Heat maps and clustering

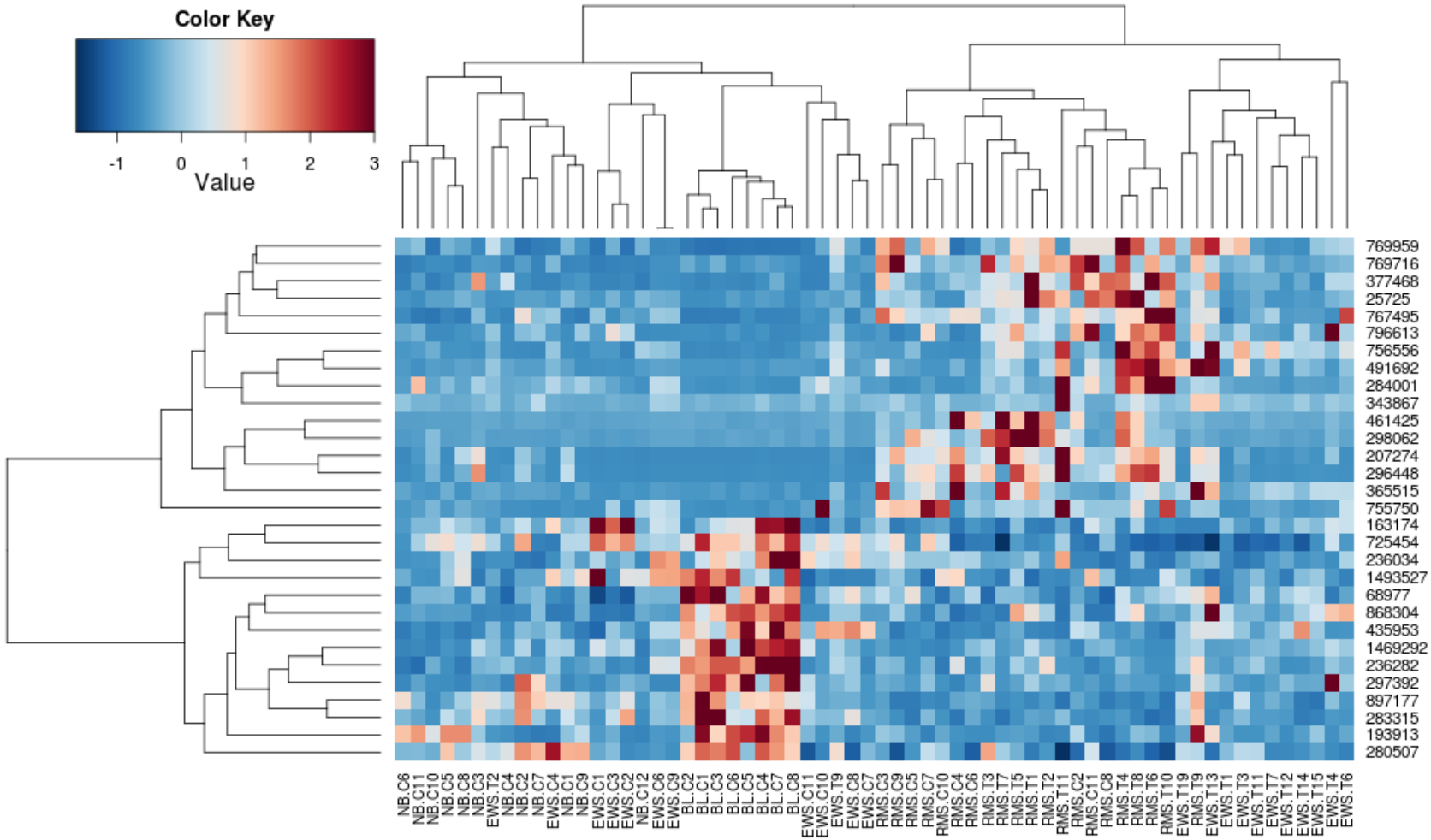


Figure 1:

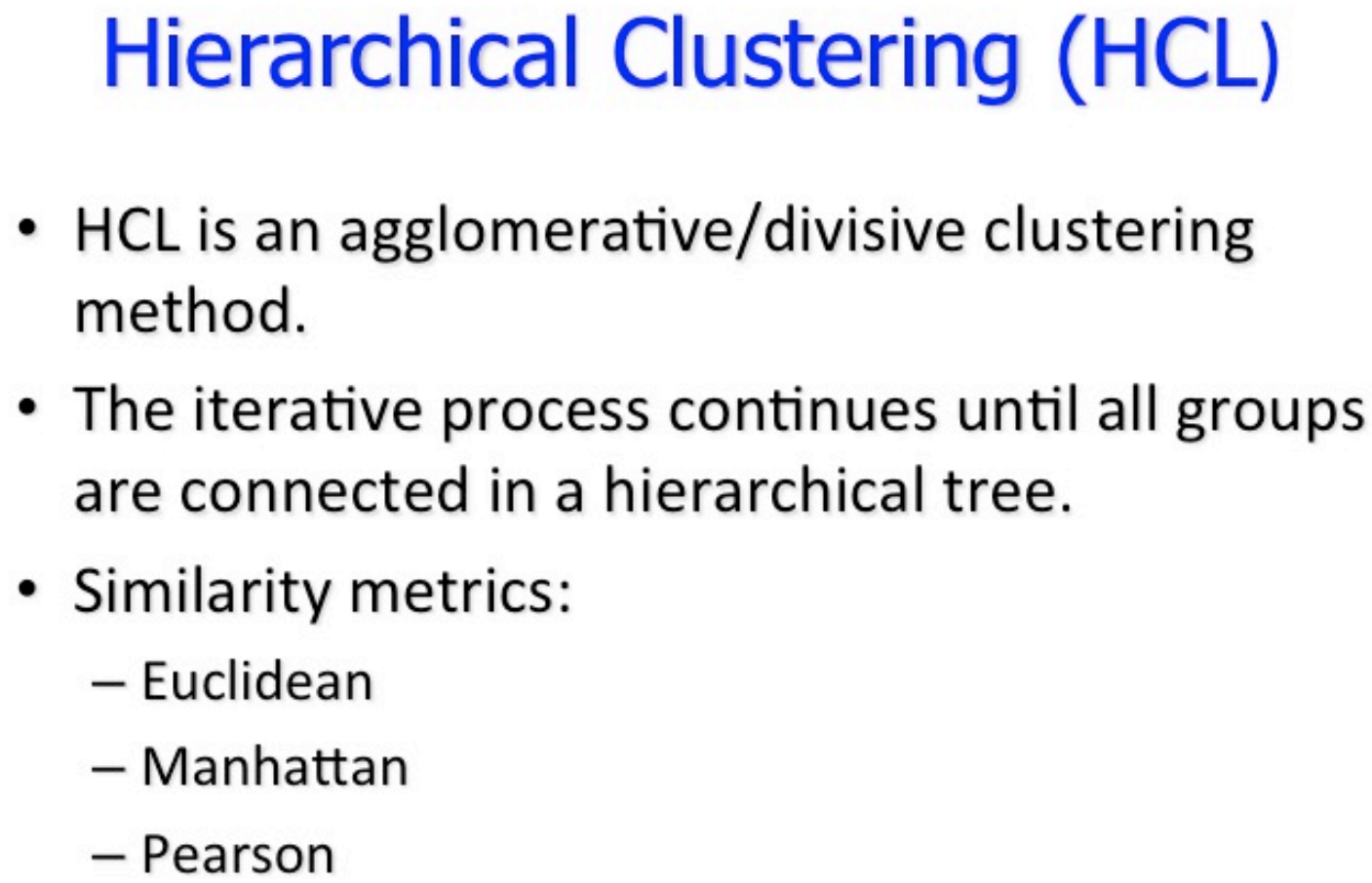

Hierarchical Clustering (HCL)
• HCL is an agglomerative/divisive clustering method.
• The iterative process continues until all groups are connected in a hierarchical tree.
• Similarity metrics:
– Euclidean
– Manhattan
– Pearson

Figure 2:

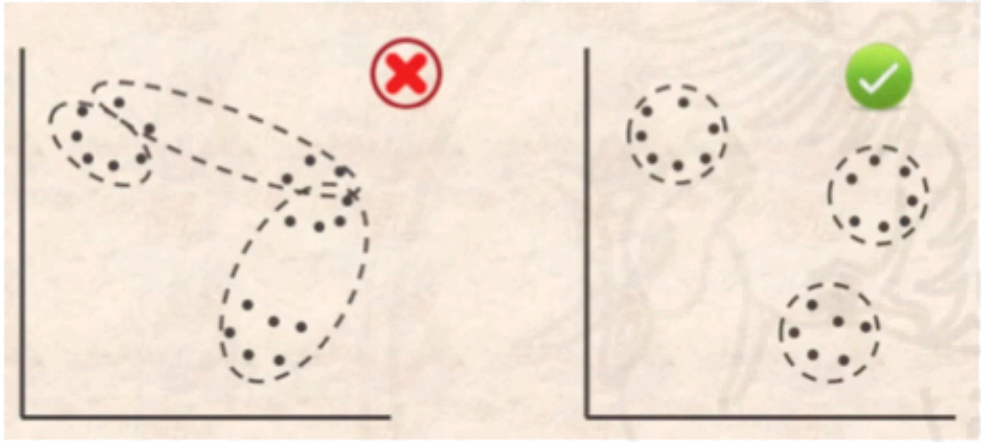
Heat maps and clustering

Clustering as optimization problem

Toward a Computational Problem

Good Clustering Principle:

Elements within the same cluster should be closer to each other than elements in different clusters.



- we define a threshold Δ then:
  - distance between elements in the same cluster must be $\leq \Delta$;
  - distance between elements in different clusters must be $> \Delta$;

Figure 3:

Heat maps and clustering

Clustering as optimization problem

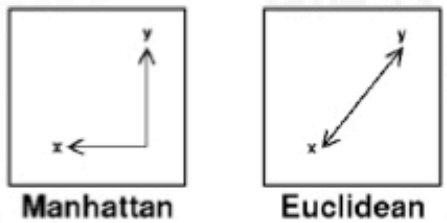
Distance from a Single DataPoint to Centers

- Different distance metrics can be used;
- The most used metrics are:
  - Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i \in m} (x_i - y_i)^2}$$

  - Manhattan distance:

$$d(x, y) = \sum_{i \in m} |(x_i - y_i)|$$



- hereafter we will use Euclidean distance, Manhattan distance works better in case of high dimensional vectors.

Figure 4:

Heat maps and clustering

Hierarchical Clustering

Stratification of Clusters

- Clusters often have **sub-cluster**, which have sub-clusters, and so on.

Figure 5:

Heat maps and clustering

Hierarchical Clustering

Stratification of Clusters

- Clusters often have **sub-cluster**, which have sub-clusters, and so on.

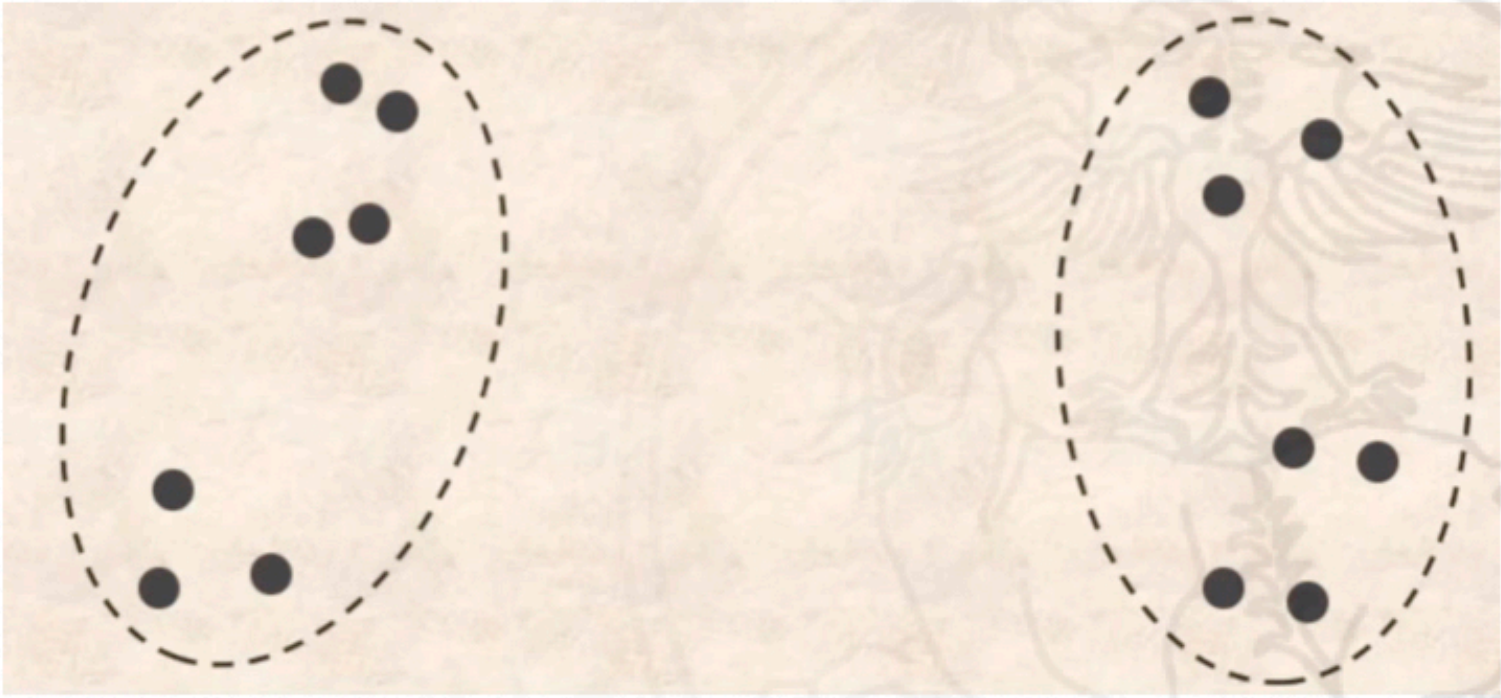


Figure 6:

Heat maps and clustering

Hierarchical Clustering

Stratification of Clusters

- Clusters often have **sub-cluster**, which have sub-clusters, and so on.

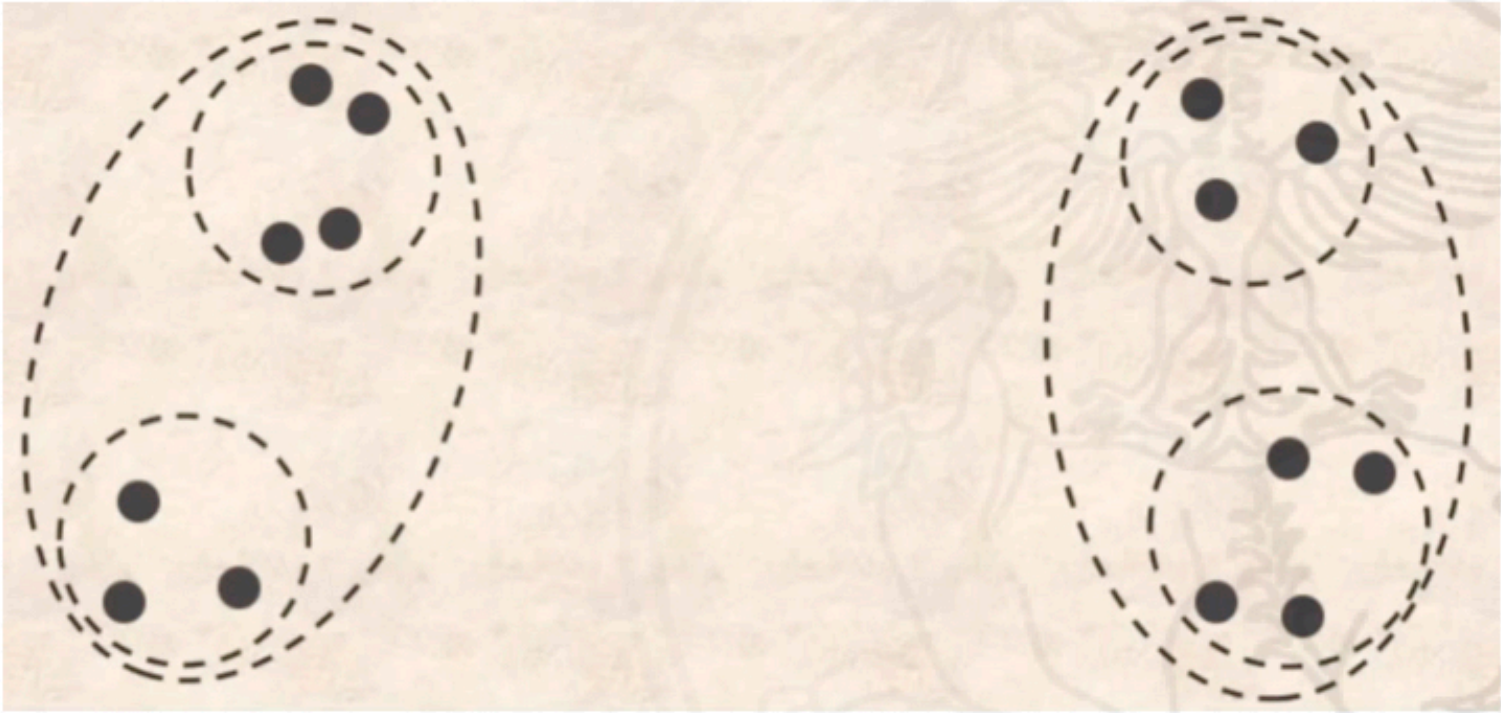


Figure 7:

Heat maps and clustering

Hierarchical Clustering

Stratification of Clusters

- Clusters often have **sub-cluster**, which have sub-clusters, and so on.

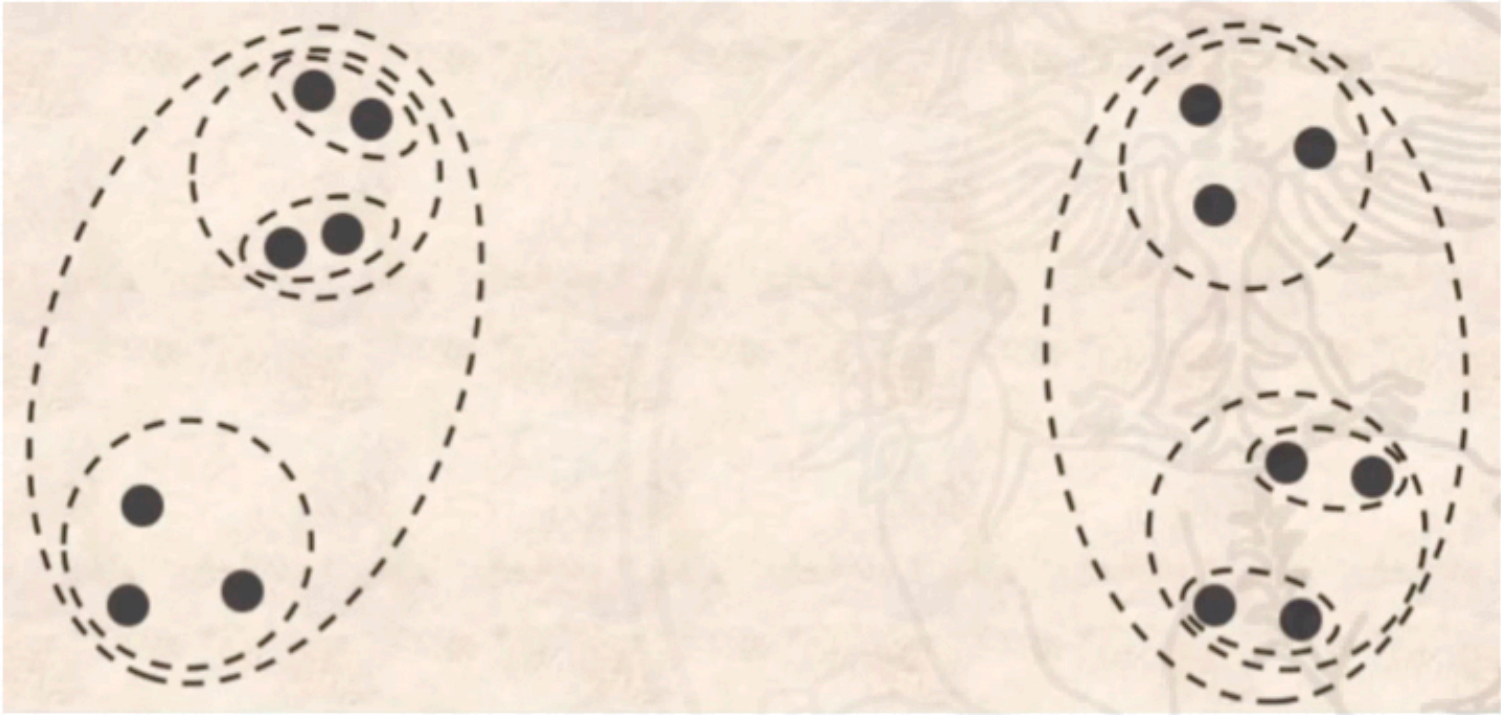


Figure 8:

Heat maps and clustering

Hierarchical Clustering

From Data to a Tree

- To capture stratification, the **hierarchical clustering** algorithm organizes n data points int a tree.

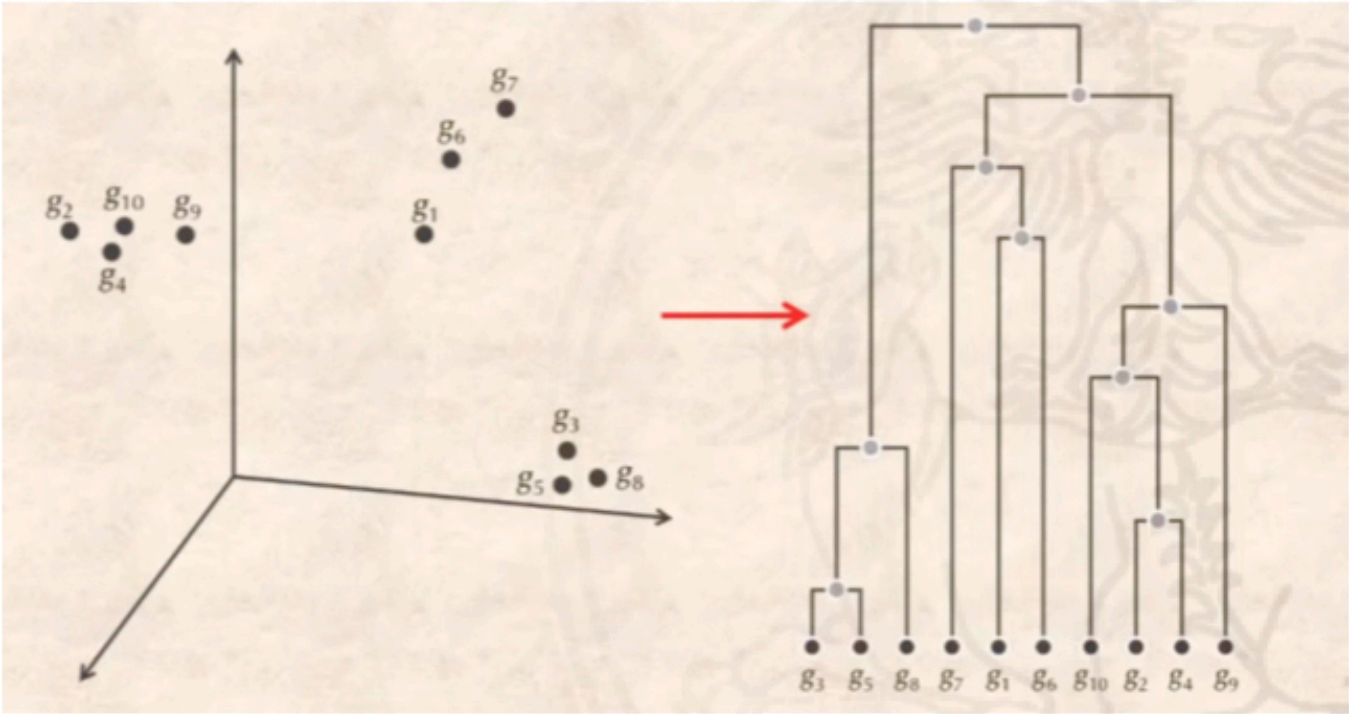


Figure 9:

Heat maps and clustering

Hierarchical Clustering

From Data to a Tree

- To capture stratification, the **hierarchical clustering** algorithm organizes n data points int a tree.

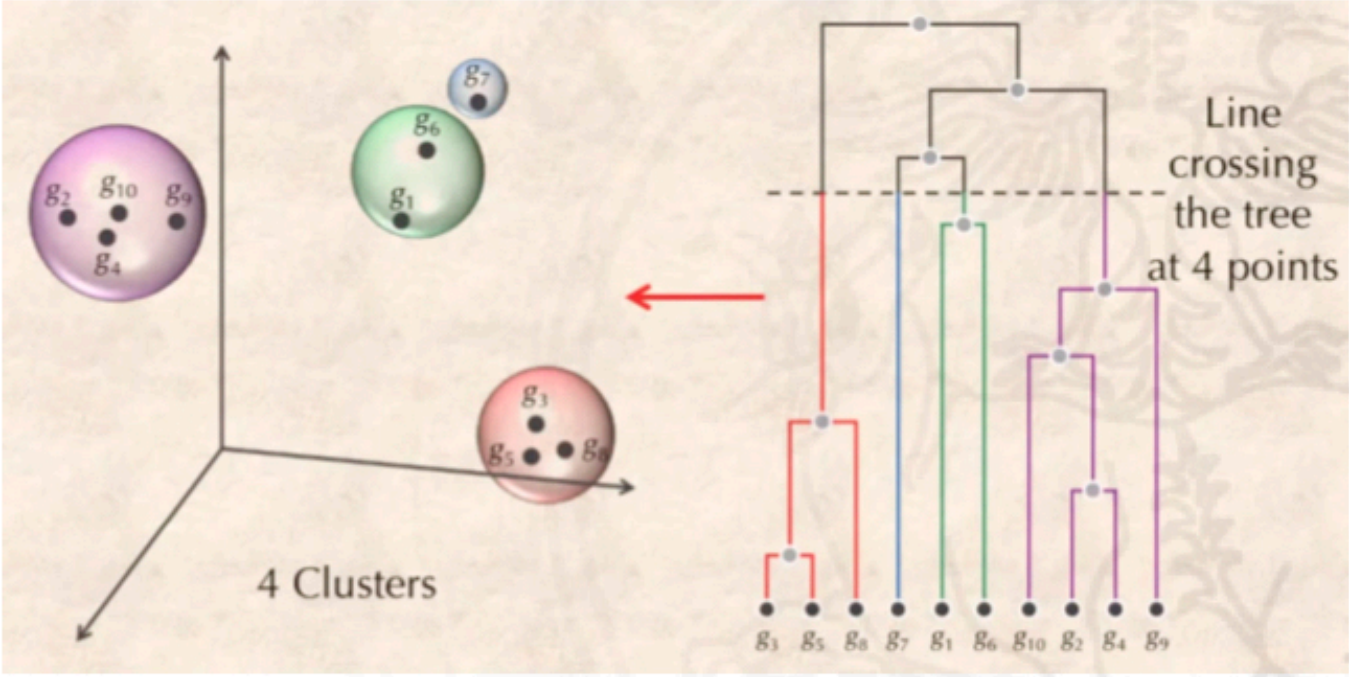


Figure 10:

Heat maps and clustering

Hierarchical Clustering

From Data to a Tree

- To capture stratification, the **hierarchical clustering** algorithm organizes n data points int a tree.

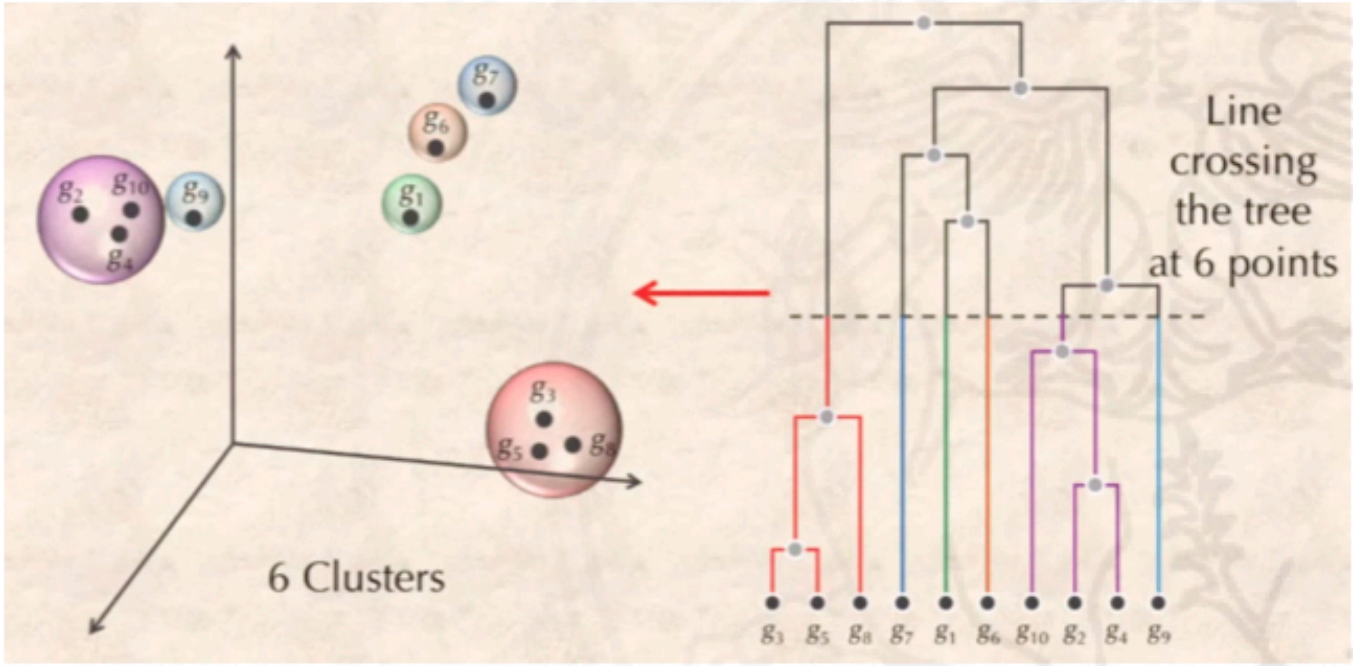


Figure 11:

Heat maps and clustering

Clustering as optimization problem

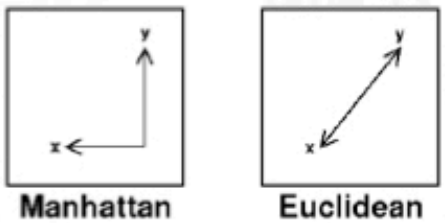
Distance from a Single DataPoint to Centers

- Different distance metrics can be used;
- The most used metrics are:
  - Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i \in m} (x_i - y_i)^2}$$

  - Manhattan distance:

$$d(x, y) = \sum_{i \in m} |(x_i - y_i)|$$

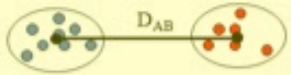
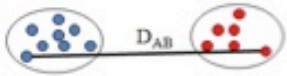


- hereafter we will use Euclidean distance, Manhattan distance works better in case of high dimensional vectors.

Figure 12:

Heat maps and clustering

Agglomerative Linkage Methods

- Linkage methods are rules or metrics that return a value that can be used to determine which elements (clusters) should be linked.
- Three linkage methods that are commonly used are:
  - Single Linkage 
  - Average Linkage 
  - Complete Linkage 

Modified by TMEV presentation (www.tigr.org)

Figure 13:

Heat maps and clustering

Hierarchical Clustering

Different distance functions

Minimum distance between elements of two clusters:

$$D_{min}(C_1, C_2) = \min_{\text{all points } i \text{ and } j \text{ in clusters } C1 \text{ and } C2, \text{ respectively}} D_{i,j}$$

Average distance between elements of two clusters:

$$D_{avg}(C_1, C_2) = \left(\sum_{\text{all points } i \text{ and } j \text{ in clusters } C1 \text{ and } C2, \text{ respectively}} D_{i,j}\right) / \left(|C_1| * |C_2|\right)$$

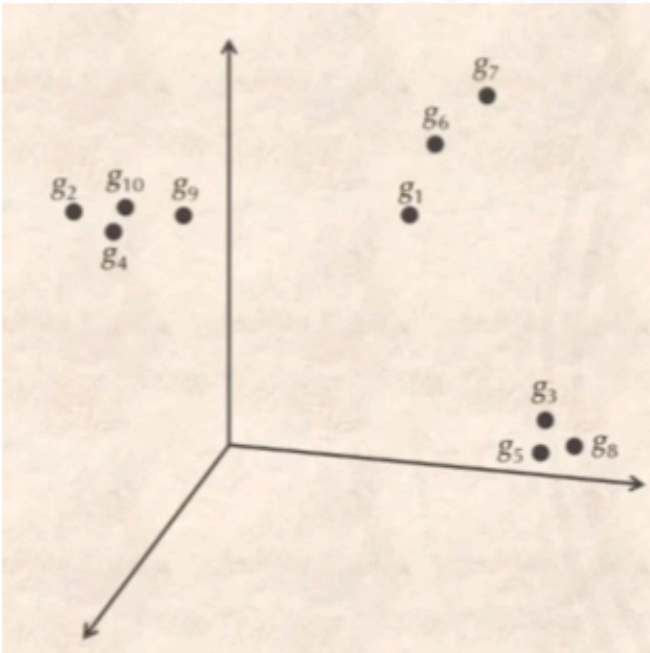
Figure 14:

Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Hierarchical clustering starts from a trasformation of $n \times m$ expression matrix into $n \times n$ **similarity matrix** or **Distance matrix**;
- it can be obtained by simply computing Euclidean/Manhattan distance between genes.



Distance Matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

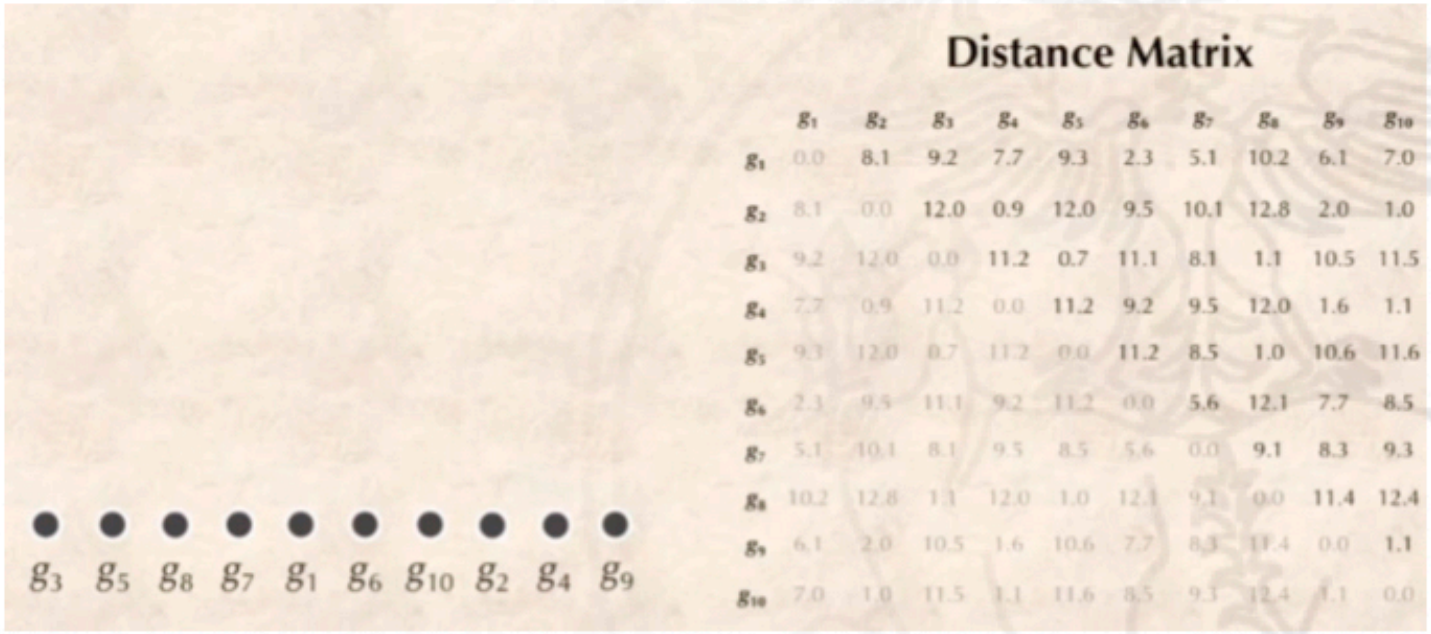
Figure 15:

Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Create a node (i.e. a single element cluster) for every gene.



Distance Matrix

	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

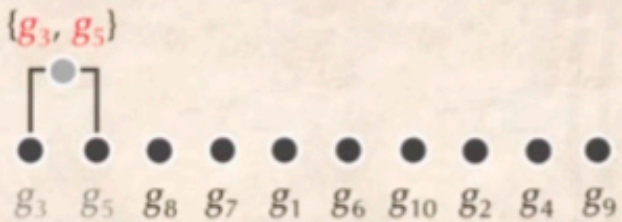
Figure 16:

Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Identify the two ***closest*** clusters and merge them.



	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	9.3	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	12.0	9.5	10.1	12.8	2.0	1.0
g_3	9.2	12.0	0.0	11.2	0.7	11.1	8.1	1.1	10.5	11.5
g_4	7.7	0.9	11.2	0.0	11.2	9.2	9.5	12.0	1.6	1.1
g_5	9.3	12.0	0.7	11.2	0.0	11.2	8.5	1.0	10.6	11.6
g_6	2.3	9.5	11.1	9.2	11.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	8.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.1	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	10.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	11.6	8.5	9.3	12.4	1.1	0.0

Figure 17:

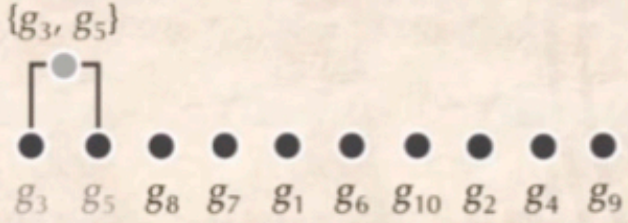
Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Recompute the distance between two clusters as the minimal distance between the elements in the clusters:

$$D(C_1, C_2) = \min_{\forall i \in C_1, j \in C_2} D_{i,j}$$



	g_1	g_2	g_3 g_5	g_4	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	9.5	10.1	12.8	2.0	1.0
g_3 g_5	9.2	12.0	0.0	11.2	11.1	8.1	1.0	10.5	11.5
g_4	7.7	0.9	11.2	0.0	9.2	9.5	12.0	1.6	1.1
g_6	2.3	9.5	11.1	9.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.0	12.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	8.5	9.3	12.4	1.1	0.0

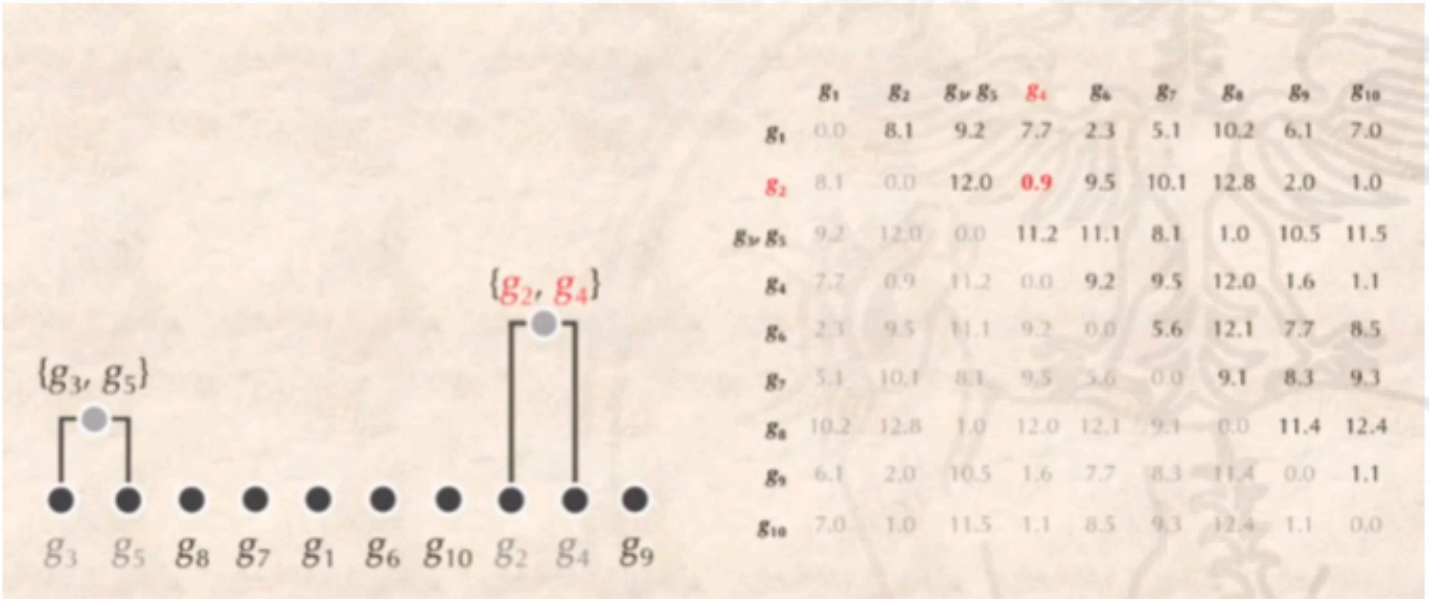
Figure 18:

Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Identify the two ***closest*** clusters and merge them.



	g_1	g_2	g_3 g_5	g_4	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	8.1	9.2	7.7	2.3	5.1	10.2	6.1	7.0
g_2	8.1	0.0	12.0	0.9	9.5	10.1	12.8	2.0	1.0
g_3 g_5	9.2	12.0	0.0	11.2	11.1	8.1	1.0	10.5	11.5
g_4	7.7	0.9	11.2	0.0	9.2	9.5	12.0	1.6	1.1
g_6	2.3	9.5	11.1	9.2	0.0	5.6	12.1	7.7	8.5
g_7	5.1	10.1	8.1	9.5	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.8	1.0	12.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	2.0	10.5	1.6	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	1.1	8.5	9.3	12.4	1.1	0.0

Figure 19:

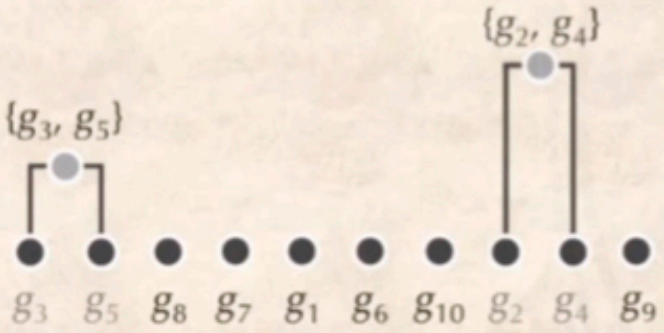
Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Recompute the distance between two clusters as the minimal distance between the elements in the clusters:

$$D(C_1, C_2) = \min_{\forall i \in C_1, j \in C_2} D_{i,j}$$



	g_1	g_2, g_4	g_3, g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	7.7	9.2	2.3	5.1	10.2	6.1	7.0
g_2, g_4	7.7	0.0	11.2	9.2	9.5	12.0	1.6	1.0
g_3, g_5	9.2	11.2	0.0	11.1	8.1	1.0	10.5	11.5
g_6	2.3	9.2	11.1	0.0	5.6	12.1	7.7	8.5
g_7	5.1	9.5	8.1	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	1.6	10.5	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	8.5	9.3	12.4	1.1	0.0

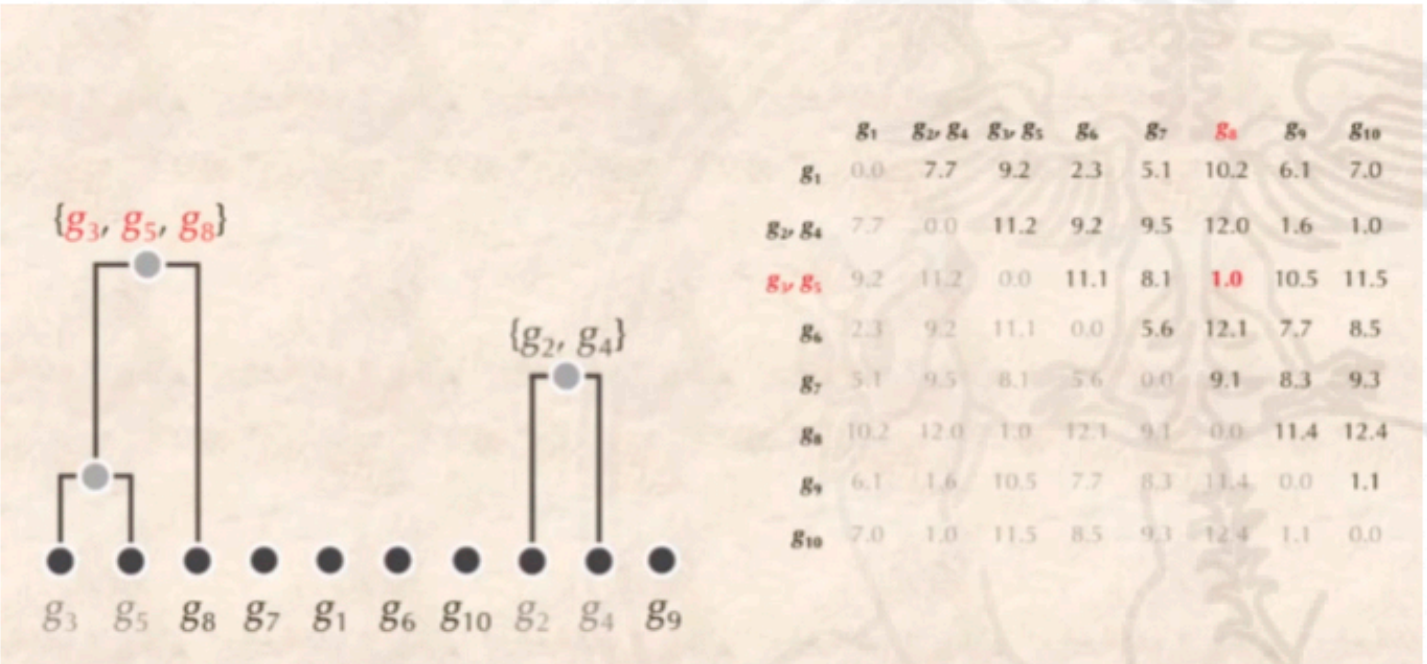
Figure 20:

Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Identify the two ***closest*** clusters and merge them.



	g_1	g_2, g_4	g_3, g_5	g_6	g_7	g_8	g_9	g_{10}
g_1	0.0	7.7	9.2	2.3	5.1	10.2	6.1	7.0
g_2, g_4	7.7	0.0	11.2	9.2	9.5	12.0	1.6	1.0
g_3, g_5	9.2	11.2	0.0	11.1	8.1	1.0	10.5	11.5
g_6	2.3	9.2	11.1	0.0	5.6	12.1	7.7	8.5
g_7	5.1	9.5	8.1	5.6	0.0	9.1	8.3	9.3
g_8	10.2	12.0	1.0	12.1	9.1	0.0	11.4	12.4
g_9	6.1	1.6	10.5	7.7	8.3	11.4	0.0	1.1
g_{10}	7.0	1.0	11.5	8.5	9.3	12.4	1.1	0.0

Figure 21:

Heat maps and clustering

Hierarchical Clustering

Constructing the Tree

- Iterate until all elements form a single cluster(i.e. ***root***).

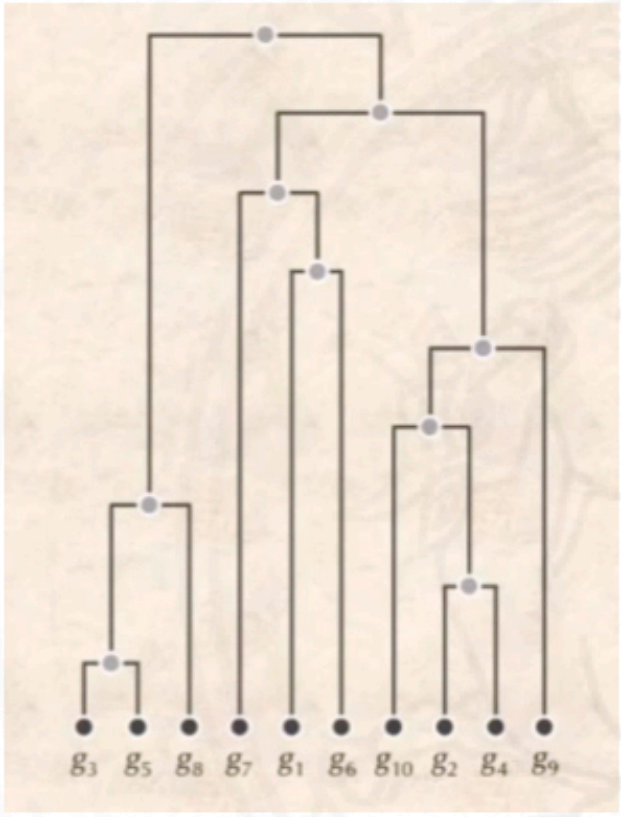


Figure 22:

How to prepare filtered tables for heatmap generation: filtered TPM/FPKM, and mean-centered filtered TPM/FPKM

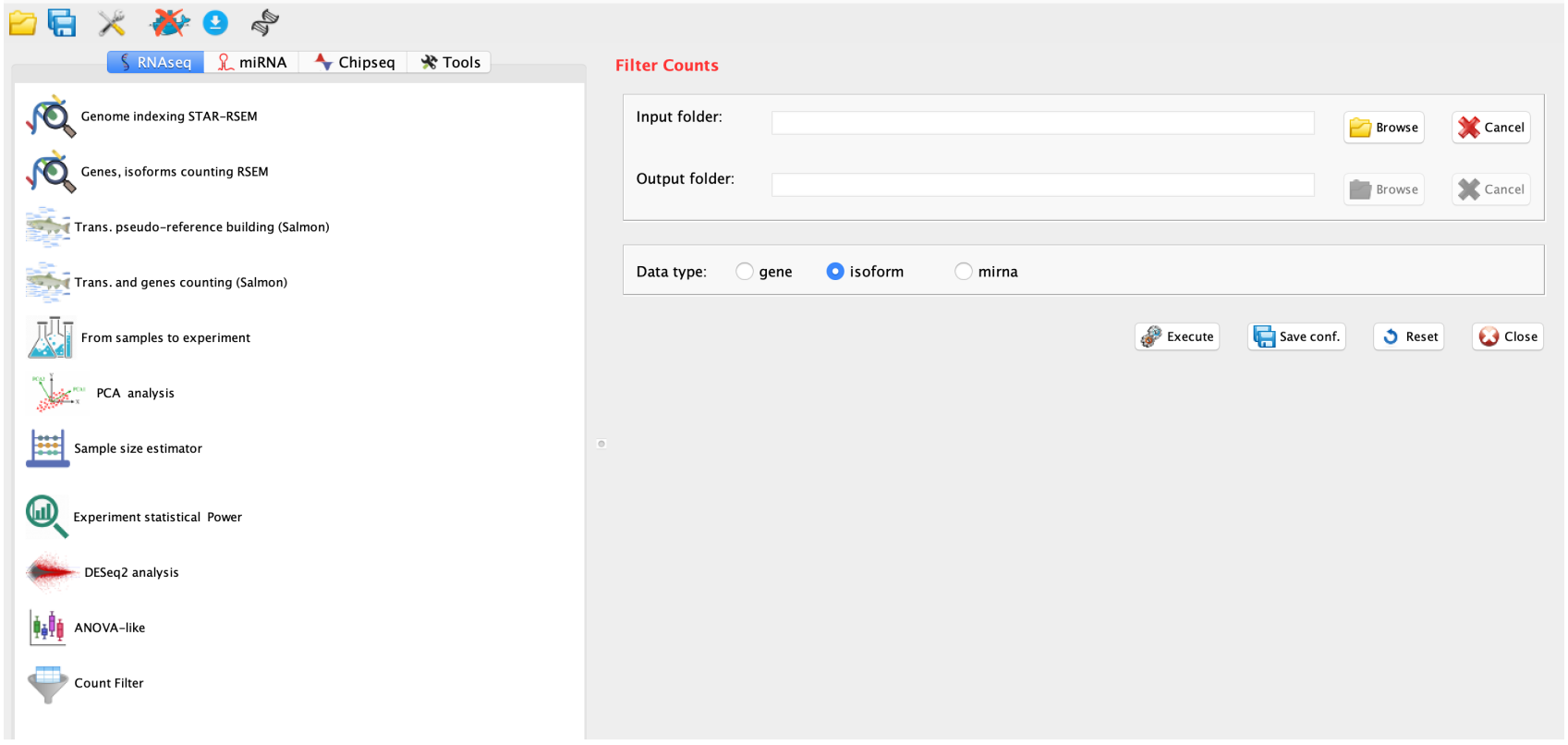


Figure 23:

Counts Filter from DESeq2/ANOVA-like

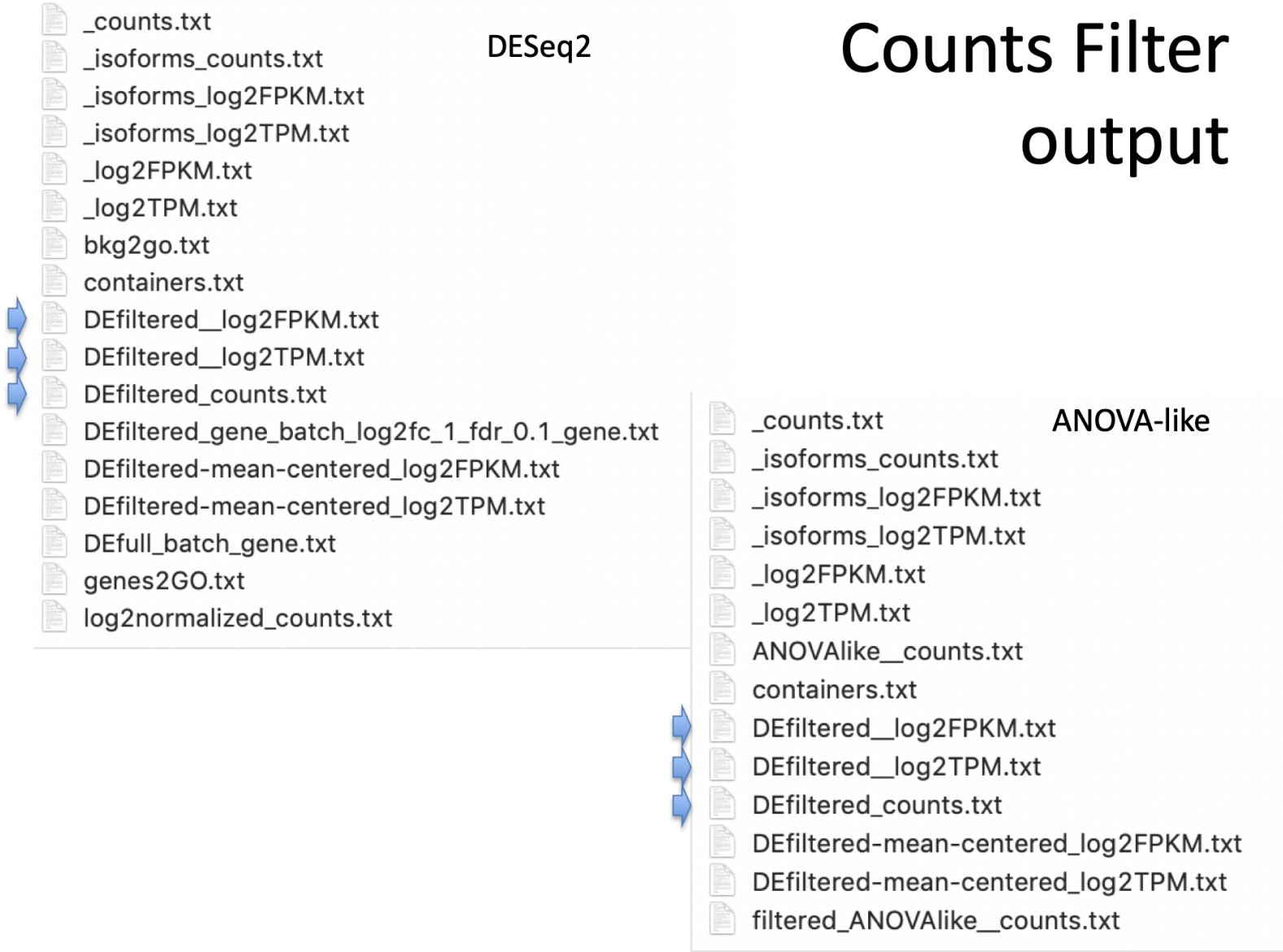


Figure 24:

Heatmaps and clustering

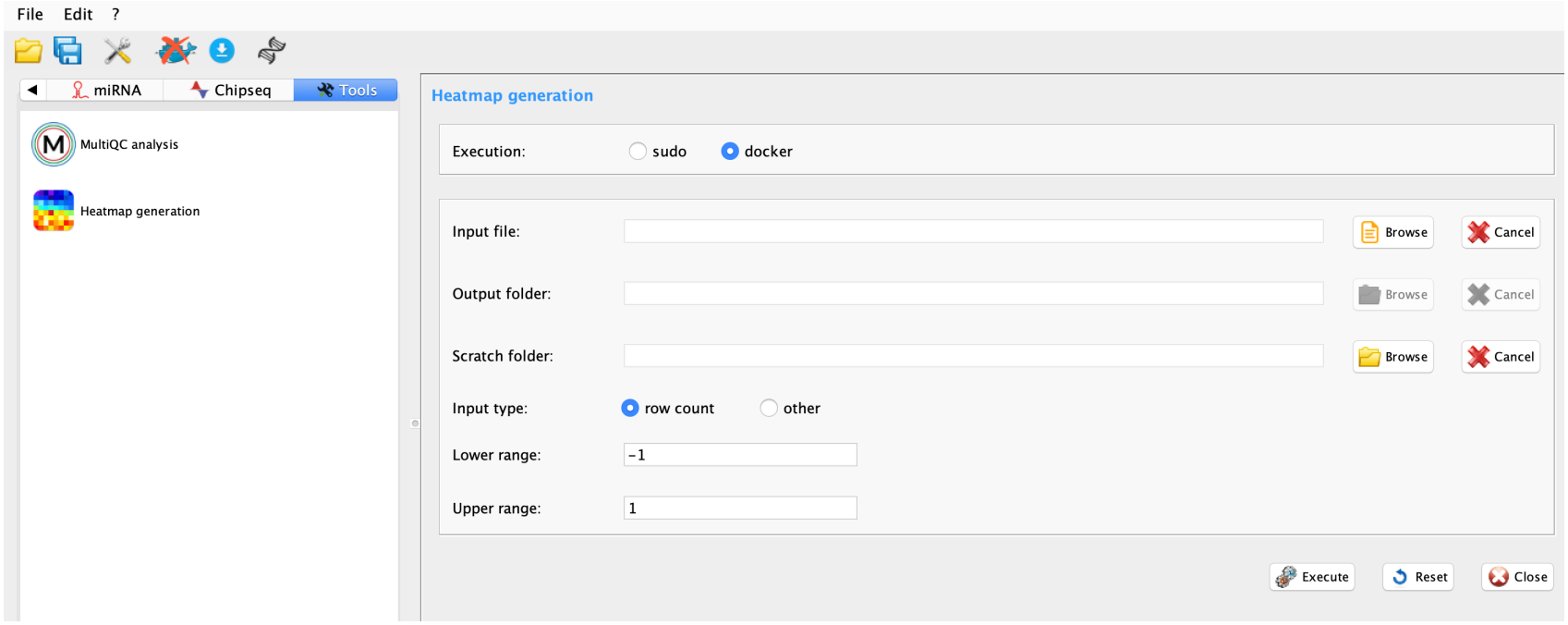


Figure 25:

Heatmaps results mrn4a set interesting genes

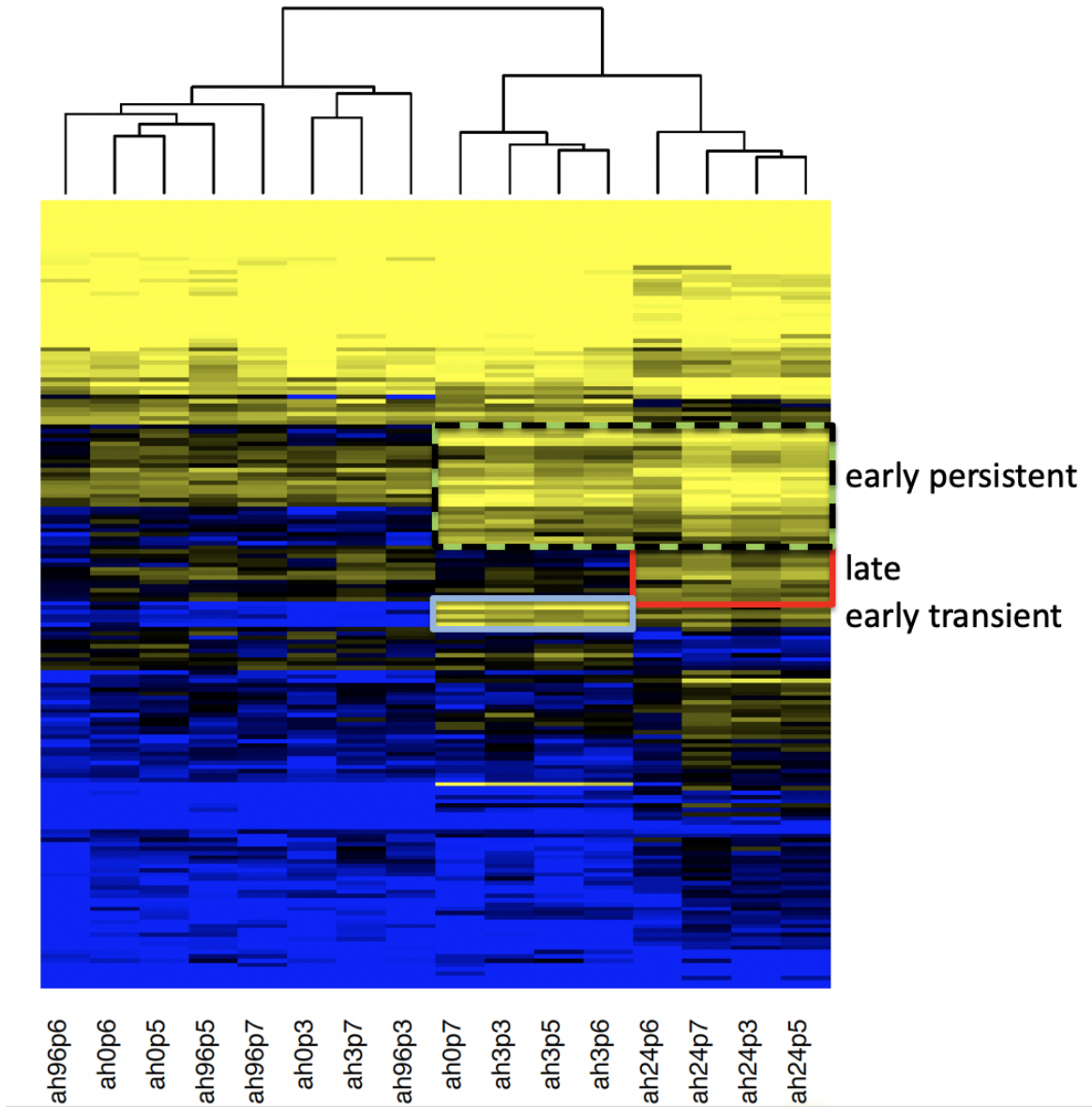


Figure 31: log2 TPM

GO enrichment and GSEA

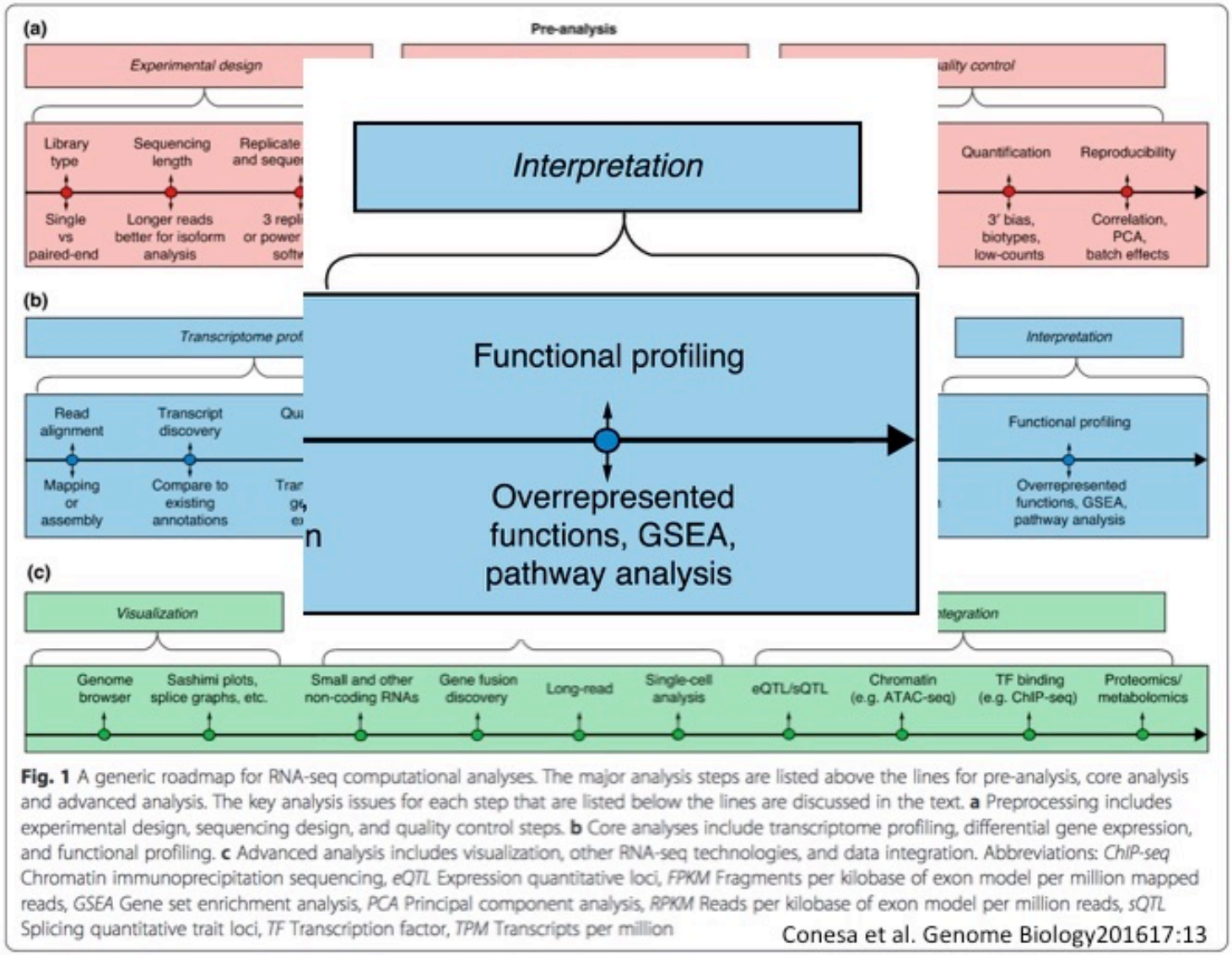


Fig. 1 A generic roadmap for RNA-seq computational analyses. The major analysis steps are listed above the lines for pre-analysis, core analysis and advanced analysis. The key analysis issues for each step that are listed below the lines are discussed in the text. **a** Preprocessing includes experimental design, sequencing design, and quality control steps. **b** Core analyses include transcriptome profiling, differential gene expression, and functional profiling. **c** Advanced analysis includes visualization, other RNA-seq technologies, and data integration. Abbreviations: *ChIP-seq* Chromatin immunoprecipitation sequencing, *eQTL* Expression quantitative loci, *FPKM* Fragments per kilobase of exon model per million mapped reads, *GSEA* Gene set enrichment analysis, *PCA* Principal component analysis, *RPKM* Reads per kilobase of exon model per million reads, *sQTL* Splicing quantitative trait loci, *TF* Transcription factor, *TPM* Transcripts per million

Conesa et al. Genome Biology201617:13

Figure 32: Functional Annotation

Ontologies

- An ontology is a specification of a conceptualization:
  - a hierarchical mapping of concepts within a given frame of reference.
- An ontology is a restricted structured vocabulary of terms that represent domain knowledge.
- An ontology specifies a vocabulary that can be used to exchange queries and assertions.
- A commitment to the use of the ontology is an agreement to use the shared vocabulary in a consistent way.

The Gene Ontology

- The goal of the Gene Ontology (GO) Consortium is to produce a controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.
  - http://www.geneontology.org/
- For genes and gene products the Gene Ontology Consortium (GO) is an initiative that is designed to address the problem of defining common set of terms and descriptions for basic biological functions.
- GO provides a restricted vocabulary as well as clear indications of the relationships between terms.

The Gene Ontology

- The Gene Ontology (GO) consortium produces three independent ontologies for gene products.
- The three ontologies are:
  - molecular function of a gene product which is defined to be biochemical activity or action of the gene product.
  - biological process interpreted as a biological objective to which the gene product contributes.
  - cellular component is a component of a cell that is part of some larger object or structure.

GO structure

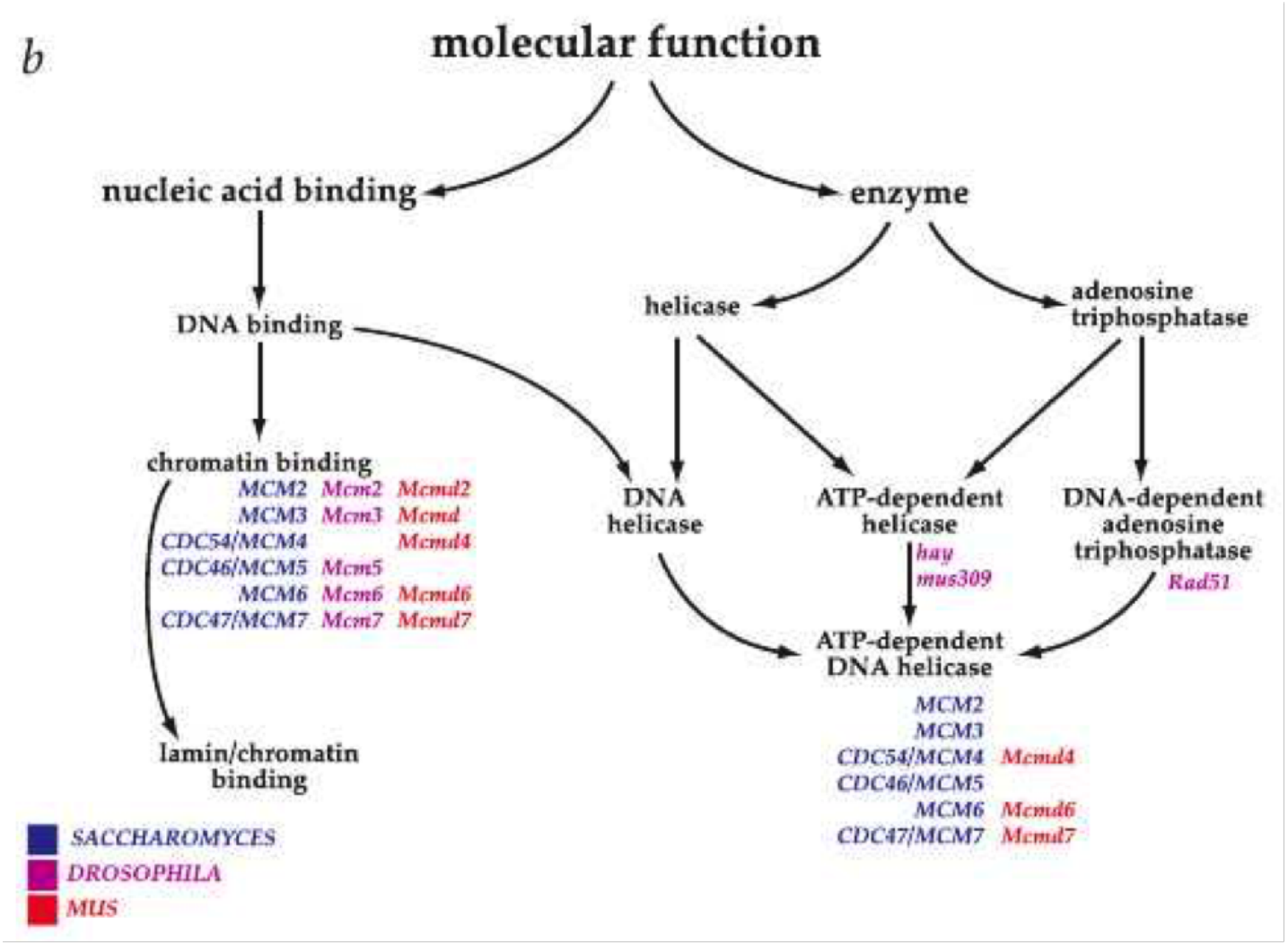


Figure 33:

GO criticalities

GOID	EVIDENCE	ONTOLOGY	ENTREZID	SYMBOL	GENENAME
GO:0030154	IEA	BP	13642	Efnb2	ephrin B2
GO:0030154	IEA	BP	14175	Fgf4	fibroblast growth factor 4
GO:0030154	IEA	BP	14367	Fzd5	frizzled homolog 5 (Drosophila)
GO:0030154	IEA	BP	15482	Hspa1l	heat shock protein 1-like
GO:0030154	IEA	BP	16413	Itgb1bp1	integrin beta 1 binding protein 1
GO:0030154	IMP	BP	16600	Klf4	Kruppel-like factor 4 (gut)
GO:0030154	IMP	BP	16923	Sh2b3	SH2B adaptor protein 3
GO:0030154	IEA	BP	17242	Mdk	midkine
GO:0030154	IEA	BP	17450	Morc1	microrchidia 1

- The *Author Statement evidence codes* used by GO are:
  - Traceable Author Statement (TAS)
  - Non-traceable Author Statement (NAS)
- The *Curatorial Statement codes* are:
  - Inferred by Curator (IC)
  - No biological Data available (ND) evidence code
- The *Automatically-Assigned evidence code* is:
  - Inferred from Electronic Annotation (IEA)

Figure 34:

GO analysis

Enrichment analysis

We consider a total population of genes, e.g. the genes expressed in a high-throughput experiment, and we are interested in the property of a **gene to belong to a specific GO category**. The aim is to establish whether the class of the DE genes presents an **enrichment and/or a depletion of the GO category of interest with respect to the total gene population.**

The **null hypothesis** that the property for a gene to belong to the GO category of interest and that to be DE are **independent**, or **equivalently that the DE genes are picked at random from the total gene population**

Figure 35:

GO analysis

Hypergeometric distribution and Fisher's test

The hypergeometric distribution is a discrete probability distribution that describes the probability of *k* successes (random draws for which the object drawn has a specified feature) in *n* draws, **without replacement**, from a **finite population** of size *N* that **contains exactly *K* objects with that feature,** wherein each **draw is either a success or a failure.**

Fisher's exact test to determine if something is enriched or not.

Figure 36:

GO analysis

Hypergeometric distribution and Fisher's test

Bag of balls

I extract 7 blue balls and 1 red

What does that say about the distributions of colours in the bag?

Do I have more blues than normal?

Can I calculate a p-value from this sample?

Figure 37:

GO analysis

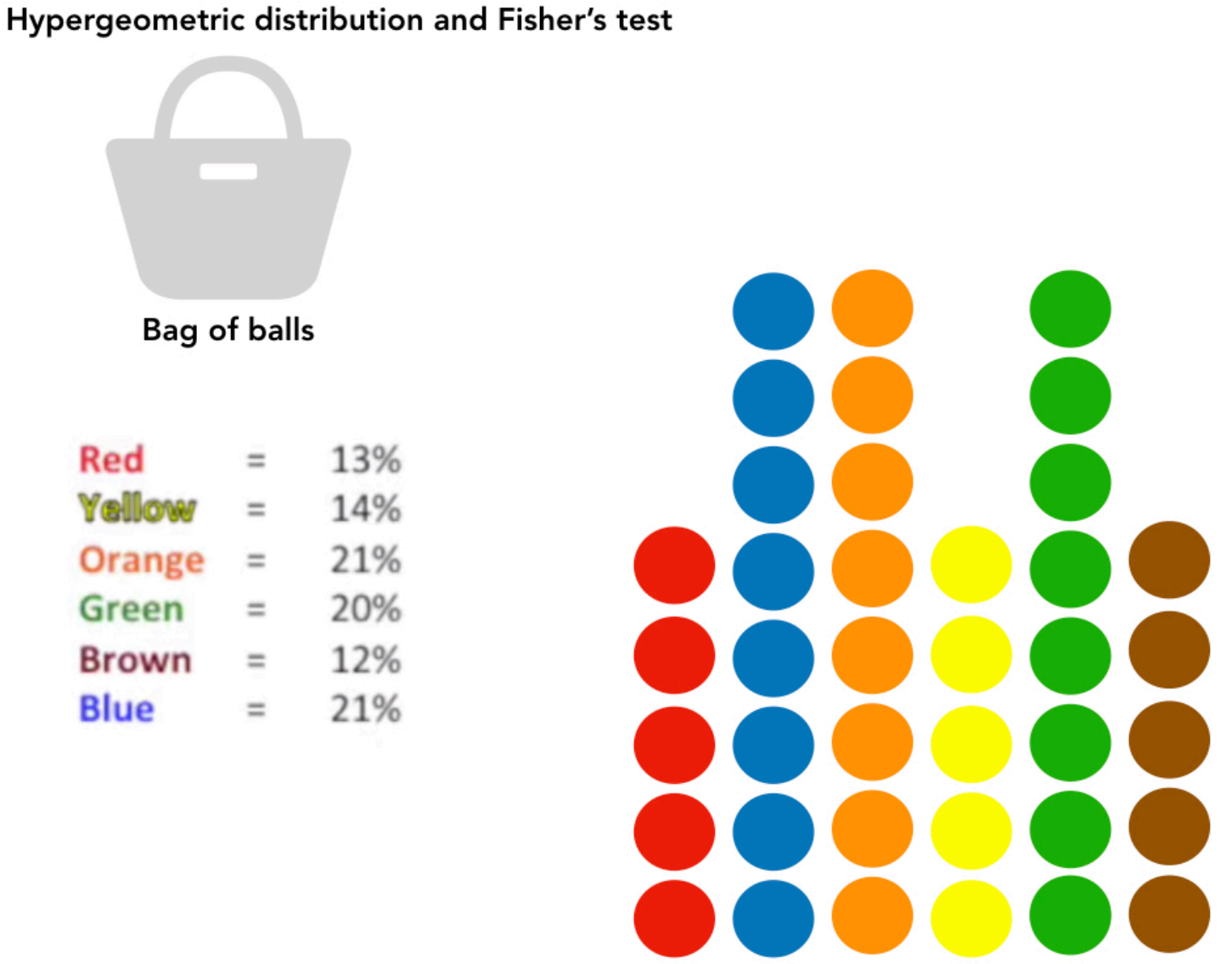
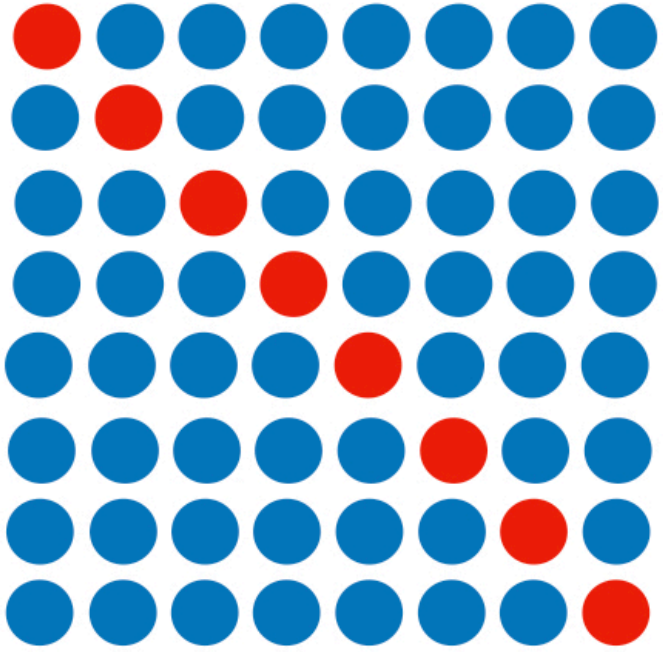


Figure 38:

GO analysis

Hypergeometric distribution and Fisher's test

Determine if the set of balls of this sample is special or not?



The order of how the balls are extracted is not important, then consider all possible ordering of the 7 blue and 1 red as legit

Figure 39:

GO analysis

Hypergeometric distribution and Fisher's test

Let's start by calculating the probability of getting 7 blues balls followed by a single red

The probability that the first ball blue is 8/40 ,

Where:

8 because there are 8 blues

40 is the total number of balls

The probability that the second ball blue is 7/39 ,

Where:

7 because there are 8 blues

39 is the total number of balls

Figure 40:

GO analysis

Hypergeometric distribution and Fisher's test

Let's start by calculating the probability of getting 7 blues balls followed by a single red

The probability that the first ball blue is 8/40 ,

Where:

8 because there are 8 blues

40 is the total number of balls

The probability that the first ball blue is 7/39 ,

Where:

7 because there are 7 blues

39 is the total number of balls

The probability that the first ball red is 5/33 ,

Where:

5 because there are 5 reds

33 is the total number of balls

Figure 41:

GO analysis

Hypergeometric distribution and Fisher's test

Let's start by calculating the probability of getting 7 blues balls followed by a single red

Multiply all those probabilities together to get the probability of getting 7 blues followed by one red is 0.000000065

The probability to obtain 7 blues and 1 red not depend by the order then, to calculate the probability of getting 7 blues and 1 red we need to consider all the probabilities of each possible ordering.

We repeat the computation of the probability considering any order and we obtain: 0.00000053

Figure 42:

GO analysis

Hypergeometric distribution and Fisher's test

We repeat the computation of the probability considering any order and we obtain:
0.00000053

Compute the p-value

Figure 43:

GO analysis

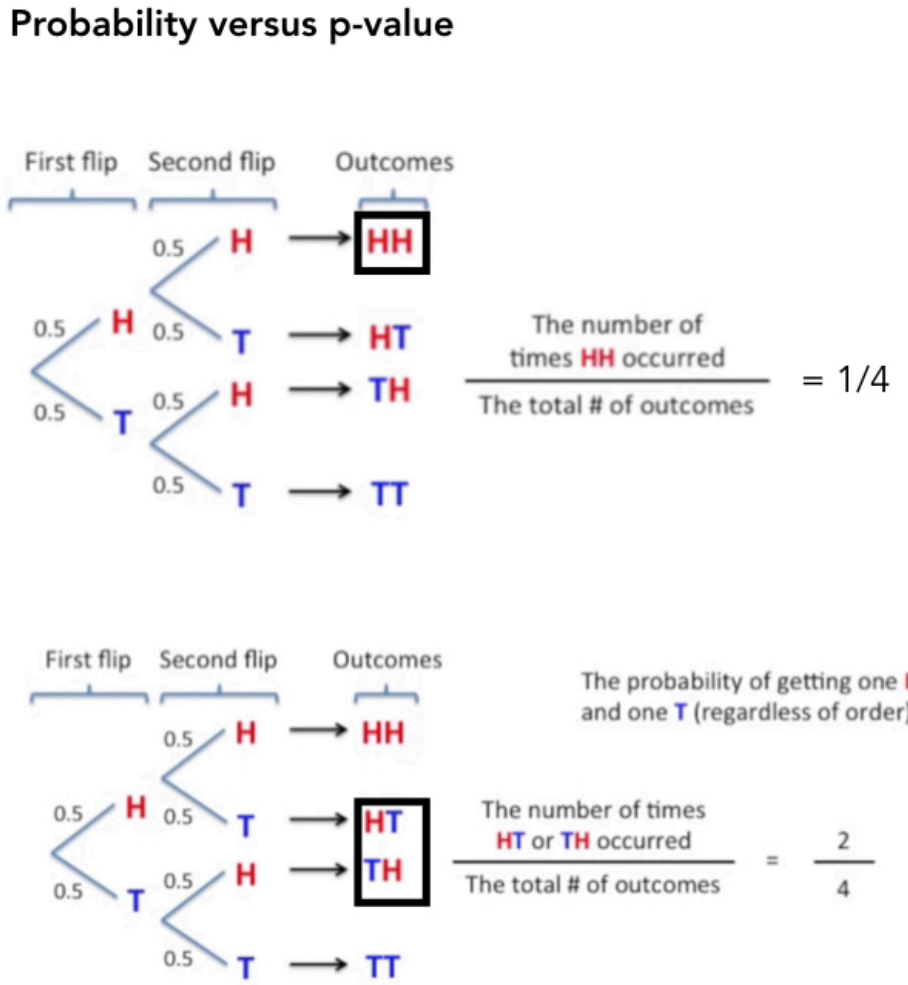


Figure 44:

GO analysis

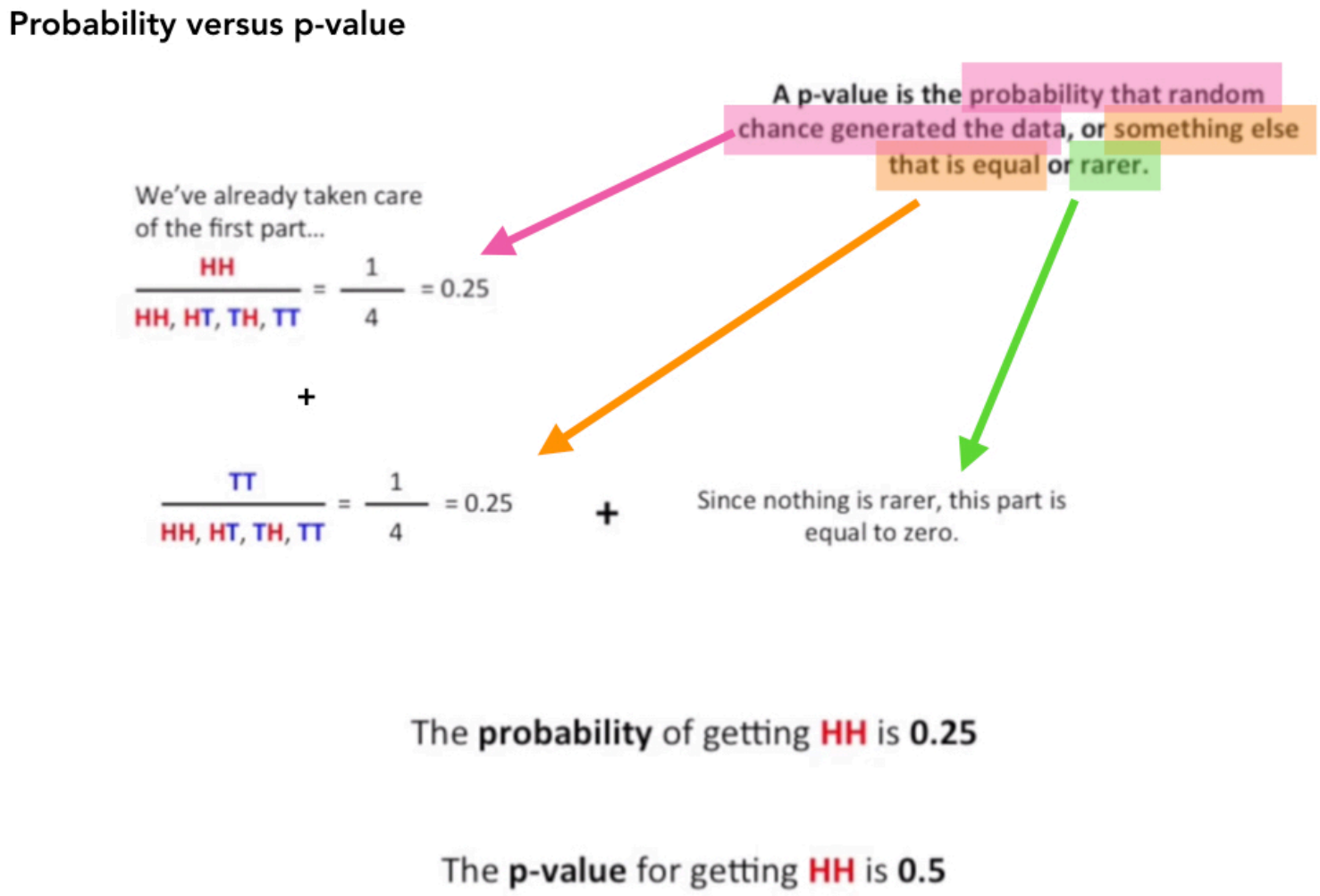


Figure 45:

GO analysis

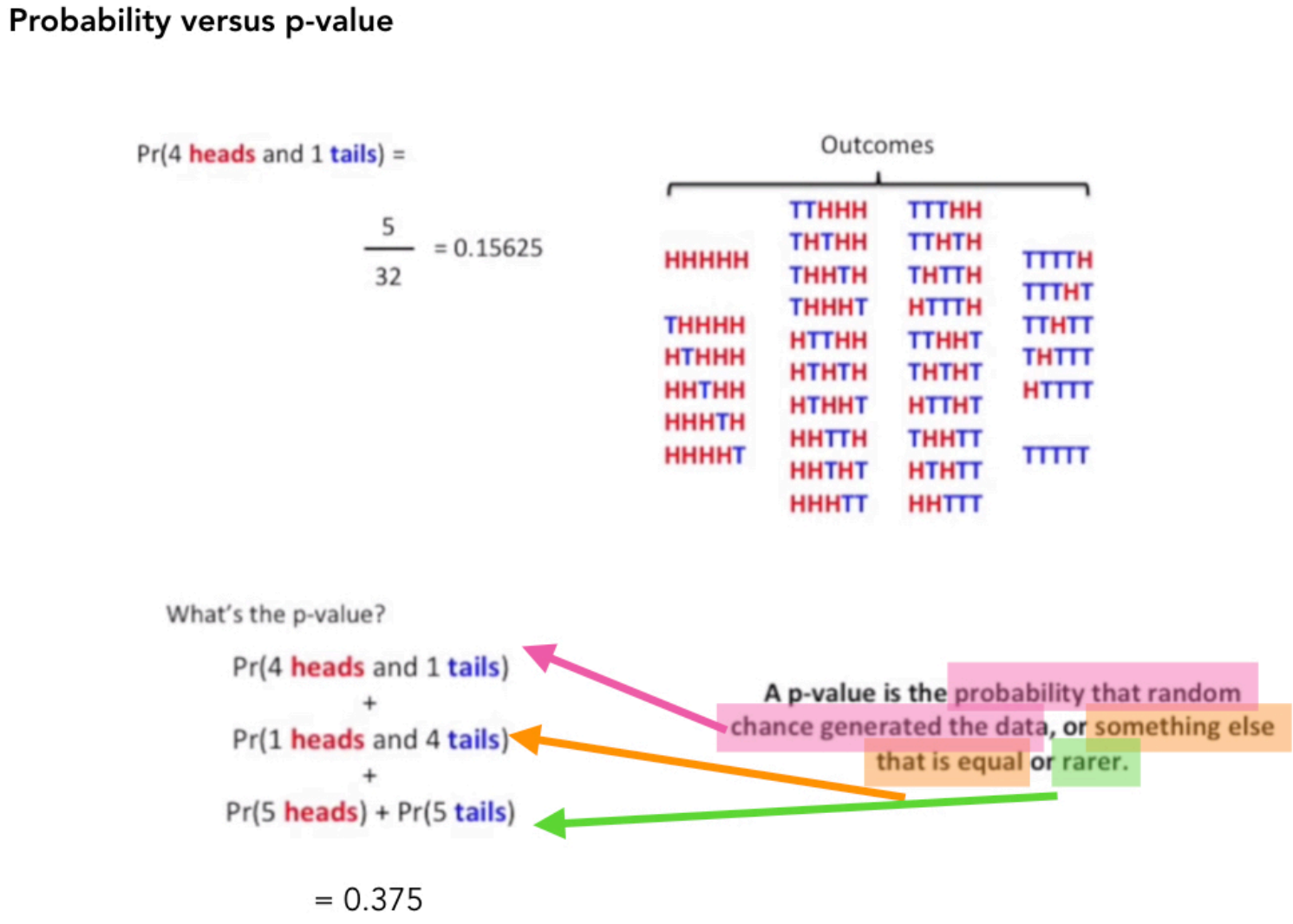


Figure 46:

GO analysis

Hypergeometric distribution and Fisher's test

We repeat the computation of the probability considering any order and we obtain:
0.00000053

The p-value is the sum of the probabilities of all things equally rare or rarer. Then compete the probability for 7 blues and 1 orange, 8 blues (as the rarer) etc.

Finally the p-values is 0.01.

This is call Fisher's exact test.

Enrichment for other things, "does this list of genes have more involved in metabolism than normal" can be answered following the same way.

Figure 47:

GO analysis

Consider a population of genes representing a diverse set of GO terms shown below as different colors.

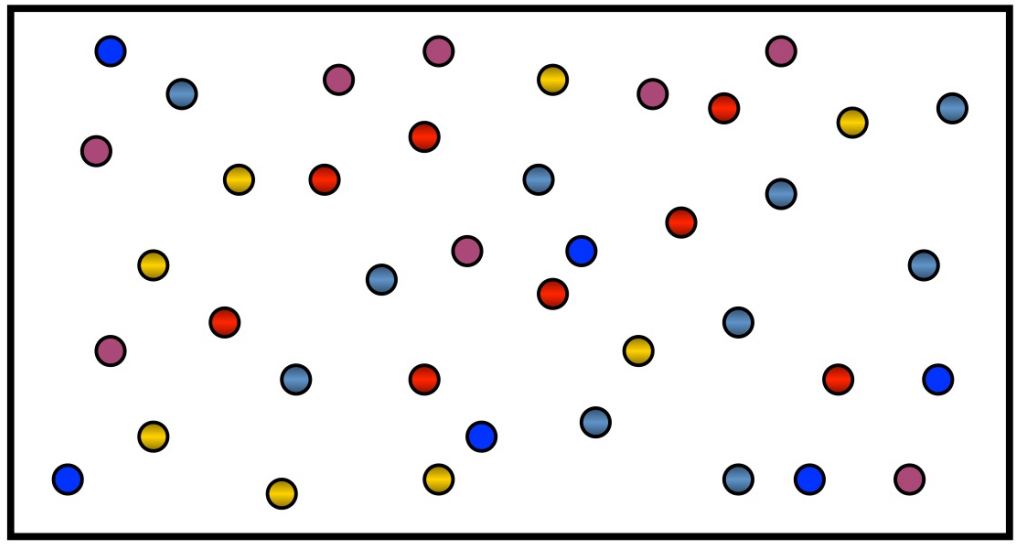


Figure 48:

GO analysis

Many methods can be used to identify a set of differentially expressed genes

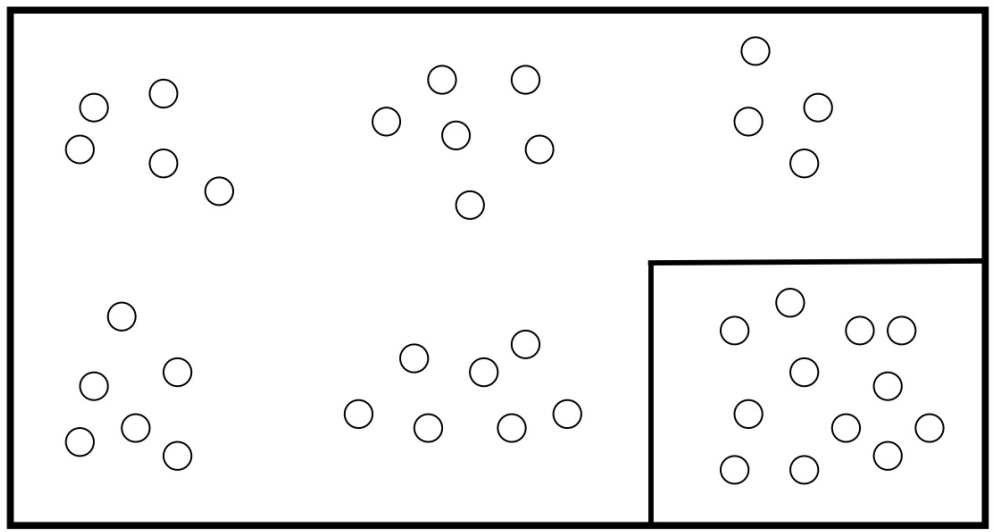


Figure 49:

GO analysis

What are the some of the predominant GO terms represented in the set of differentially expressed genes and how should significance be assigned to a discovered GO term?

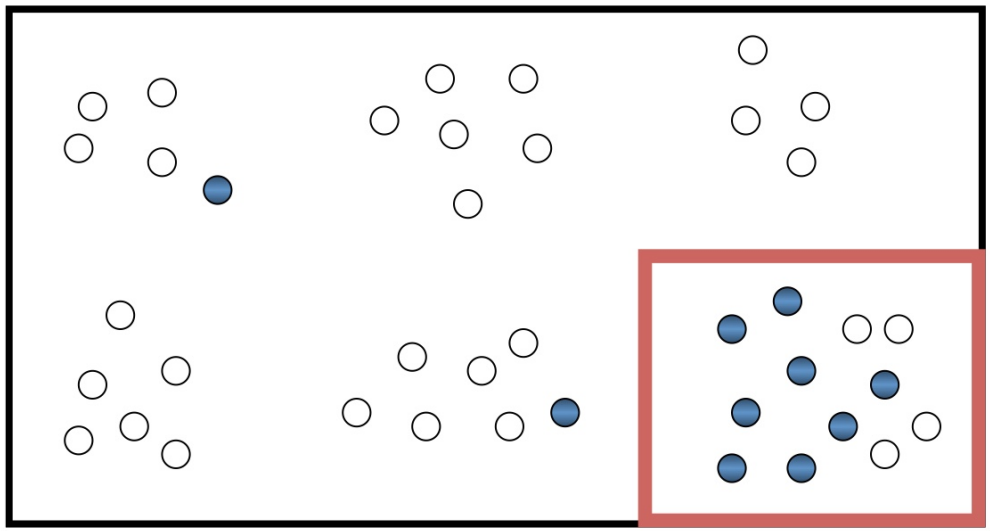


Figure 50:

GO analysis

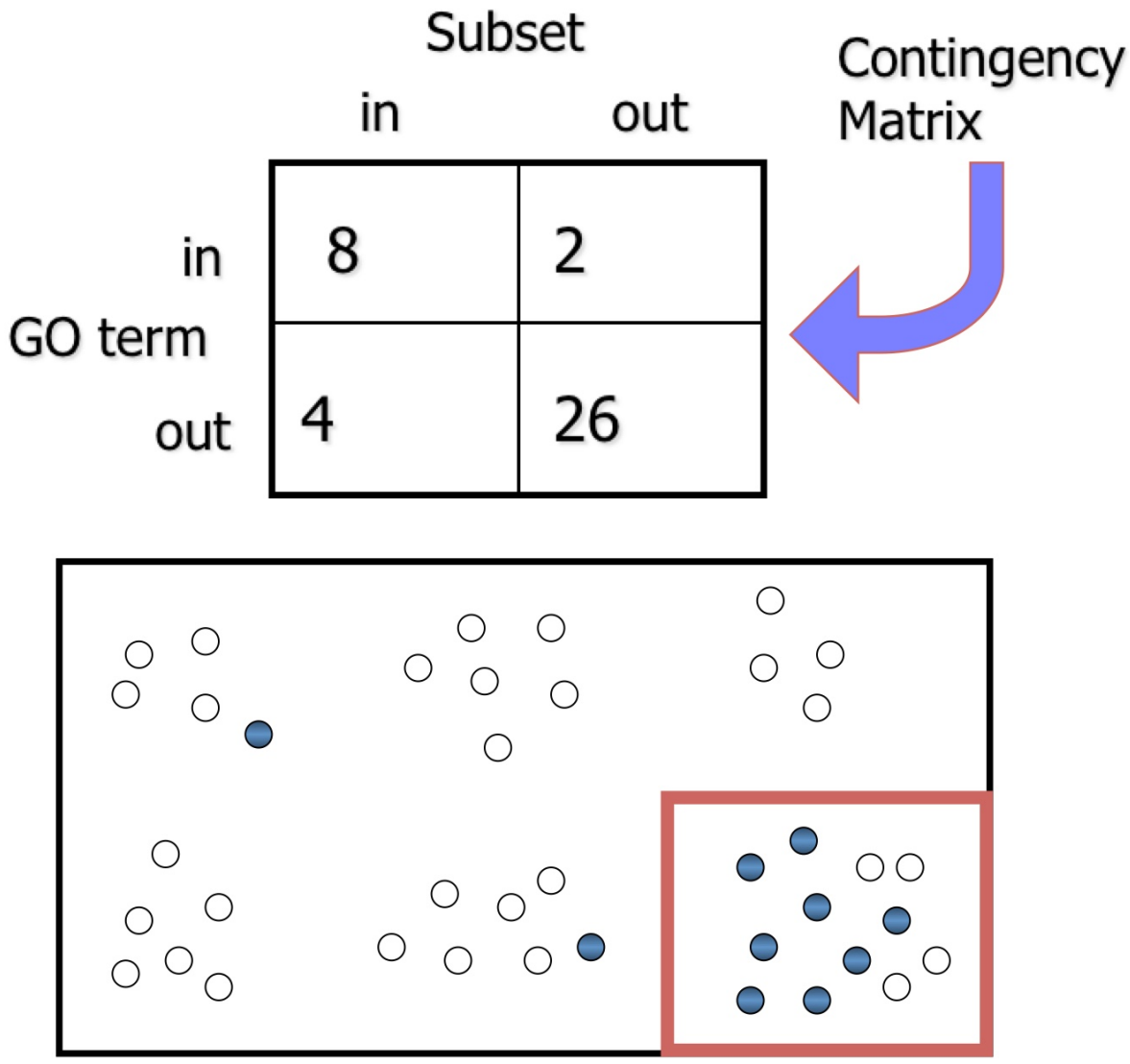


Figure 51:

GO analysis

a	b	a+b
c	d	c+d
a+c	b+d	

The probability of any **particular** matrix occurring by random selection, given no association between the two variables, is given by the **hypergeometric rule**.

$$\frac{\frac{(a+c)!}{a!c!}\times\frac{(b+d)!}{b!d!}}{\frac{n!}{(a+b)!(c+d)!}}=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

Figure 52:

GO analysis

The **HyperGeometric Test** permits us to determine if there are non-random associations between the two variables, differential expression membership and membership to a particular Gene Ontology term.

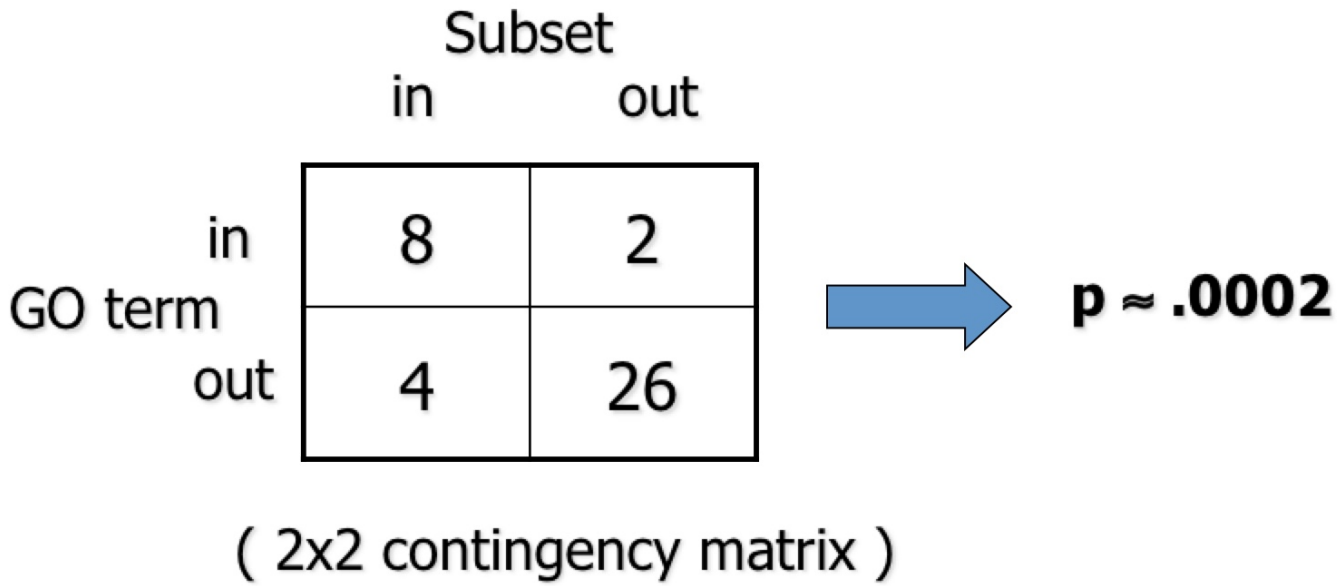


Figure 53:

Gene Ontology tools

- DAVID (https://david.ncifcrf.gov/)
- GORILLA (http://cbl-gorilla.cs.technion.ac.il/)