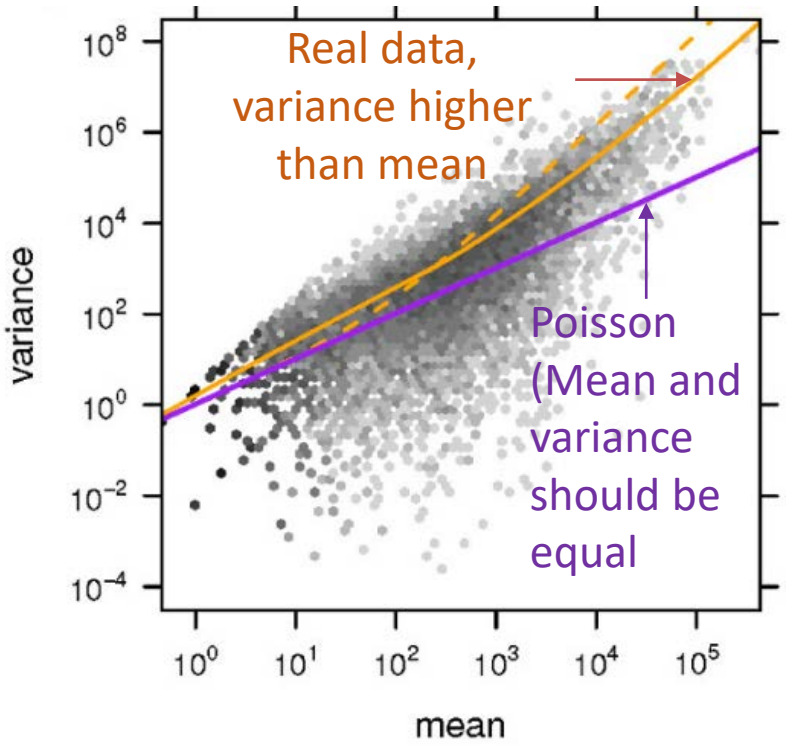
- Actual variance:

$v = \mu + \alpha \mu^2$

Slide adapted from Priya Pantham, PhD

Dispersion matters!

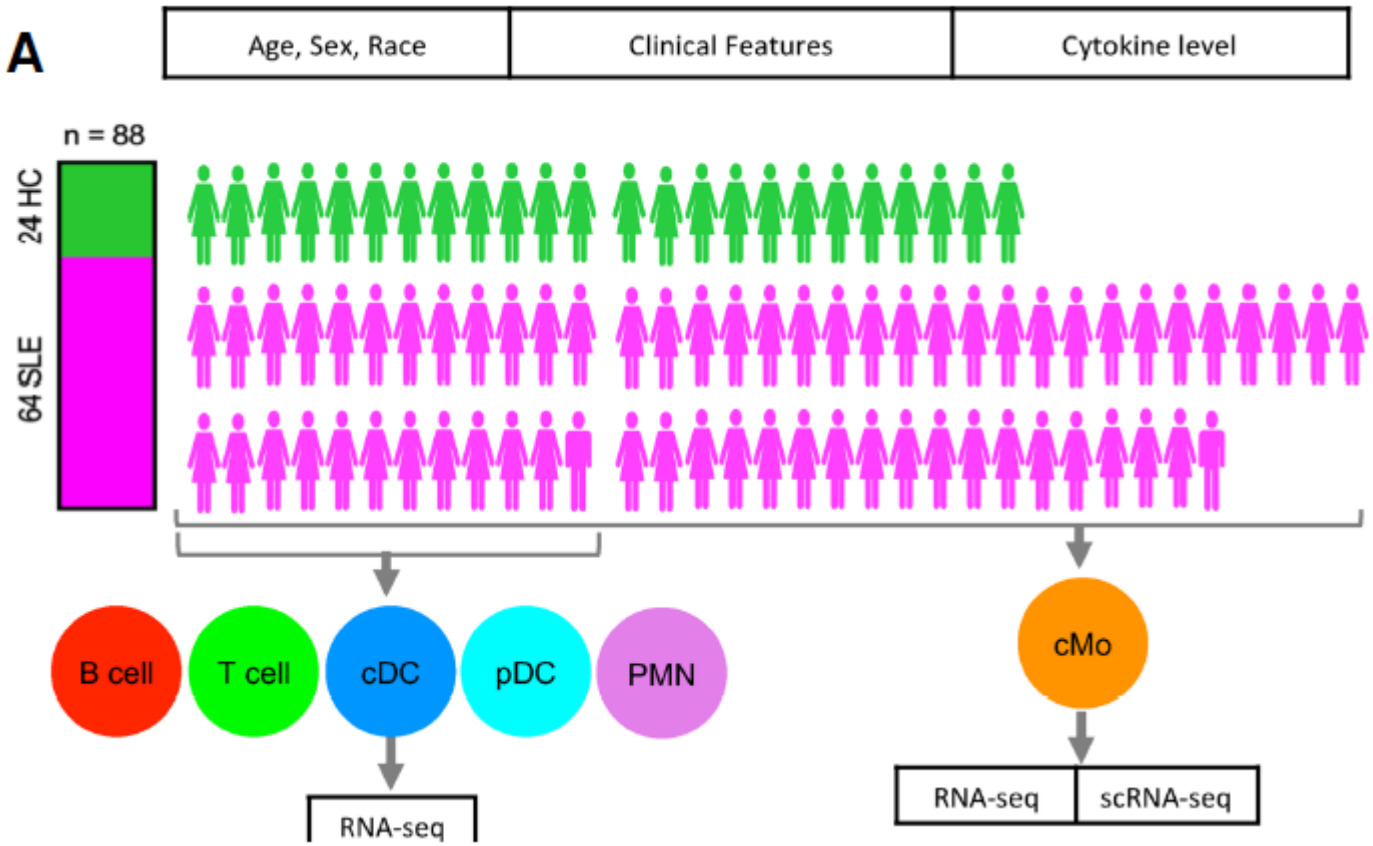
- - : estimate using edgeR
——: fit to real RNA- Seq data
——: Poisson variance for each mean

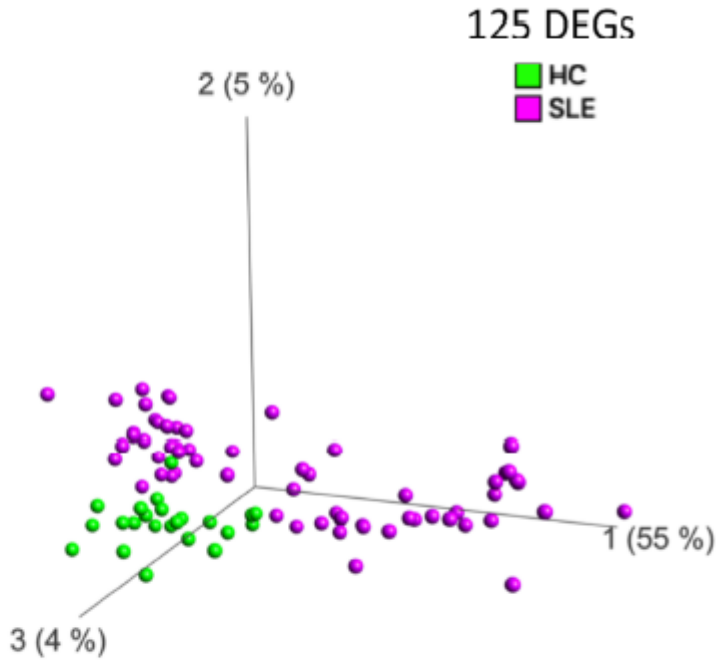


- α : "dispersion" $\alpha = (\mu - v) / \mu^2$
  (squared coefficient of variation of extra-Poisson variability)
- Dispersion is a measure of the spread or variability in the data
- Biological Data is often 'overdispersed'. With increasing mean the variance grows disproportionally
- Negative binomial model can account for this overdispersion

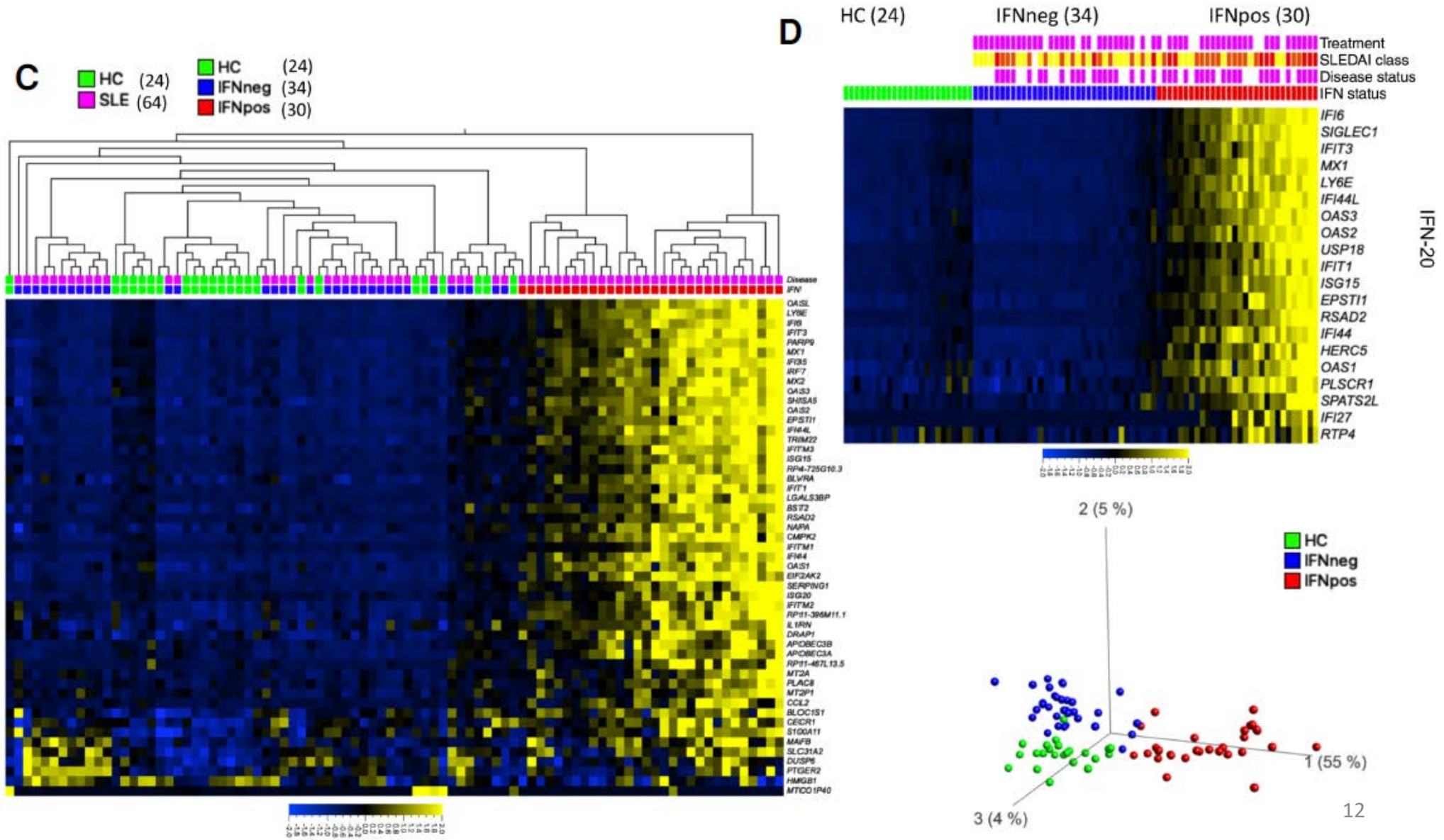
Slide adapted from Priya Pantham, PhD

SLE paper


A
Age, Sex, Race
Clinical Features
Cytokine level
n = 88
24 HC
64 SLE
B cell
T cell
cDC
pDC
PMN
cMo
RNA-seq
RNA-seq
scRNA-seq


B
125 DEGs
HC
SLE
2 (5 %)
1 (55 %)
3 (4 %)

SLE paper


C
HC (24)
SLE (64)
HC (24)
IFNneg (34)
IFNpos (30)
Disease
IFN
OASL
LY6E
IFI6
IFIT3
PARP9
MX1
IFI35
IRF7
MX2
OAS3
SHISA5
OAS2
EPSTI1
IFI44L
TRIM22
IFITM3
ISG15
RP4-725G10.3
BLVRA
IFIT1
LGALS3BP
BST2
RSAD2
NAPA
CMPK2
IFITM1
IFI44
OAS1
EIF2AK2
SERPING1
ISG20
IFITM2
RP11-396M11.1
IL1RN
DRAP1
APOBEC3B
APOBEC3A
RP11-467L13.5
MT2A
PLAC8
MT2P1
CCL2
BLOC1S1
CECR1
S100A11
MAFB
SLC31A2
DUSP6
PTGER2
HMGB1
MTCO1P40
D
HC (24)
IFNneg (34)
IFNpos (30)
Treatment
SLEDAI class
Disease status
IFN status
IFI6
SIGLEC1
IFIT3
MX1
LY6E
IFI44L
OAS3
OAS2
USP18
IFIT1
ISG15
EPSTI1
RSAD2
IFI44
HERC5
OAS1
PLSCR1
SPATS2L
IFI27
RTP4
IFN-20
2 (5 %)
HC
IFNneg
IFNpos
1 (55 %)
3 (4 %)

Homework #1

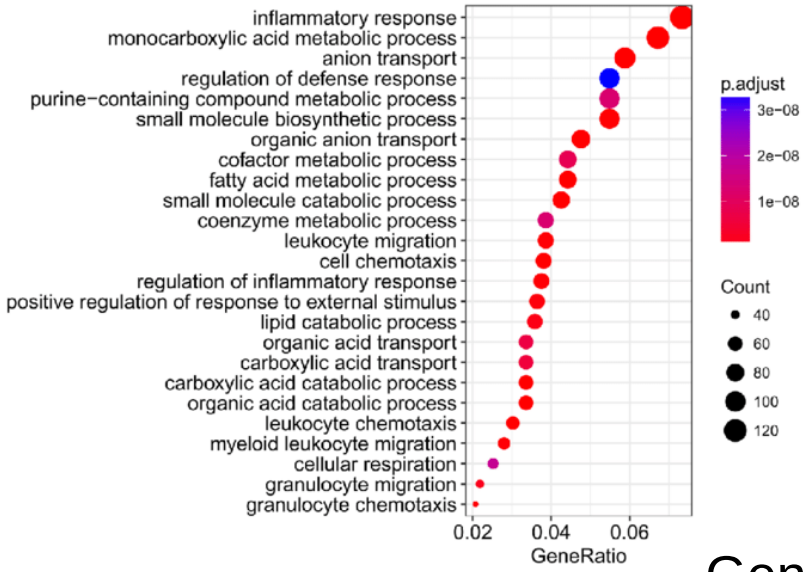
SLE mini project

- Do the same exercise (SLE.qmd) for another cell type and answer the questions below.
  - How many genes were differentially expressed at adjusted p-value cutoff of 5%? how may up and down-regulated?
  - How many genes remain when you filter with log2FC greater than 1 versus absolute log2FC greater than 1?
  - Write the resulting short list of genes (p.adj <0.05 and log2FC >1) in a csv file.
  - From that short list, select one gene and write 3-4 sentences about how this gene in this specific cell type may be relevant to SLE. Ask Google and ChatGPT for help if you like.

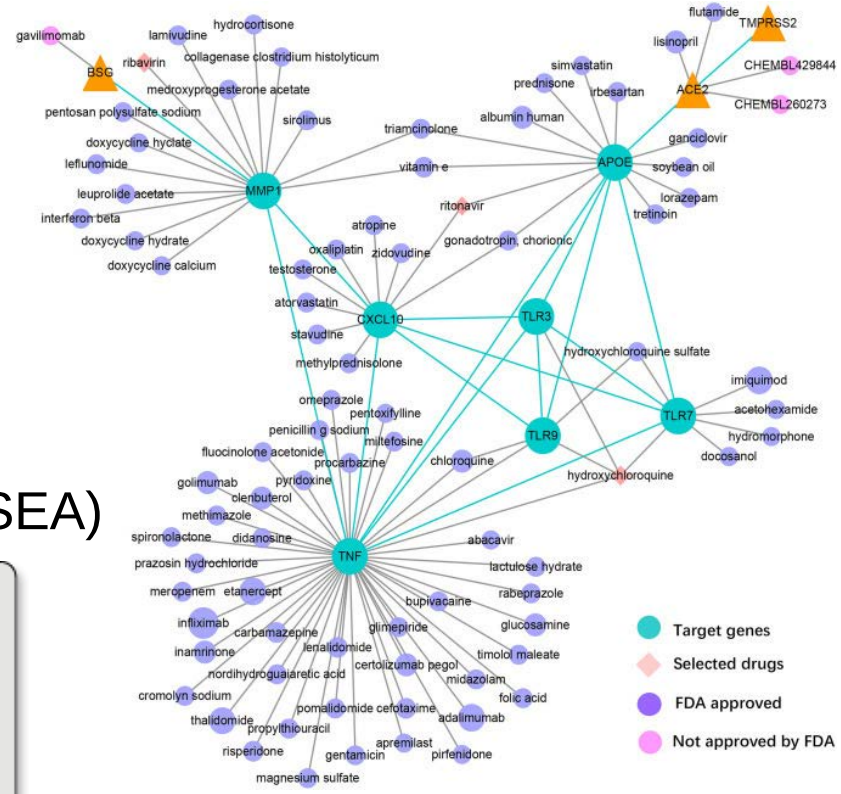
Questions?

Functional analysis of gene sets

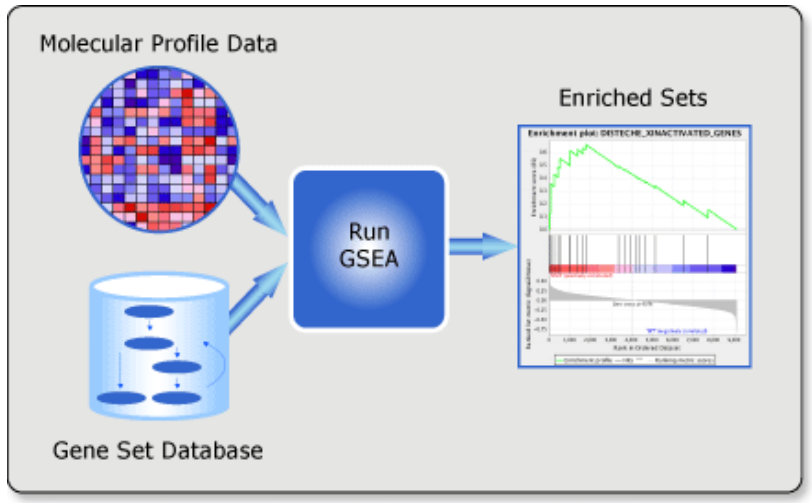
GO Term Enrichment Analysis



Gene Co-expression Networks



Gene Set Enrichment Analysis (GSEA)



Slide adapted from Priya Pantham, PhD

Homework #2

GO Term enrichment

- In the enrichGO function, try setting universe = names(sig_genes) instead of universe = names(all_genes_list). What happened? How many terms are statistically significant now?

- In the enrichGO function, set ont = "CC" rather than ont = "BP". What did this do? Do you believe BP or CC will be more relevant for most use-cases?

- Go to the NYU link and select one other visualization you like to use. Add a code chunk that generates this visualization. https://learn.gencore.bio.nyu.edu/rna-seq-analysis/over-representation-analysis/

Thank You!

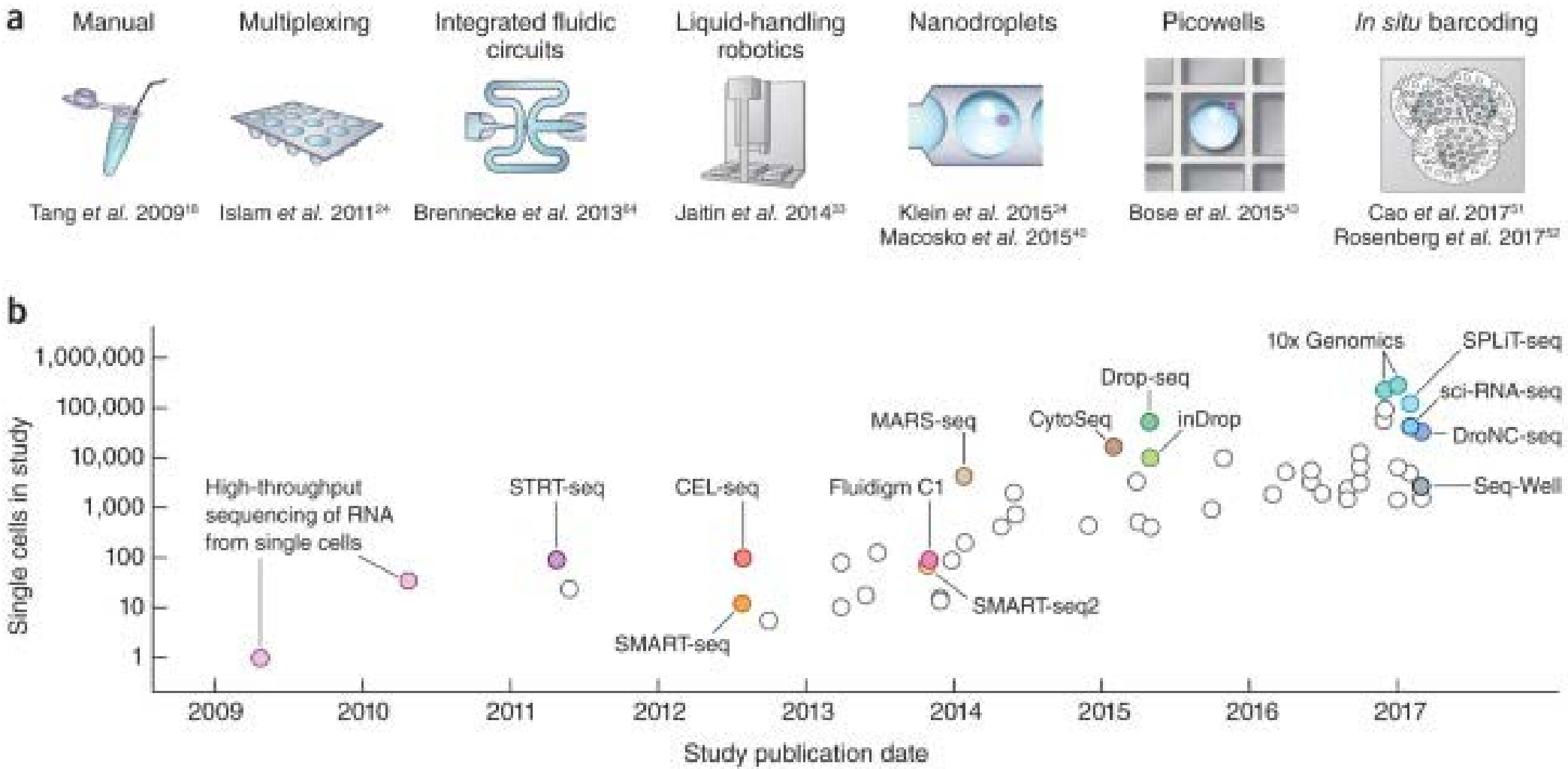
- Paramita Dutta - LJI
- Priya Pantham - UCSD
- Barry Grant - UCSD

Resources

- https://learn.gencore.bio.nyu.edu/rna-seq-analysis/over-representation-analysis/
- https://allisonhorst.com/r-packages-functions
- http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
- http://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html

single-cell!

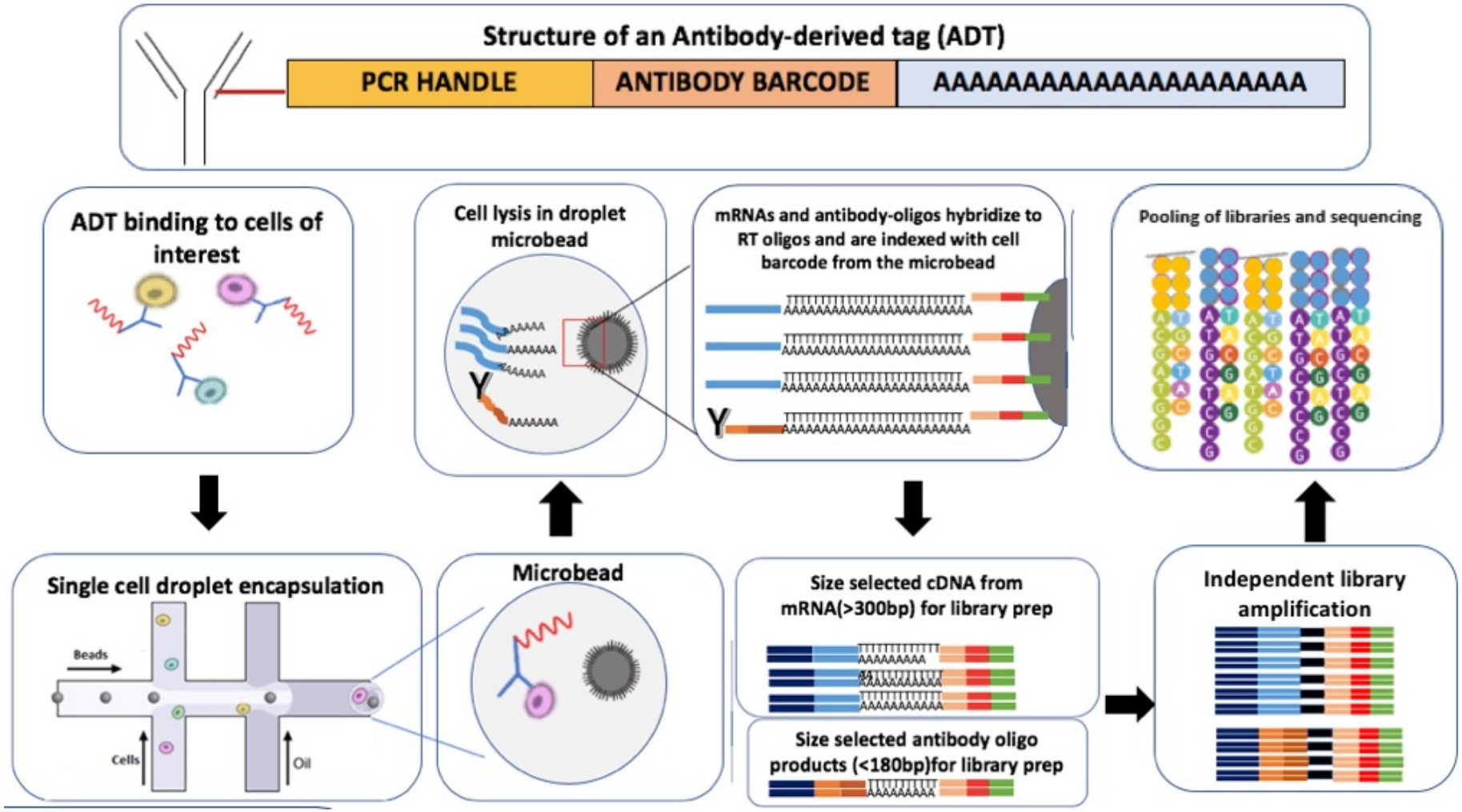
Evolution of single cell techniques (till 2018)



Valentine Svensson, Roser Vento-Tormo & Sarah A Teichmann

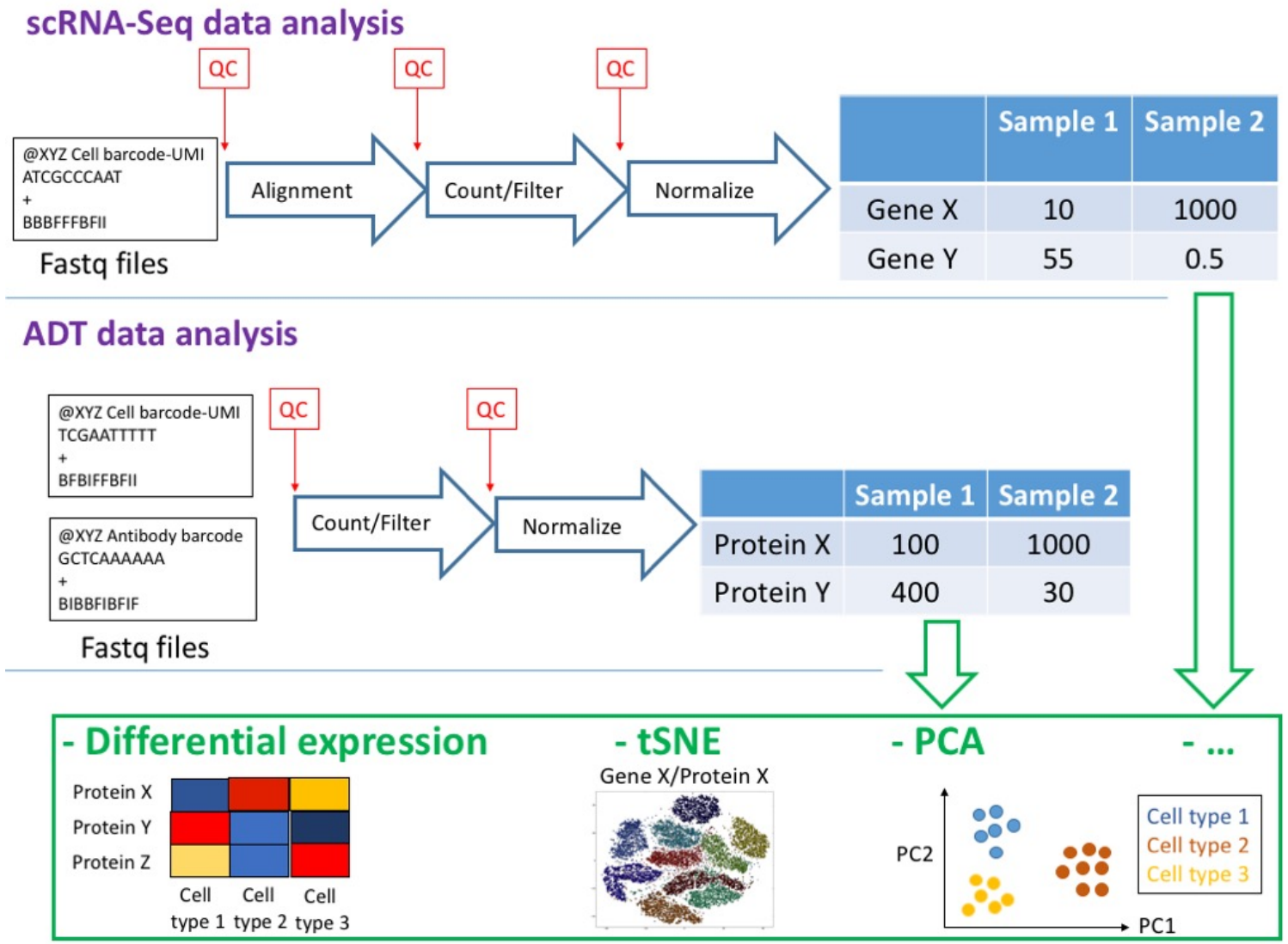
Nature Protocols **13**, 599–604 (2018) | Cite this article

CITE-seq workflow - wet



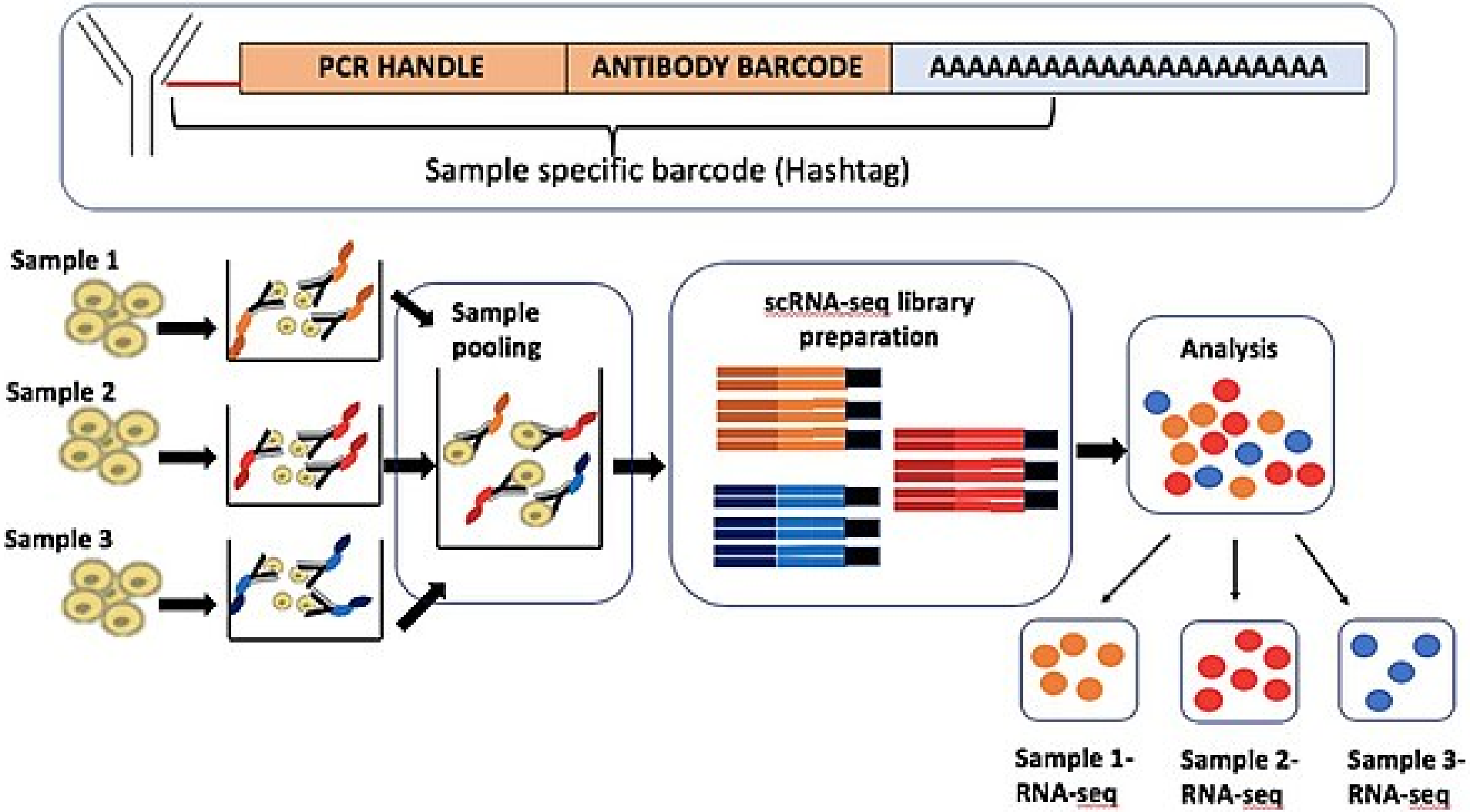
Nature Methods **volume 14**, pages865–868 (2017)

CITE-seq workflow - dry

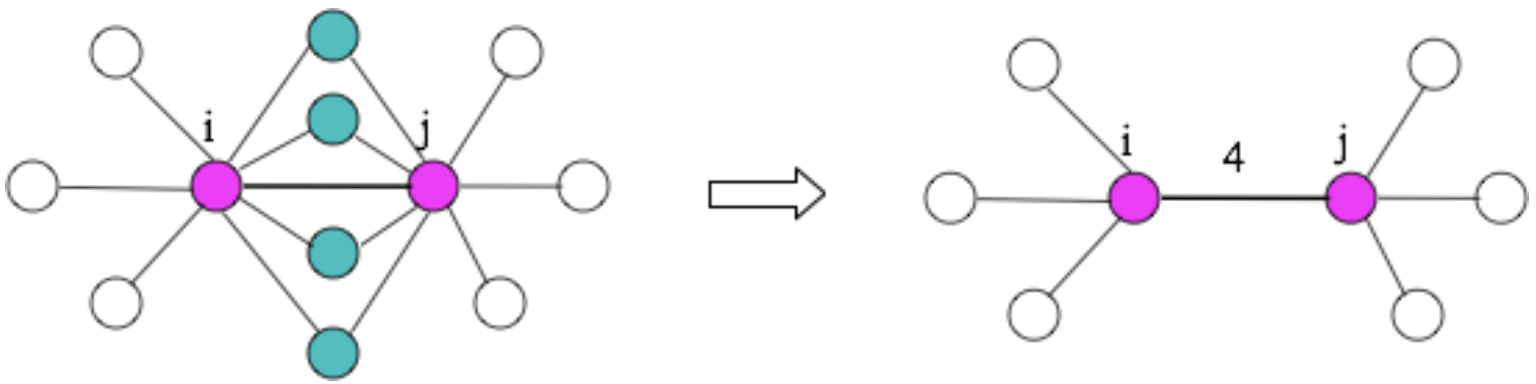


Nature Methods **volume 14**, pages865–868 (2017)

Things you can do with ADTs



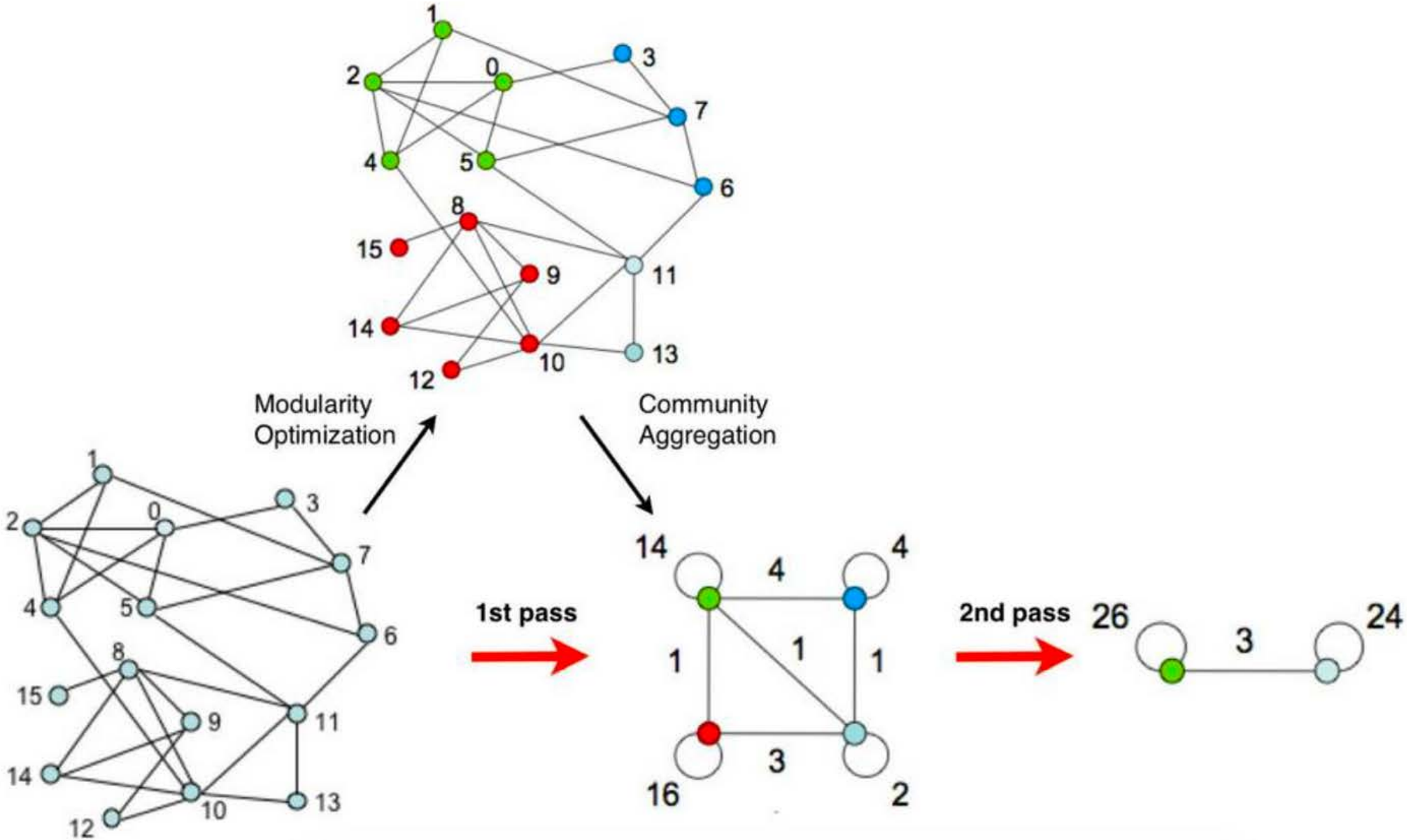
Shared nearest neighbor graph (SNN)



Louvain algorithm

- Is a greedy algorithm
- Weighted graphs
- This algorithm has been widely utilized in many application domains because of:
  - Its rapid convergence properties
  - High modularity
  - Hierarchal partitioning

Louvain algorithm



Louvain algorithm

Algorithm [edit]

The value to be optimized is modularity, defined as a value in the range $[-1/2, 1]$ that measures the density of links inside communities compared to links between communities.[1] For a weighted graph, modularity is defined as:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j),$$

where

- A_{ij} represents the edge weight between nodes i and j;
- k_i and k_j are the sum of the weights of the edges attached to nodes i and j, respectively;
- m is the sum of all of the edge weights in the graph;
- c_i and c_j are the communities of the nodes; and
- δ is Kronecker delta function ($\delta(x, y) = 1$ if $x = y$, 0 otherwise).

Based on the above equation, the modularity of a community c can be calculated as:

$$Q_c = \frac{\Sigma_{in}}{2m} - (\frac{\Sigma_{tot}}{2m})^2,$$

where

- Σ_{in} is the sum of edge weights between nodes within the community c (each edge is considered twice); and
- Σ_{tot} is the sum of all edge weights for nodes within the community (including edges which link to other communities).

Resources

- Review: https://www.nature.com/articles/nprot.2017.149
- 10x datasets: https://www.10xgenomics.com/resources/datasets
- CITE-seq: https://www.nature.com/articles/nmeth.4380
- Ab-seq: https://www.nature.com/articles/srep44447
- Louvain algorithm: https://towardsdatascience.com/louvain-algorithm-93fde589f58c