RNA-Seq Differential Expression Analysis

Identify patterns that are biologically meaningful

Most common questions asked from RNA-Seq data?

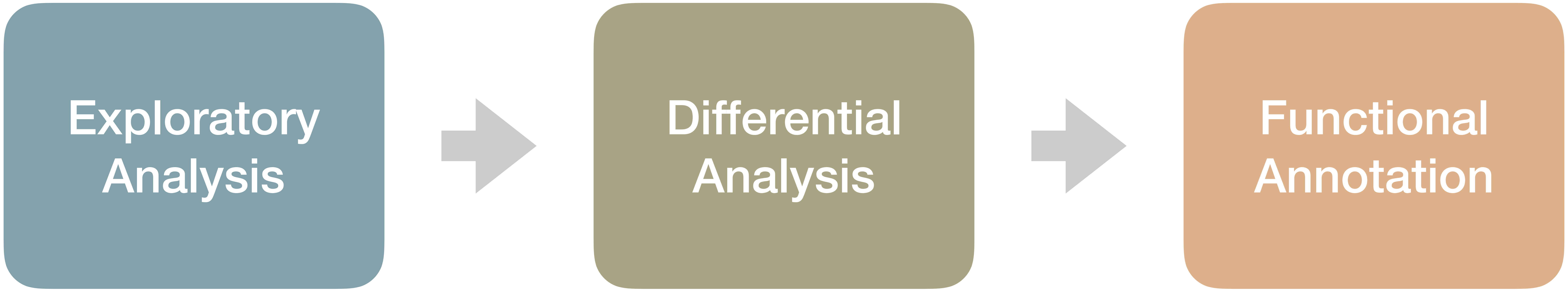
Is there biases affecting the results?

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 Strategy



Exploratory Analysis

Differential Analysis

Functional Annotation

- Unsupervised analysis
- Do not test hypothesis
- Use to discover biases and unexpected variability

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

- Interpret the biology
- Find molecular functions or pathway affected by different conditions

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?

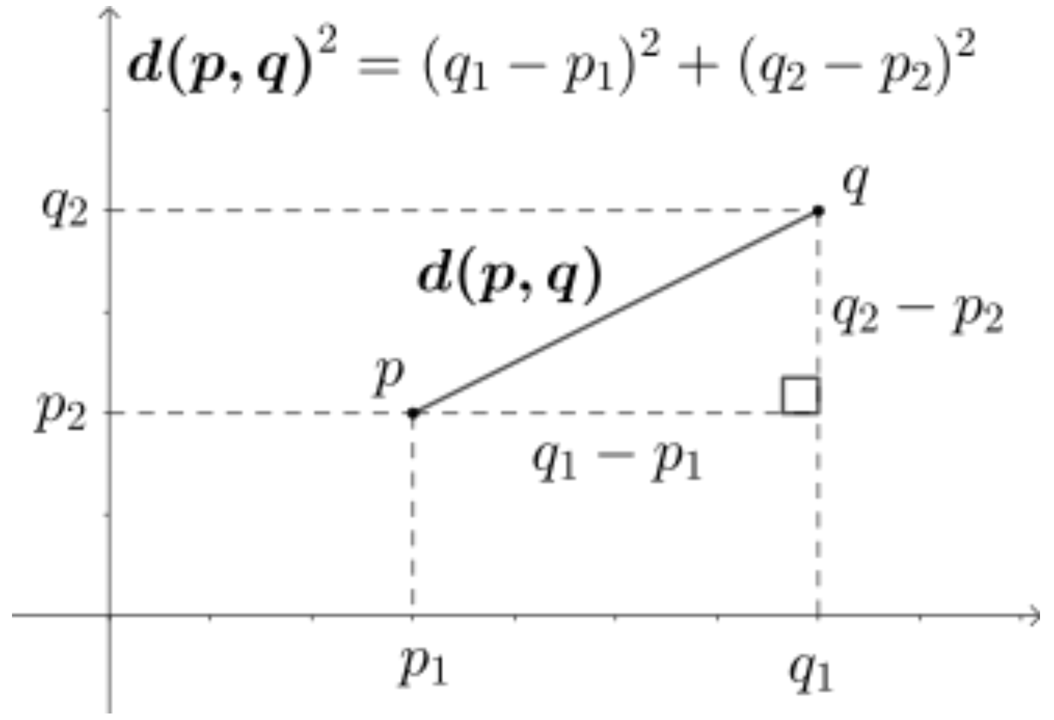
	G1	G2	G3	G4	G5	G6	…	Gi
S1	3	3	9	13	4	5	…	…
S2	4	6	6	6	11	11	…	…
Distance	1	3	3	7	7	6	…	…

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

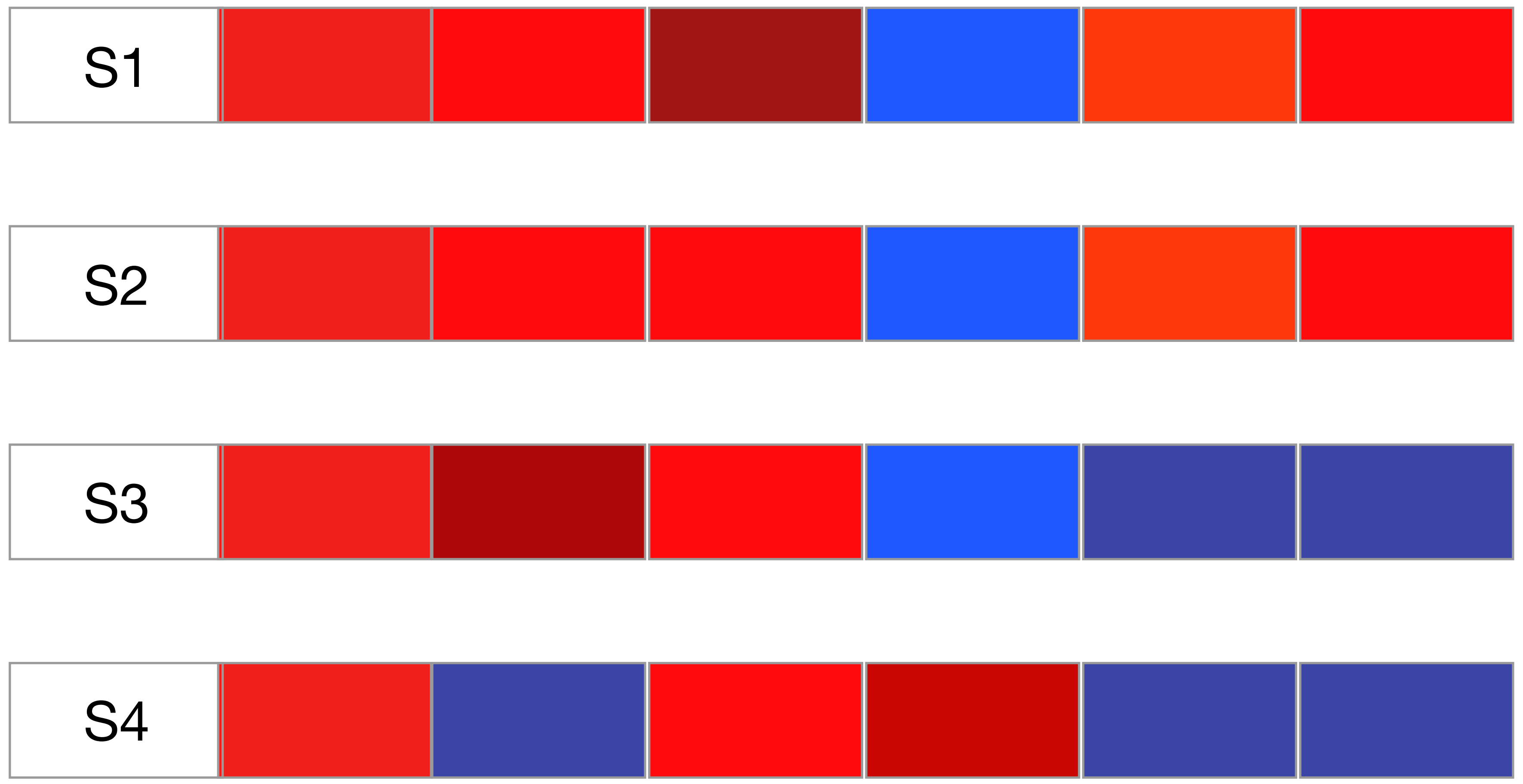
	G1	G2	G3	G4	G5	G6	…	Gi
S1	3	3	9	13	4	5	…	…
S2	4	6	6	6	11	11	…	…
Distance	1	3	3	7	7	6	…	…

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



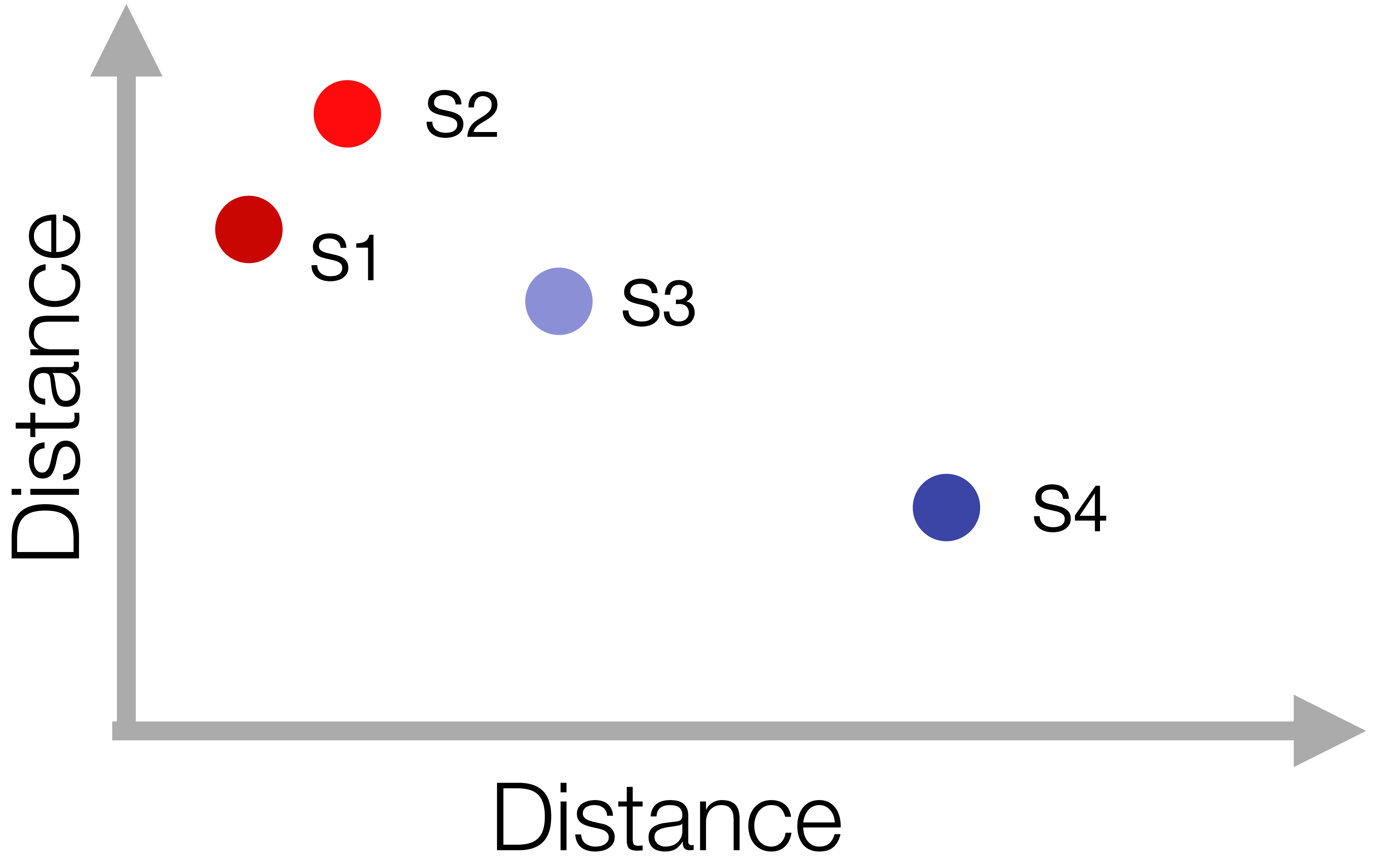
Compute pairwise distances

	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

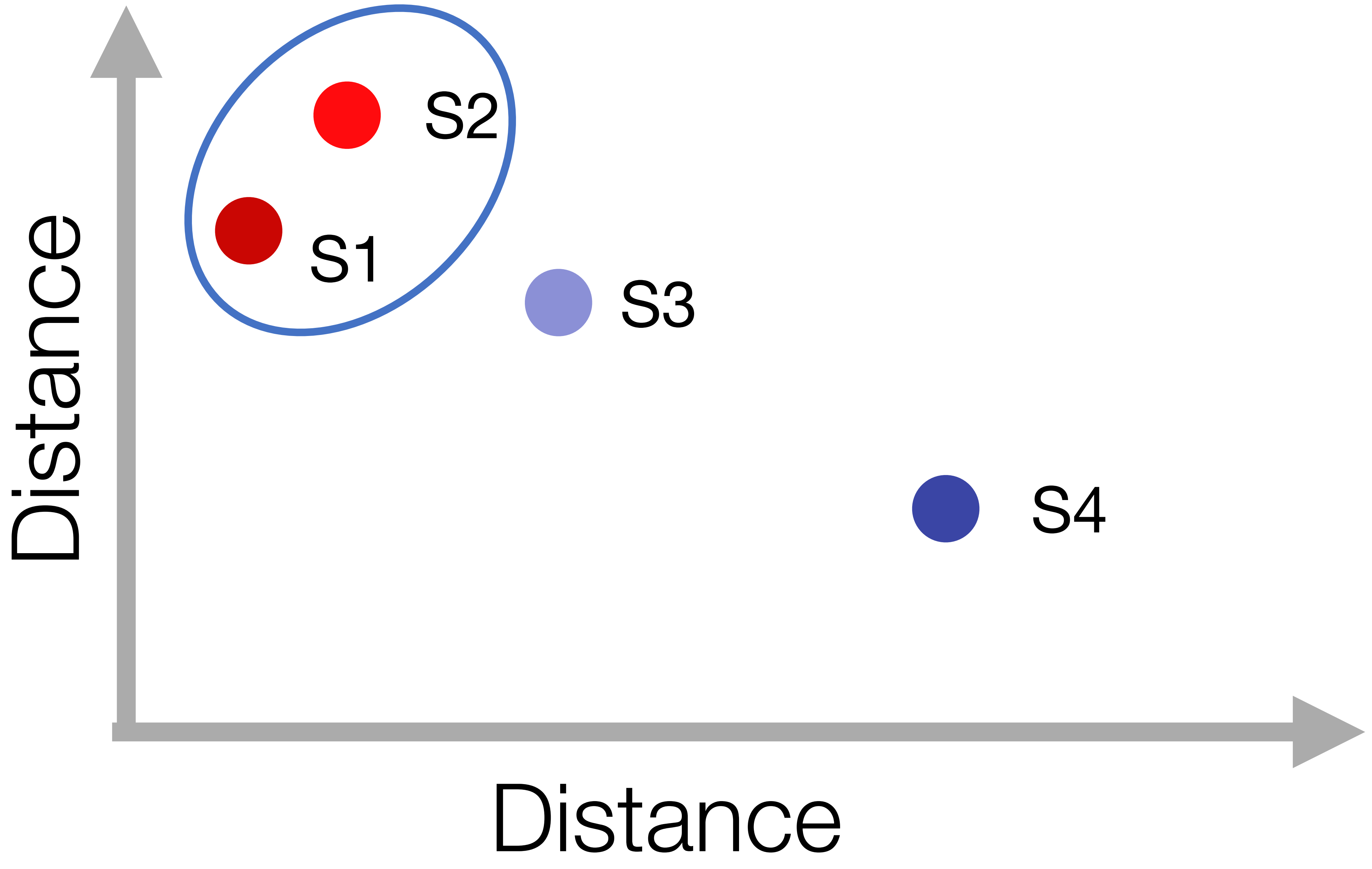
Hierarchical Clustering Tree

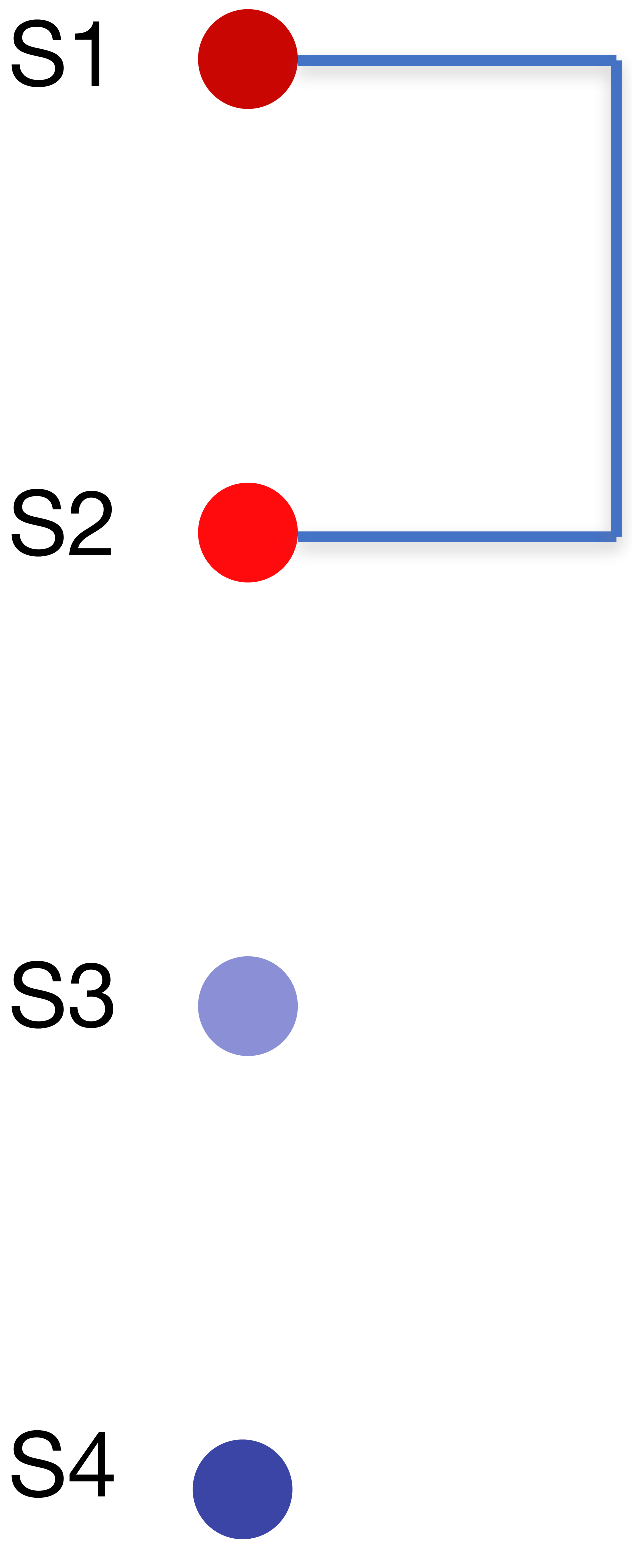
Goal: partition the samples into homogeneous groups such that the within group similarities are large



Hierarchical Clustering Tree

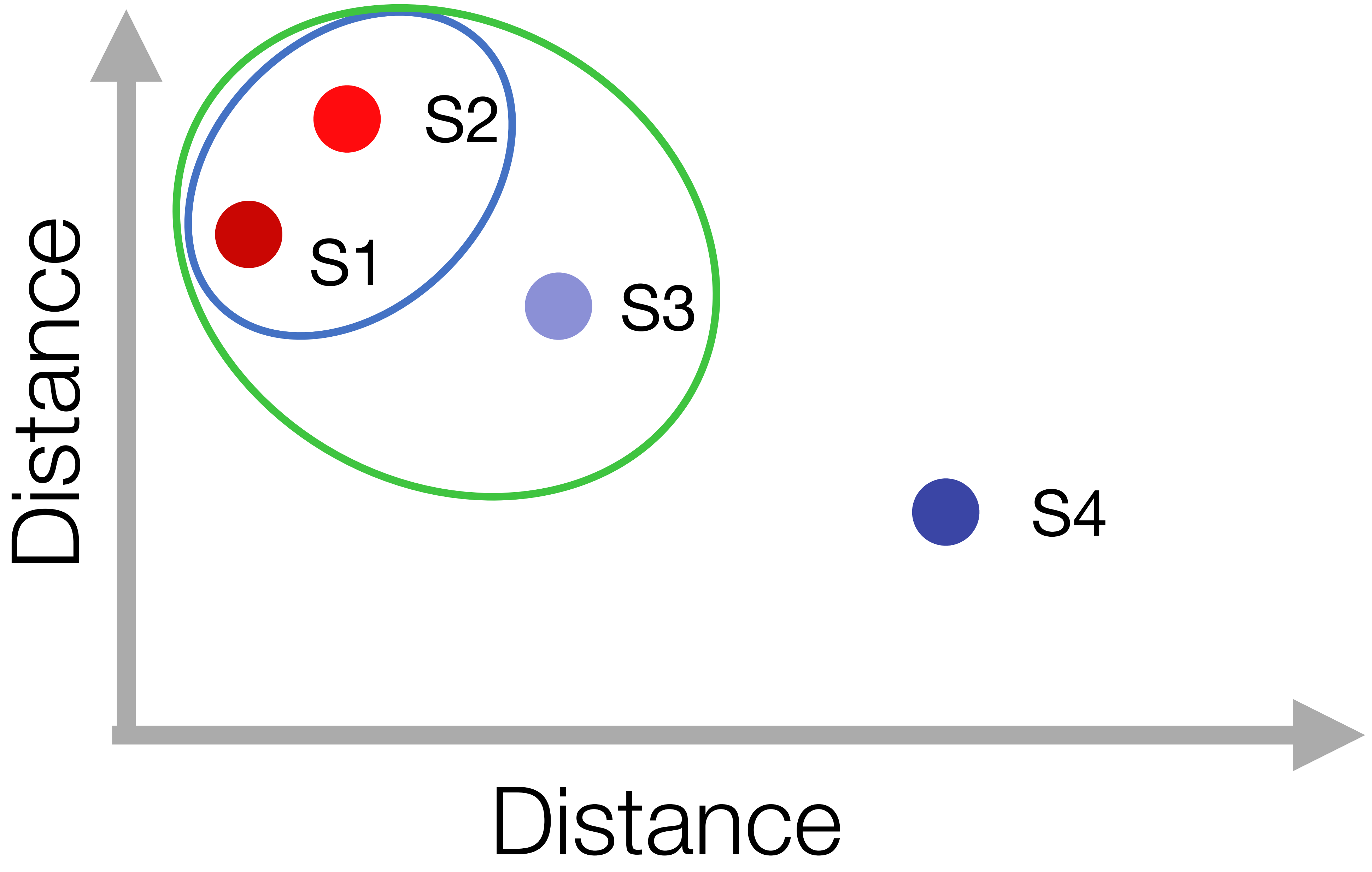
Goal: partition the samples into homogeneous groups such that the within group similarities are large

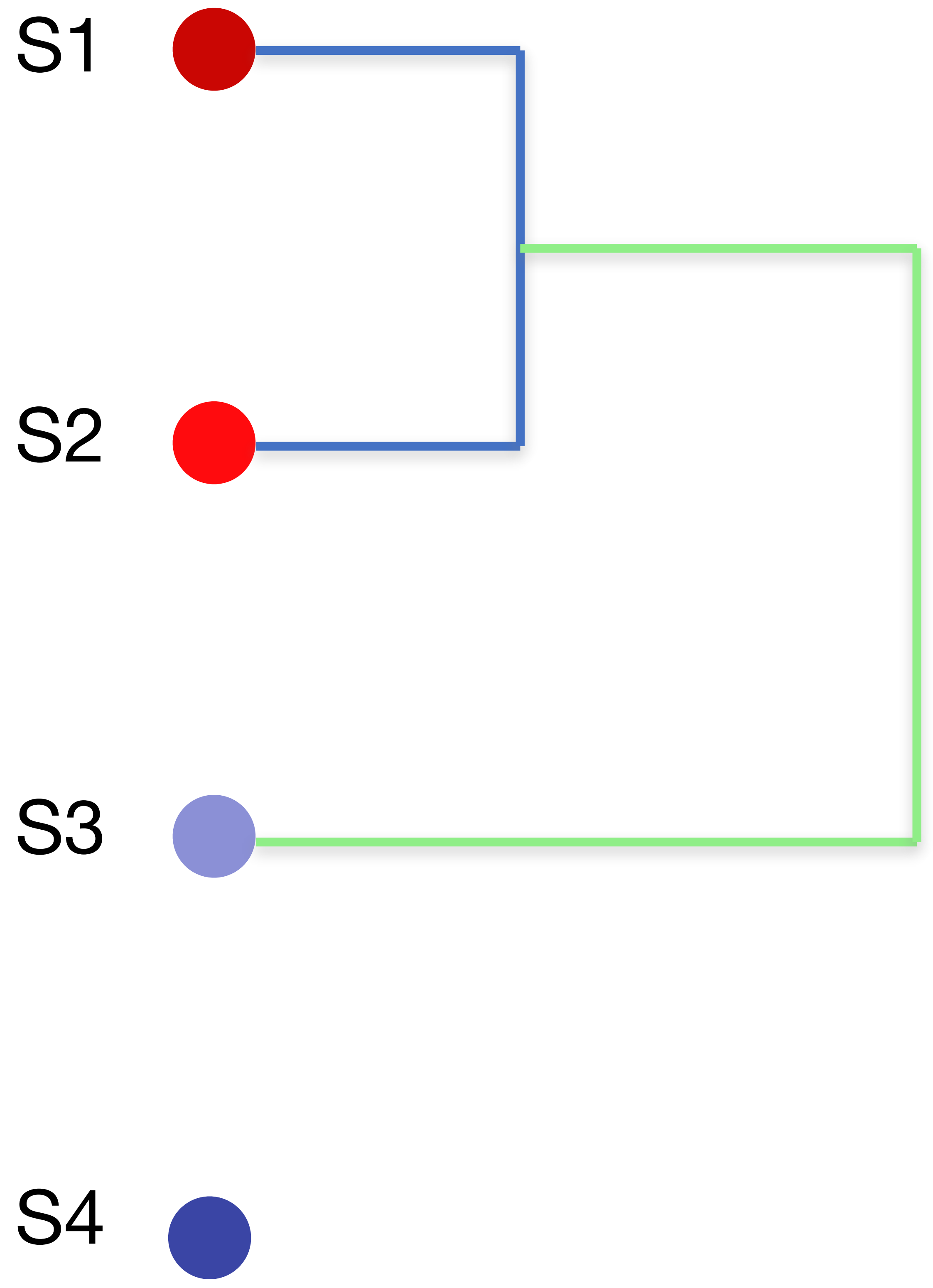




Hierarchical Clustering Tree

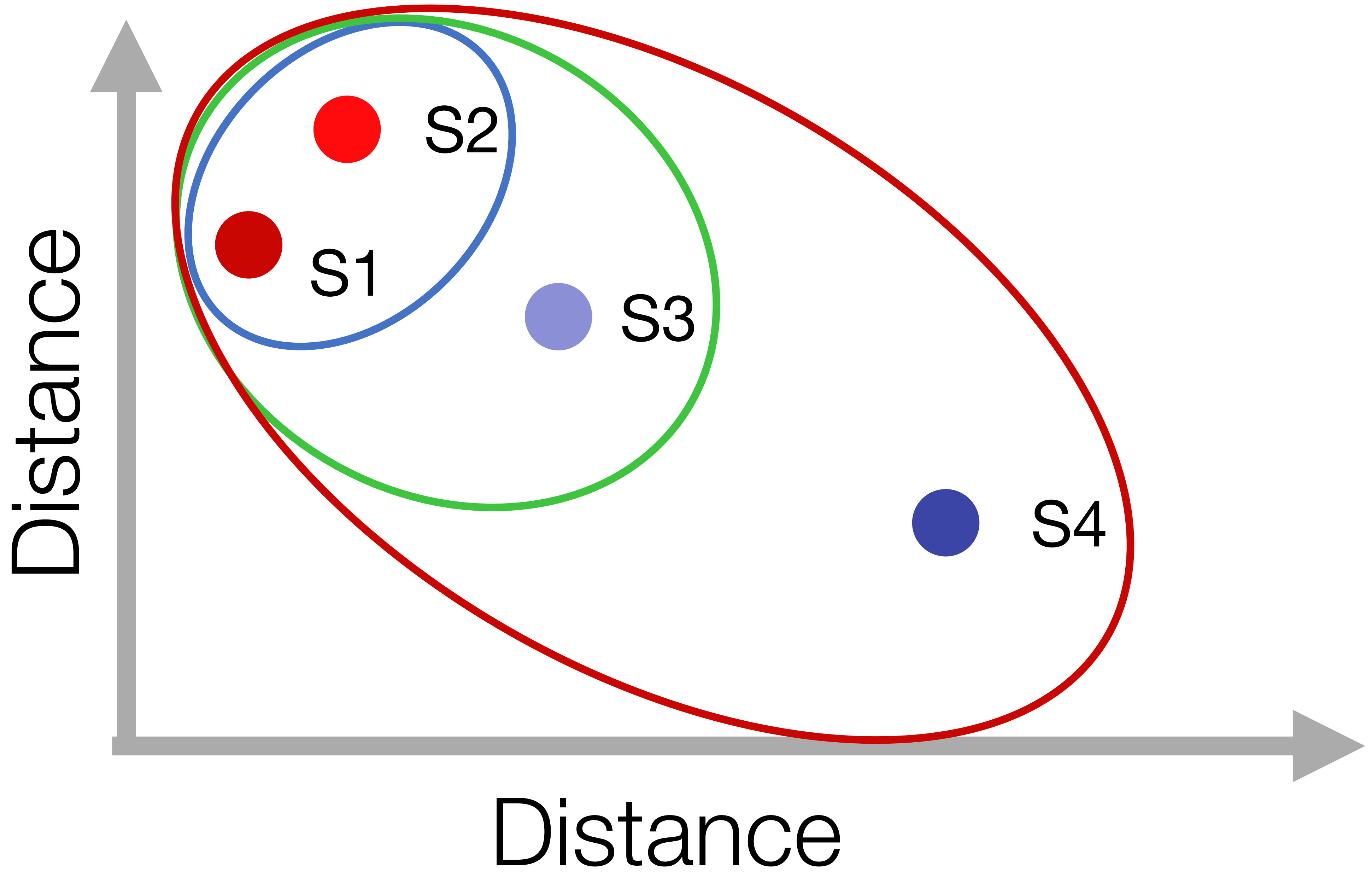
Goal: partition the samples into homogeneous groups such that the within group similarities are large

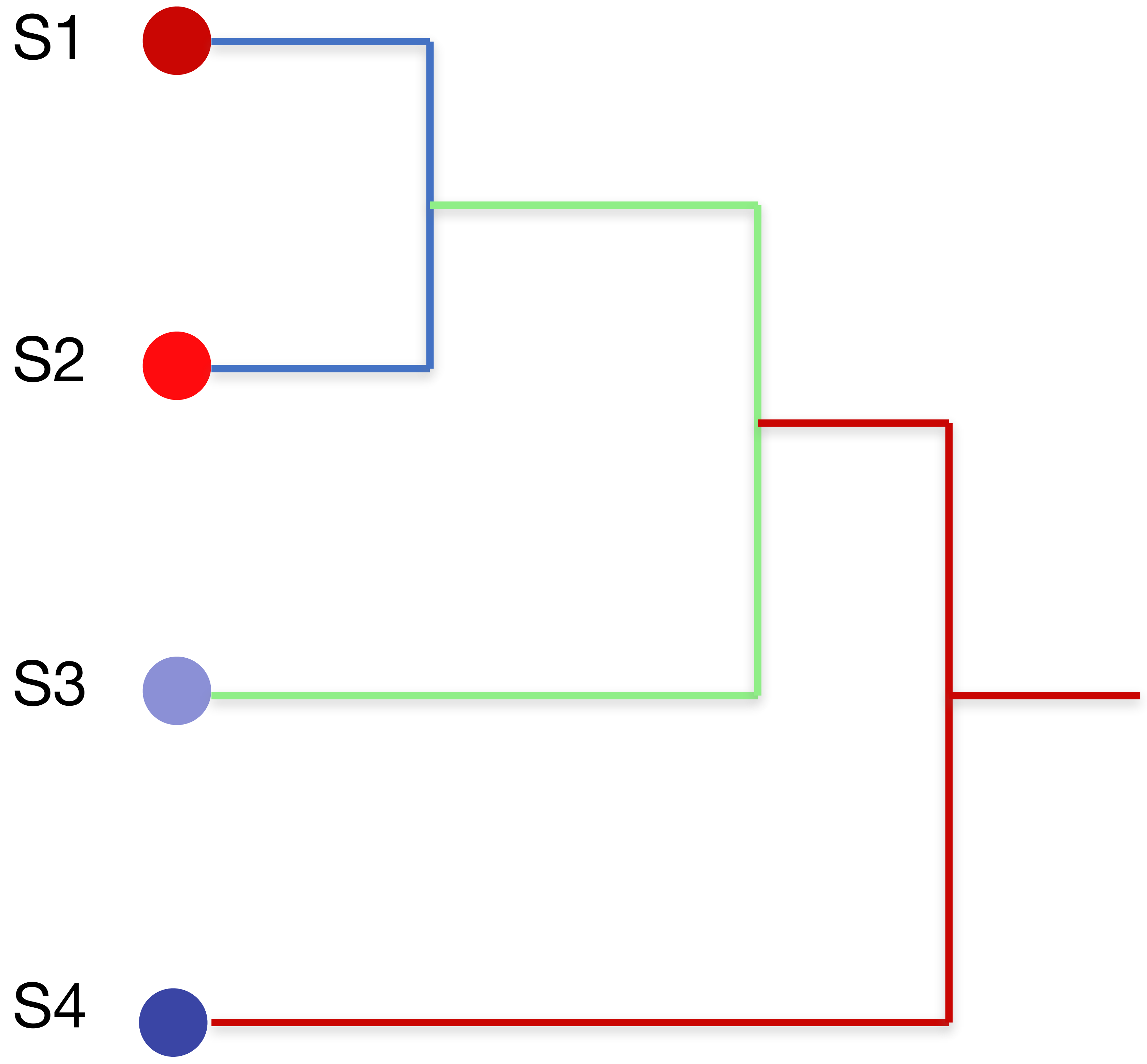




Hierarchical Clustering Tree

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

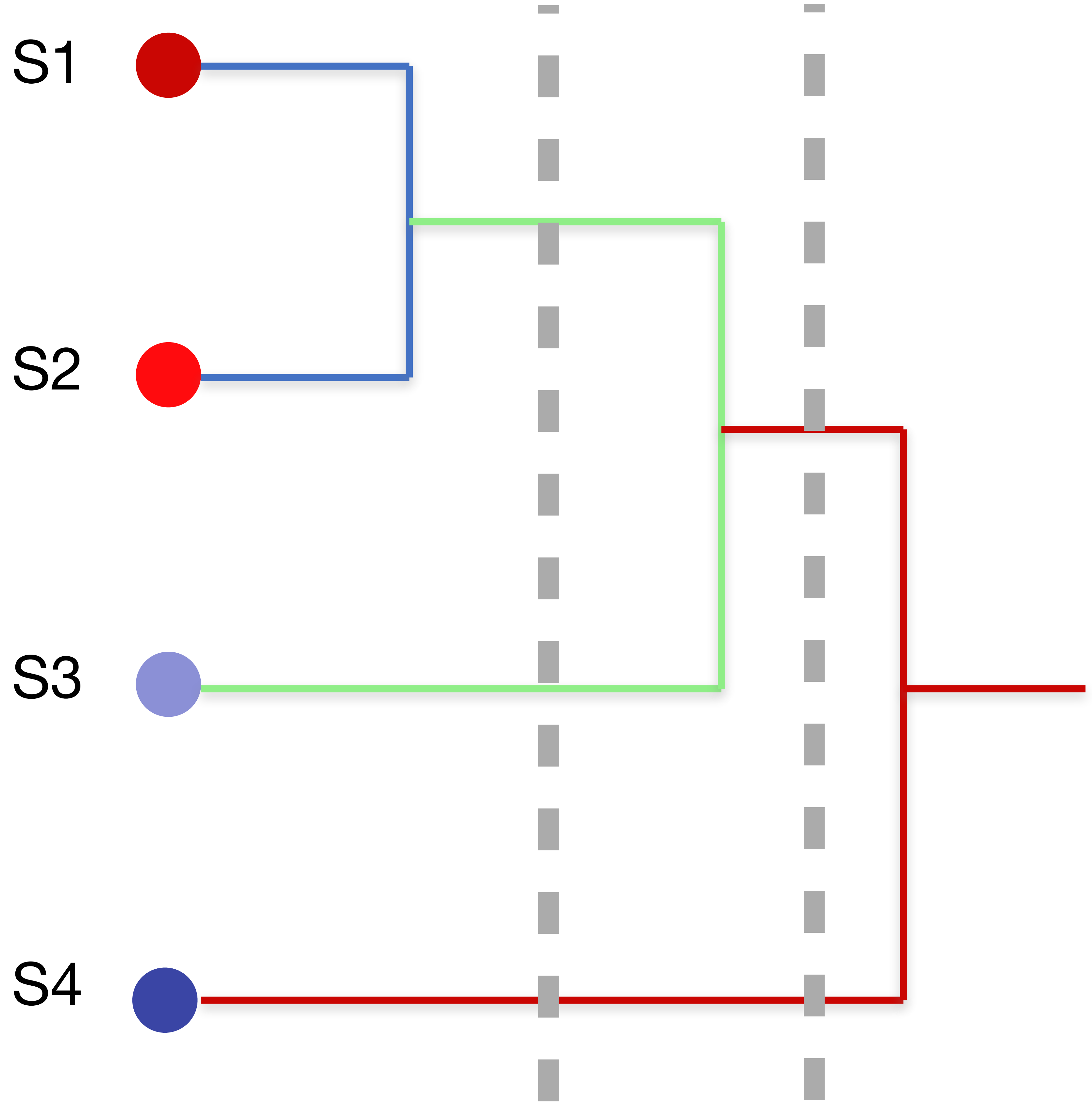




Hierarchical Clustering Tree

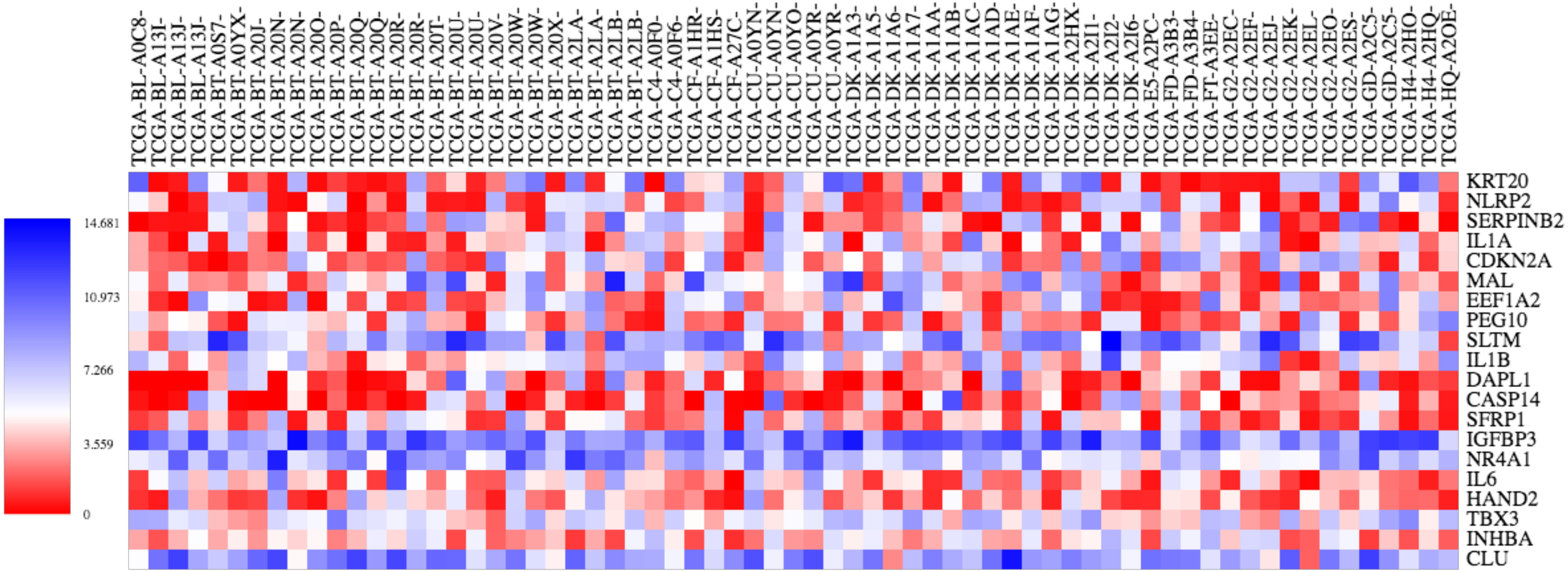
Goal: partition the samples into homogeneous groups such that the within group similarities are large

- Determine pairwise distance between all samples with each sample being its own cluster
- Connect closest pair of cluster until there is only one
- Cutting the dendrogram at a desired level to obtain desired number of clusters



Feature Reduction Technique

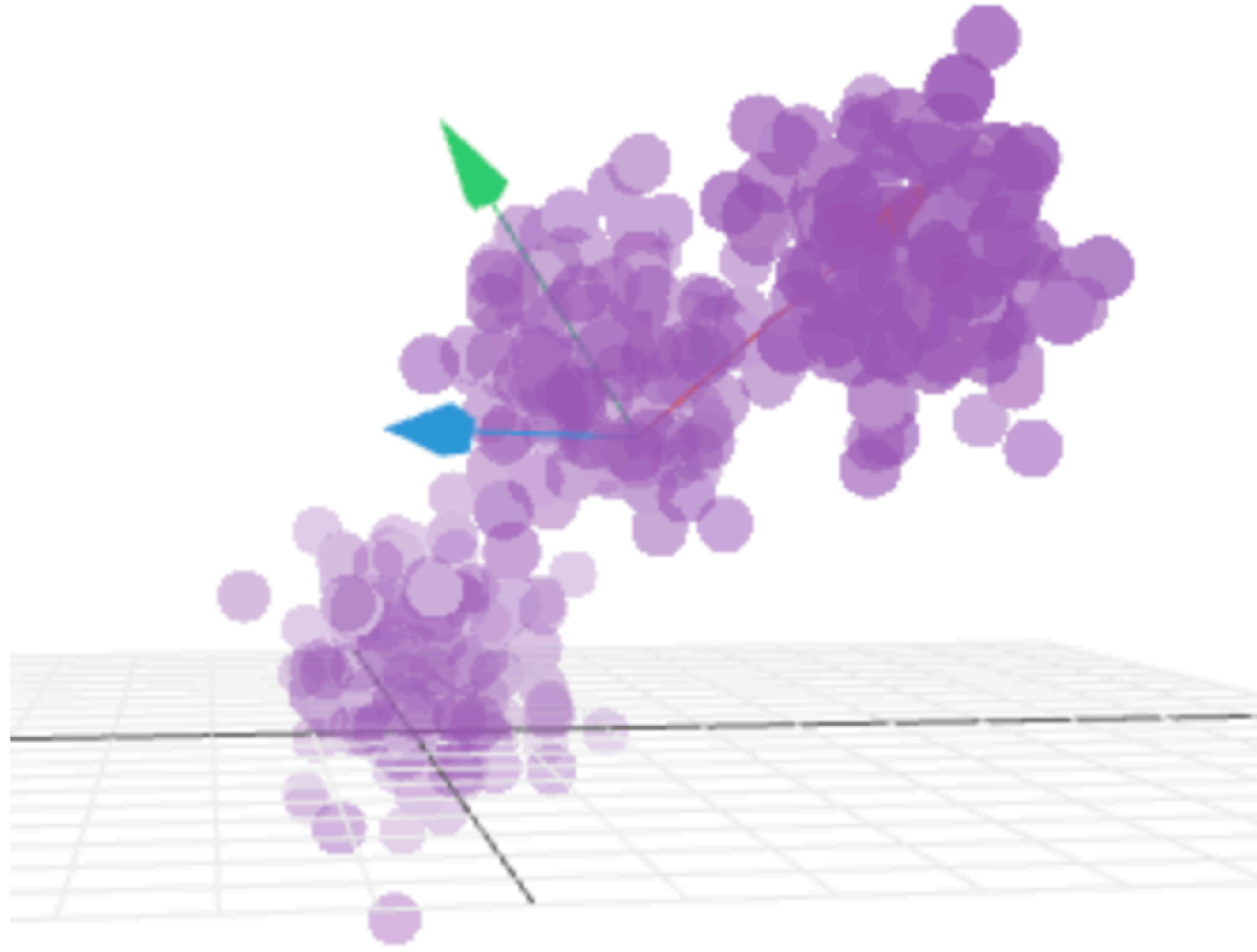
Goal: Reduce the dataset to fewer dimensions yet approx. preserve the distance between the individual samples



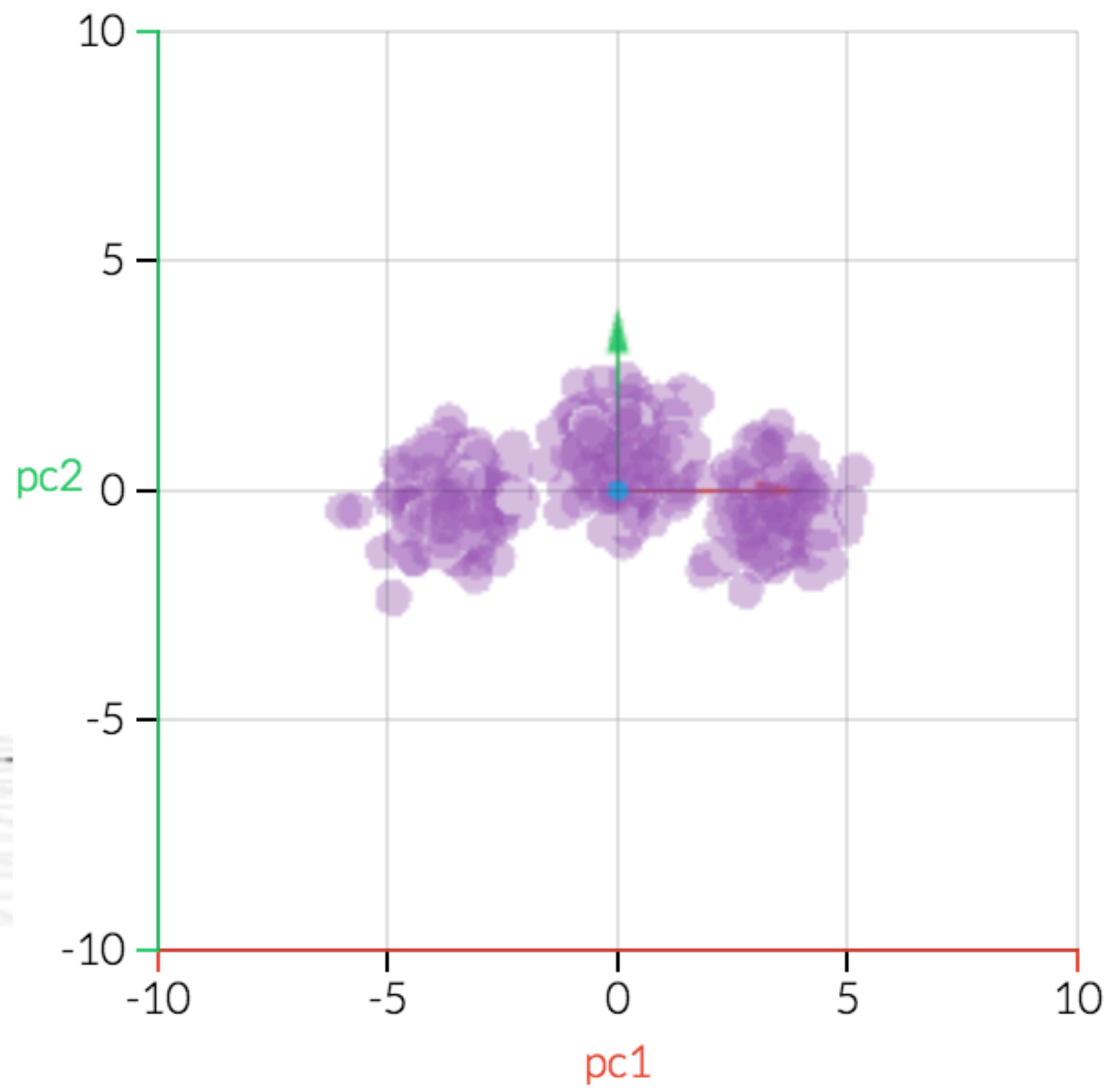
http://mev.tm4.org, TCGA-BRCA

Feature Reduction Technique

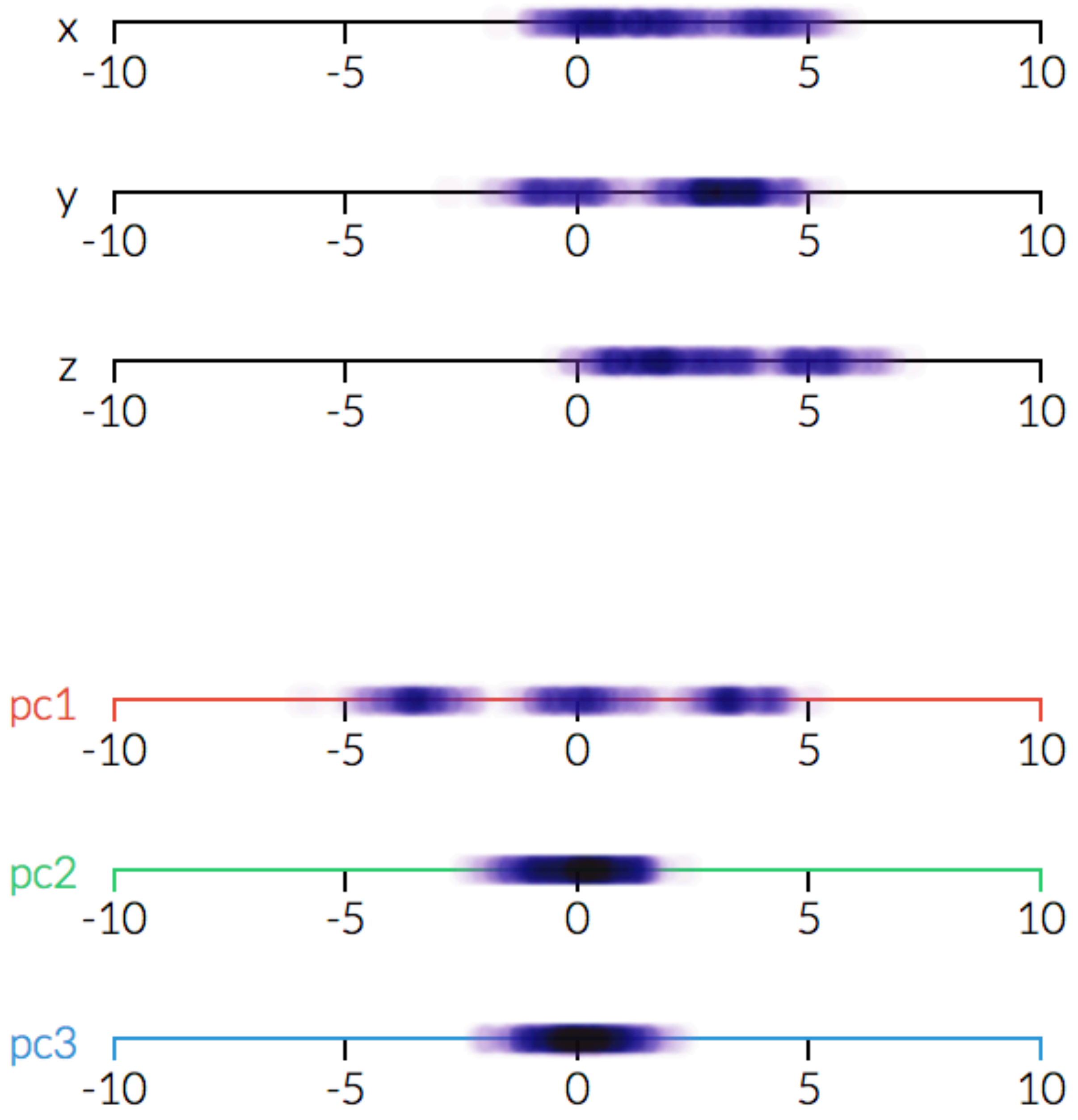
Principle Component Analysis (PCA) convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables (Principle Component or PC’s)



3-D



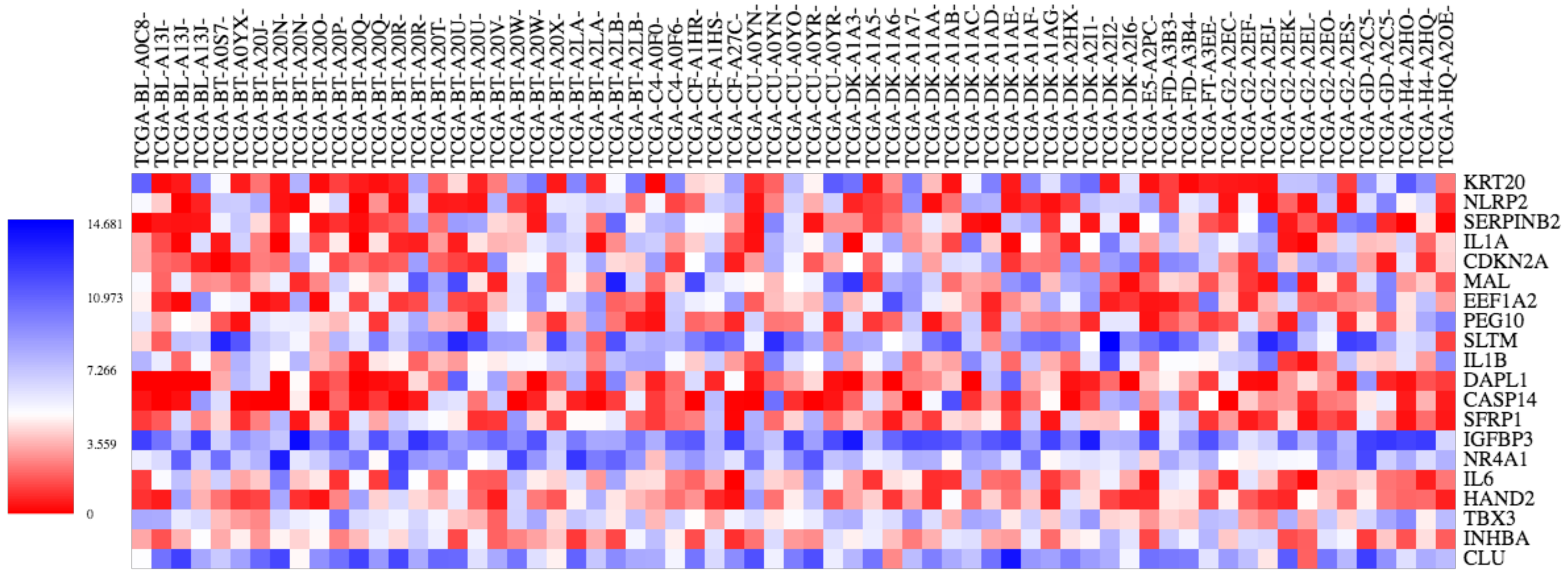
2-D



1-D

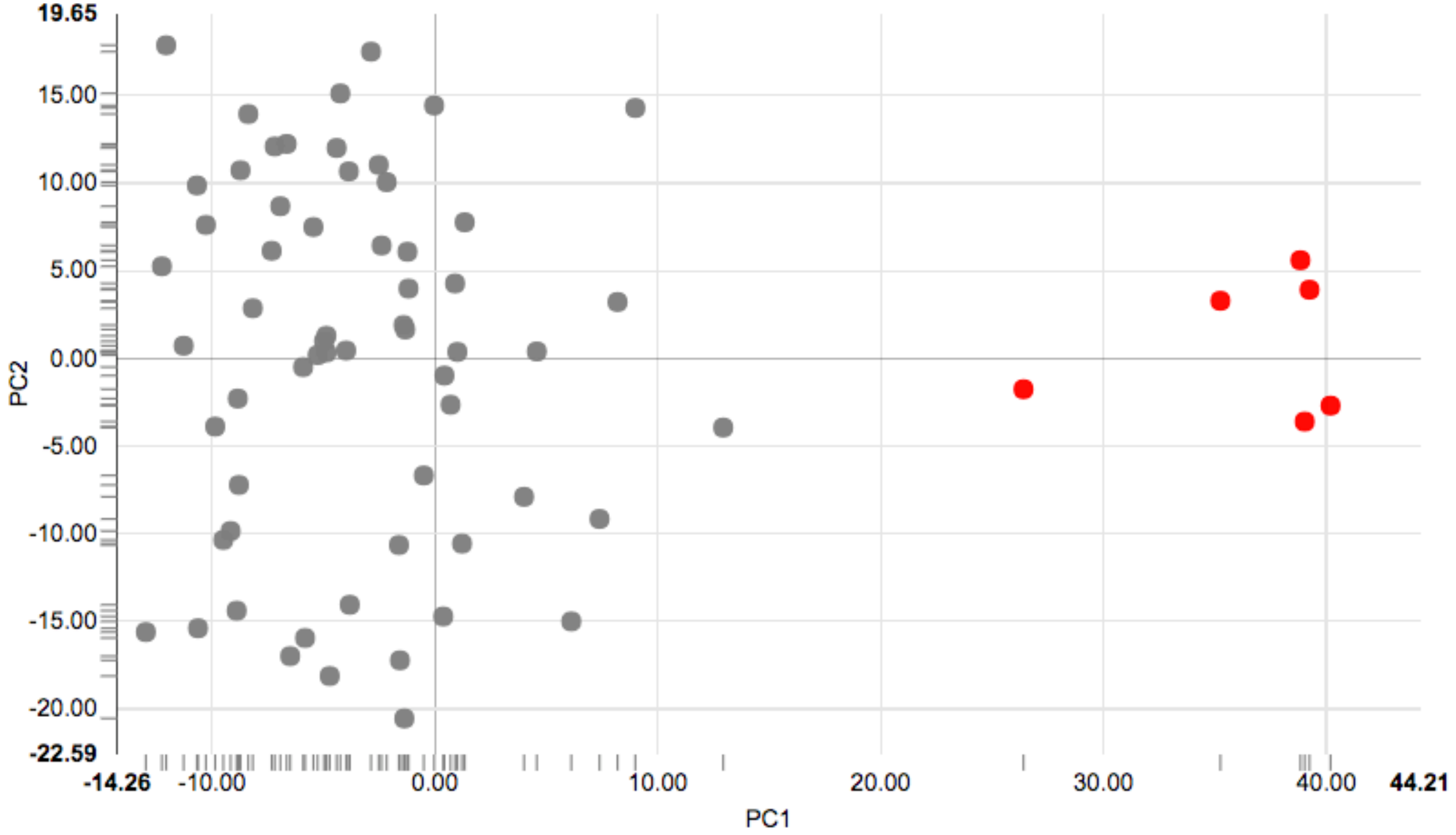
http://setosa.io/ev/principal-component-analysis/

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

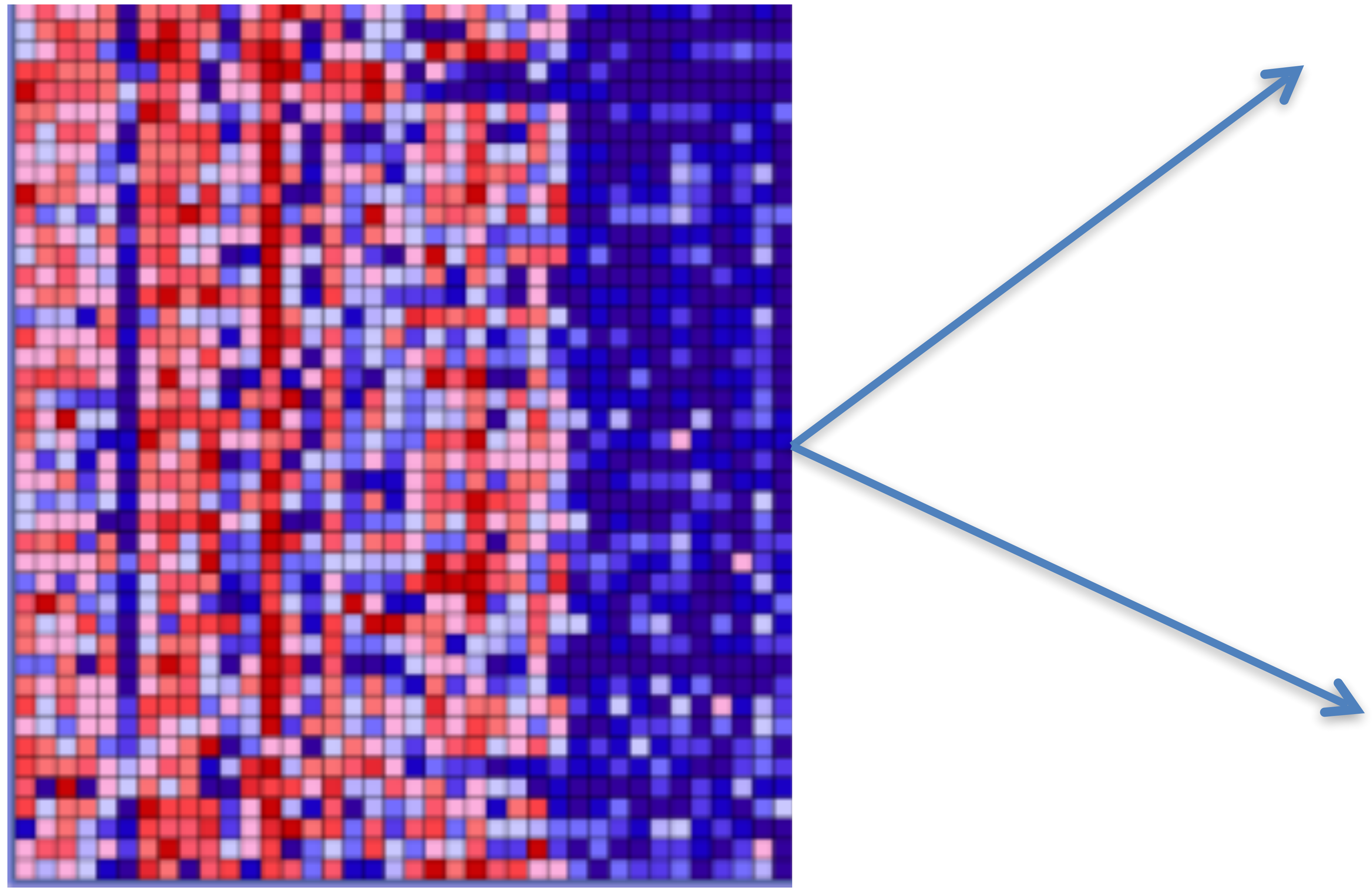


Selected	PC1	PC2
TCGA-BL-A13J-11A-13R-A10U-07	39.2507	3.9165
TCGA-BT-A20N-11A-11R-A14Y-07	40.1933	-2.6946
TCGA-BT-A20Q-11A-11R-A14Y-07	38.8414	5.5994
TCGA-BT-A20R-11A-11R-A16R-07	39.0328	-3.6043
TCGA-CU-A0YN-11A-11R-A10U-07	35.2515	3.2868
TCGA-CU-A0YR-11A-13R-A10U-07	26.4164	-1.7572

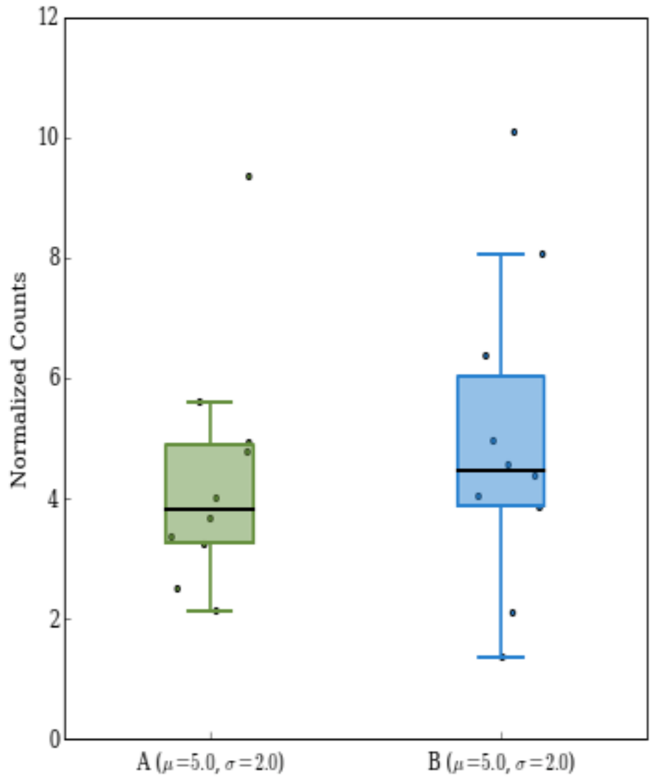
Differential gene expression analysis

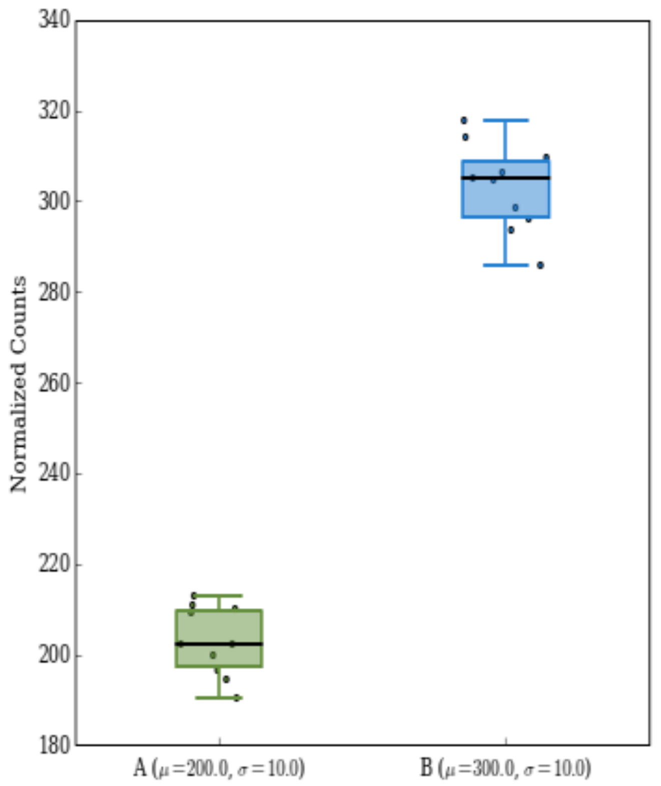
Identify genes with statistically significant expression differences between samples of different conditions

Modeling for Differential Gene Expression



1 test per gene!!!





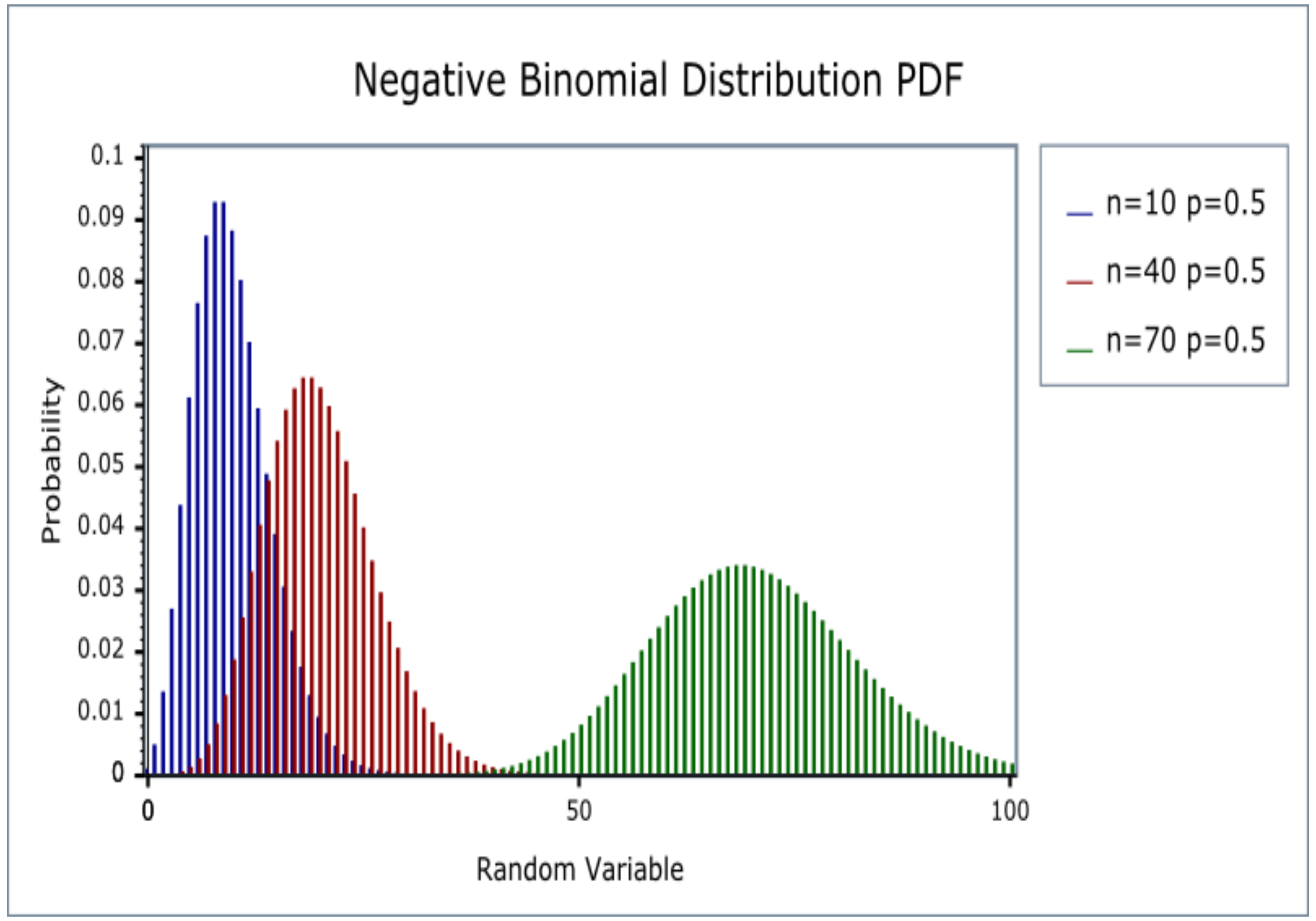
1. Estimate **magnitude** of DGE
   - Report as LogFC (log fold change)

2. Estimate the **significance** of
   - (adjusted) p-values that account for performing thousands of tests

H0: no difference in the read distribution between

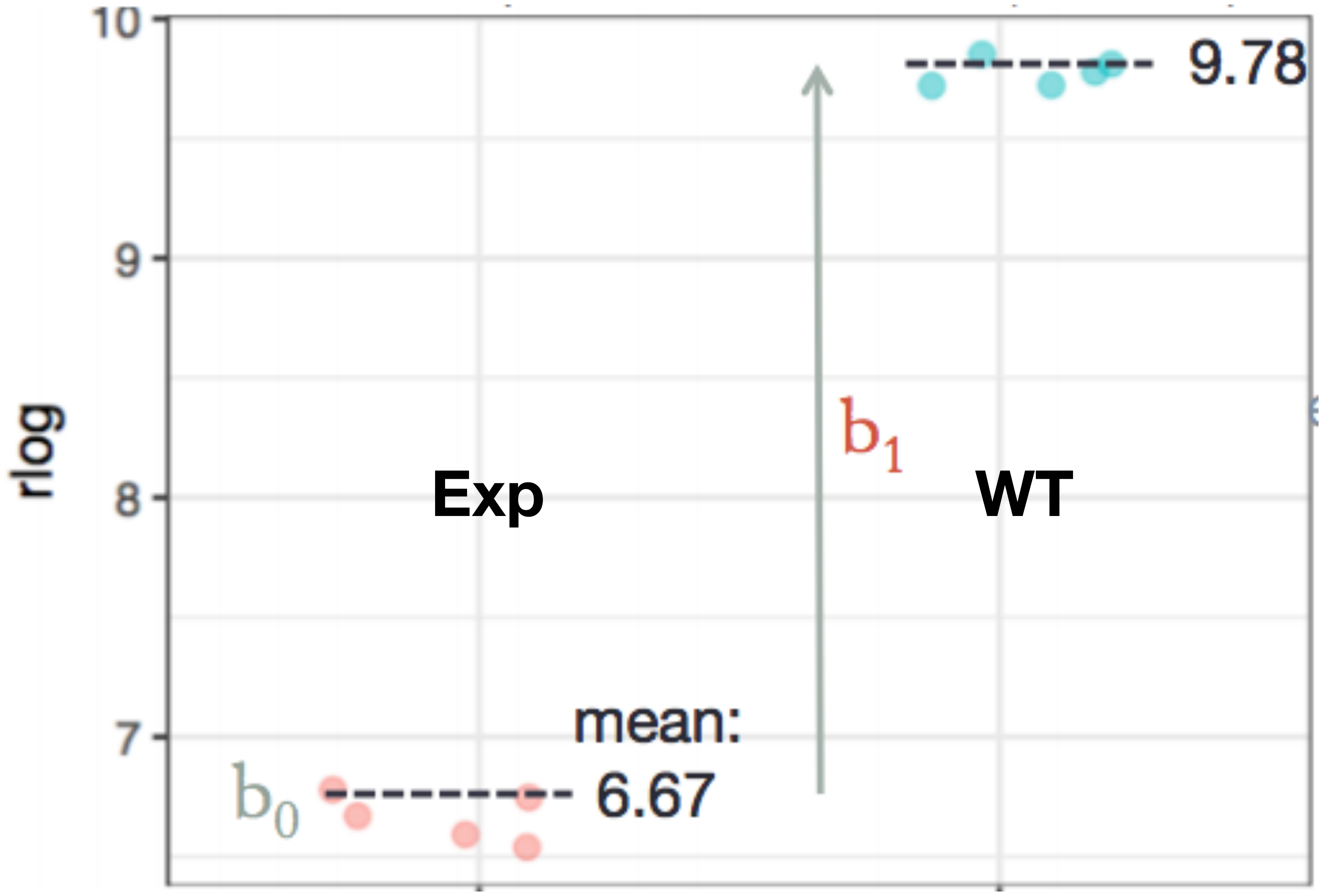
Modeling for Differential Gene Expression

1. Fit a statistical model



Empirically fit a distribution to estimate read count properties by **negative binomial distribution**

2. Estimate difference



Estimate the difference between groups using a linear model

$$Y = b_0 + b_1 * x + e$$

DGE Results

Gene	baseMean	baseMeanA	baseMeanB	foldChange	log2FC	pval	padj
FTL2	94.324	2.319	186.329	80.318	6.327	7.97E-44	2.89E-40
REC8	120.143	229.661	10.626	0.0462	-4.433	4.05E-38	9.32E-35
DLK2	626.928	1026.15	227.706	0.221	-2.171	1.18E-18	1.87E-15
…	…	…	…	…	…	…	…
PDE6b	430.808	301.37	560.239	1.858	0.894	0.328	0.765
LEPREL4	495.854	532.61	459.092	0.862	-0.214	0.328	0.765
NLRP12	4.009	5.466	2.535	0.463	-1.108	0.329	0.766

Commonly used methods DESeq2, edgeR, limma all produce results in similar format

Controlling false-positives by multiple comparisons

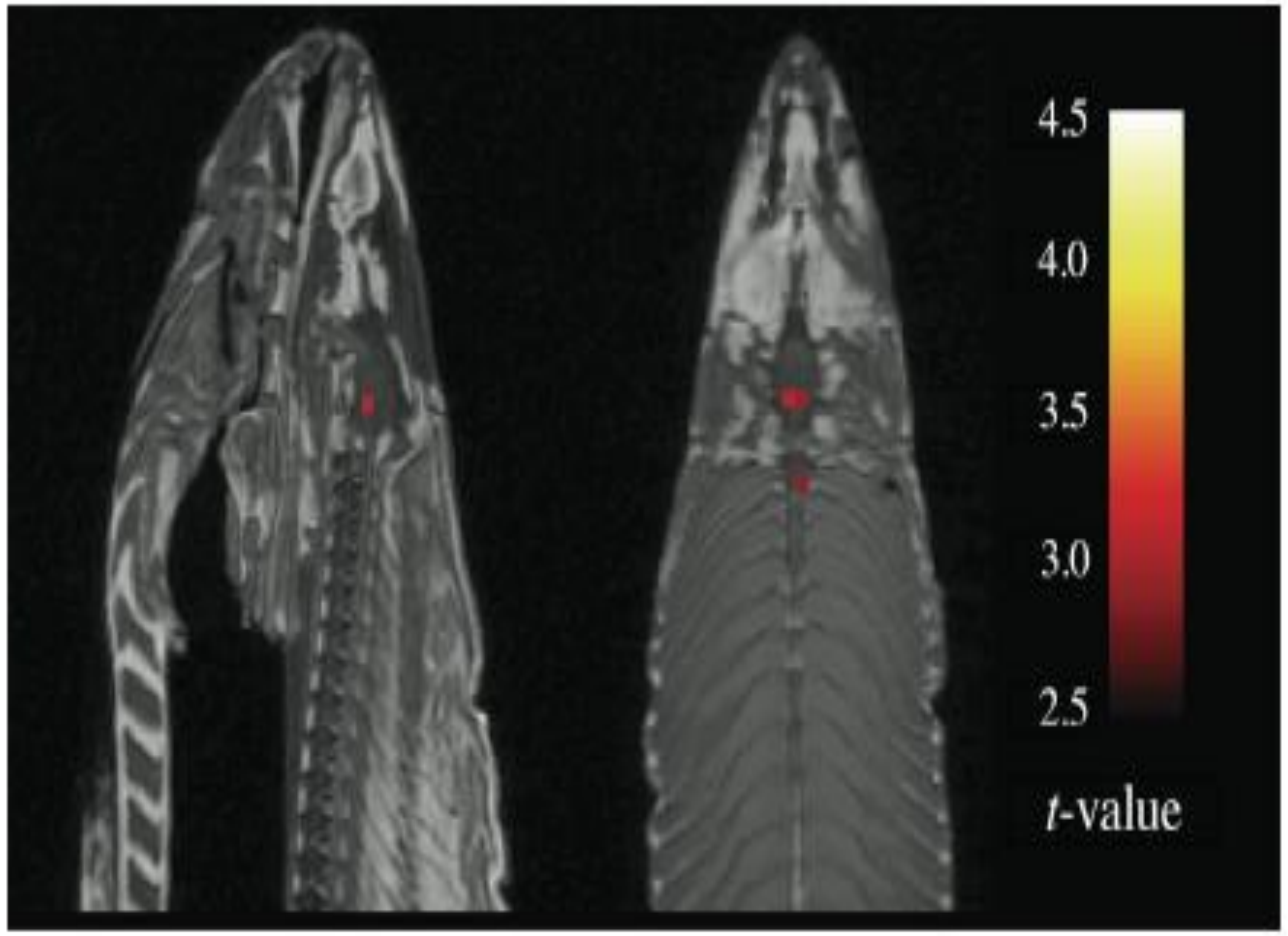
- When the same question is asked thousands of times, some will show up as significant by random

- Most commonly used method for RNASeq is False Discovery Rate (FDR) by Benjamini-Hochberg

$$FDR = Q_e = E[V/(V + R)]$$

V = False Positives
R = True Positives + False Positives

Ask a dead salmon a series of questions...



fMRI with many statistical tests performed (just like testing differential expression on many genes!)

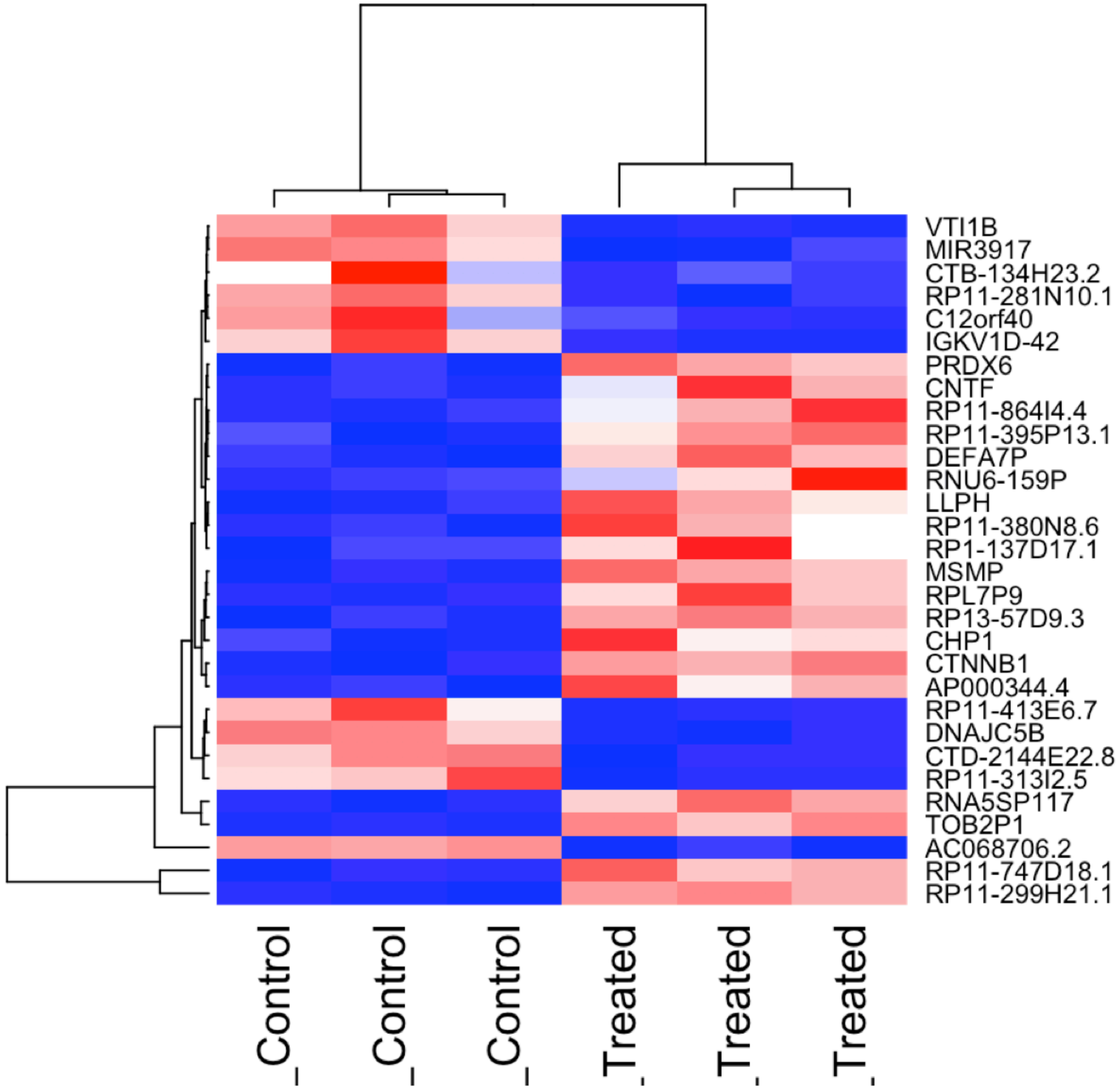
Bennett (2009)

DGE Results Examination

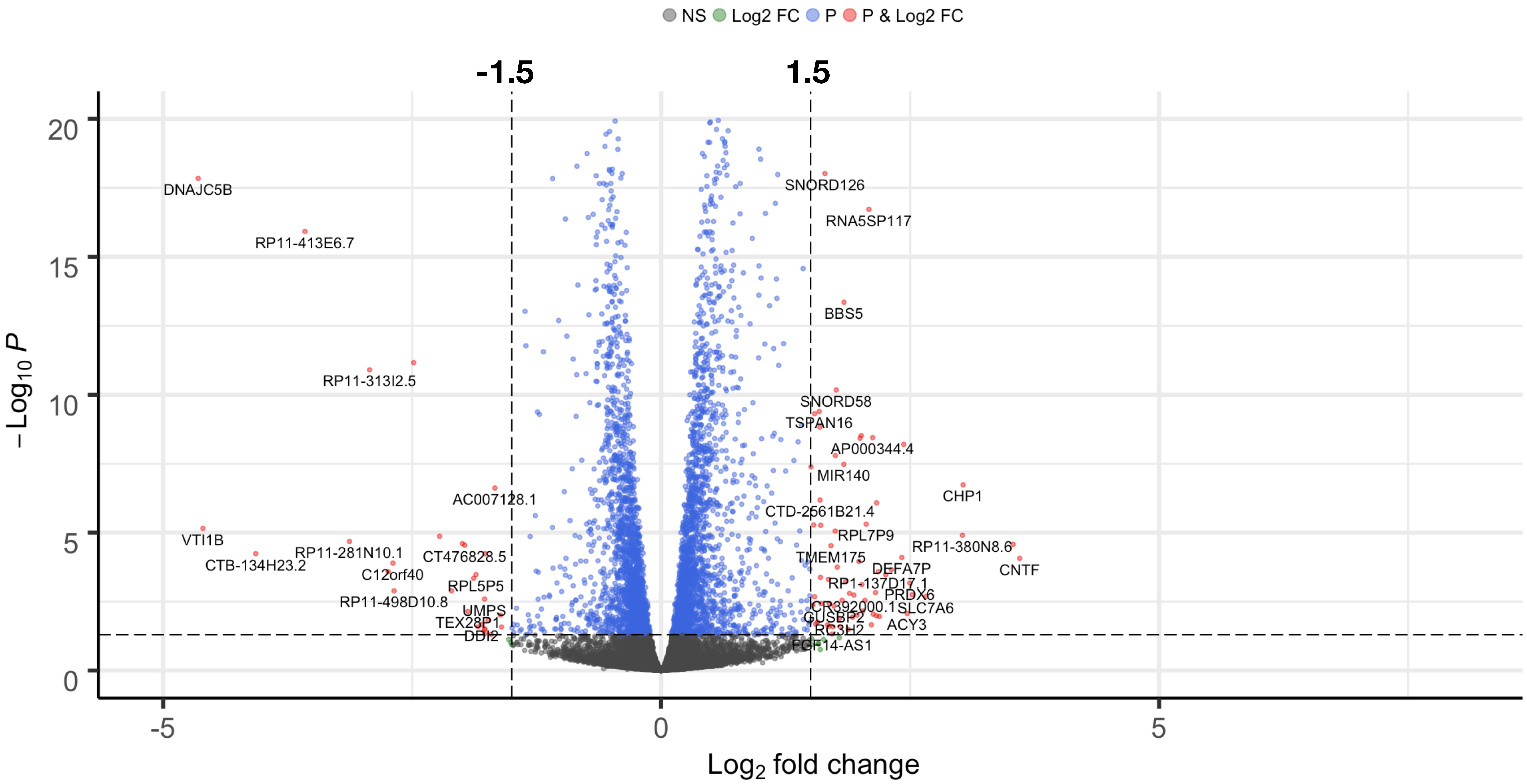
Gene	baseMean	baseMeanA	baseMeanB	foldChange	log2FC	pval	padj
FTL2	94.324	2.319	186.329	80.318	6.327	7.97E-44	2.89E-40
REC8	120.143	229.661	10.626	0.0462	-4.433	4.05E-38	9.32E-35
DLK2	626.928	1026.15	227.706	0.221	-2.171	1.18E-18	1.87E-15
…	…	…	…	…	…	…	…
PDE6b	430.808	301.37	560.239	1.858	0.894	0.328	0.765
LEPREL4	495.854	532.61	459.092	0.862	-0.214	0.328	0.765
NLRP12	4.009	5.466	2.535	0.463	-1.108	0.329	0.766

Filter DGE result table by Log2FC (usually > 1.2) or adjusted P-value

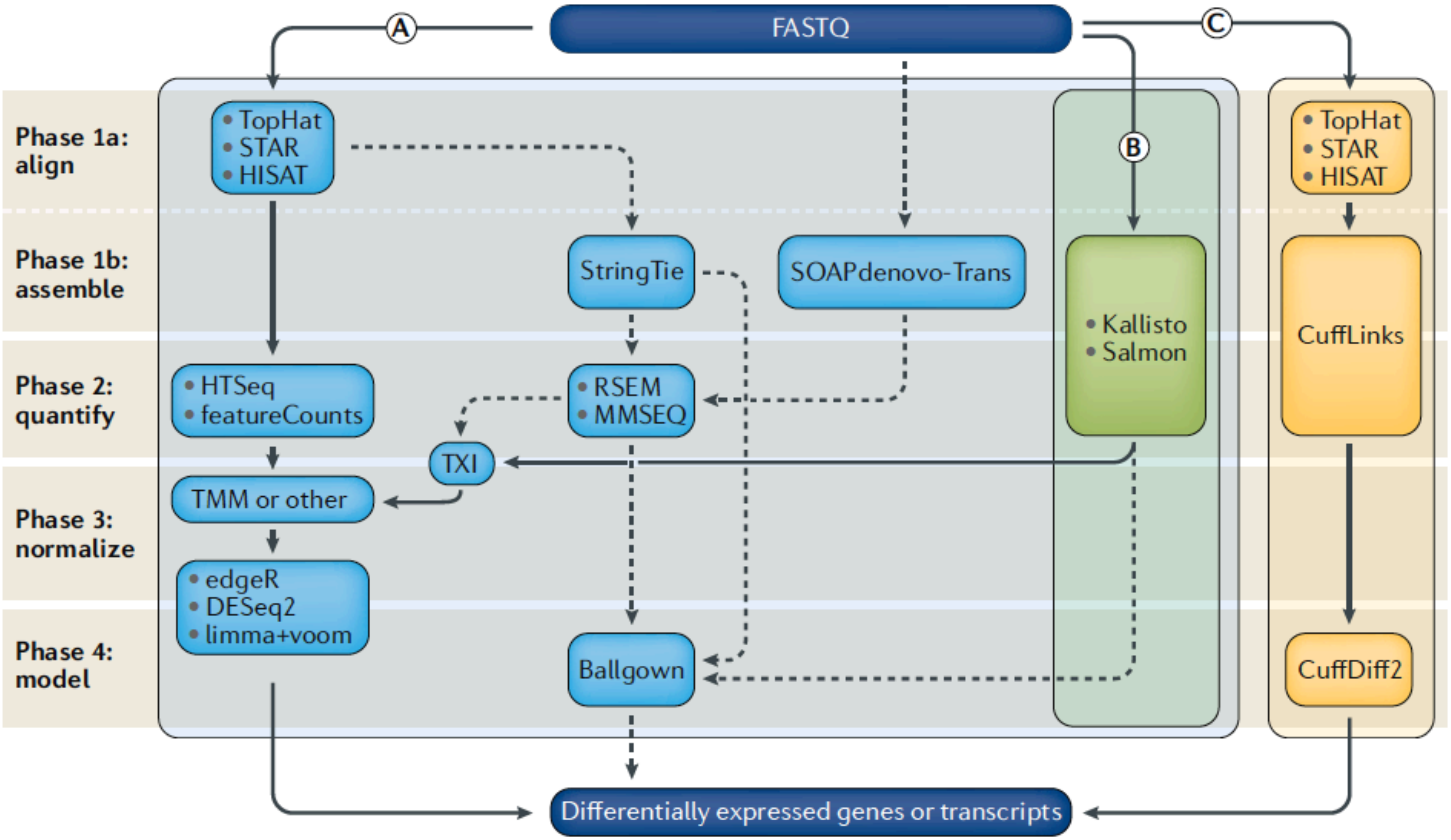
DGE Results: Heatmap



DGE Results - Volcano Plots



RNA-Seq Workflow Summary



Stark, et.al, Nat. Rev. Gen. (2019)

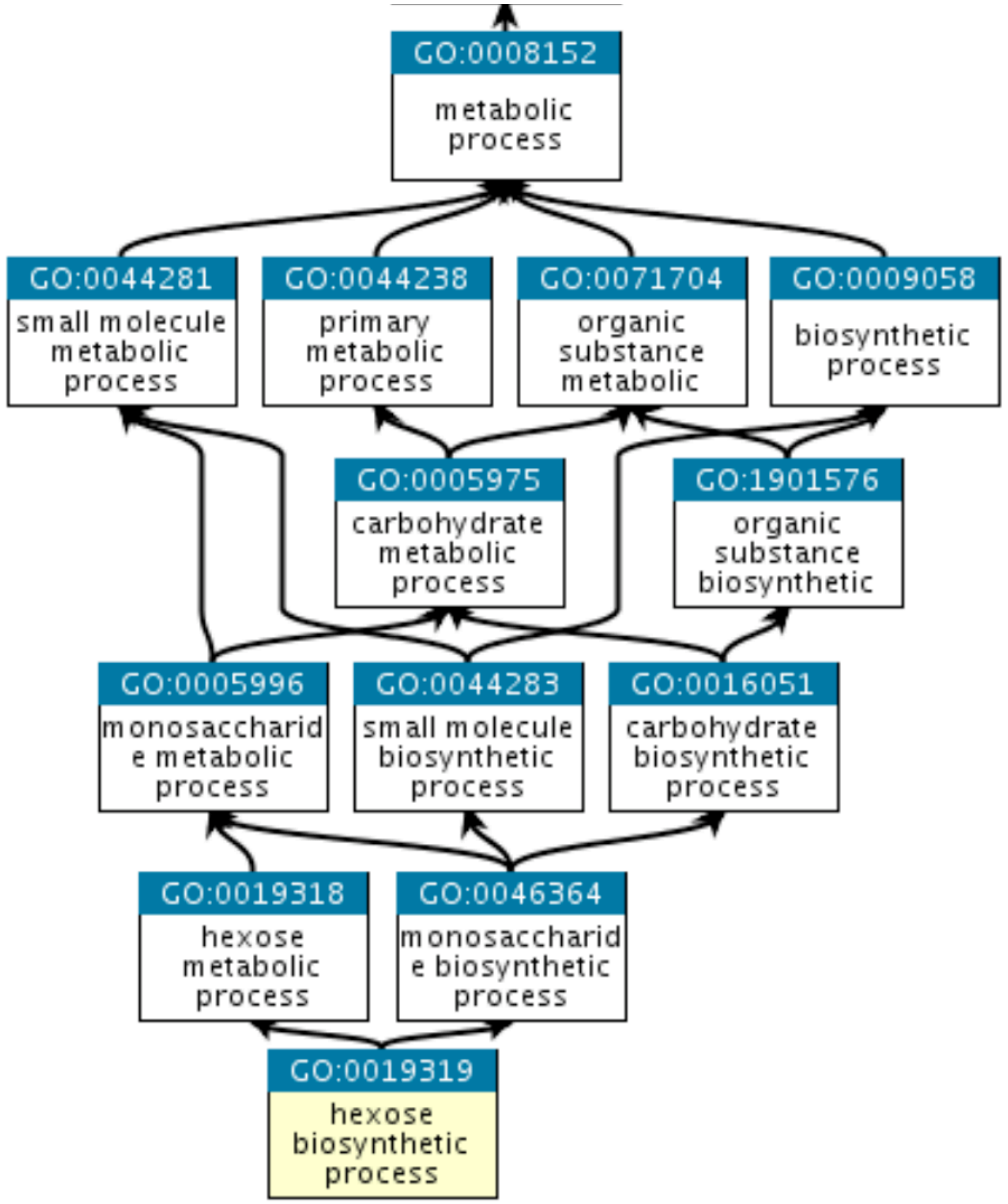
Functional Enrichment Analysis

Putting differential expressed genes into biological context using gene annotation databases

GENEONTOLOGY
Unifying Biology

Gene Ontology Database (GO) - terms that group genes into sets of classes by their annotations

1. **Molecular Function:** Molecular-level activities performed by gene products, such as “catalysis” or “transport”
2. **Cellular Component:** The locations relative to cellular structures in which a gene product performs a function (e.g. “mitochondrion”, “ribosome”)
3. **Biological Process:** The larger processes, or ‘biological programs’ accomplished by multiple molecular activities (e.g. “DNA repair”, “signal transduction”)


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

Loosely hierarchical GO Term structure

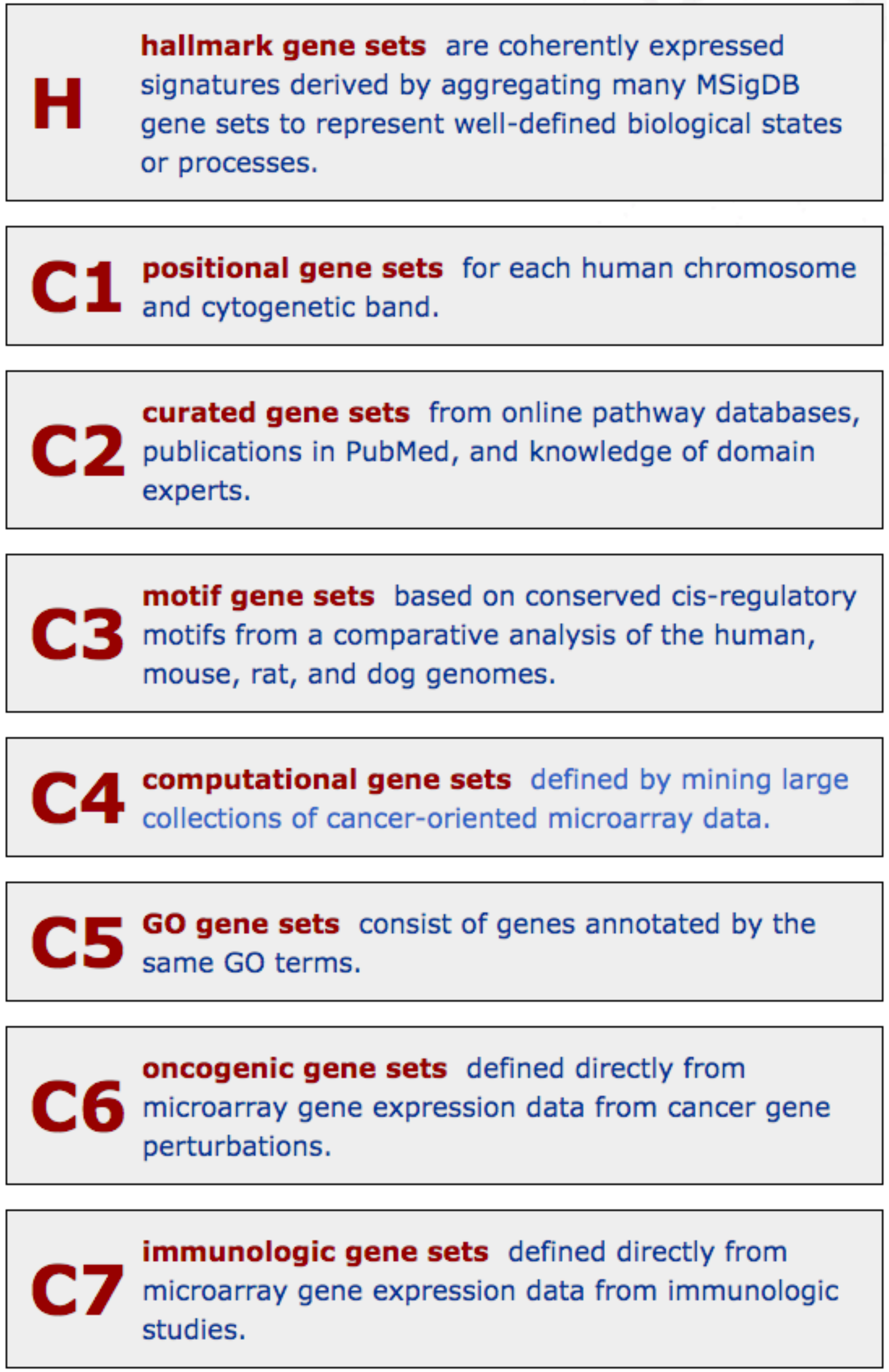
http://geneontology.org/

Ashburner et al. Nat. Genet. (2000)

MSigDB
Molecular Signatures
Database

The Molecular Signatures Database (MSigDB) is a collection of annotated gene sets

- Curated Gene Sets from literatures, such as functional pathway (KEGG), gene functional groups. **Most commonly used gene set class**
- Contain domain specific gene sets (H, C6, C7)
- Human genome location (C1) Predicted gene sets (C2, C4)
- GO term (C5)



H **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.

C1 **positional gene sets** for each human chromosome and cytogenetic band.

C2 **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.

C3 **motif gene sets** based on conserved cis-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.

C4 **computational gene sets** defined by mining large collections of cancer-oriented microarray data.

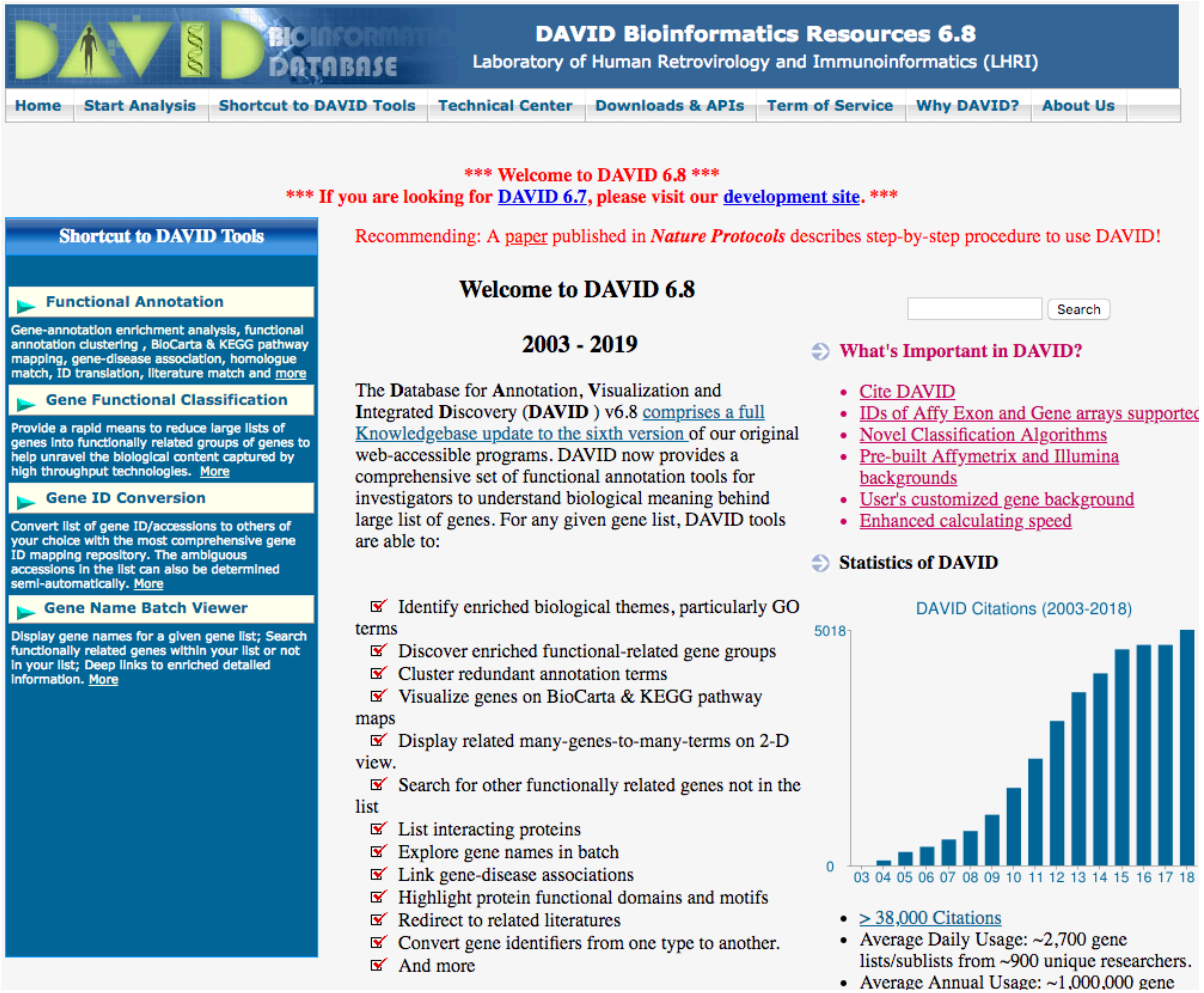
C5 **GO gene sets** consist of genes annotated by the same GO terms.

C6 **oncogenic gene sets** defined directly from microarray gene expression data from cancer gene perturbations.

C7 **immunologic gene sets** defined directly from microarray gene expression data from immunologic studies.

http://software.broadinstitute.org/gsea/msigdb/index.jsp *Tamayo, et al. PNAS (2005)*

Functional Enrichment Analysis with DAVID



DAVID Bioinformatics Resources 6.8

Laboratory of Human Retrovirology and Immunoinformatics (LHRI)

Home | Start Analysis | Shortcut to DAVID Tools | Technical Center | Downloads & APIs | Term of Service | Why DAVID? | About Us

*** Welcome to DAVID 6.8 ***

*** If you are looking for DAVID 6.7, please visit our development site. ***

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

Recommending: A paper published in *Nature Protocols* describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.8

2003 - 2019

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.8 comprises a full Knowledgebase update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

☑ Identify enriched biological themes, particularly GO terms
☑ Discover enriched functional-related gene groups
☑ Cluster redundant annotation terms
☑ Visualize genes on BioCarta & KEGG pathway maps
☑ Display related many-genes-to-many-terms on 2-D view.
☑ Search for other functionally related genes not in the list
☑ List interacting proteins
☑ Explore gene names in batch
☑ Link gene-disease associations
☑ Highlight protein functional domains and motifs
☑ Redirect to related literatures
☑ Convert gene identifiers from one type to another.
☑ And more

Search

What's Important in DAVID?

- Cite DAVID
- IDs of Affy Exon and Gene arrays supportec
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID

DAVID Citations (2003-2018)

- > 38,000 Citations
- Average Daily Usage: ~2,700 gene lists/sublists from ~900 unique researchers.
- Average Annual Usage: ~1,000,000 gene

Use a modified **Fisher Exact Test** to determine if there is enrichment

Confusion Matrix	Number of genes is DGE	Number of genes is not DGE
Number of genes in pathway *y*	76	20
Number of genes not in pathway *y*	2	29920

p<0.00001!!!

Conclusion: Pathway *y* is differentially regulated

https://david.ncifcrf.gov/

Sherman, et. al, Nat. Prot.(2008)

Detecting modest but coordinate changes

The goal of GSEA is to detect modest but coordinated changes in pre-specified sets of related genes by using all genes and their statistical variation values

Step 0: Sort Genes into a ranked gene list

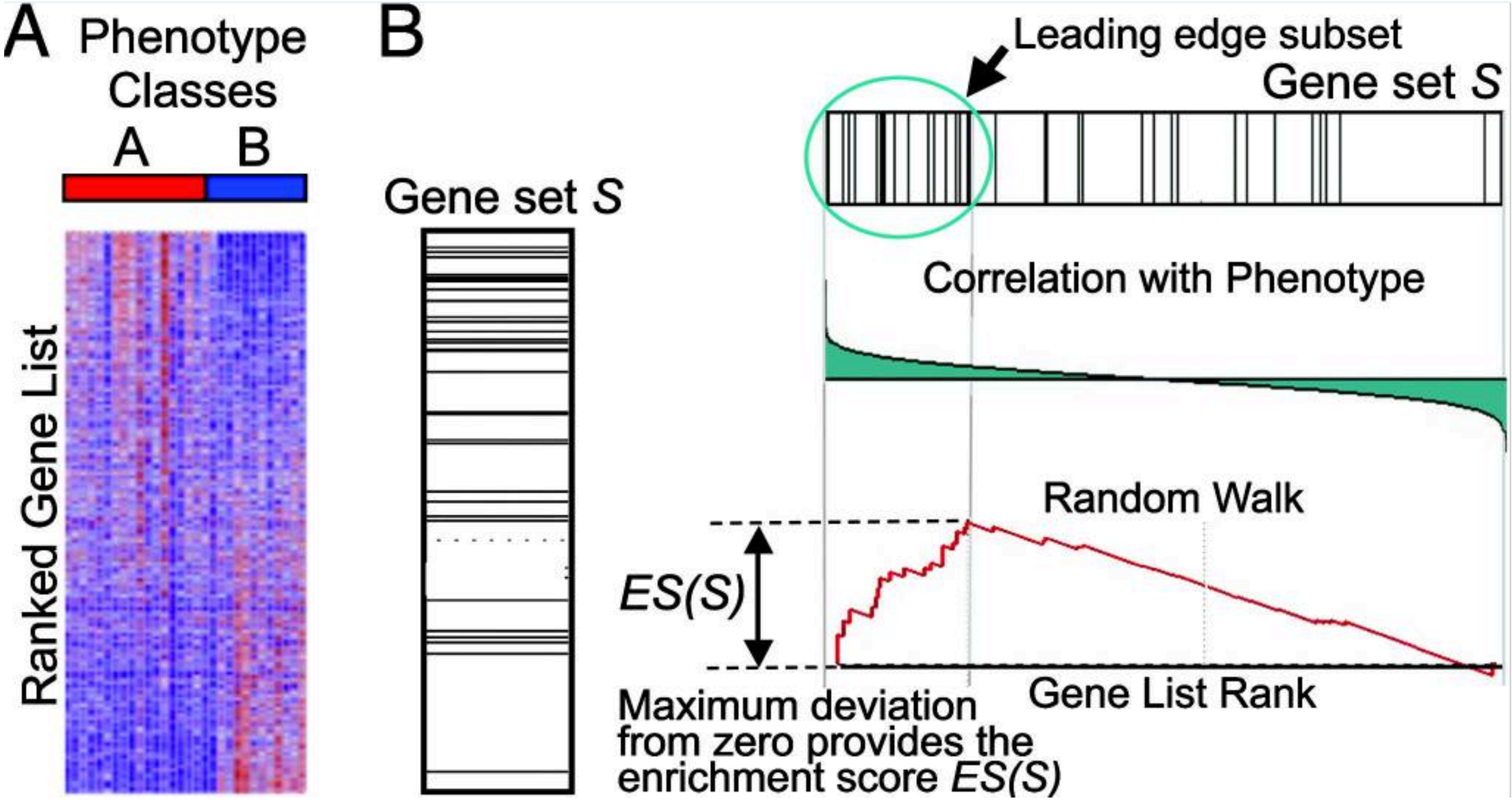
Step 1: Calculate Enrichment Score: Compute cumulative sum over ranked genes by summing statistics of gene in a set, and subtracting statistics of genes outside of the set

Step 2: Assess significance using Permutation Test: permute sample phenotype labels

Step 3: Adjust for multiple hypothesis testing: using FDR correction

http://software.broadinstitute.org/gsea/index.jsp *Tamayo, et al. PNAS (2005)*

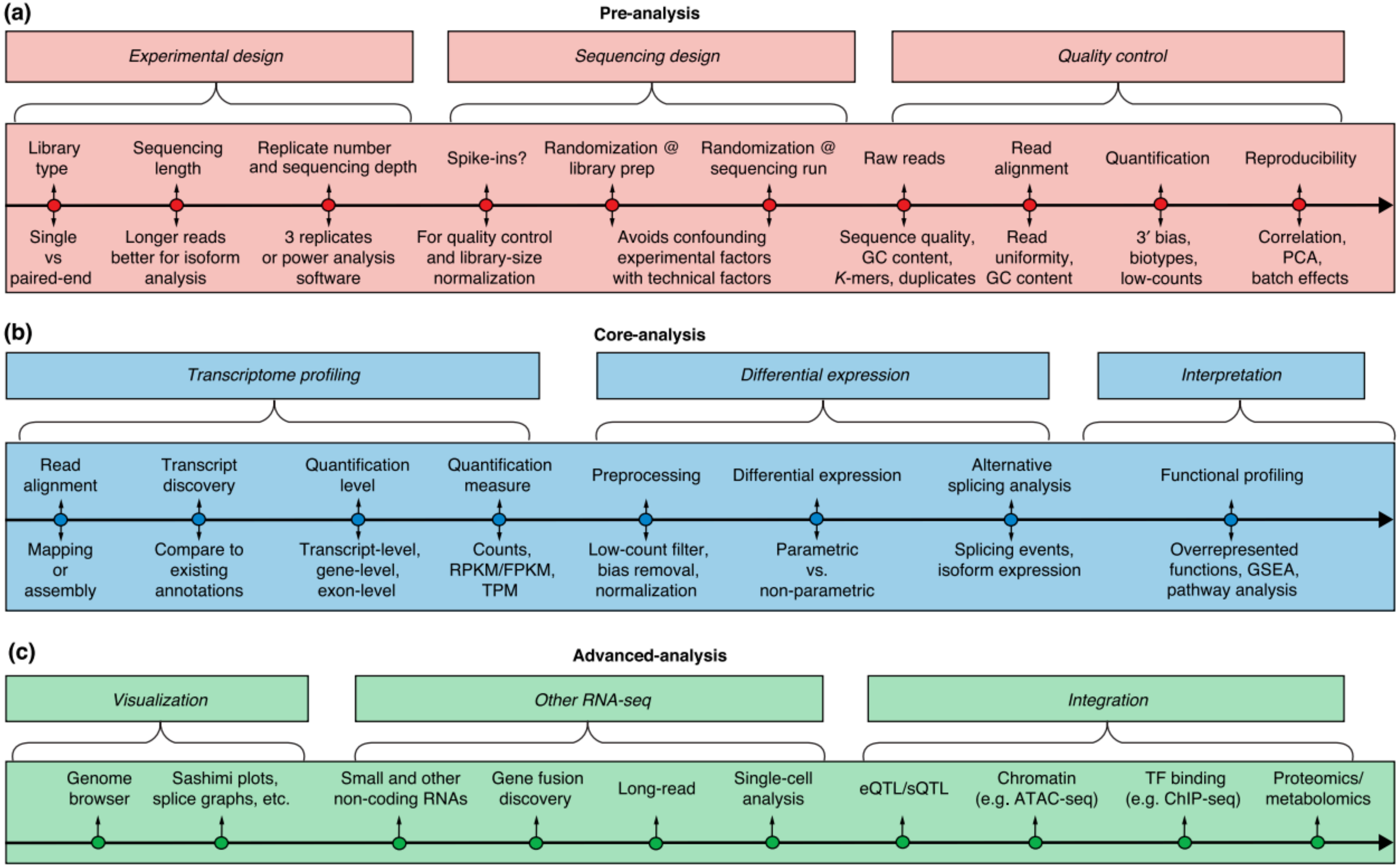
Detecting modest but coordinate changes



http://software.broadinstitute.org/gsea/index.jsp

Tamayo, et al. PNAS (2005)

RNA-Seq Analysis Road Map



(a) **Pre-analysis**

Experimental design

Library type	Single vs paired-end
Sequencing length	Longer reads better for isoform analysis
Replicate number and sequencing depth	3 replicates or power analysis software

Sequencing design

Spike-ins?	For quality control and library-size normalization
Randomization @ library prep	Avoids confounding experimental factors with technical factors
Randomization @ sequencing run	

Quality control

Raw reads	Sequence quality, GC content, *K*-mers, duplicates
Read alignment	Read uniformity, GC content
Quantification	3′ bias, biotypes, low-counts
Reproducibility	Correlation, PCA, batch effects

(b) **Core-analysis**

Transcriptome profiling

Read alignment	Mapping or assembly
Transcript discovery	Compare to existing annotations
Quantification level	Transcript-level, gene-level, exon-level
Quantification measure	Counts, RPKM/FPKM, TPM

Differential expression

Preprocessing	Low-count filter, bias removal, normalization
Differential expression	Parametric vs. non-parametric
Alternative splicing analysis	Splicing events, isoform expression

Interpretation

Functional profiling	Overrepresented functions, GSEA, pathway analysis

(c) **Advanced-analysis**

Visualization

- Genome browser
- Sashimi plots, splice graphs, etc.

Other RNA-seq

- Small and other non-coding RNAs
- Gene fusion discovery
- Long-read
- Single-cell analysis

Integration

- eQTL/sQTL
- Chromatin (e.g. ATAC-seq)
- TF binding (e.g. ChIP-seq)
- Proteomics/ metabolomics

Conesa, et al, Geno. Bio, 2016