Basic RNASeq DGE analysis using R
HSPH-IID Virtual Workshop

Quantitative Biomedical Research Center (QBRC)
email: qbrc@hsph.harvard.edu

Basic RNASeq Data exploratory Analysis

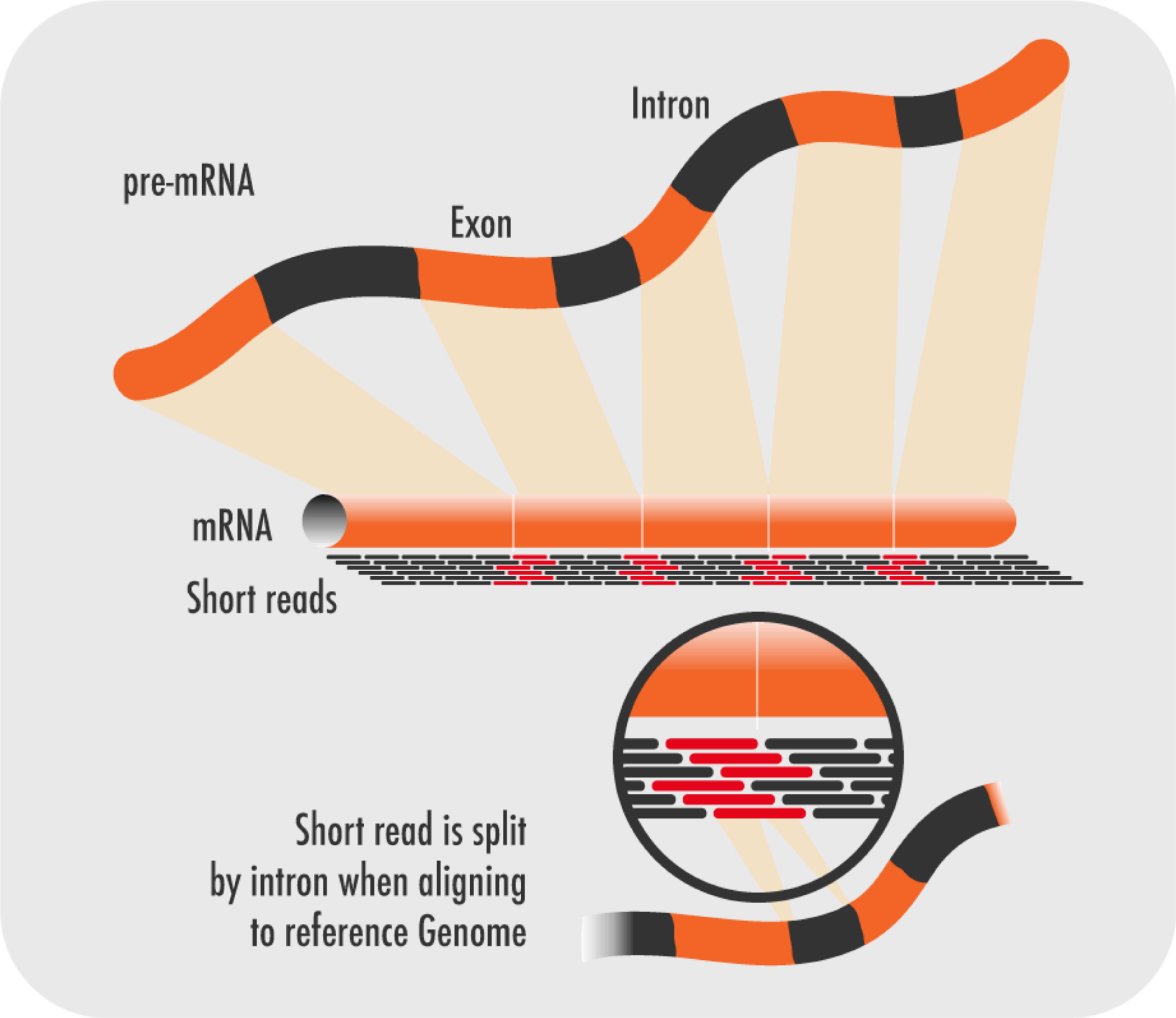
Identify patterns that are biologically meaningful

Objectives

- Basic bulk RNASeq experiment workflow
- RNASeq read data processing, quantification, and normalization
- Basic methods in exploratory analysis

What is RNA-Seq

- Using NGS technology to sequence RNA transcripts
- Typically refers to the sequencing of mRNA
- Different RNA species (i.e. miRNA, snoRNA, tRNA) require different preparation protocol
- Any type of RNA from any sample sources, such as cell, body fluid, stool, water, etc. can be the sequenced
- Sample from different sample sources, such as cell, body fluid, stool, water, etc, require different extraction method
-

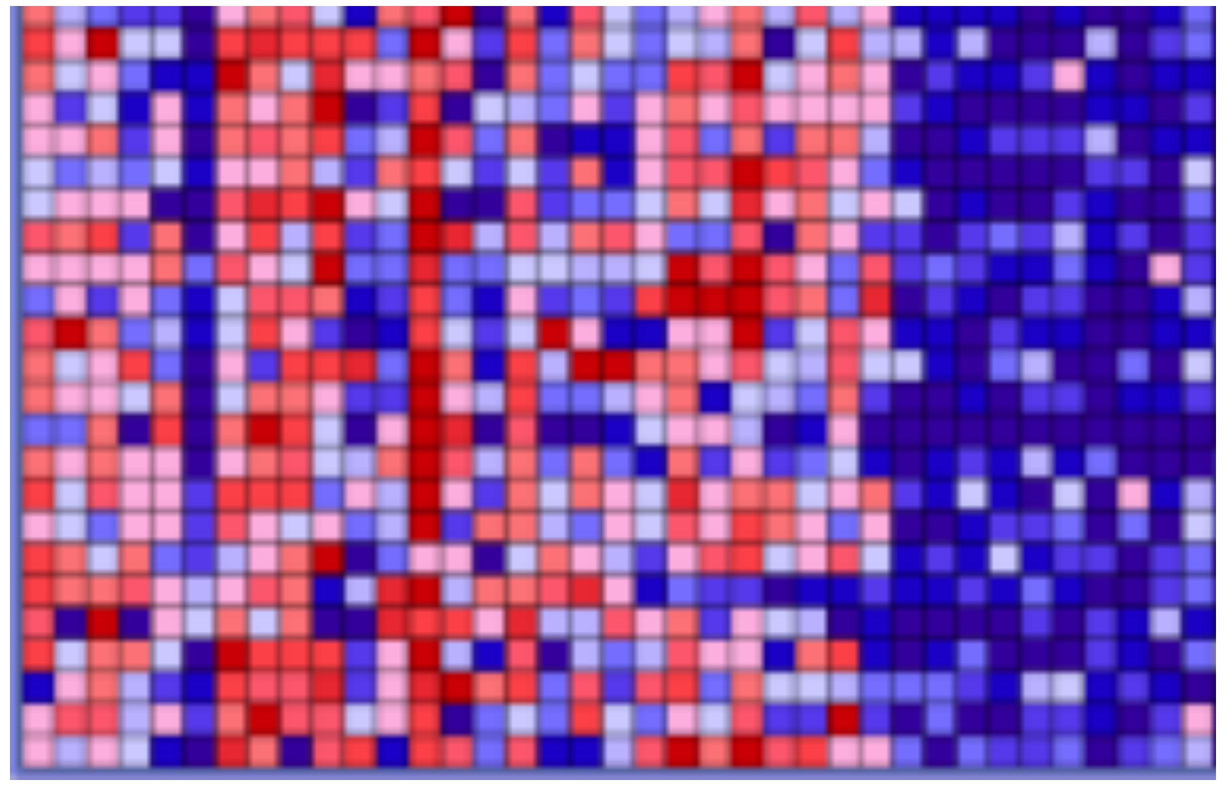


MacKenzie, 2018, Tech. Network.com

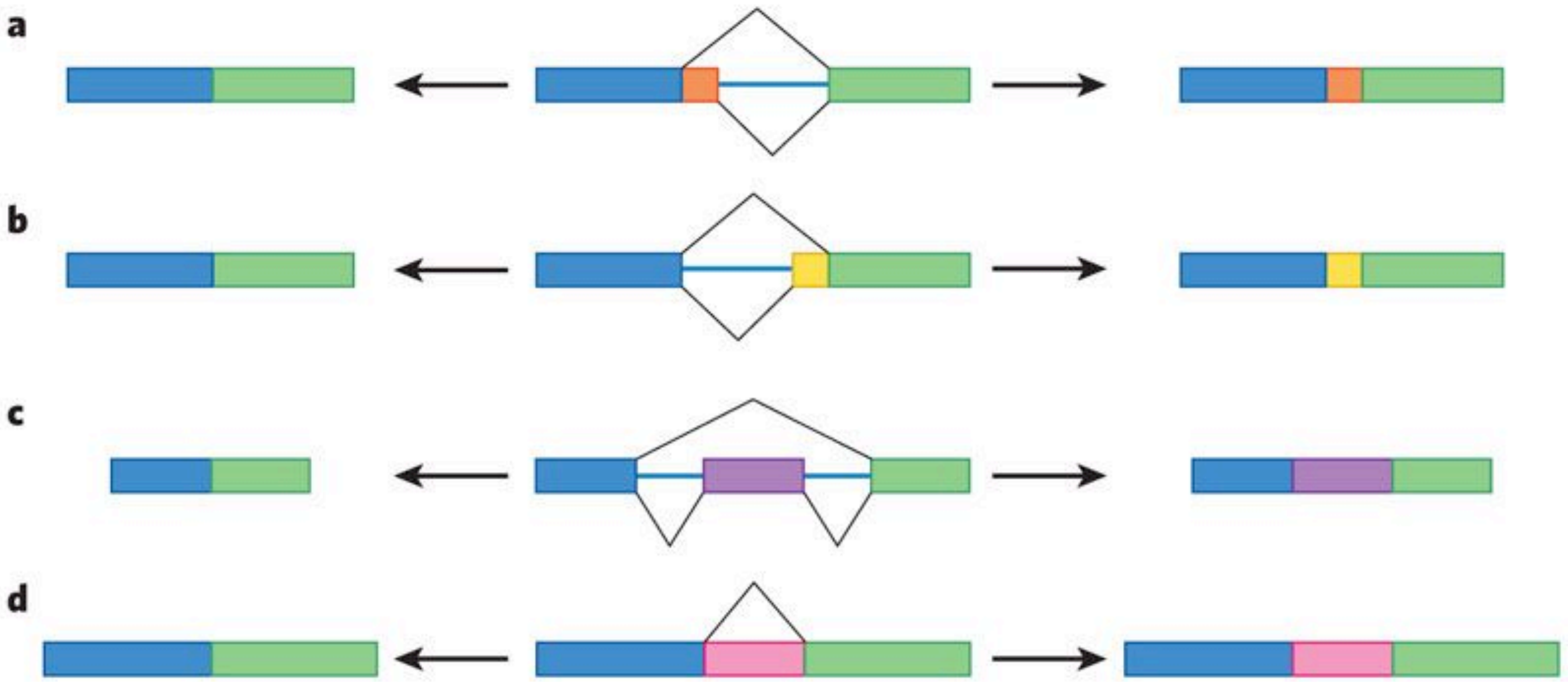
Why do RNA-Seq?

- Which genes are differentially expressed in different conditions?
- Are genes being transcribed in alternatively spliced transcript isoforms?
- Are there mutations being transcribed such as insertions, deletions, or novel isoforms?

Transcriptomic Profiling

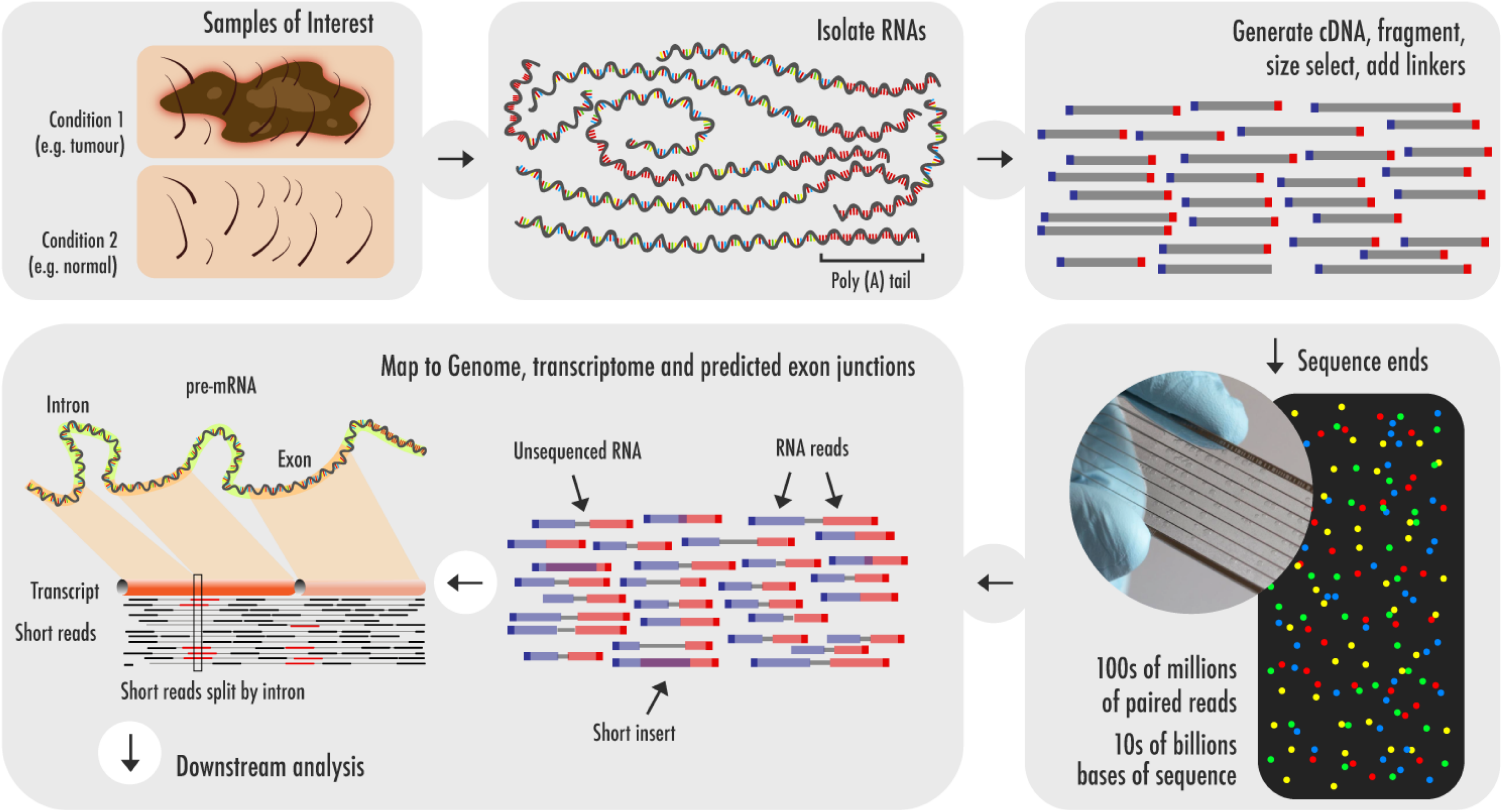


Basic types of alternative splicing



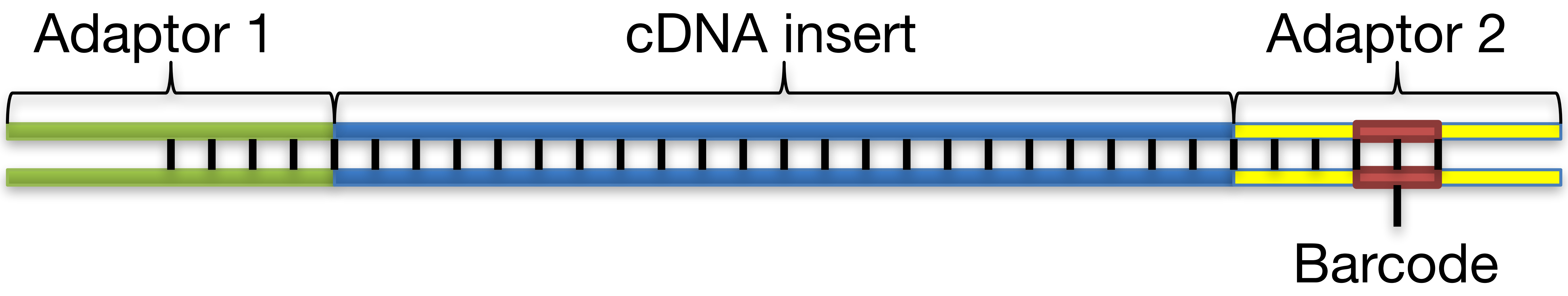
Nielsen and Graveley, 2010, Nature

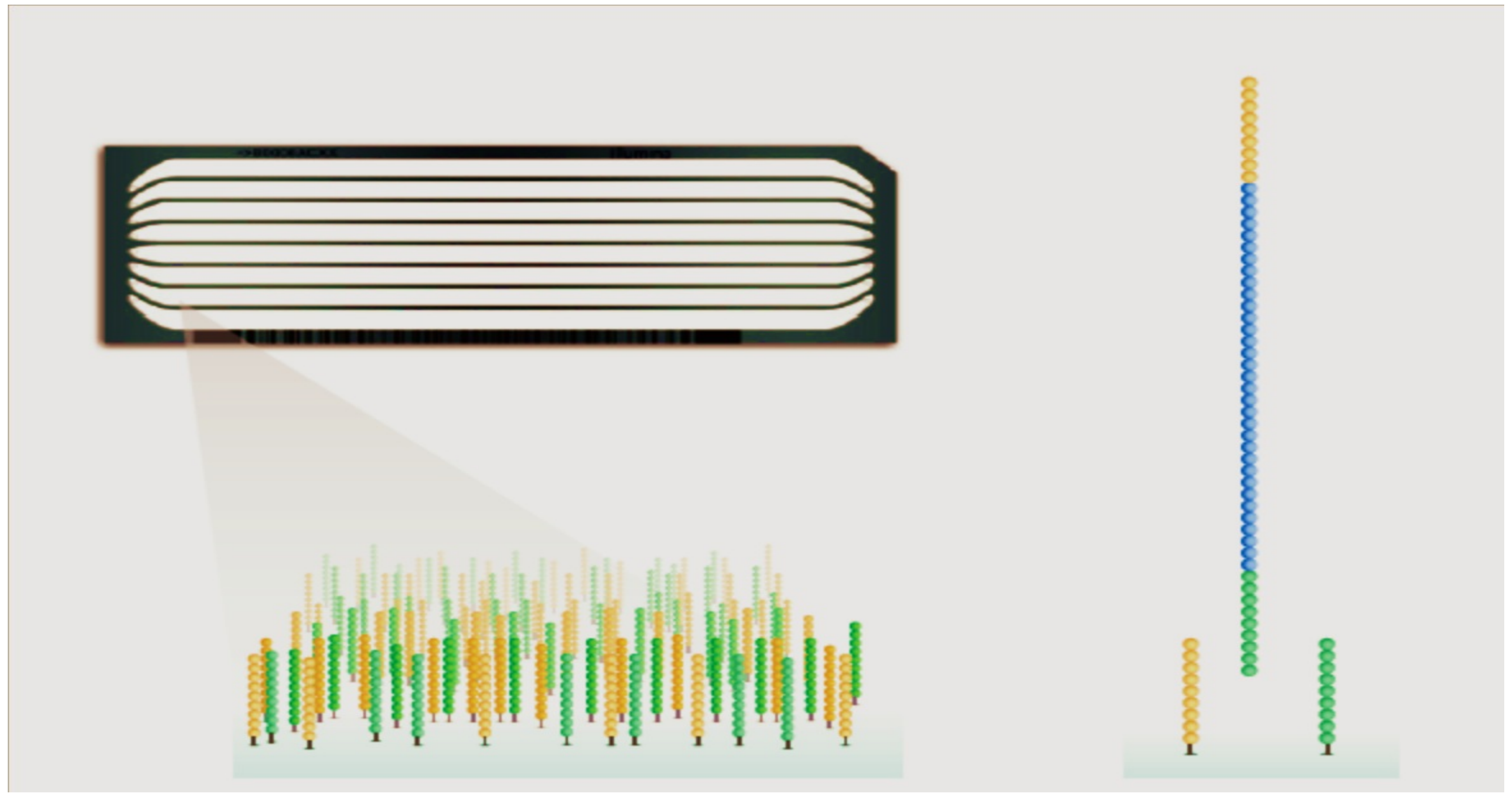
RNA-Seq Experiment Workflow



MacKenzie, 2018, Tech. Network.com

Sequencing Library Structure





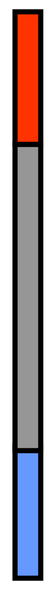
Adaptor – 58 bp nucleotide sequence to fix sequence library onto flow cell

Barcode – optional index sequence for sample multiplexing

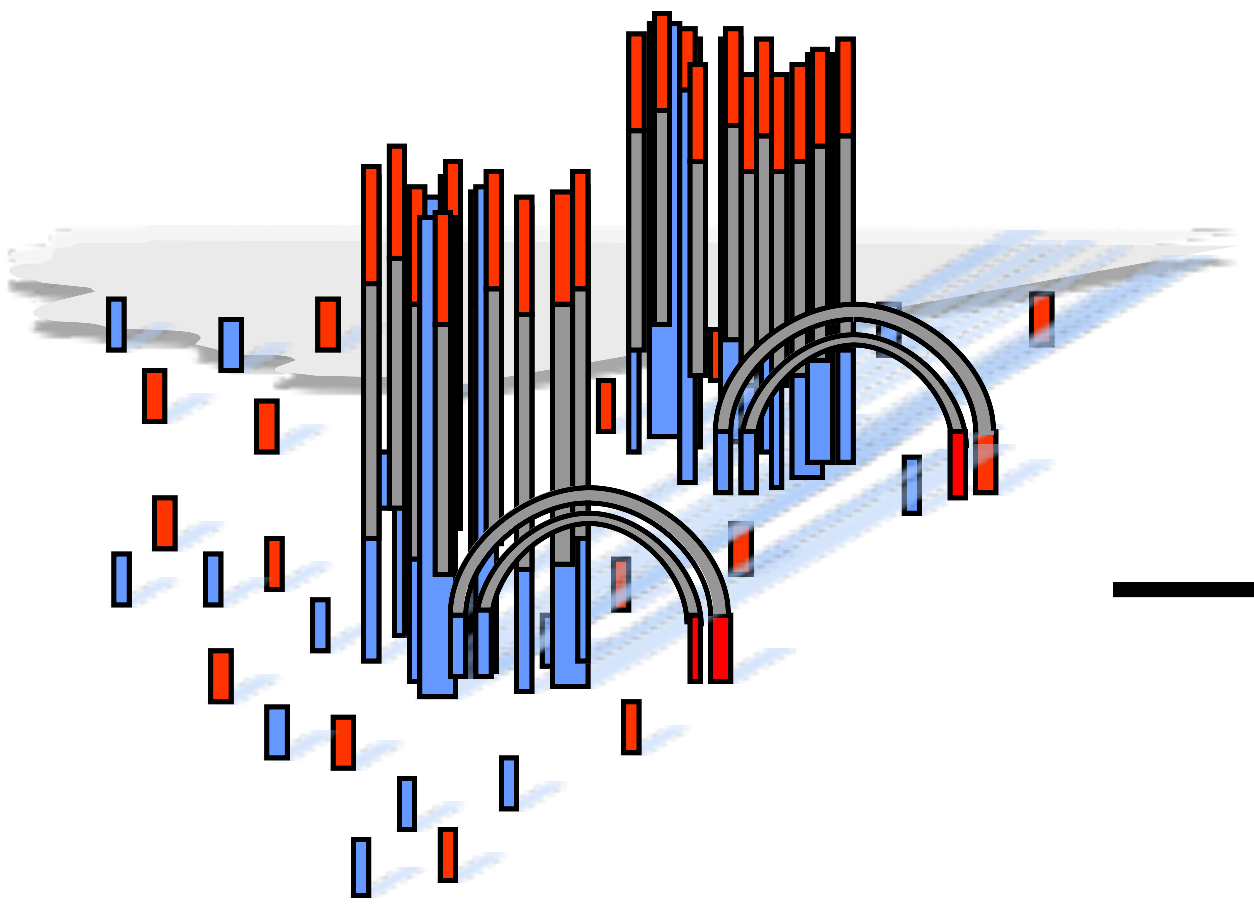
cDNA insert – fragmented cDNA sequence generated from mRNA of interest. The insert typically range between 300-500bp for mRNA

Sequence by Synthesis (SBS)

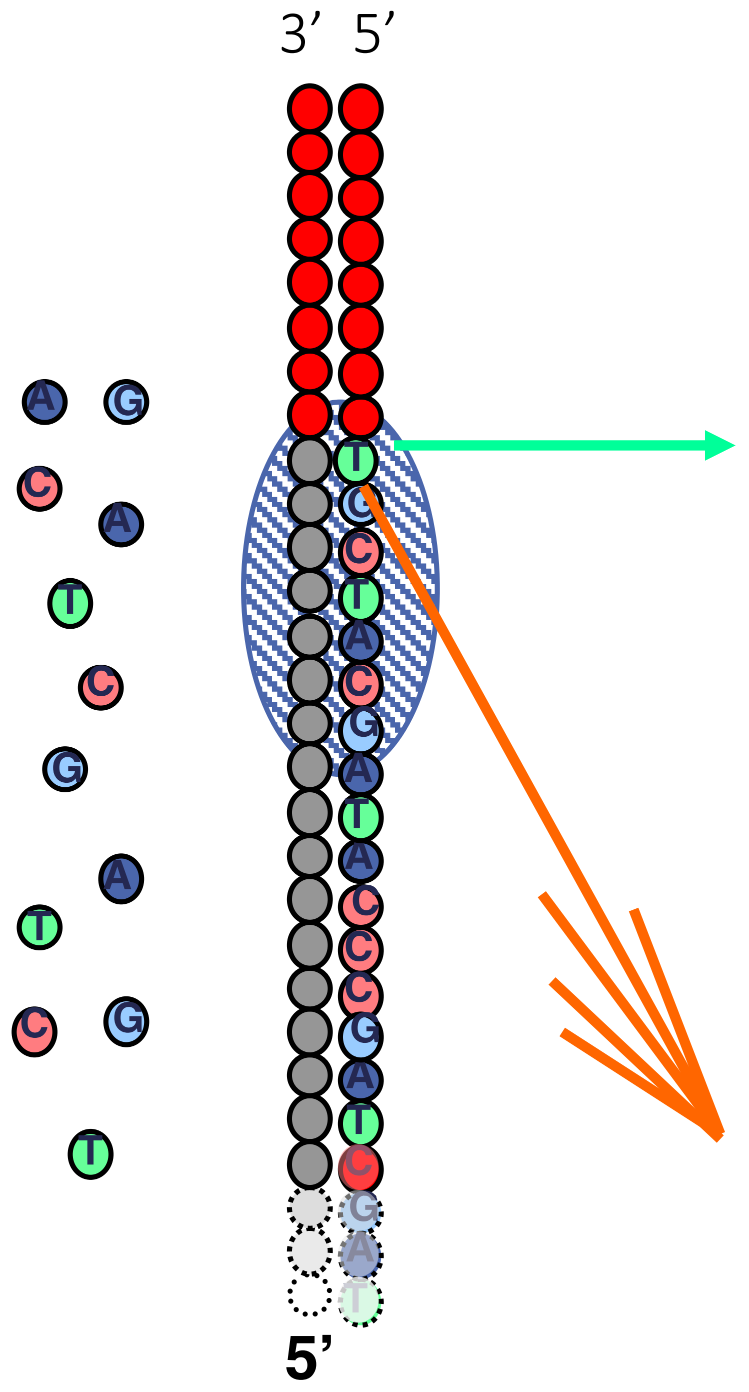
cDNA



DNA library with
ligated Adaptors



Cluster growth by Bridge
Synthesis



Sequence DNA by using
nucleotide bases with color dye

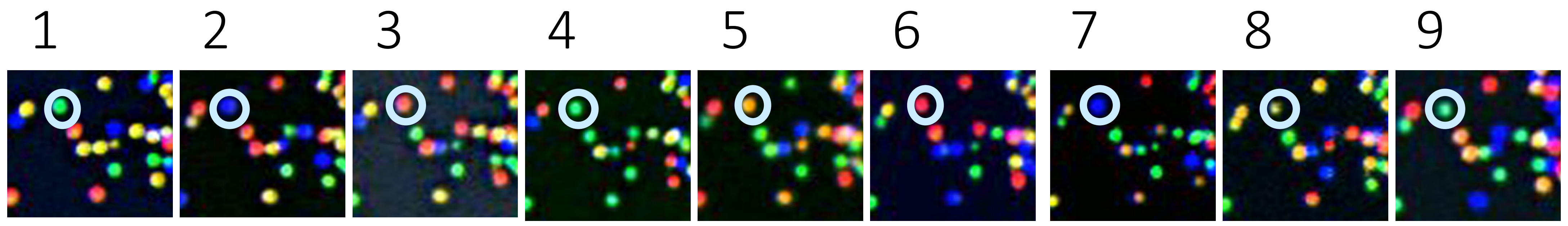
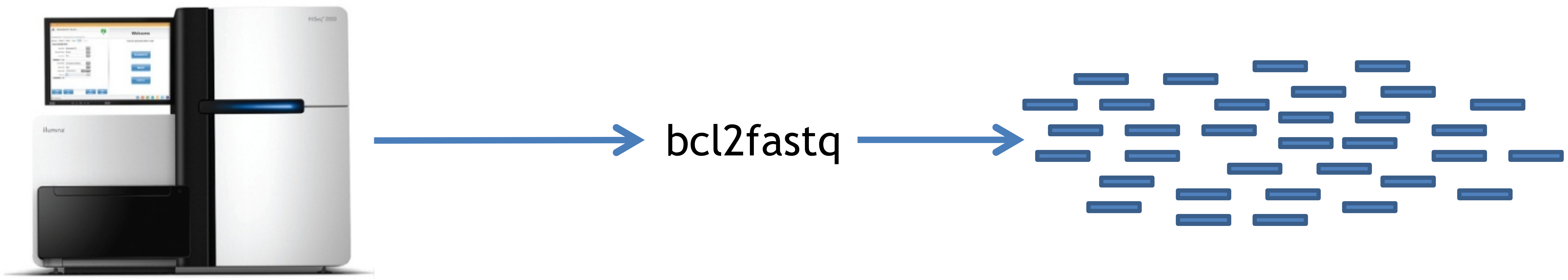


Image acquisition

TGCTACGAT...

Base calling

Reads are ready.



bcl2fastq

Big Fastq files (2-30Gb)

- Reads represent real biology.
- More reads corresponding to a transcript indicate higher abundance of that transcript.
- Reads may represent novel transcripts or novel arrangements of exons that are not present in any known reference genome.

Most common questions asked from RNA-Seq data?

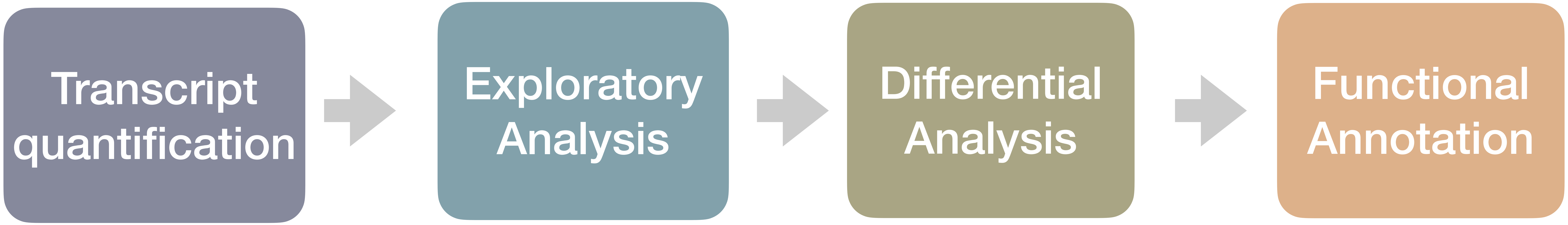
Is there biases affecting the results?

What samples are similar/different ?

What genes are differentially over/under-expressed ?

What are the functional pathways affected by these genes ?

RNA-Seq Differential Expression Analysis Workflow



Transcript quantification → Exploratory Analysis → Differential Analysis → Functional Annotation

Transcript quantification

- Quantify genes and transcripts
- Computationally intensive
- Analytical services are available*

Exploratory Analysis

- Unsupervised analysis
- Use to discover biases and unexpected variability
- **PCA and Hierarchical Clustering**

Differential Analysis

- Guided analysis
- Test experimental hypothesis
- Identify important
- features/genes
- **DESeq2**

Functional Annotation

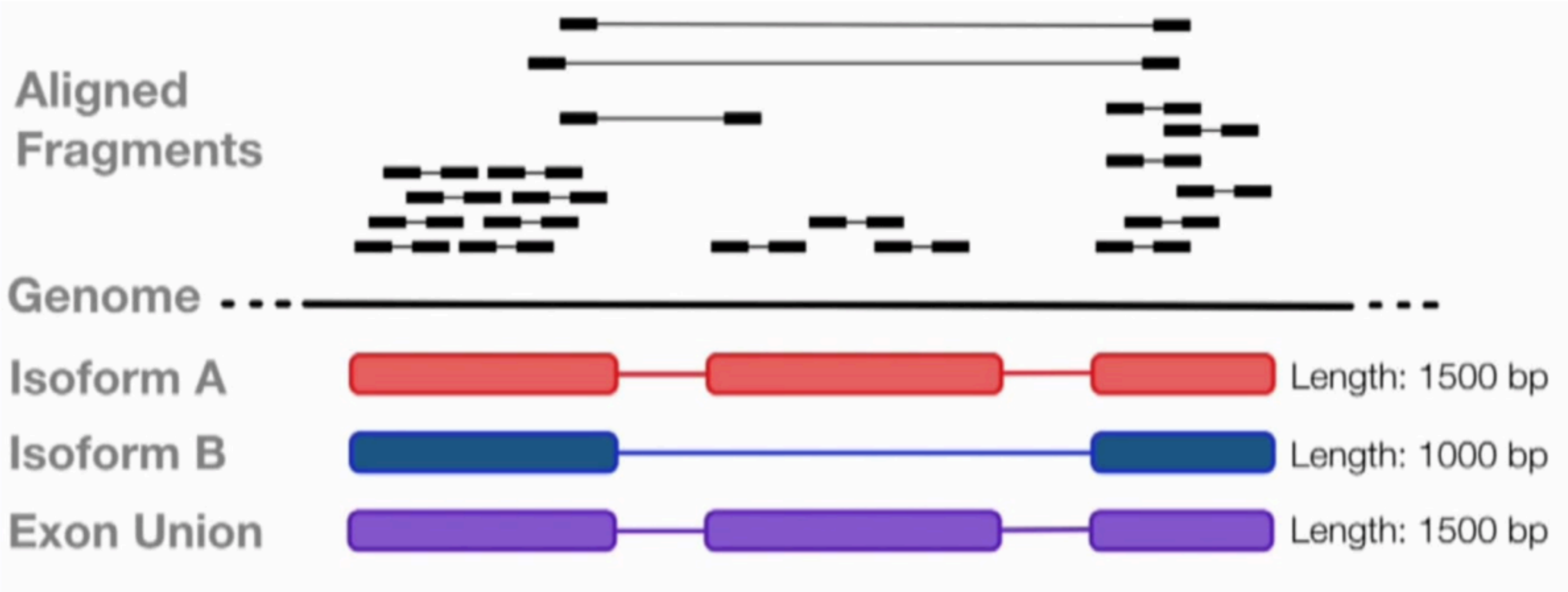
- Interpret the biology
- Find molecular functions or pathway affected by different conditions
- **MSigDB** and **GSEA**

* https://www.hsph.harvard.edu/qbrc/cloud-services/cnap/

Transcript Quantification

Counting reads and quantifying gene expression across different samples for comparison

Aligning to Transcript model



Different philosophies of transcript quantification

- **Alignment-based:** Sequence alignment followed by counting of reads overlapping with a given annotated gene

- **Pseudo-Alignment:** Sometimes called ‘Alignment Free’, Quantify the number of reads that are consistent with a given transcript (the exact location within the transcript is ignored)

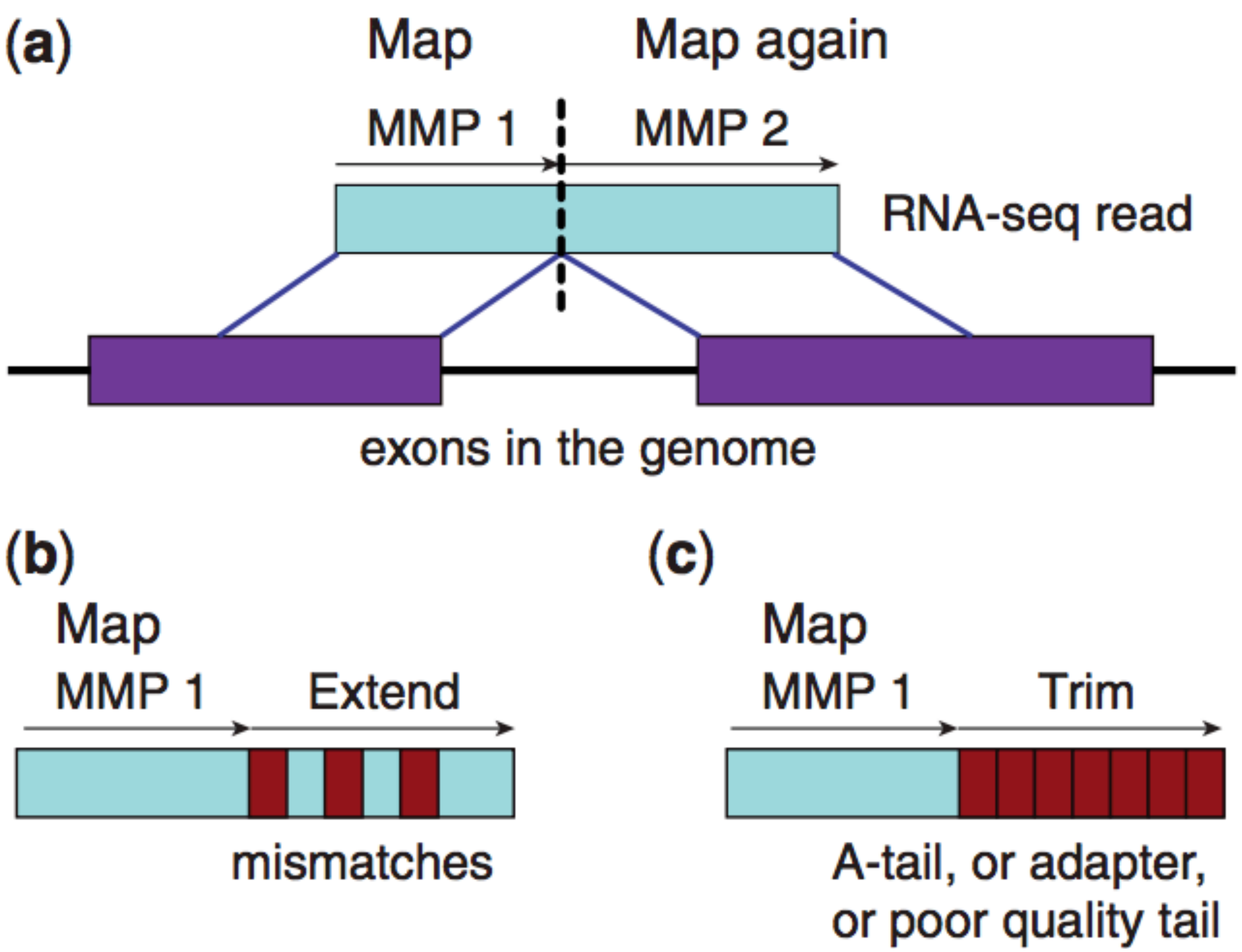
Alignment-based: STAR Aligner

Rationale: Mapping of split reads is computationally slow.

Solution:

1. Use K-mer index
2. look for maximal mappable prefix/ maximum matching portion (MMP) by extension
3. Split only when extension is not possible
4. Pieces back the split reads

Results: Much faster and more sensitive algorithm in detecting transcript

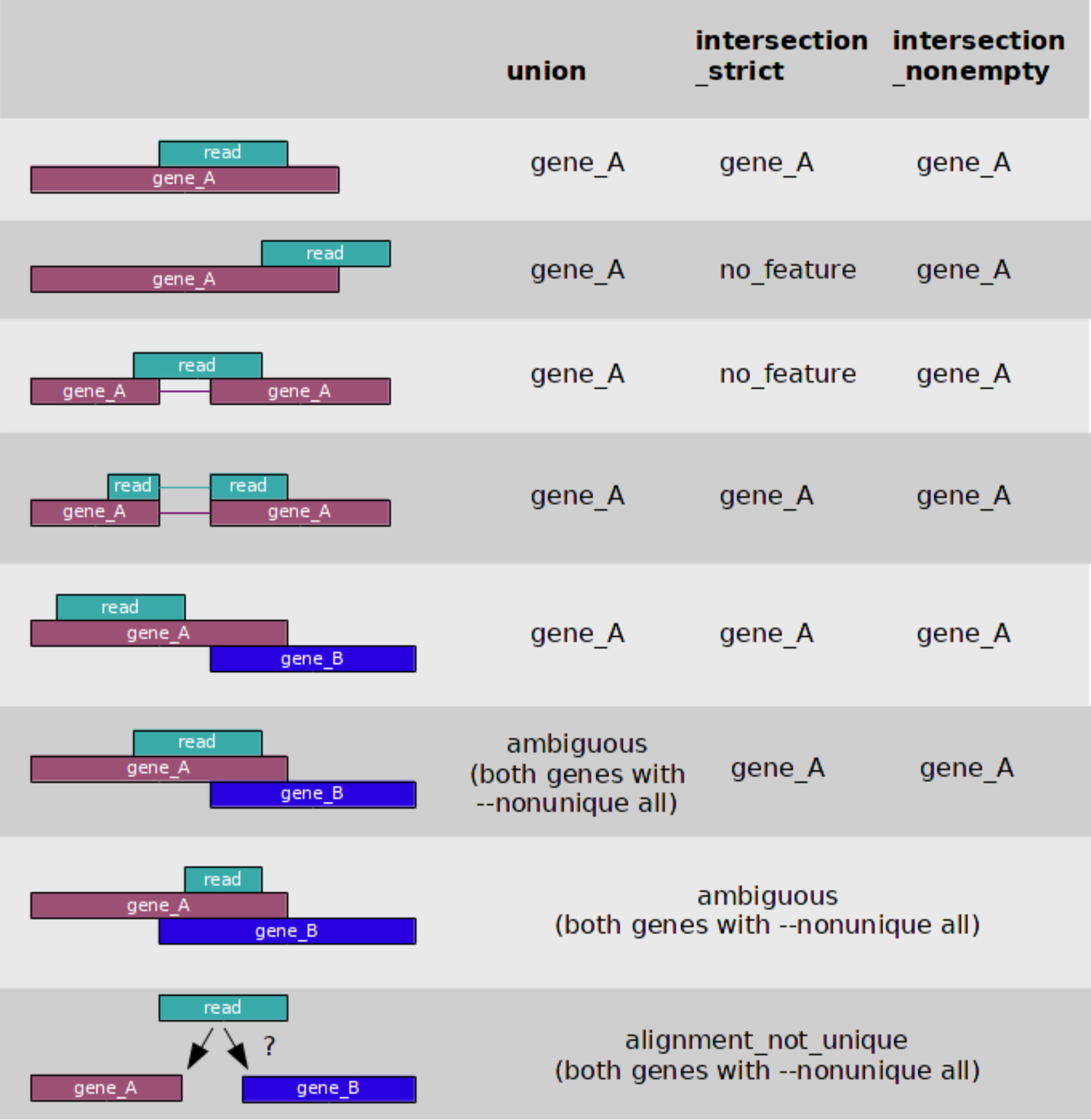


https://github.com/alexdobin/STAR

Dobin, et al. (2013) Bioinfo.

Counting Mapped Reads by HTSeq

- Operate on aligned BAM files
- Total number of reads aligned to a gene is used as surrogate for gene expression level (called ‘raw read counts’)
- However, how reads are being counted is very much user defined !!!



	union	intersection _strict	intersection _nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read gene_A gene_B	ambiguous (both genes with --nonunique all)		
read ? gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

https://htseq.readthedocs.io/en/release_0.11.1/count.html

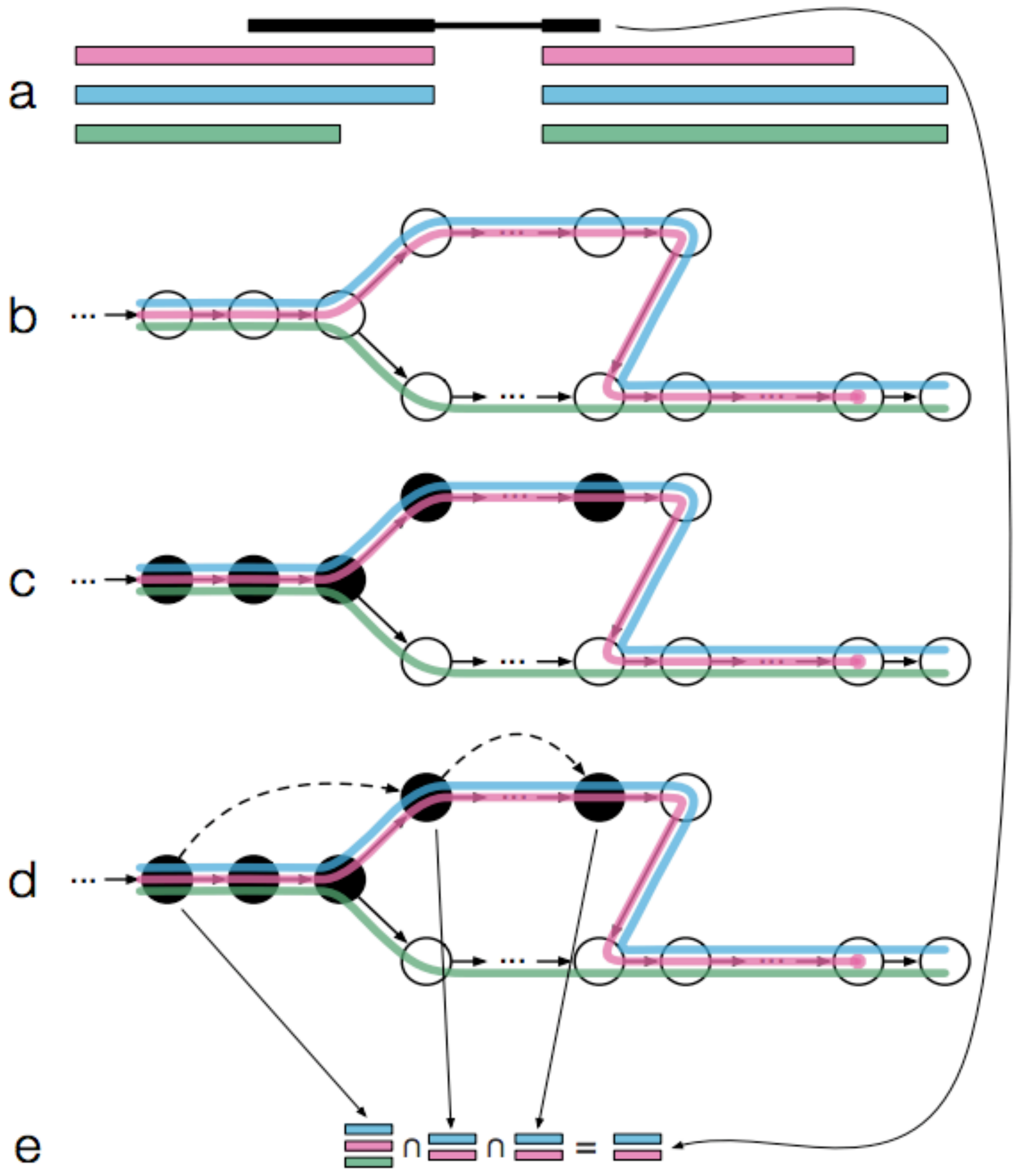
Pseudo-Align: Kallisto

Rationale:

1. Alignment based transcript reconstruction is computationally expensive
2. Read counting method introduces ambiguity

Kallisto Algorithm:

1. Construct a graphical model (de Brujin) for all know transcript isoforms
2. Allocate reads onto the graph model
3. Use Expectation-Maximization (EM) algorithm to estimate the number of transcripts



Bray, et al, Nat. Biotech (2016)

Summary

Alignment-Based

Pros:

- Generate BAM files can be used for extensive QC and visualization

Cons:

- Computationally expensive
- Maybe less accurate for transcript quantification

Pseudo-Alignment

Pros:

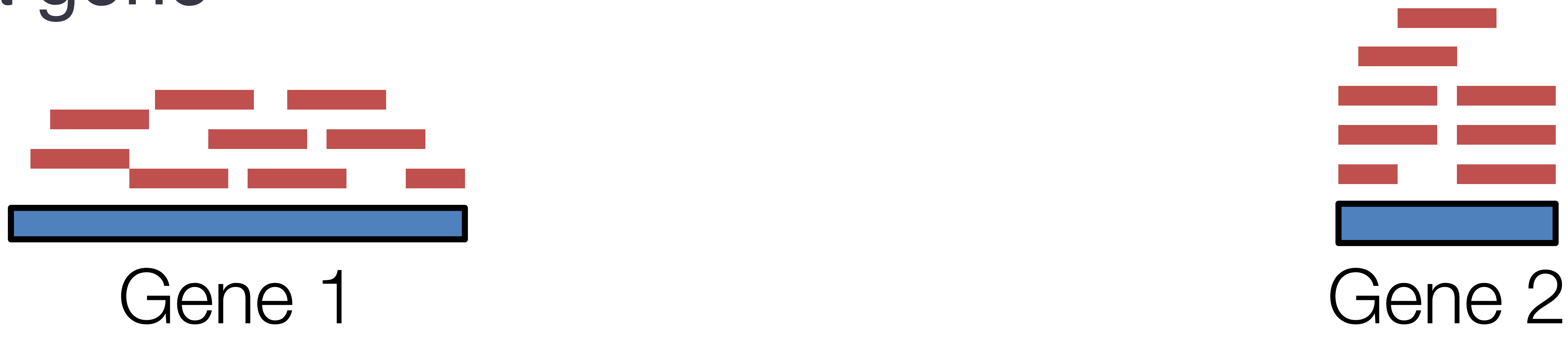
- Very fast and likely more accurate in most cases
- Computationally cheaper

Cons:

- Does not generate BAM file
- Cannot visualize on IGV
- Less robust with low quality sequence data

Gene Quantification and Normalization

Raw counts: number of reads (or fragments) overlapping with the union of exons of a gene



Raw Count != Expression level

Raw Count is strongly influenced by:

- gene length
- transcript sequence (% GC)

may cause variations for different genes expressed at the same level

- sequencing depth
- expression of all other genes

in the same sample

may cause variations for same genes in different sample

Common Normalized Measurements

- **Raw Counts:** number of reads/fragments overlapping all the exons of a gene
- **RPKM/FPKM*:** Reads/fragments per kilobase of gene per million reads mapped
- **TPM:** transcripts per million reads mapped=[gene / read count per bp/all gene count per all gene bp]
- **rlog**: log2-transformed count data normalized for small counts and library size. There are many variations:
  - Trimmed mean of M values (TMM)
  - DESeq2
  - Upper-Quartile (UQ)

$$X_i$$

$$RPKM_i = \frac{X_i}{(\frac{l_i}{10^3})(\frac{N}{10^6})}$$

$$TPM_i = \frac{X_i}{l_i} * \frac{10^6}{\sum \frac{X_j}{l_k}}$$

* Zhao et al, RNA, (2020)

Exploratory Analysis

Discover sample groups from global gene expression pattern without prior knowledge

How similar are the samples?

	G1
S1	3
S2	4
Distance	1

How to quantitatively measure how similar are two samples?

How similar are the samples?

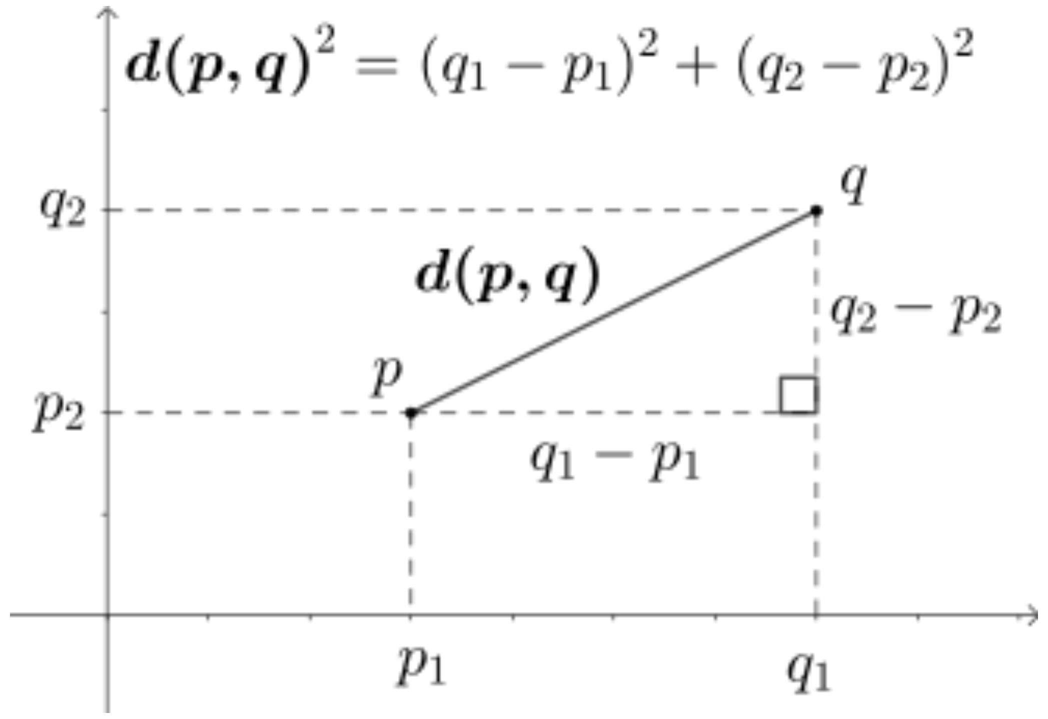
	G1	G2	G3	G4	G5	G6	…	Gi
S1	3	3	9	13	4	5	…	…
S2	4	6	6	6	11	11	…	…
Distance	1	3	3	7	7	6	…	…

How to quantitatively measure how similar are two samples?

Distance between samples

Euclidean distance:

$$d(q,p) = \sqrt{\sum_{n=0}^{i} (q_i - p_i)^2}$$



Pythagorean theorem

Pearson’s distance:

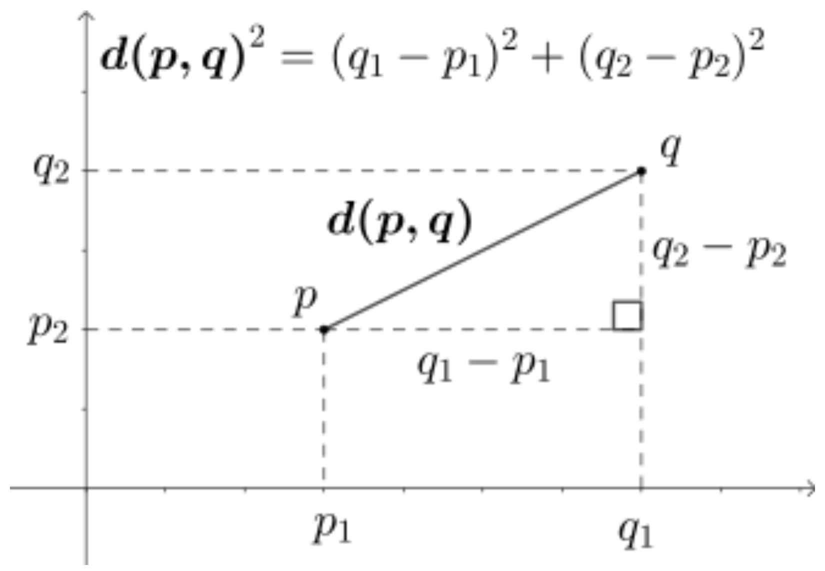
$$d(q,p) = 1 - \rho_{q,p}$$

Where $\rho_{q,p}$ is Pearson correlation coefficient between q, p

Distance between samples

	G1	G2	G3	G4	G5	G6	…	Gi
S1	3	3	9	13	4	5	…	…
S2	4	6	6	6	11	11	…	…
Distance	1	3	3	7	7	6	…	…

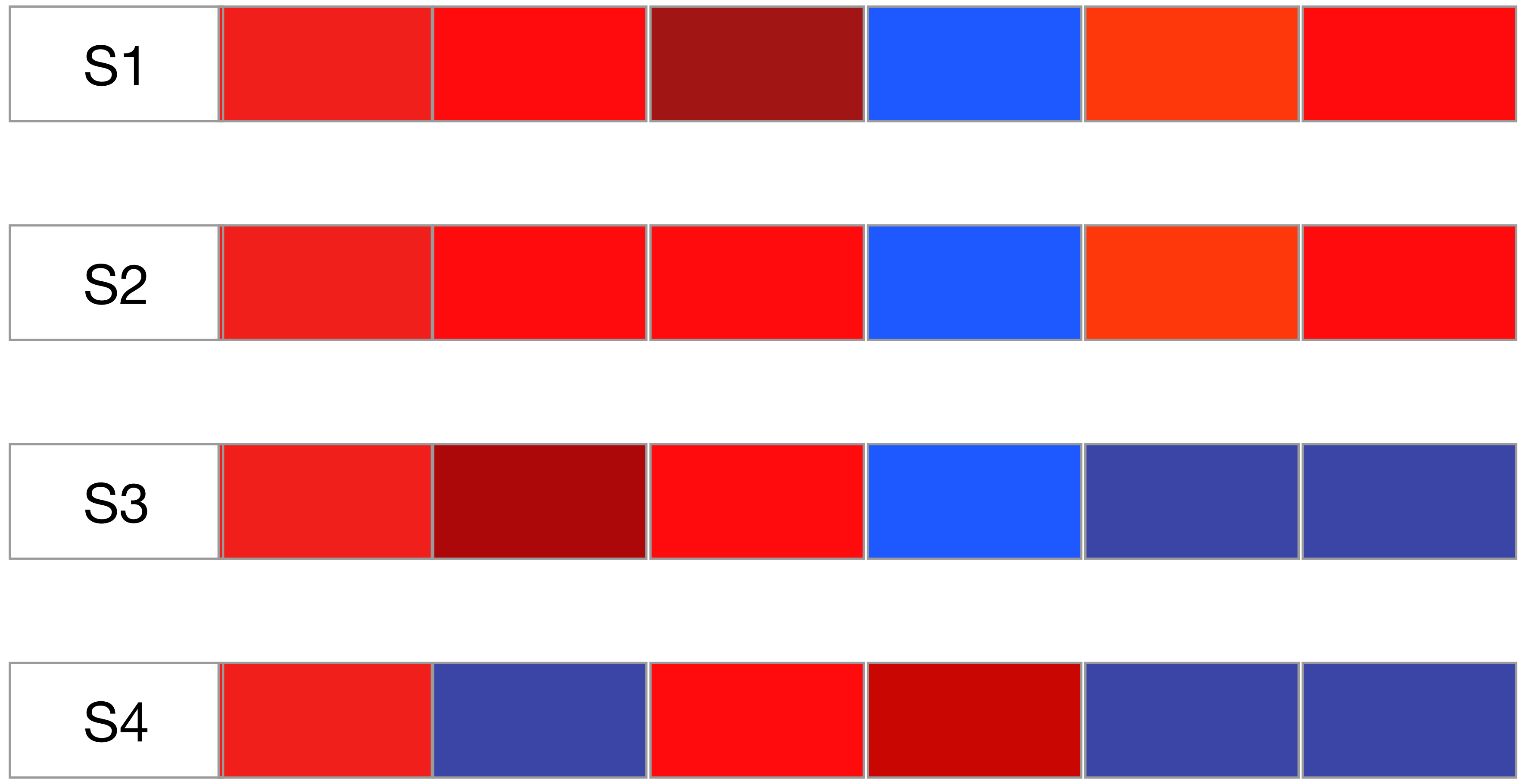
Euclidean distance:



$$d(S1,S2) = \sqrt{\sum_{n=0}^{i} (S1_i - S2_i)^2} = \sqrt{1^2 + 3^2 + 3^2 + 7^2 + 7^2 + 6^2} = 76.5$$

Distance between samples

Gene Expression



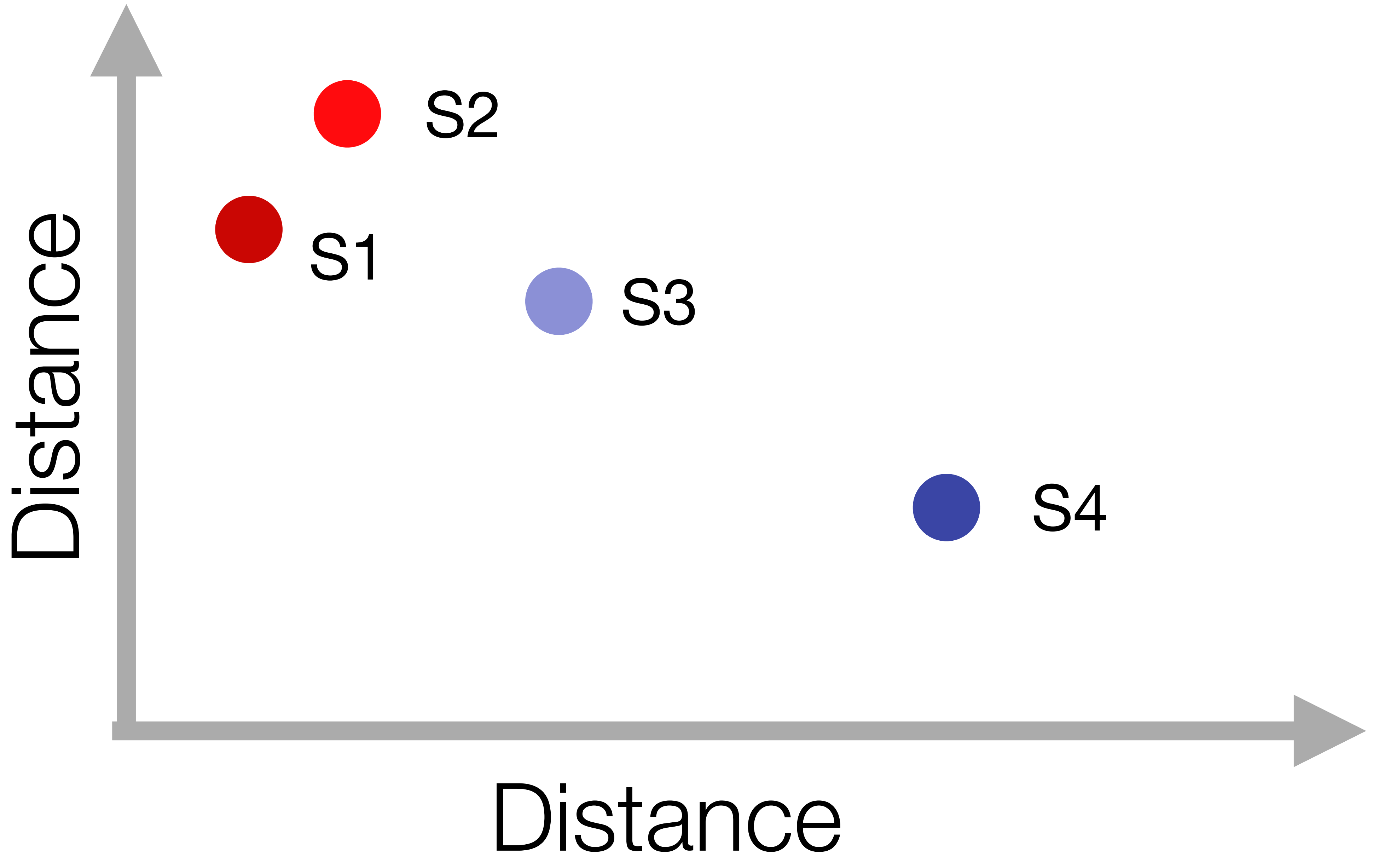
Compute pairwise distances

	S1	S2	S3	S4
S1	0	76	120	220
S2	76	0	96	198
S3	120	96	0	132
S4	220	198	132	0

Similarity Distance Matrix

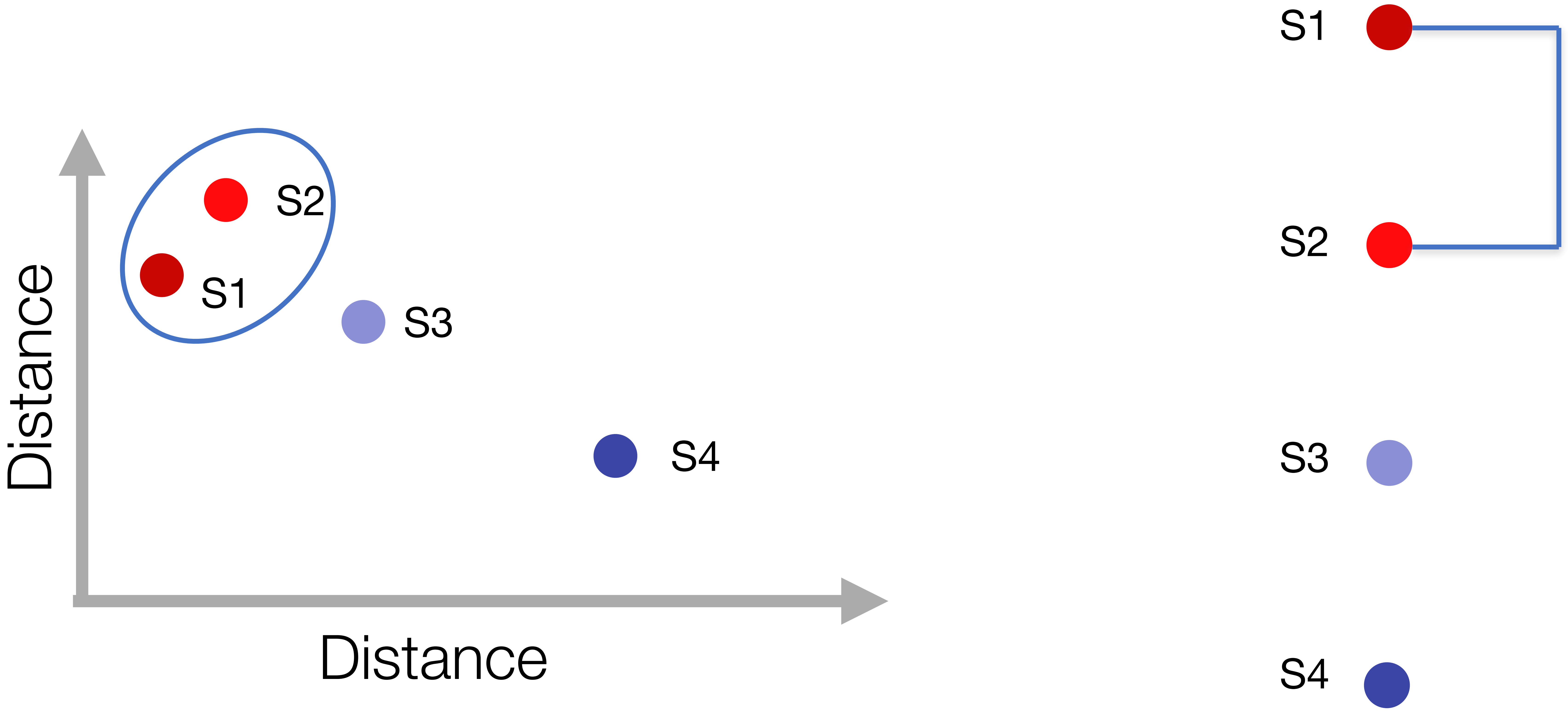
Hierarchical Clustering Tree

Goal: partition the samples into homogeneous groups such that the within group similarities are large



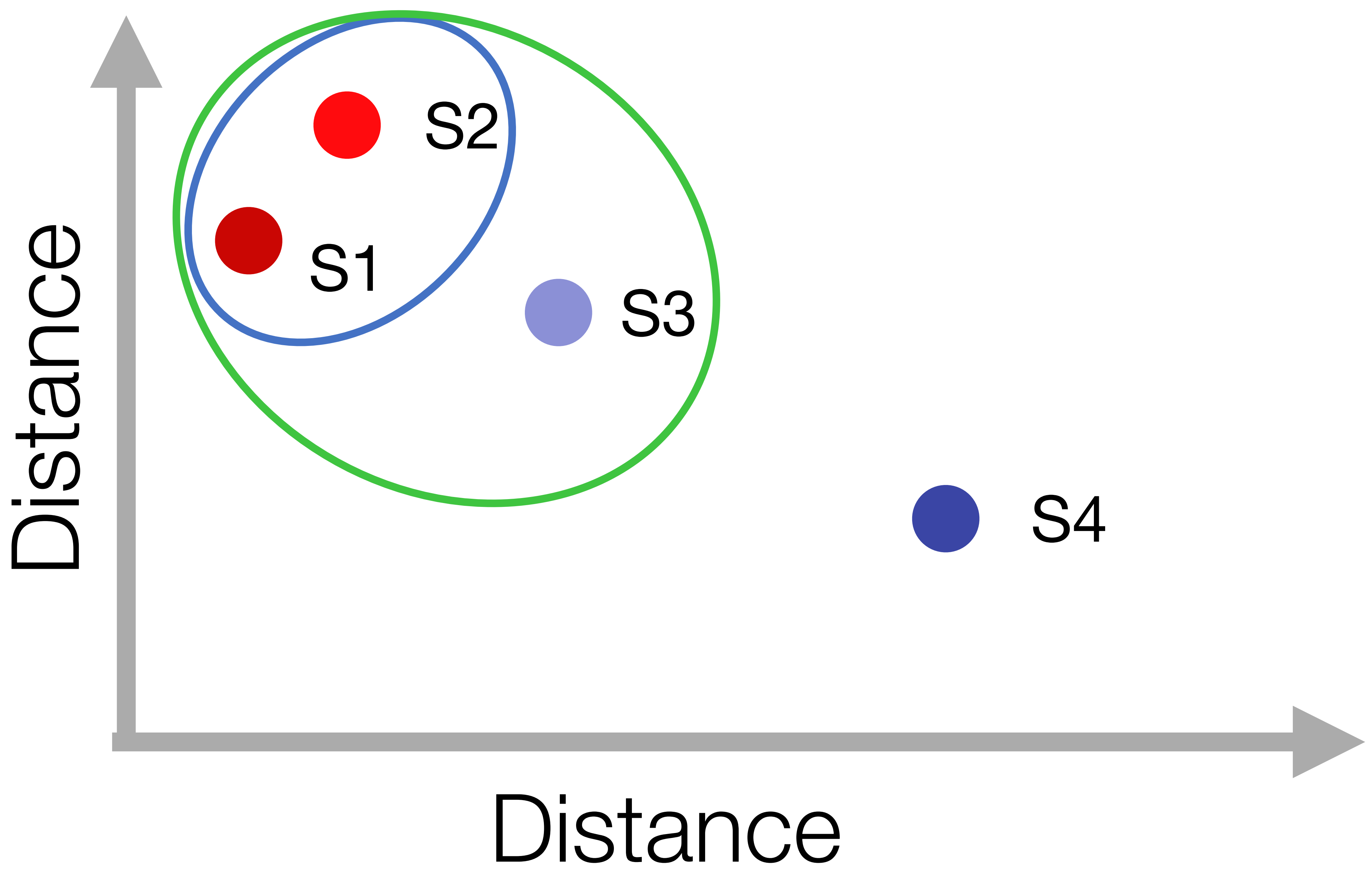
Hierarchical Clustering Tree

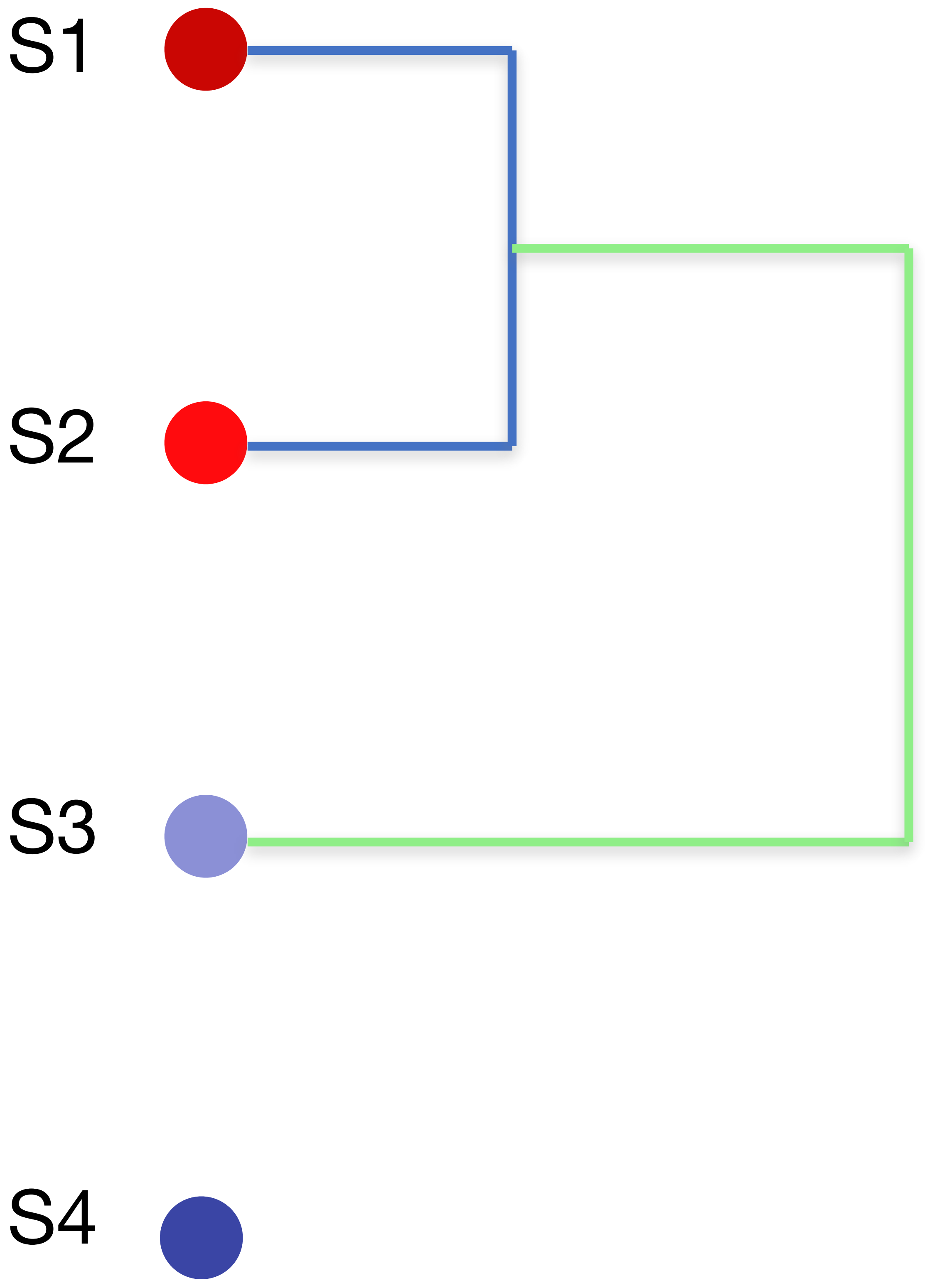
Goal: partition the samples into homogeneous groups such that the within group similarities are large



Hierarchical Clustering Tree

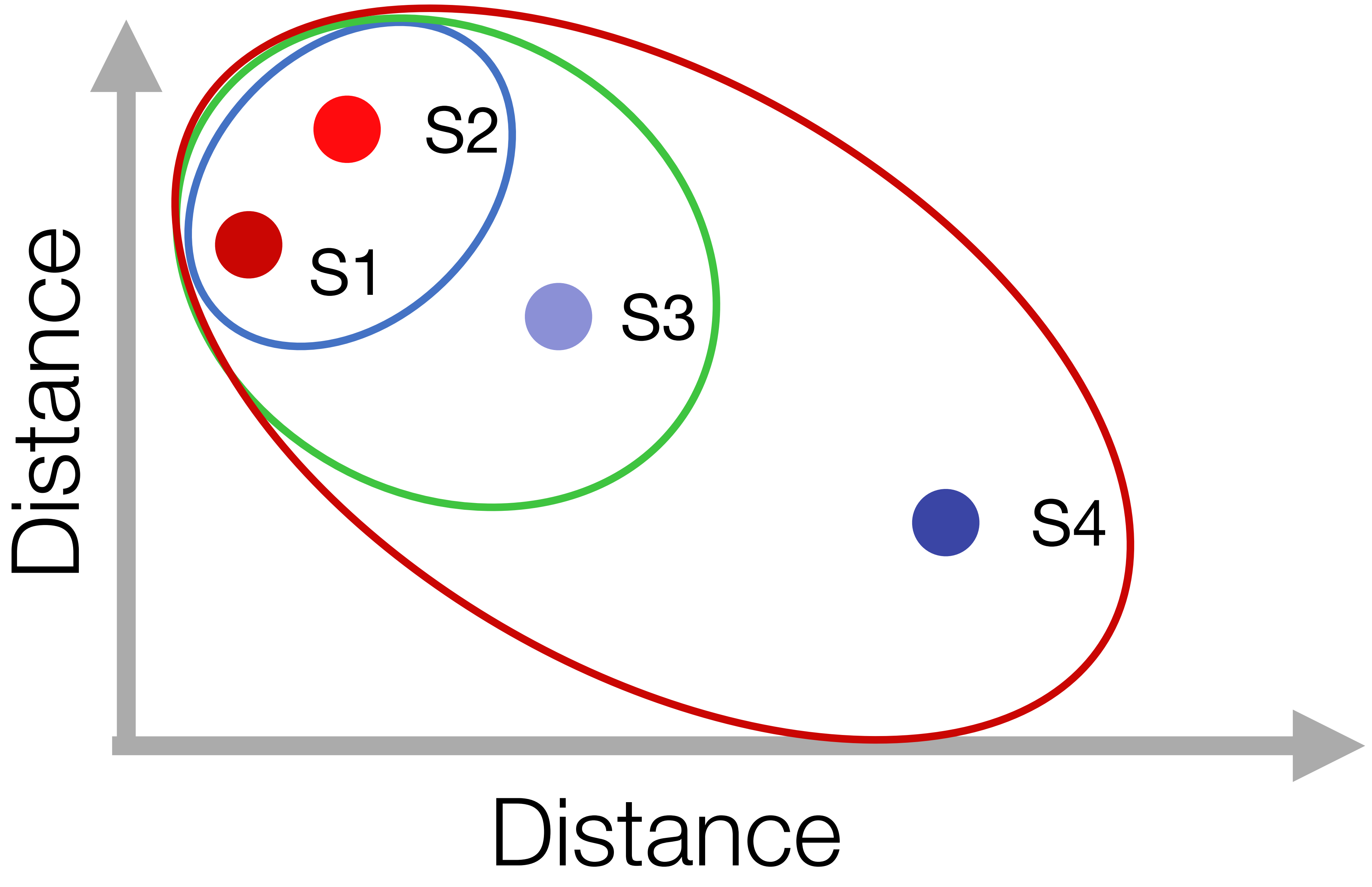
Goal: partition the samples into homogeneous groups such that the within group similarities are large

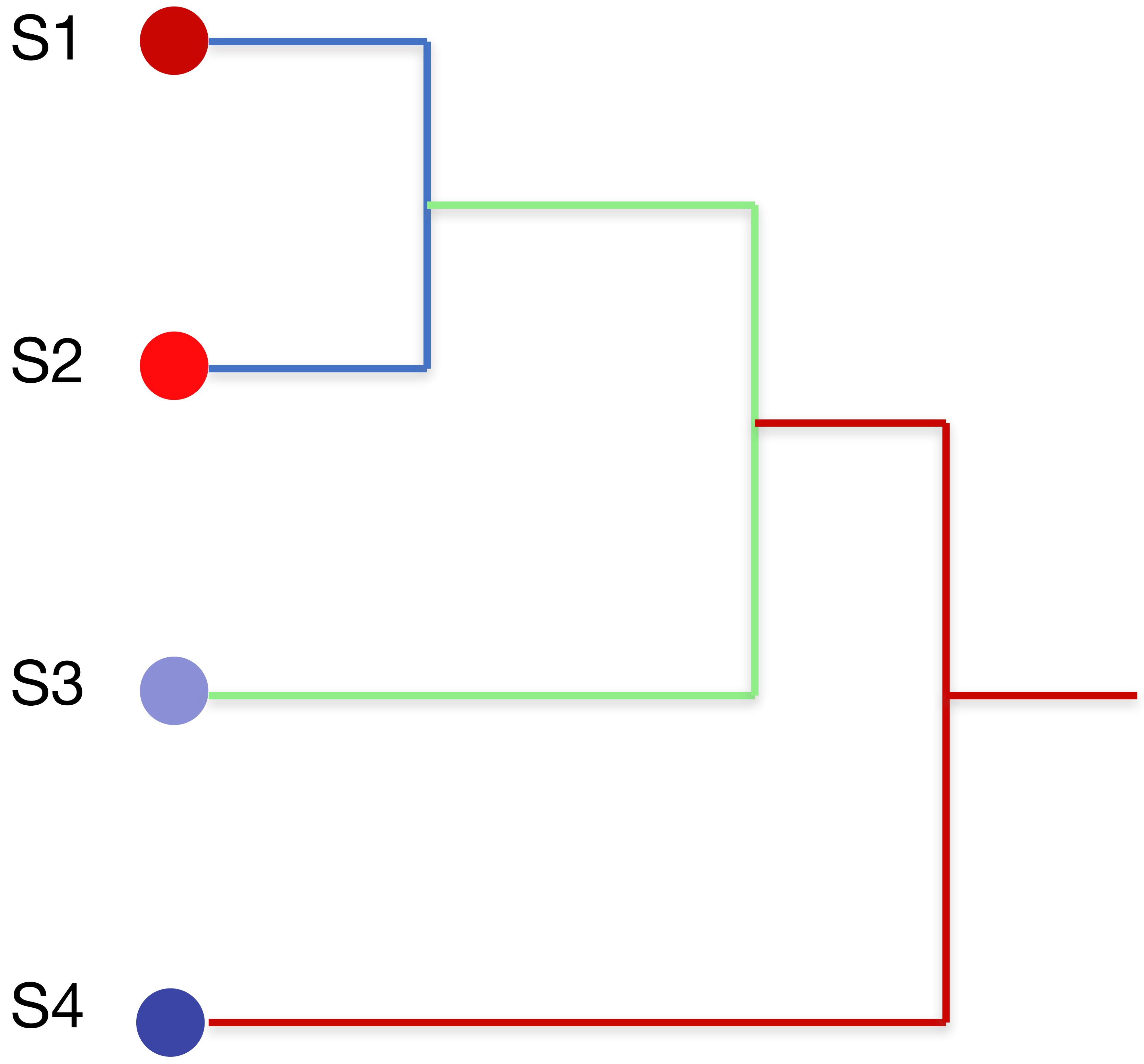




Hierarchical Clustering Tree

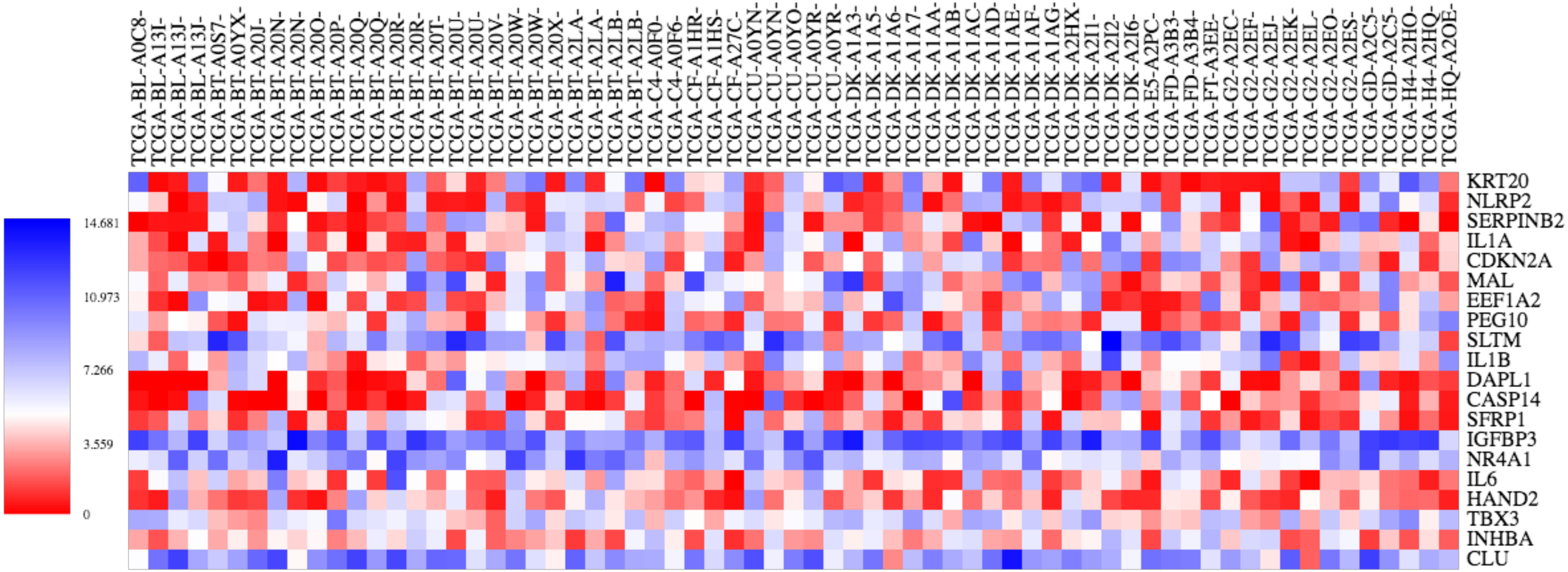
Goal: partition the samples into homogeneous groups such that the within group similarities are large





Feature Reduction Technique

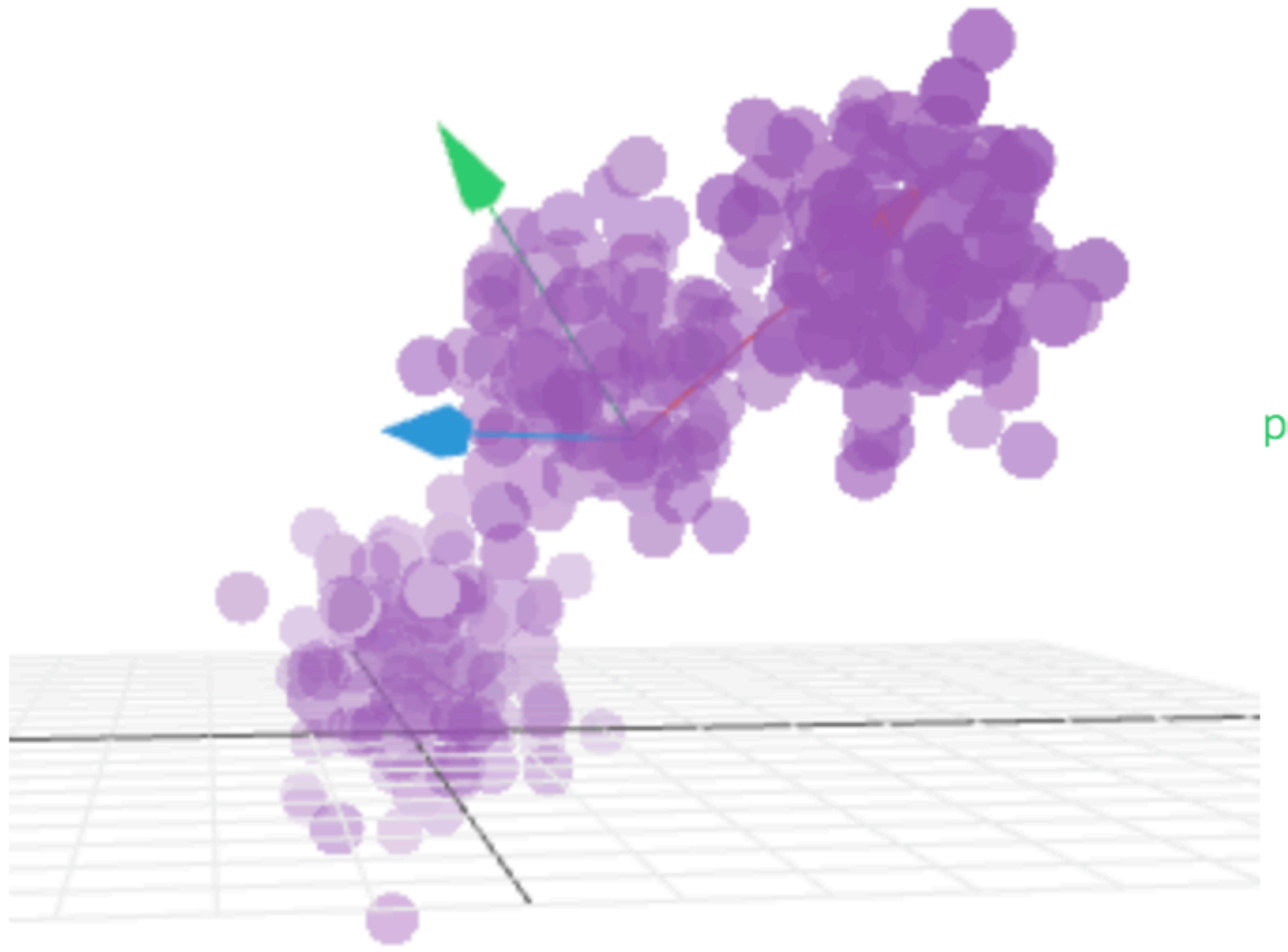
Goal: Reduce the dataset to fewer dimensions yet approx. preserve the distance between the individual samples



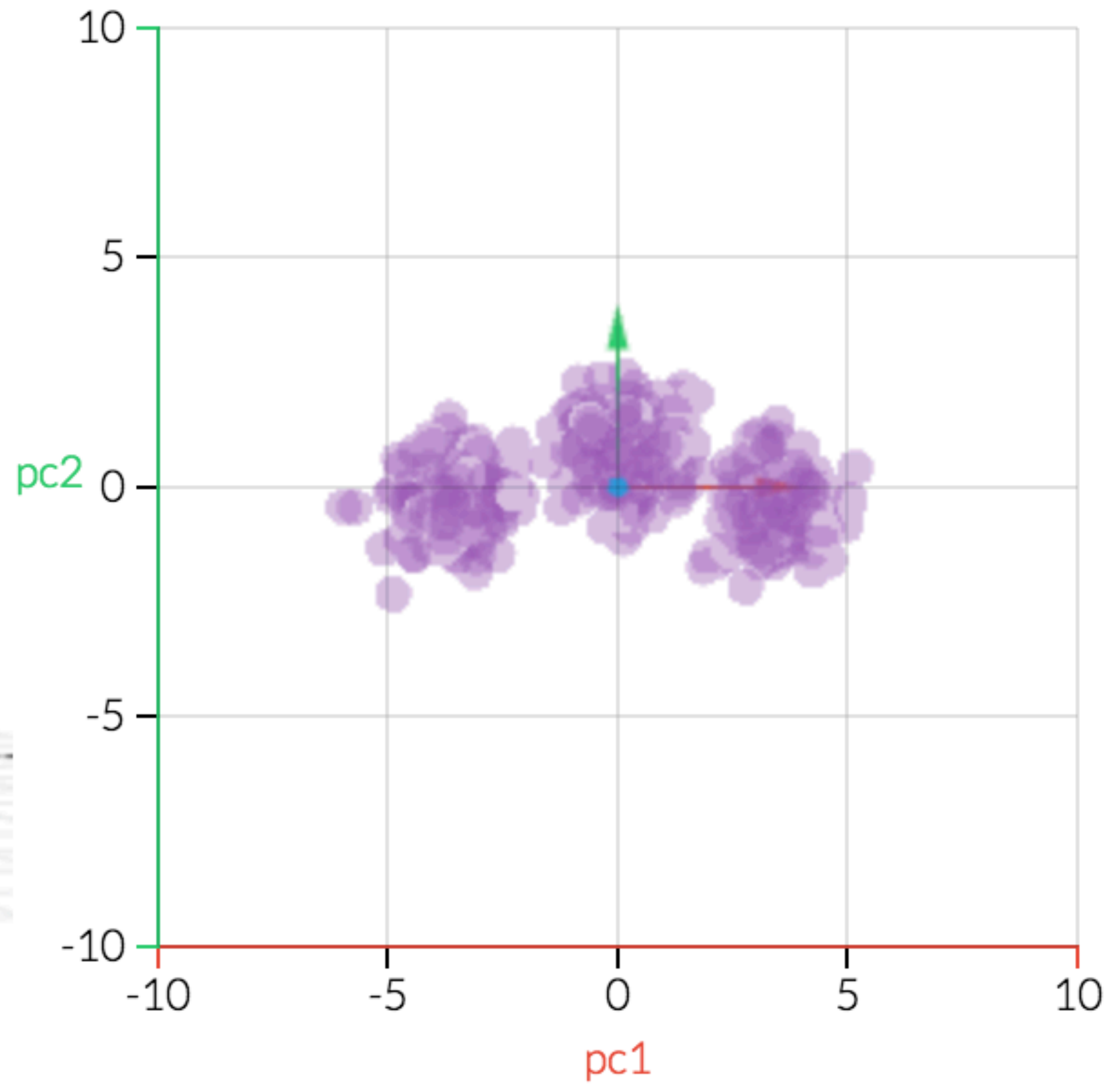
http://mev.tm4.org, TCGA-BRCA

Principle Component Analysis

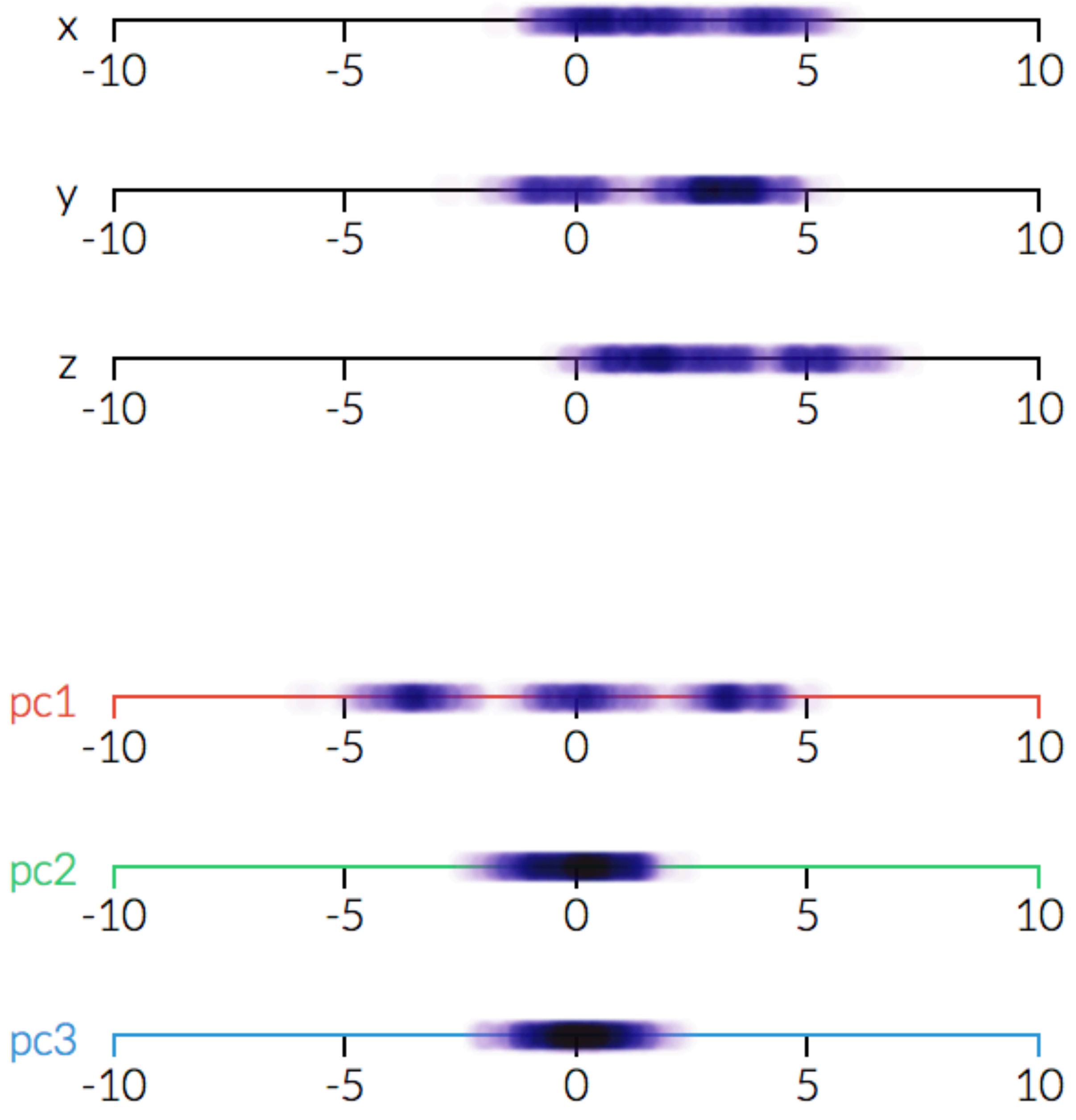
Principle Component Analysis (PCA) convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (Principle Component or PC's)



3-D



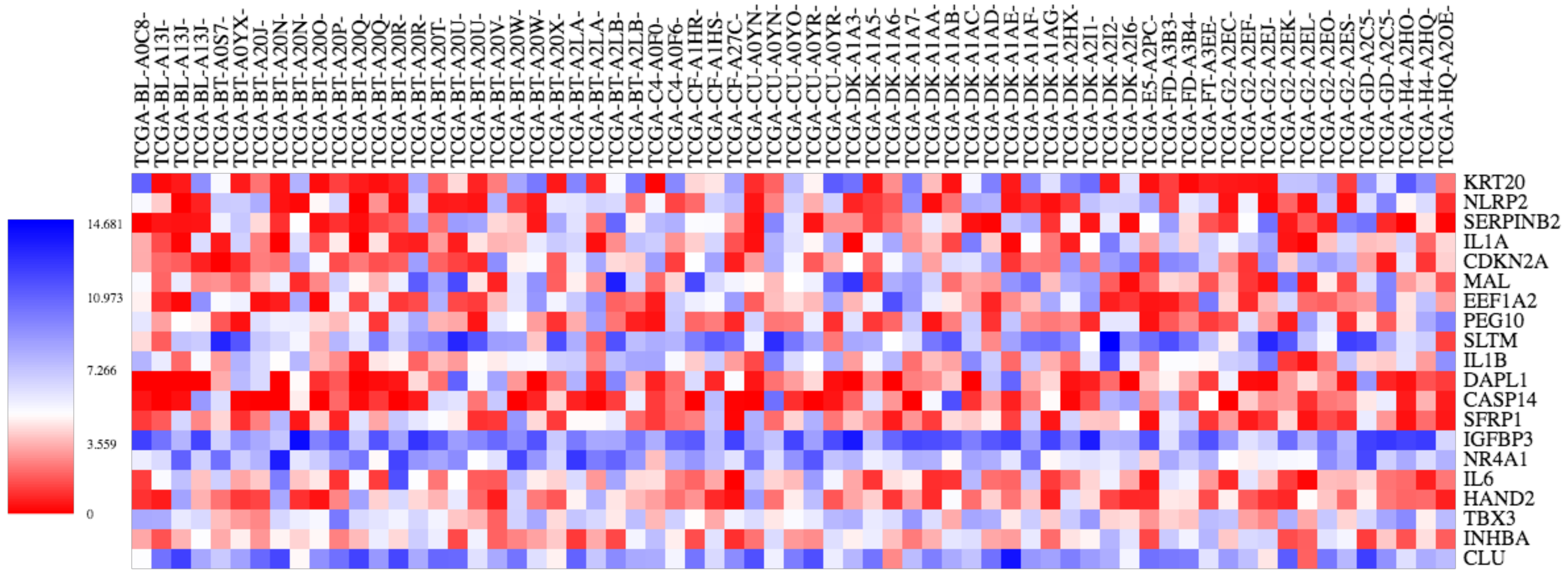
2-D



1-D

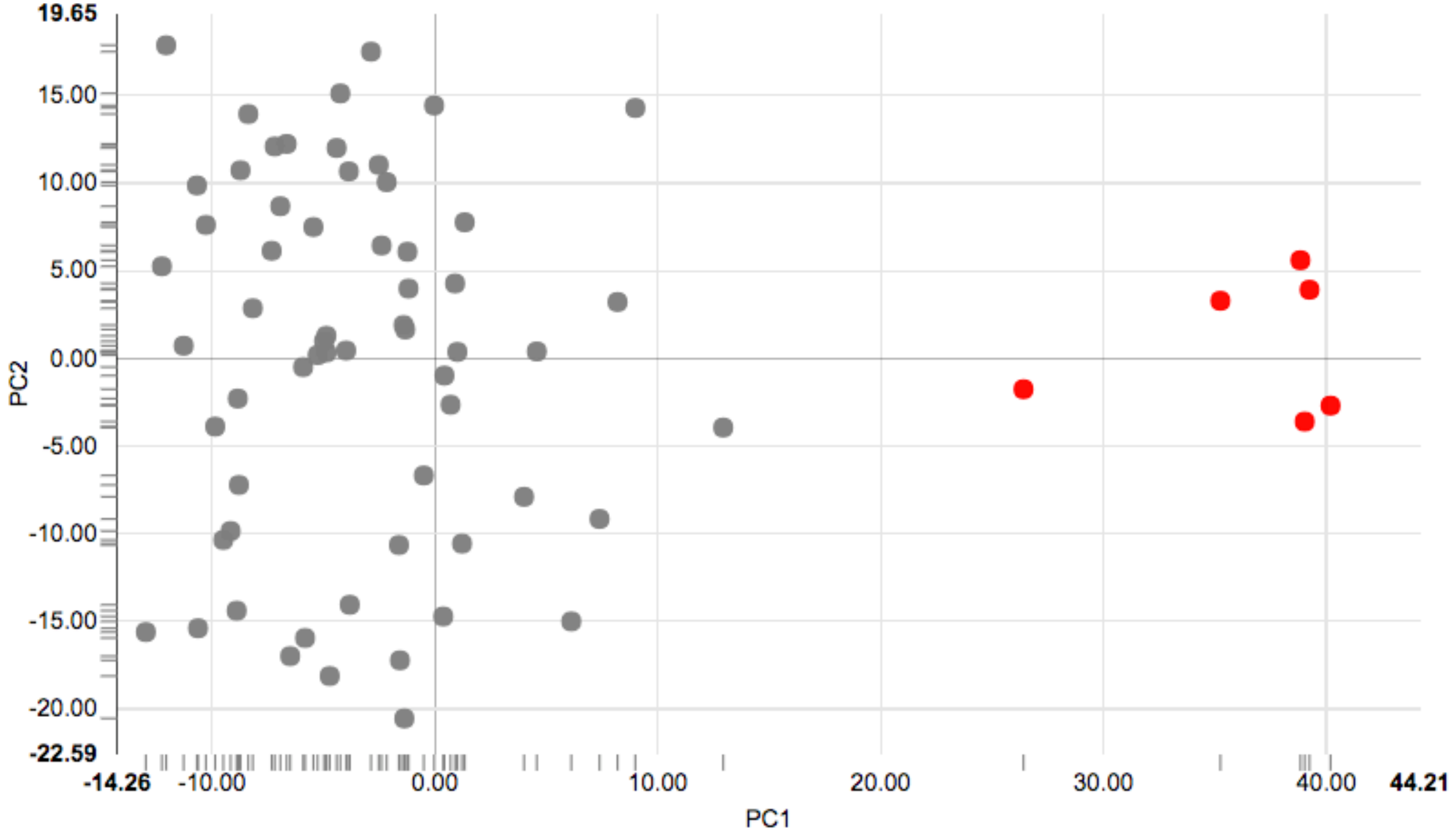
http://setosa.io/ev/principal-component-analysis/

Principle Component Analysis and Visualization



starting point: matrix with expression values per gene and sample, e.g. 22,100 genes x 67 samples

- 22,100 Principle components x 67 Samples
- PC1-3 usually sufficient to capture the major trend



Selected	PC1	PC2
TCGA-BL-A13J-11A-13R-A10U-07	39.2507	3.9165
TCGA-BT-A20N-11A-11R-A14Y-07	40.1933	-2.6946
TCGA-BT-A20Q-11A-11R-A14Y-07	38.8414	5.5994
TCGA-BT-A20R-11A-11R-A16R-07	39.0328	-3.6043
TCGA-CU-A0YN-11A-11R-A10U-07	35.2515	3.2868
TCGA-CU-A0YR-11A-13R-A10U-07	26.4164	-1.7572

Recap

- RNASeq experiment results in short reads (75-150bp) data
- RNASeq data can be quantified by either alignment-based or pseudo-alignment based methods
- Gene counts need to be normalized to remove experimental variation
- Selection of normalization method can affect down stream results
- Unsupervised analysis such as PCA and hierarchical clustering are used for first-pass data exploration