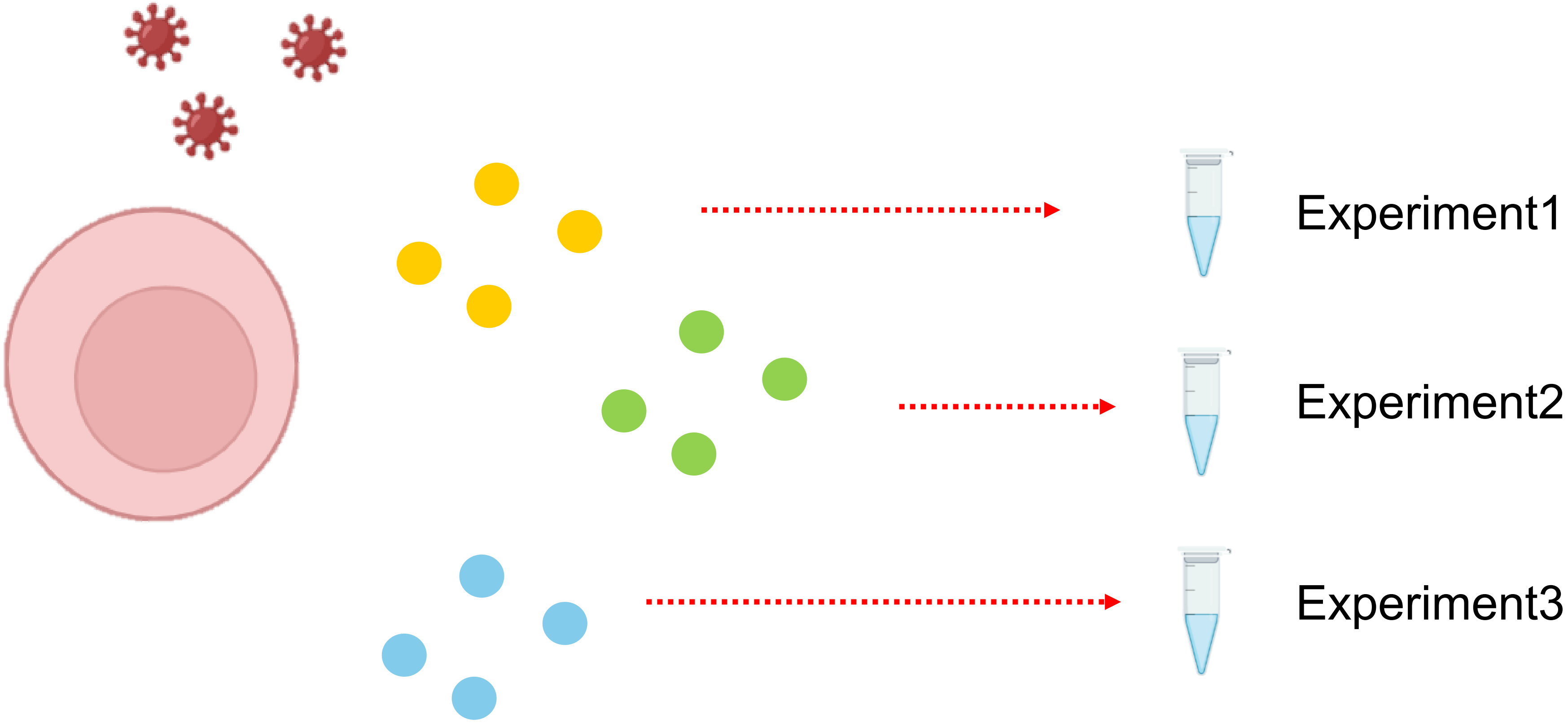
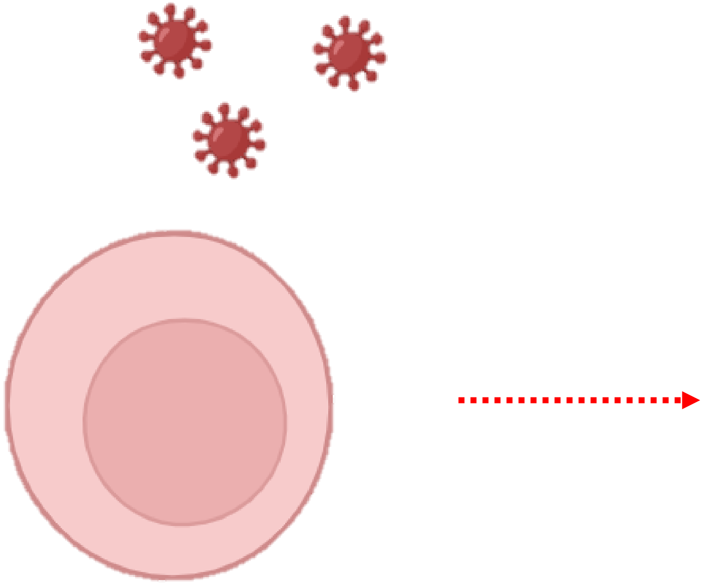
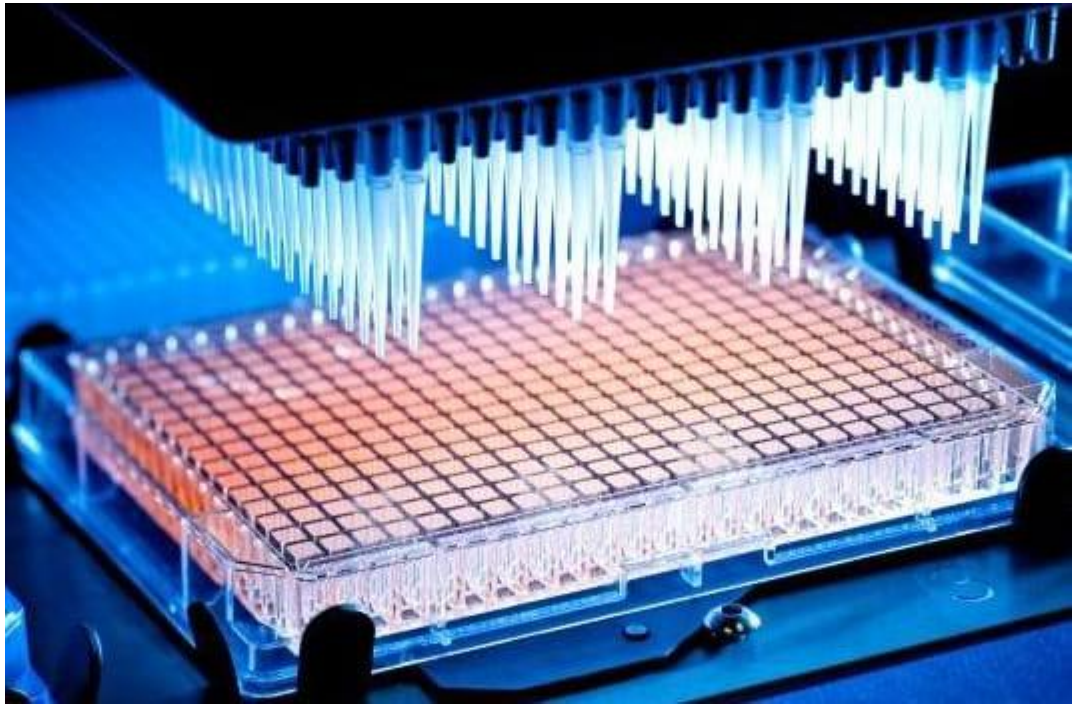
Bulk RNA-sequencing

- Data acquisition

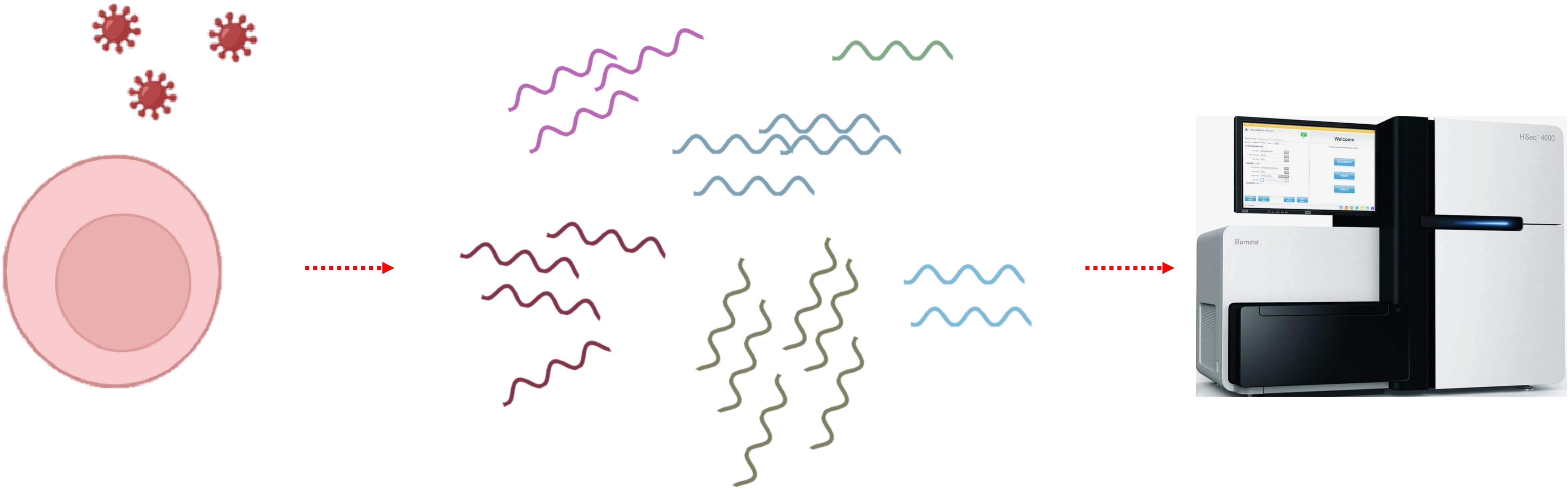


- High throughput data





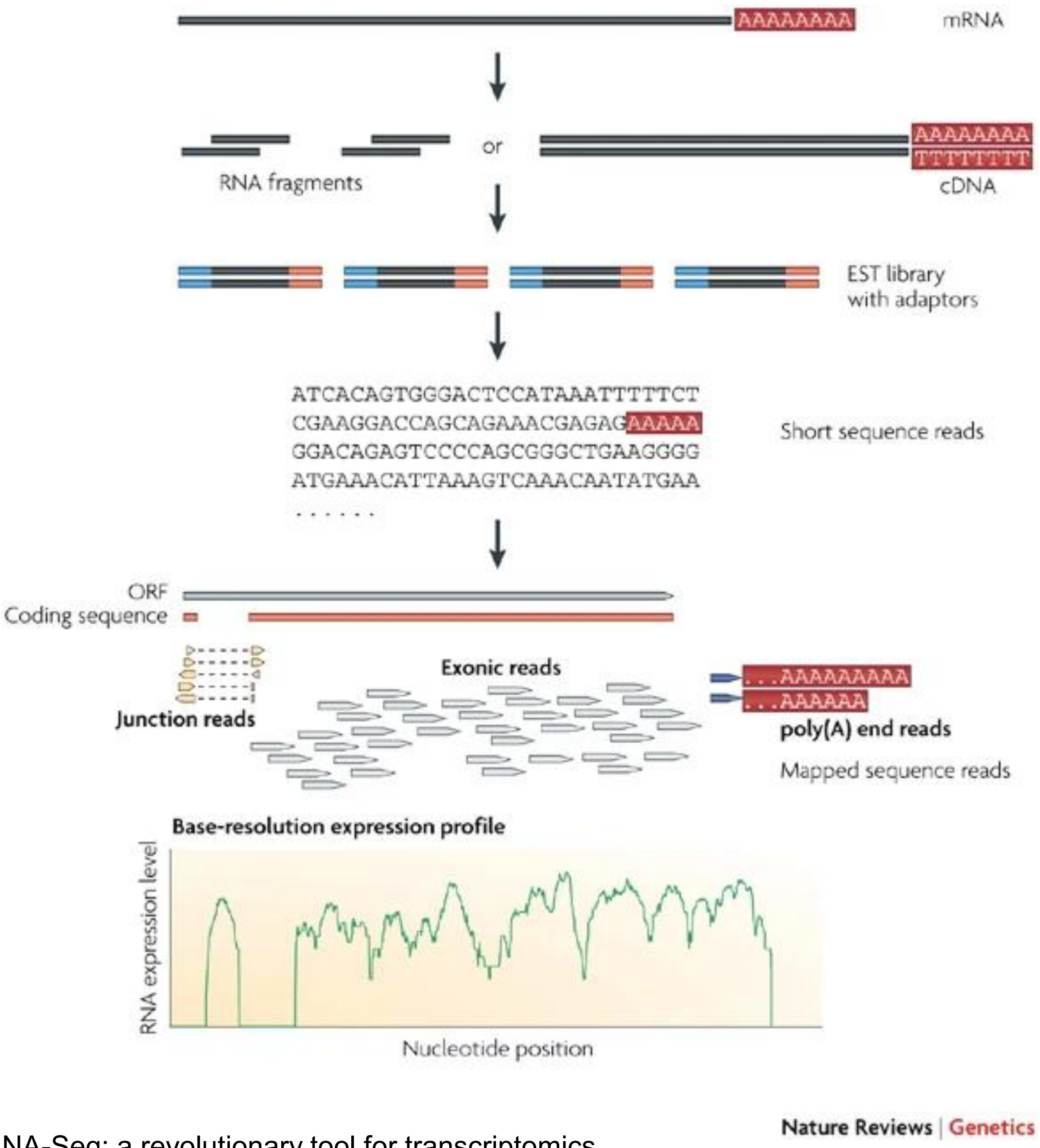
- High throughput data



Whole transcriptome

RNA-sequencing

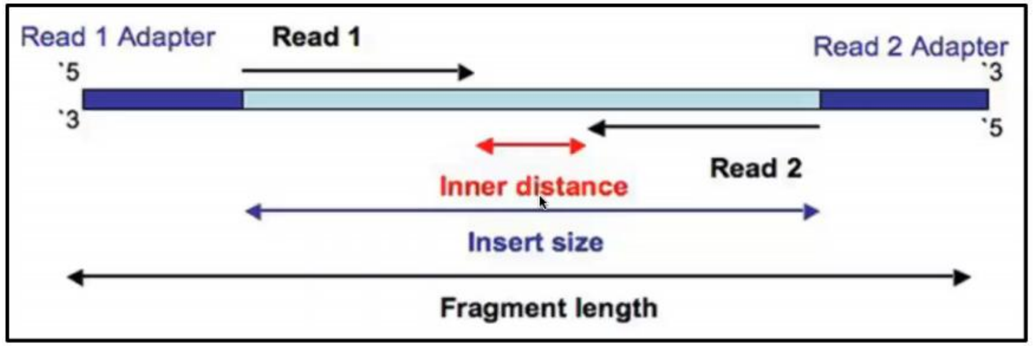
- RNA-sequencing



RNA-Seq: a revolutionary tool for transcriptomics

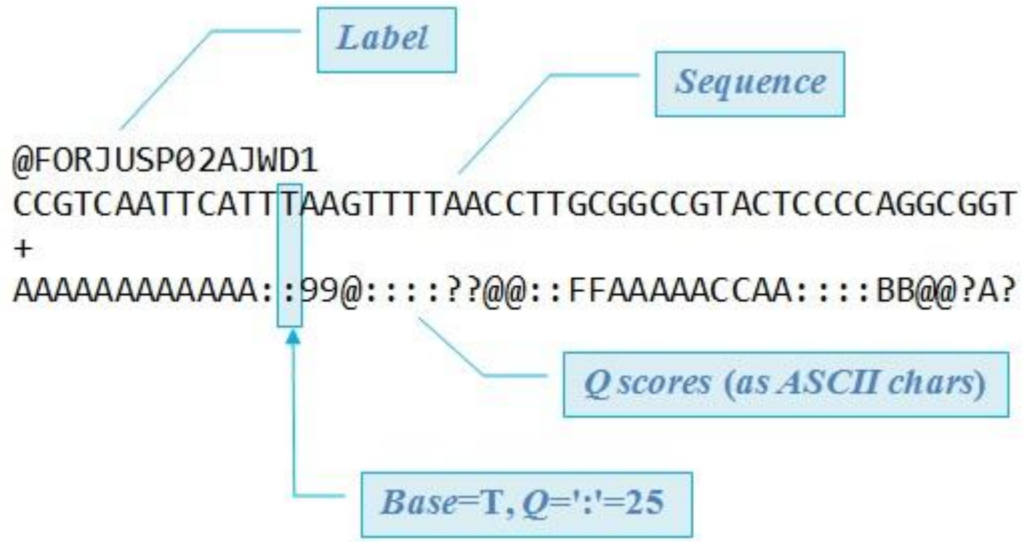
- Raw mRNA
- RNA fragment chopping
- Reverse transcriptase
- Adaptors → Obtain the sequence

Read (single-end) (100 ~ 200 bp)
Or
Fragment (paired-end): high confidence

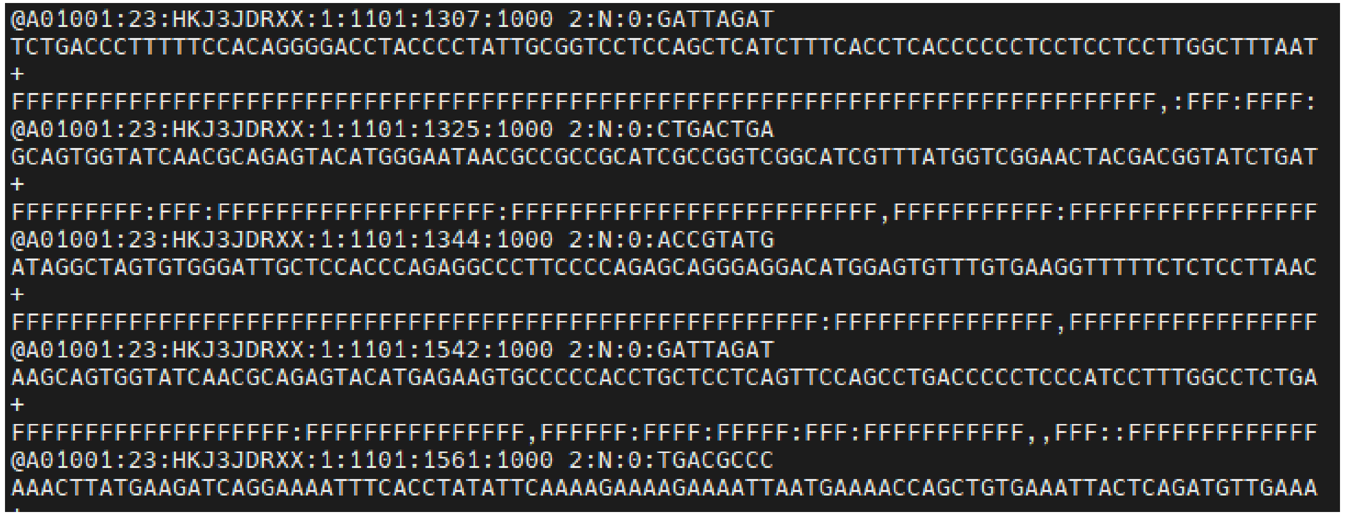


- # RNA-sequencing

Data format: fastq file



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

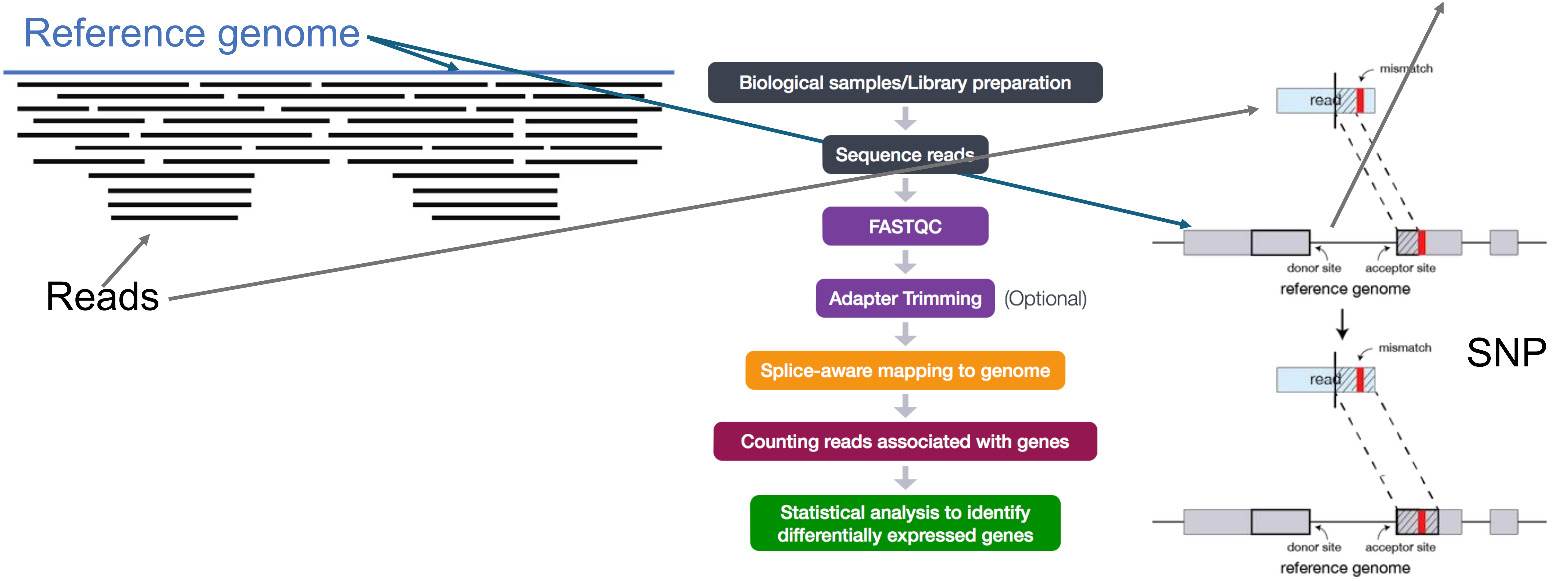


```
@A01001:23:HKJ3JDRXX:1:1101:1307:1000 2:N:0:GATTAGAT
TCTGACCCTTTTTCCACAGGGGACCTACCCCTATTGCGGTCCTCCAGCTCATCTTTCACCTCACCCCCCTCCTCCTCCTTGGCTTTAAT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,:FFF:FFFF:
@A01001:23:HKJ3JDRXX:1:1101:1325:1000 2:N:0:CTGACTGA
GCAGTGGTATCAACGCAGAGTACATGGGAATAACGCCGCCGCATCGCCGGTCGGCATCGTTTATGGTCGGAACTACGACGGTATCTGAT
+
FFFFFFFFF:FFF:FFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFF:FFFFFFFFFFFFFFFFF
@A01001:23:HKJ3JDRXX:1:1101:1344:1000 2:N:0:ACCGTATG
ATAGGCTAGTGTGGGATTGCTCCACCCAGAGGCCCTTCCCCAGAGCAGGGAGGACATGGAGTGTTTGTGAAGGTTTTTCTCTCCTTAAC
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFF
@A01001:23:HKJ3JDRXX:1:1101:1542:1000 2:N:0:GATTAGAT
AAGCAGTGGTATCAACGCAGAGTACATGAGAAGTGCCCCCACCTGCTCCTCAGTTCCAGCCTGACCCCCTCCCATCCTTTGGCCTCTGA
+
FFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF,FFFFFF:FFFF:FFFFF:FFF:FFFFFFFFFFF,,FFF::FFFFFFFFFFFFF
@A01001:23:HKJ3JDRXX:1:1101:1561:1000 2:N:0:TGACGCCC
AAACTTATGAAGATCAGGAAAATTTCACCTATATTCAAAAGAAAAGAAAATTAATGAAAACCAGCTGTGAAATTACTCAGATGTTGAAA
```

- RNA-sequencing (Alignment)

Fastq file → somehow **gene by count** matrix (it could also be a transcript or isoform)

1) Alignment (ex: **STAR** → Genome, Kallisto → Transcriptome)
- Each reads → where in the genome



Biological samples/Library preparation

Sequence reads

FASTQC

Adapter Trimming (Optional)

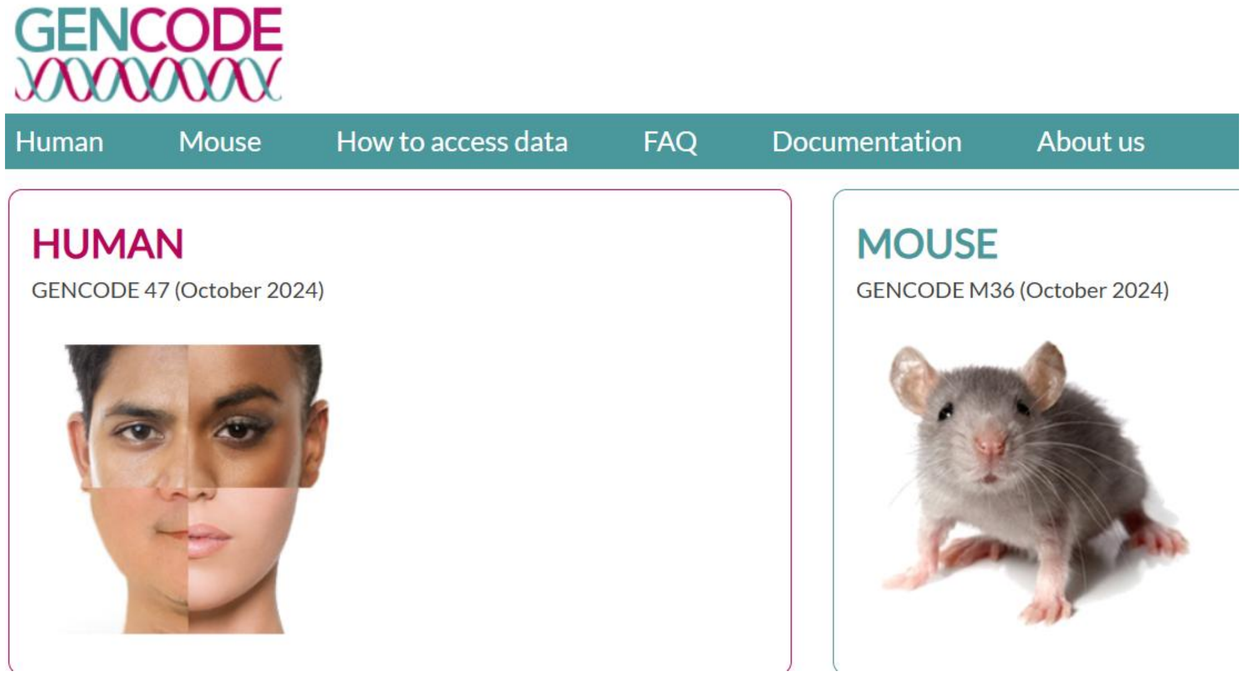
Splice-aware mapping to genome

Counting reads associated with genes

Statistical analysis to identify differentially expressed genes

- RNA-sequencing (Alignment)

- Genome: different genome assembly version
Human: GRCh37.## (hg19), GRCh38.## (hg38)
Mouse: GRCm38, GRCm39
Different versions cannot be used together
→ Different nucleotide locus
→ Same genome build but different version (or release): compatible with each other (mostly gap-filling)

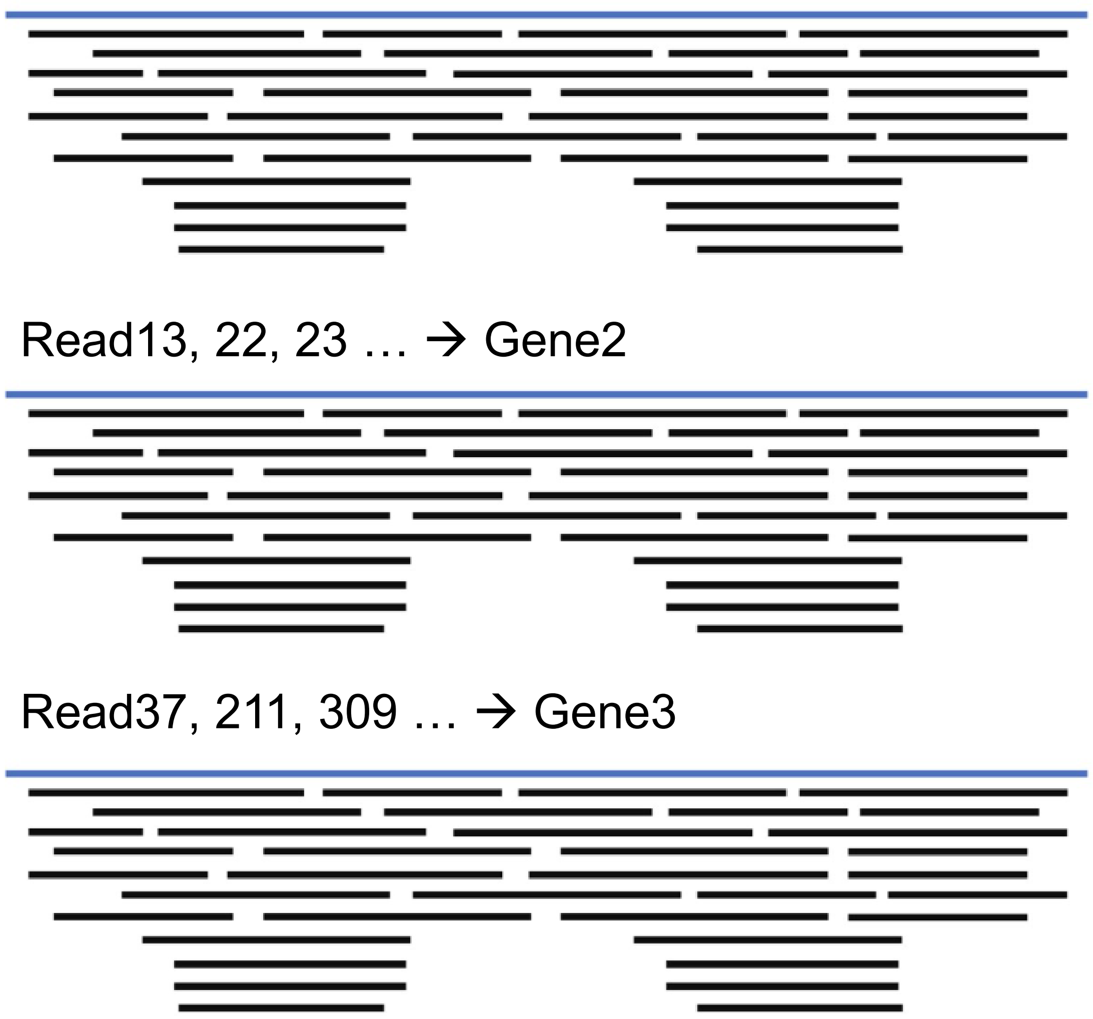


HUMAN
GENCODE 47 (October 2024)

MOUSE
GENCODE M36 (October 2024)

- Human genome size: 3.1 Gbase pairs, Mouse genome size: 2.7Gbp → mapping is not trivial
Genome building: Making a dictionary for boosting the mapping time

- RNA-sequencing (Alignment)

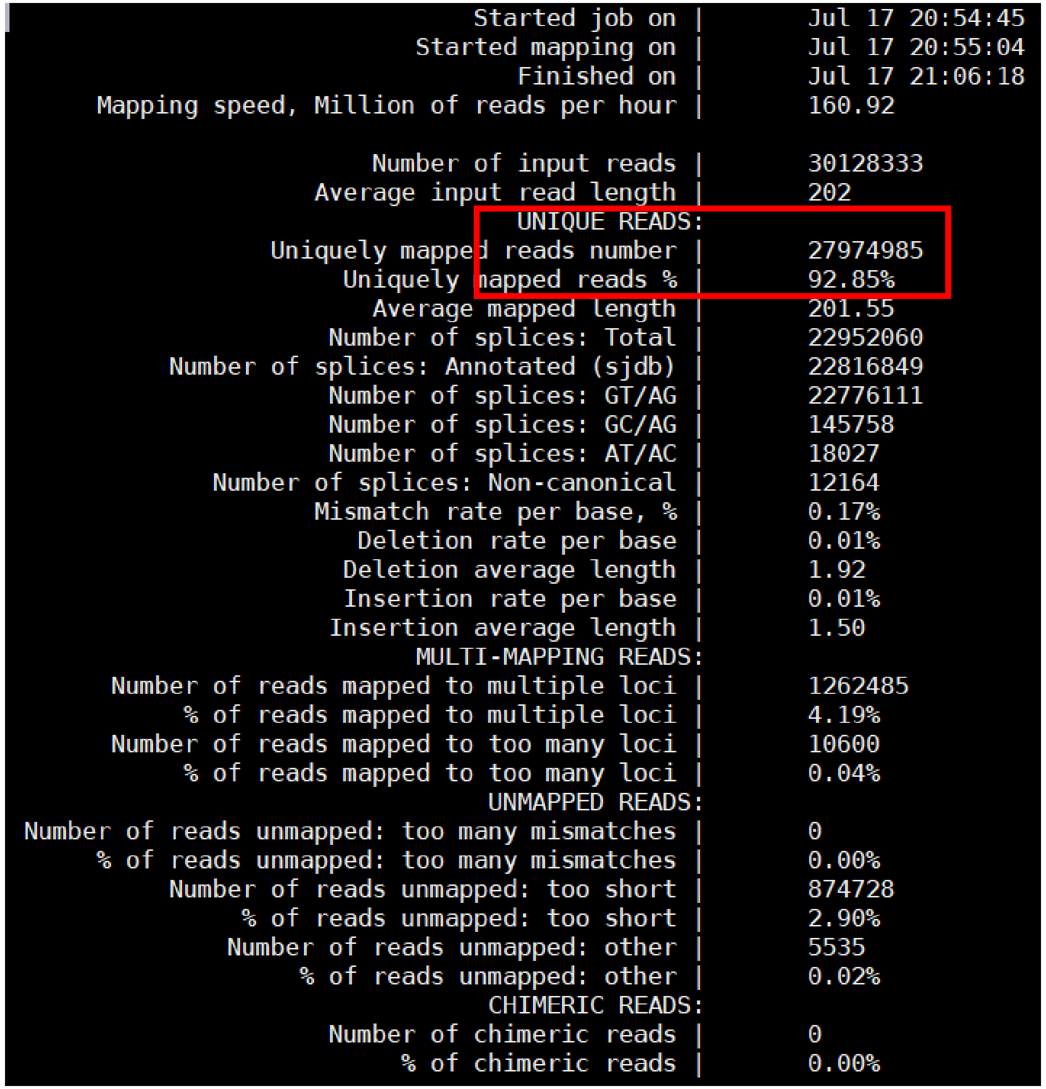


Read1, 2, 3 … → Gene1

Read13, 22, 23 … → Gene2

Read37, 211, 309 … → Gene3

Output: SAM or BAM file (BAM: binary file)

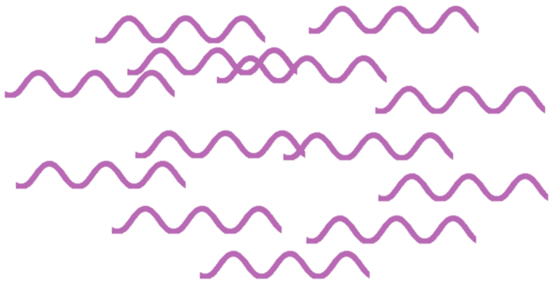


```
                               Started job on |       Jul 17 20:54:45
                           Started mapping on |       Jul 17 20:55:04
                                  Finished on |       Jul 17 21:06:18
     Mapping speed, Million of reads per hour |       160.92

                        Number of input reads |       30128333
                    Average input read length |       202
                                  UNIQUE READS:
                 Uniquely mapped reads number |       27974985
                      Uniquely mapped reads % |       92.85%
                        Average mapped length |       201.55
                     Number of splices: Total |       22952060
          Number of splices: Annotated (sjdb) |       22816849
                     Number of splices: GT/AG |       22776111
                     Number of splices: GC/AG |       145758
                     Number of splices: AT/AC |       18027
             Number of splices: Non-canonical |       12164
                    Mismatch rate per base, % |       0.17%
                       Deletion rate per base |       0.01%
                      Deletion average length |       1.92
                      Insertion rate per base |       0.01%
                     Insertion average length |       1.50
                           MULTI-MAPPING READS:
      Number of reads mapped to multiple loci |       1262485
           % of reads mapped to multiple loci |       4.19%
      Number of reads mapped to too many loci |       10600
           % of reads mapped to too many loci |       0.04%
                                UNMAPPED READS:
Number of reads unmapped: too many mismatches |       0
     % of reads unmapped: too many mismatches |       0.00%
          Number of reads unmapped: too short |       874728
               % of reads unmapped: too short |       2.90%
              Number of reads unmapped: other |       5535
                   % of reads unmapped: other |       0.02%
                                CHIMERIC READS:
                     Number of chimeric reads |       0
                          % of chimeric reads |       0.00%
```

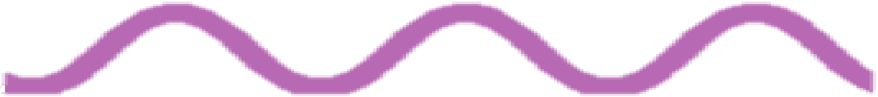
- RNA-sequencing (Alignment)

Unique mapped read → what we use for data analysis
What are other reads?



Human read

Microbiome reads, contamination …

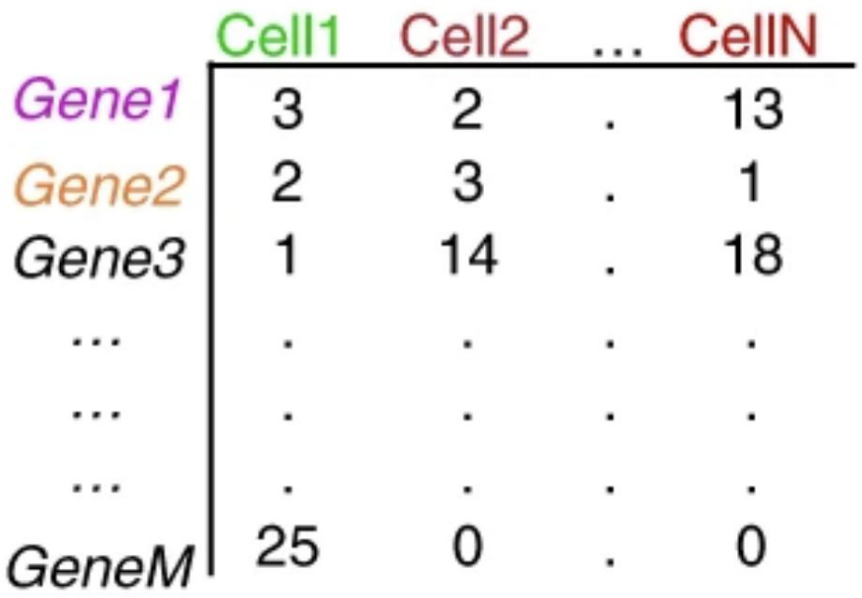


Long read: High specificity → won't map to other regions
Ex): PacBio; high error rate

Short read: Multiple mapping to many regions

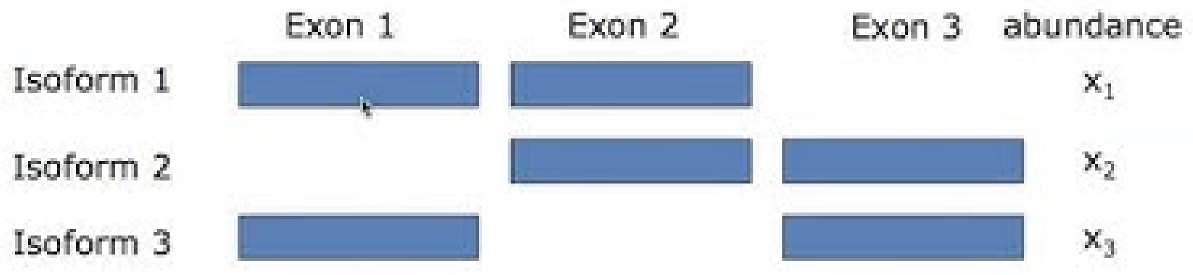
- RNA-sequencing (Quantification)

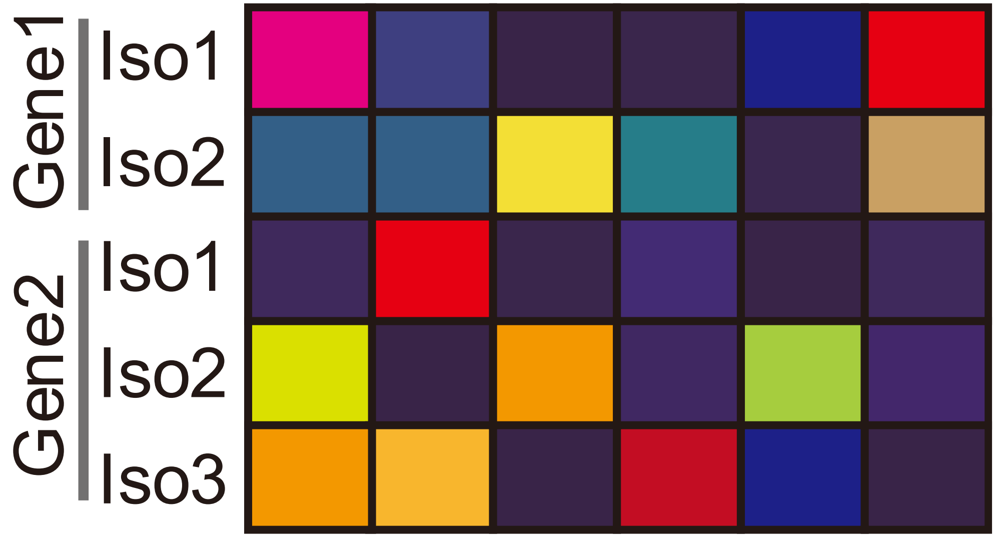
- Mapped reads → count matrix (per gene)
Counting reads mapped to a given gene
Ex) FeatureCounts (gene), Kallisto (transcript)



	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

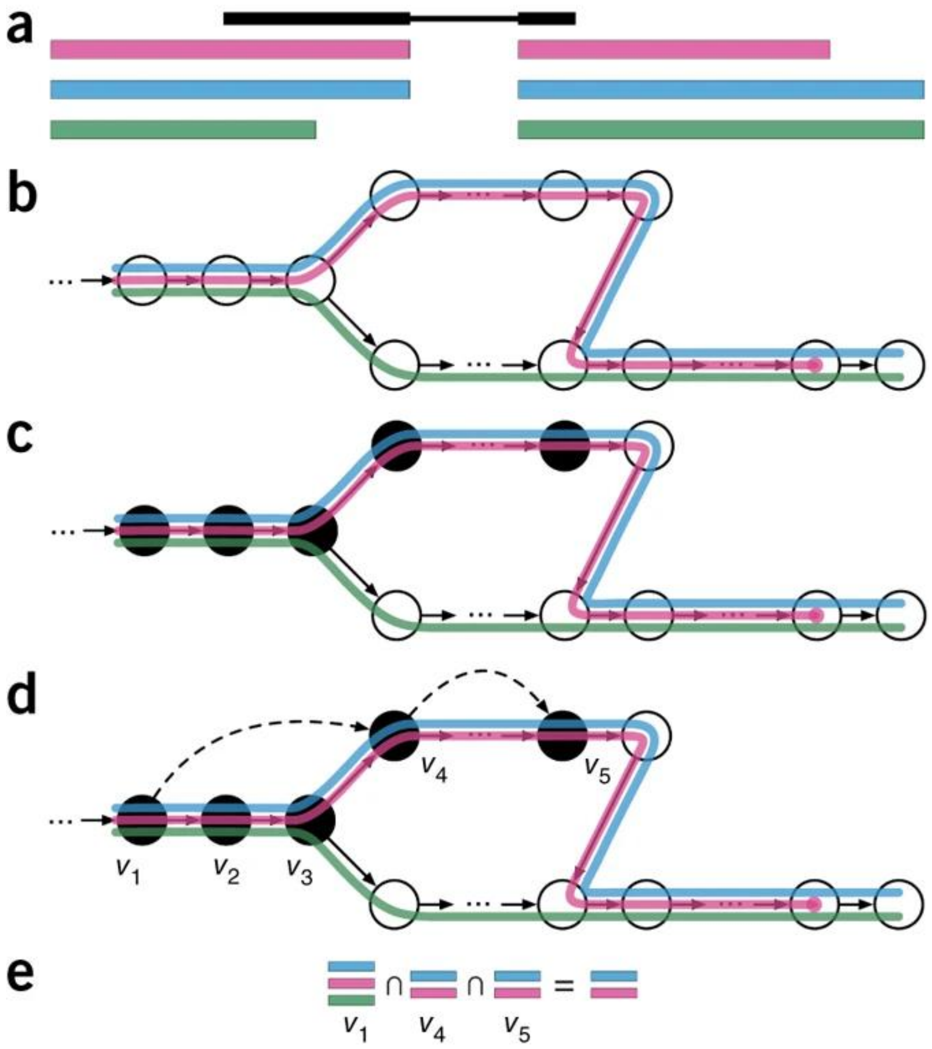
- Mapped reads → count matrix (per isoform)
Unique exon count, EM (expectation maximization) algorithm
Ex) RSEM or longread sequencing





• RNA-sequencing (Kallisto; Alignment)

Kallisto: align to transcriptome (less reference size compared to the genome) → super fast



-black line: read (fragment)
-pink, blue, green: potential transcript (from the reference)

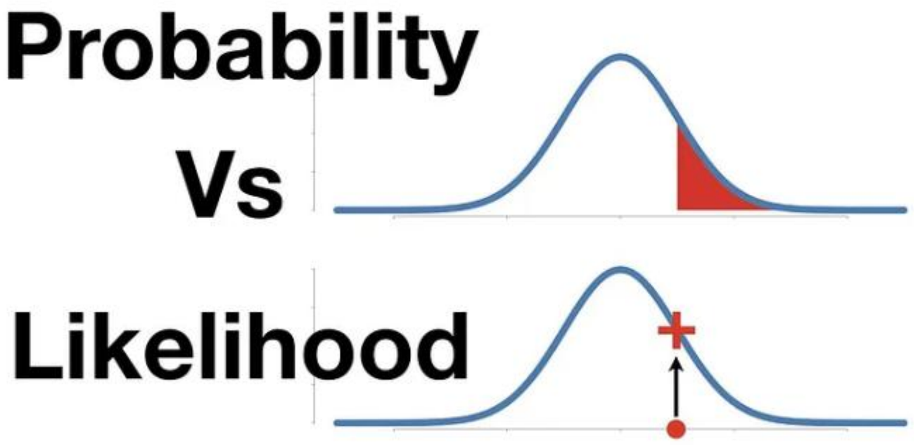
-de Bruijn graph (T-DBG) formation
-O: k-mer (hashed by indexing → fast search)
-●: compatible node between read and the reference

d: first search: v1 → skip until v4 (non-overlapping)
→ fast search
→ Align only to the possible transcripts

- RNA-sequencing (Kallisto; Quantification)

Quantification: Expectation Maximization (EM) algorithm
→ Find a variable to maximize the Likelihood → first-derivative = 0

$$L(\alpha) \propto \prod_{f\in F} \sum_{t\in T} \mathbf{y}_{f,t} \frac{\alpha_t}{l_t} = \prod_{e\in E} \left(\sum_{t\in e} \frac{\alpha_t}{l_t} \right)^{c_e}$$



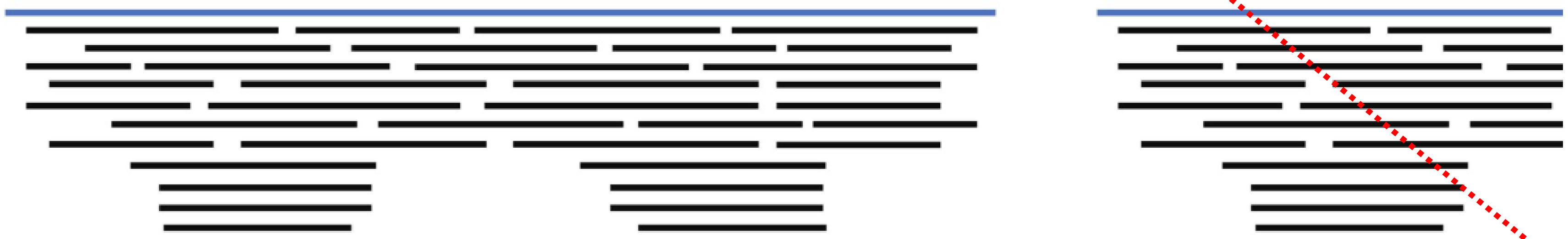
L(a): likelihood function for the alignment
a: probability of selecting fragments from transcripts
l: effective transcript length
F: set of read
T: set of transcript (reference)
Y: alignment matrix (above slide): 0 or 1
C: number of counts from equivalence class (of k-mer) e

→ What is the a to maximize L(a)?

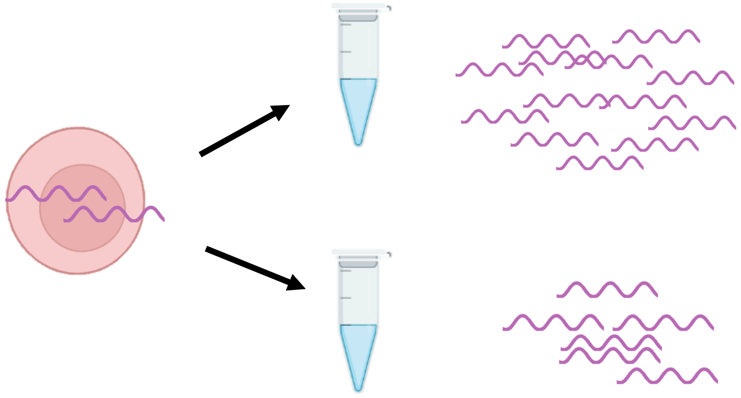
• RNA-sequencing (Normalization)

- Why?

1) Gene length normalization: Different probability of “read” capture rate during sequencing
→ longer gene has a higher chance of being mapped



2) Total read count (= read-depth) normalization: Adjust read-depth between different samples



Should be the same

3) RPKM (Reads Per Kilobase per Millions mapped reads), FPKM (Fragment), TPM (Transcrpit)

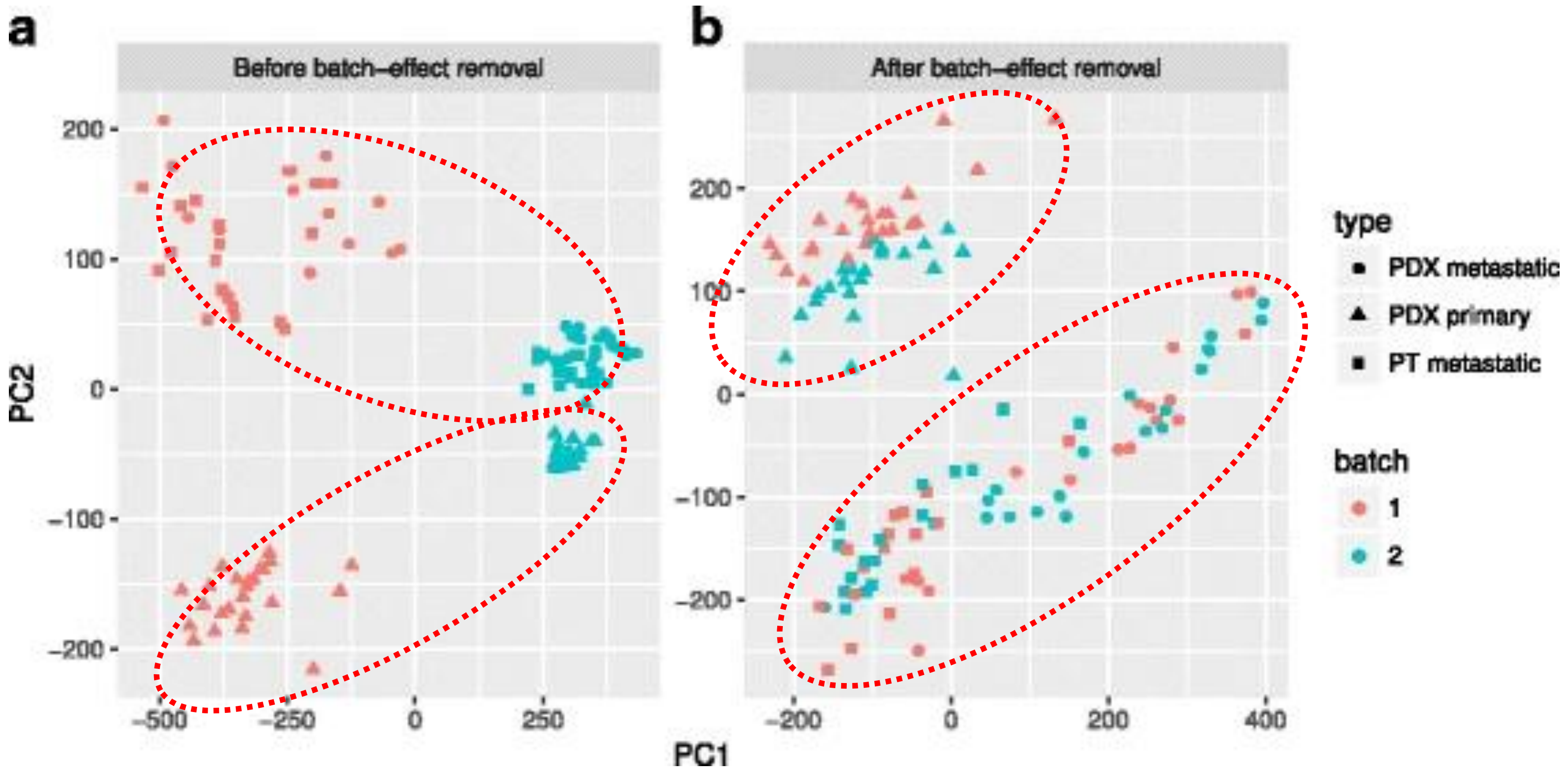
$$\text{RPKM} = \frac{\text{Total fragments}}{\text{Mapped reads (millions)} * \text{exon length (KB)}} = \frac{\text{Number of reads of the region}}{\frac{\text{Total reads}}{1{,}000{,}000} * \frac{\text{Region length}}{1{,}000}}$$

TPM: Total reads normalization → total read after gene length normalization

- RNA-sequencing (Batch correction)

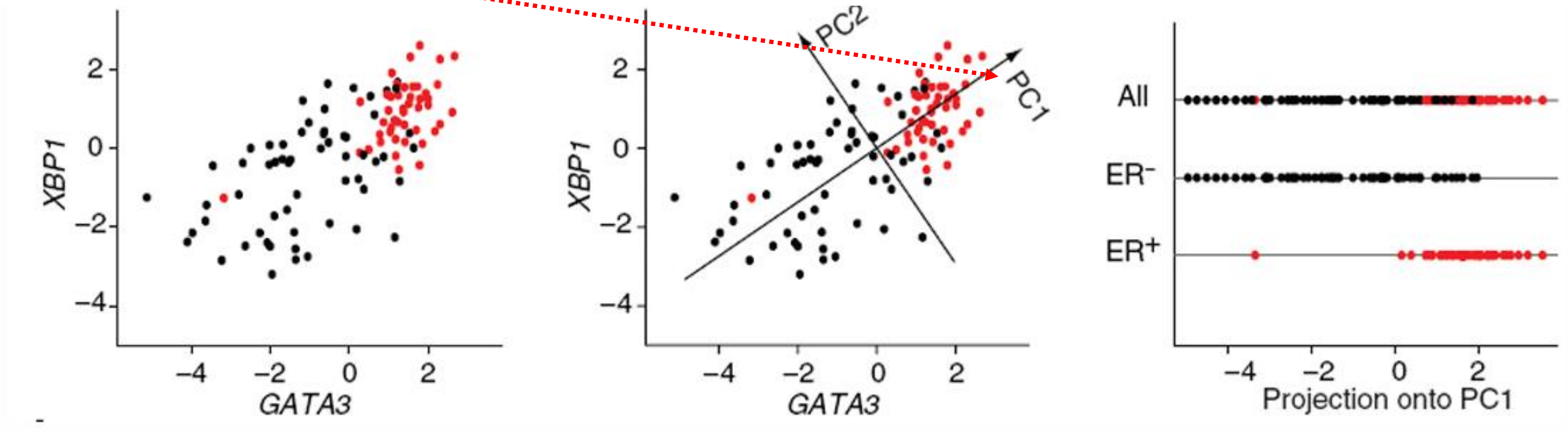
- Why?
Same sample → but, a technical confounding effect
(My experiment and your experiment should be the same!)

*Limma: linear regression-based, Combat: negative binomial distribution, DESeq2: scaling factors

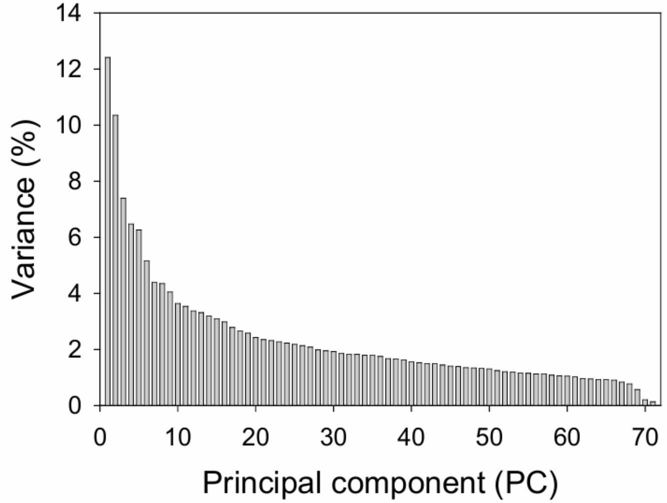


- PCA (principal component analysis)

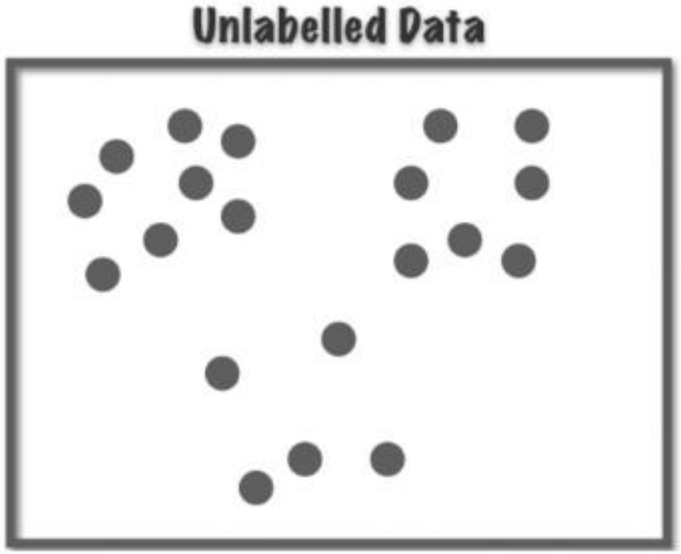
- Find a PC axis to maximize the variance of the data → Distribute samples to maximize the variance



- Number of PC == Number of features or <= sample size
→ Usually, we have more features (~20k)
→ Each PCs: explain the variance of the data
→ Using only a few PCs: Dimension reduction
Advantage: among confounding effects, noise, we can solely capture the biological difference
(or meaningful information)
+ data representation & visualization (20K genes → 2 PCs)



- Clustering



- Distance measurement for clustering

➢ **Euclidean distance** (a *straight line* distance in *n*-dimensional space)

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

$$d = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}$$

,if you don't want to increase the distance with the addition of more dimensions.

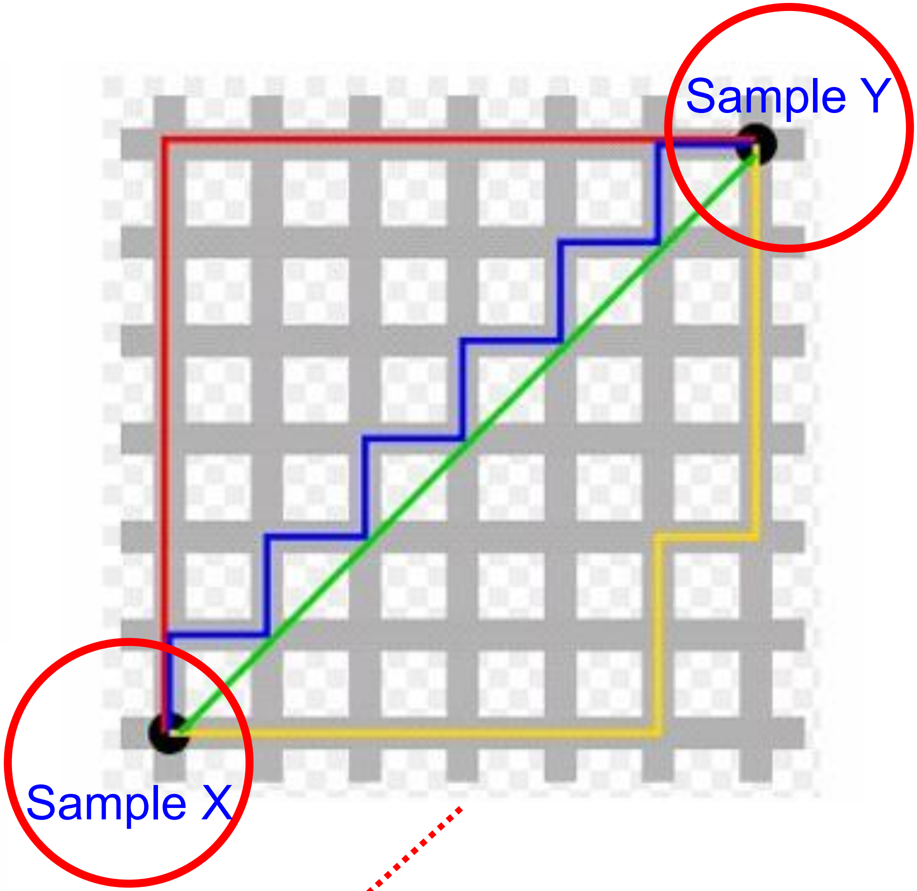
- Euclidean distances **may underestimate** join differences such as differences in two correlated expression .
- Therefore, use Euclidean distance **if** you believe that your **dimensions (variables) are not independent**.

➢ **Manhattan distance** (= city block distance, L1 distance, rectilinear distance, taxicab metric): *distances measured parallel to dimensional axes*. After NY city's grid like street pattern.

$$d = \sum_{i=1}^{n}|x_i - y_i|$$

$$d = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|$$

,if you don't want to increase the distance with the addition of more dimensions.

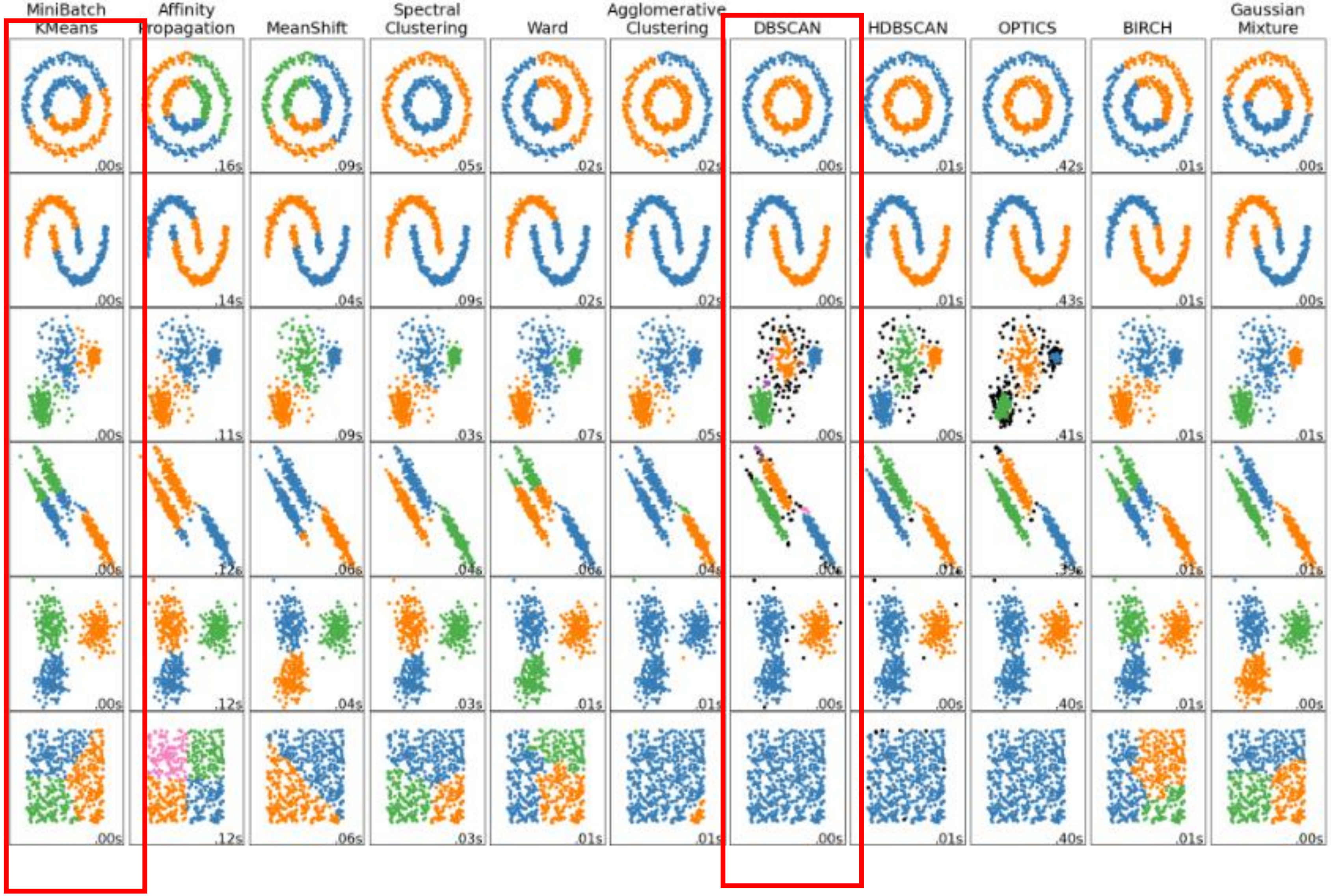


Gene expression space

i: each gene

n: the number of genes

- Clustering

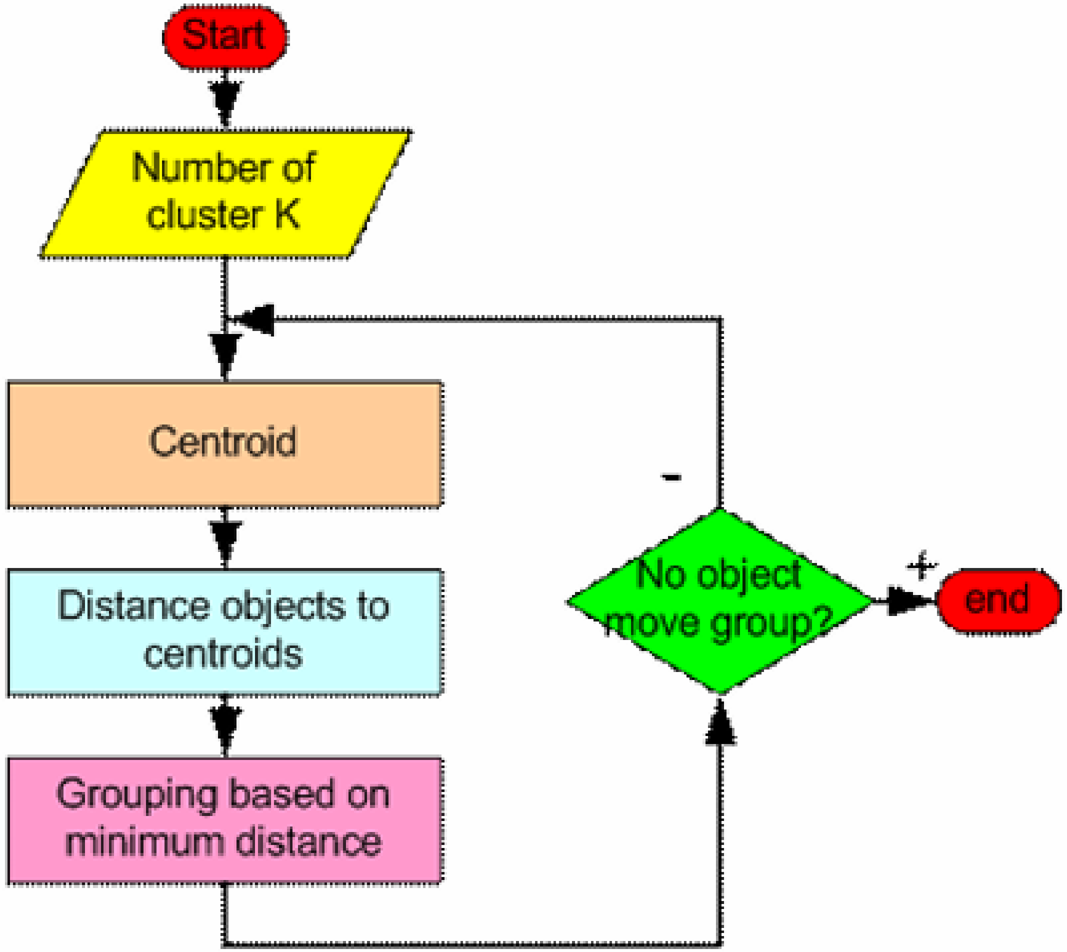

MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s
.16s
.09s
.05s
.02s
.02s
.00s
.01s
.42s
.01s
.00s

- K-mean clustering

K-mean clustering (linear approach)

➢ Algorithm

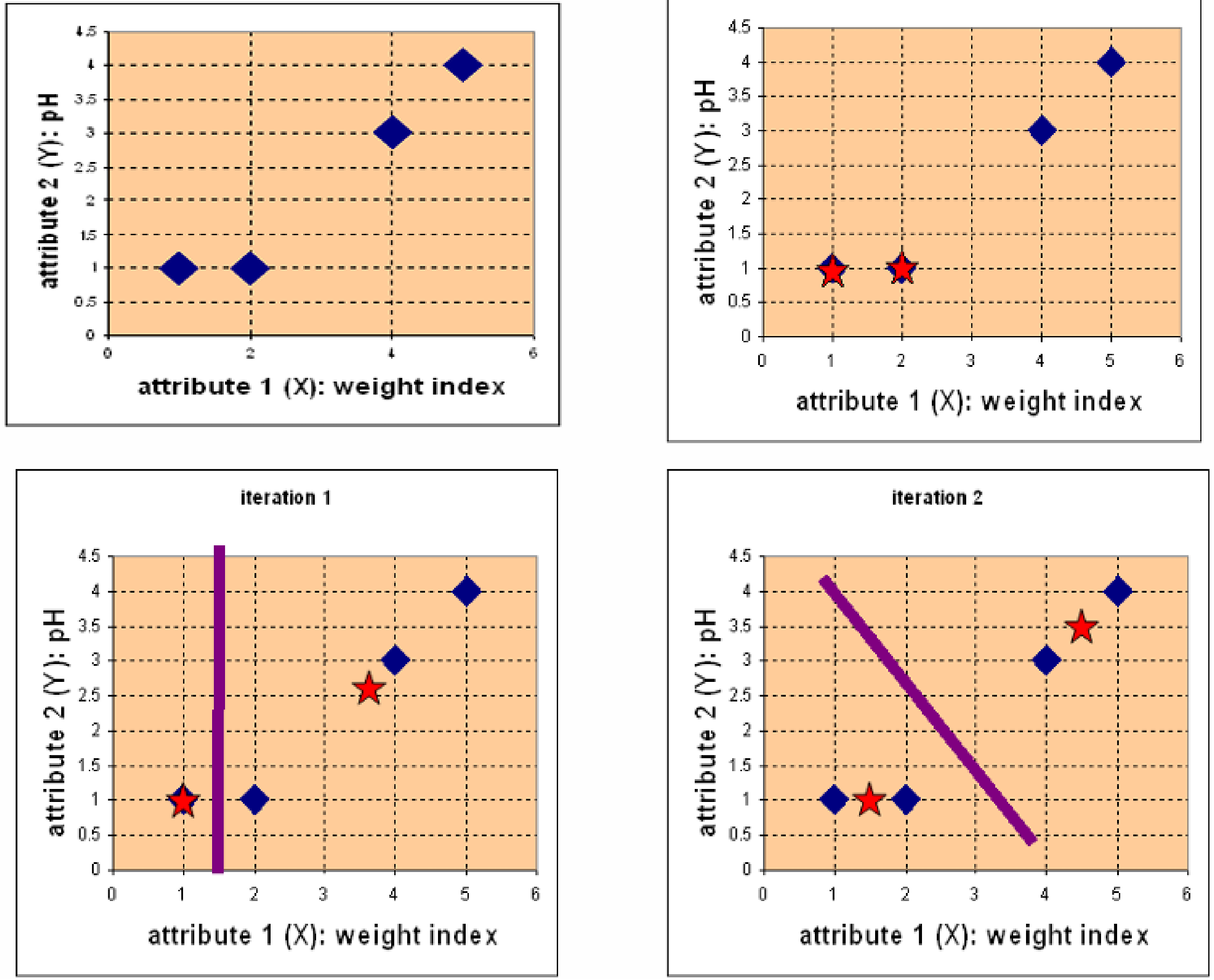
1. ***Initial k***: Decide the number of cluster (*k*)
2. ***Initial value of centroids***: Choose *k* centroids randomly.
3. ***Objects-centroids distance***: Calculate the distance between cluster centroid to each object (with any distance metric).
4. ***Objects clustering***: Assign each object based on minimum distance.
5. ***Determine new centroids***: New centroid moves to mean value of all member objects.
6. ***Iterate steps 3-5***: Until no object change centroid membership.

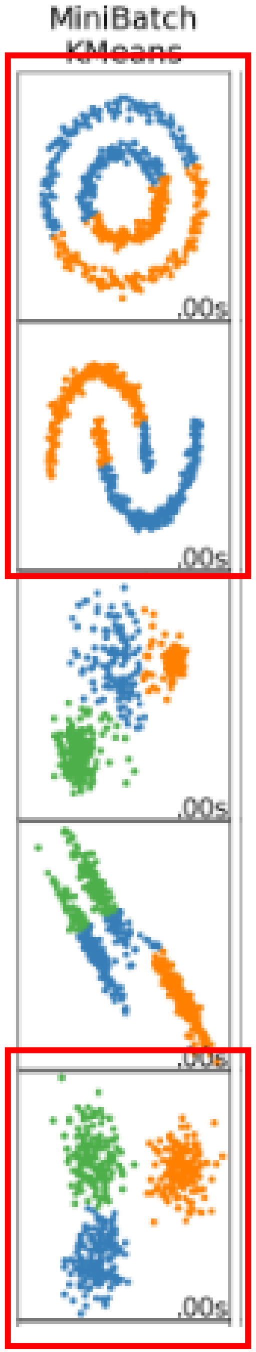

Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

- K-mean clustering

Demonstration

Diamond: object
Star: centroid (*k*=2)

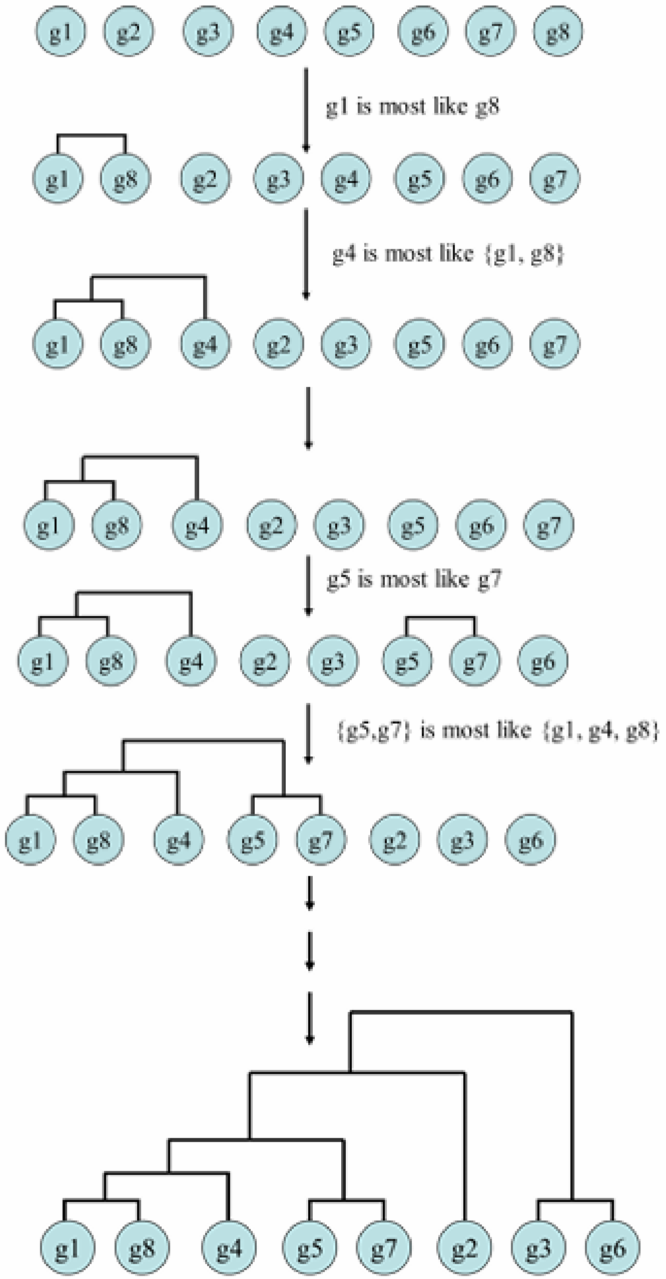




- Hierarchical clustering

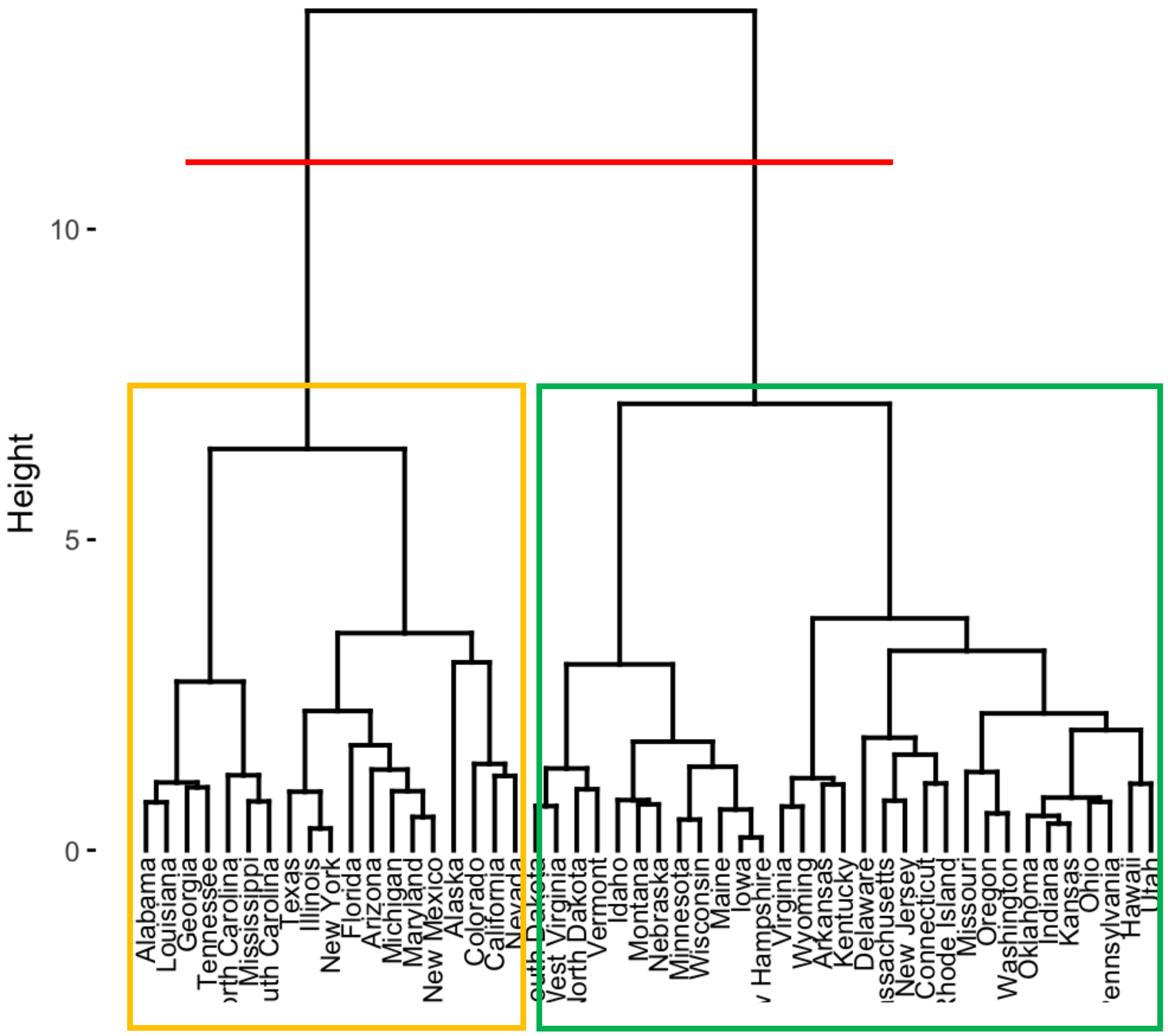
➢ Algorithm

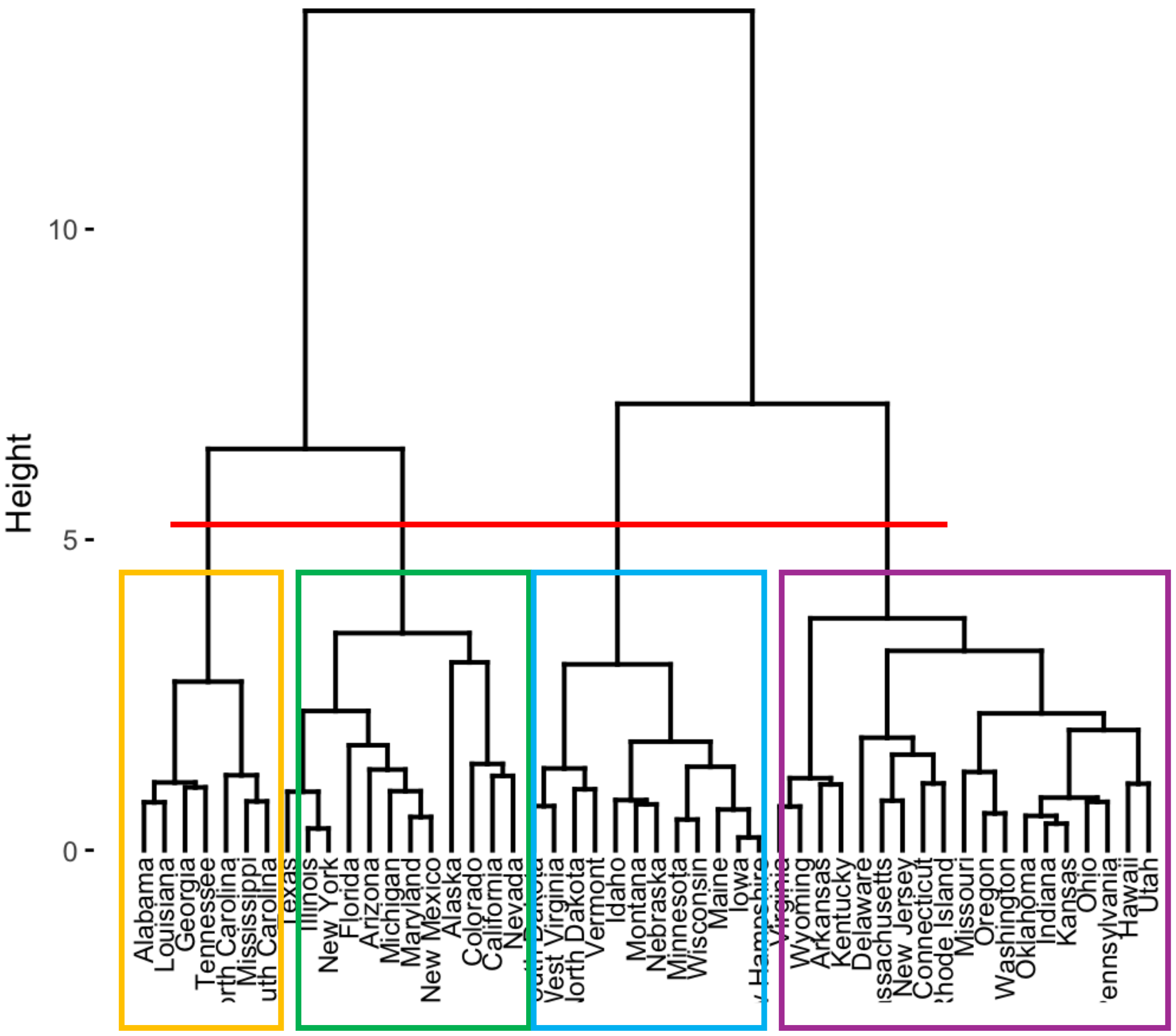
1. Initialize each single gene as a cluster
2. The pairwise distance matrix is calculated for all of the genes to be clustered.
3. The distance matrix is searched for the two most similar genes or clusters.
4. The two selected clusters are merged to produce a new cluster that now contains at least two objects.
5. The distances are calculated between this new cluster and all other clusters. There is no need to calculate all distances as only those involving the new cluster have changed.
6. Steps 3-5 are repeated until all objects are in one cluster.



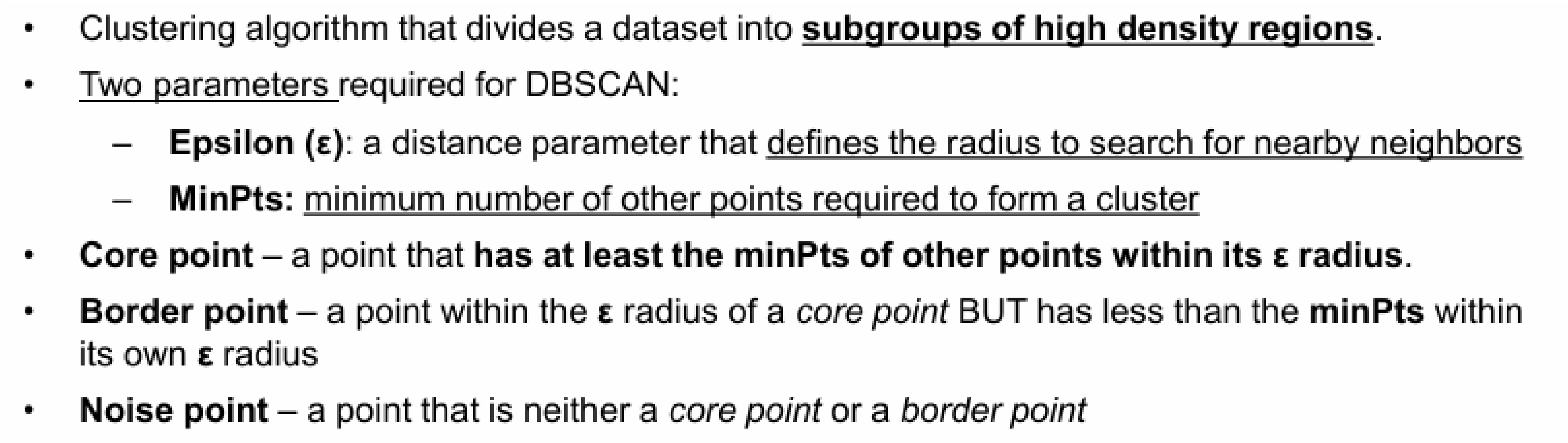
- Hierarchical clustering

Tree cutting

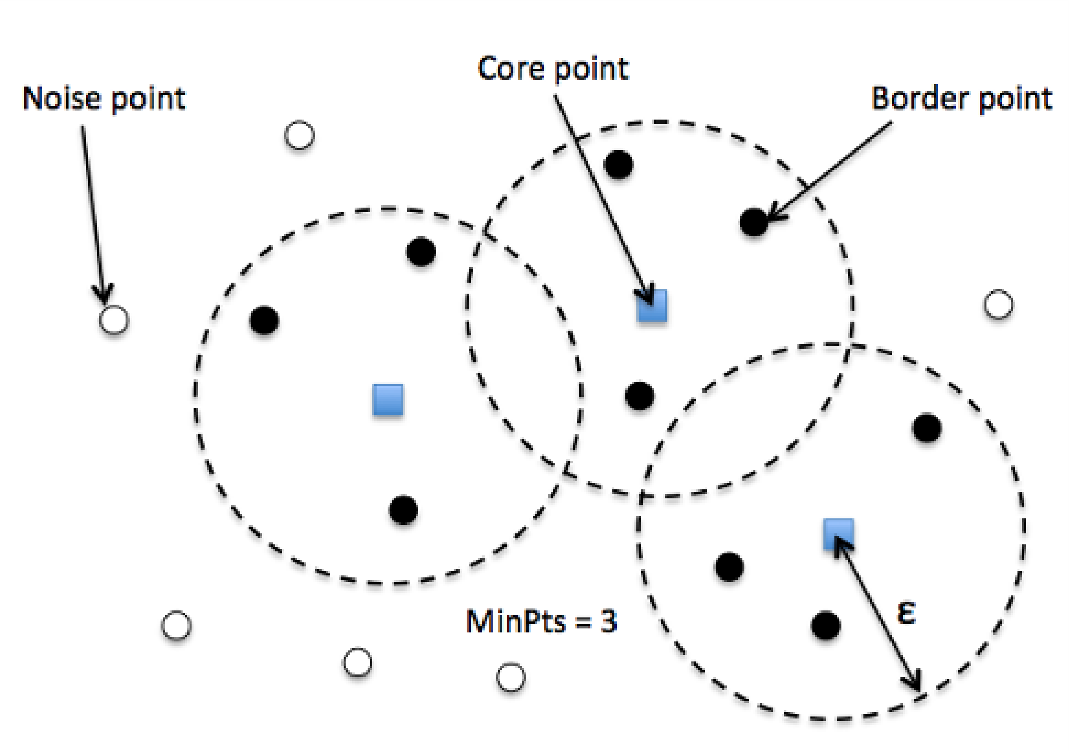




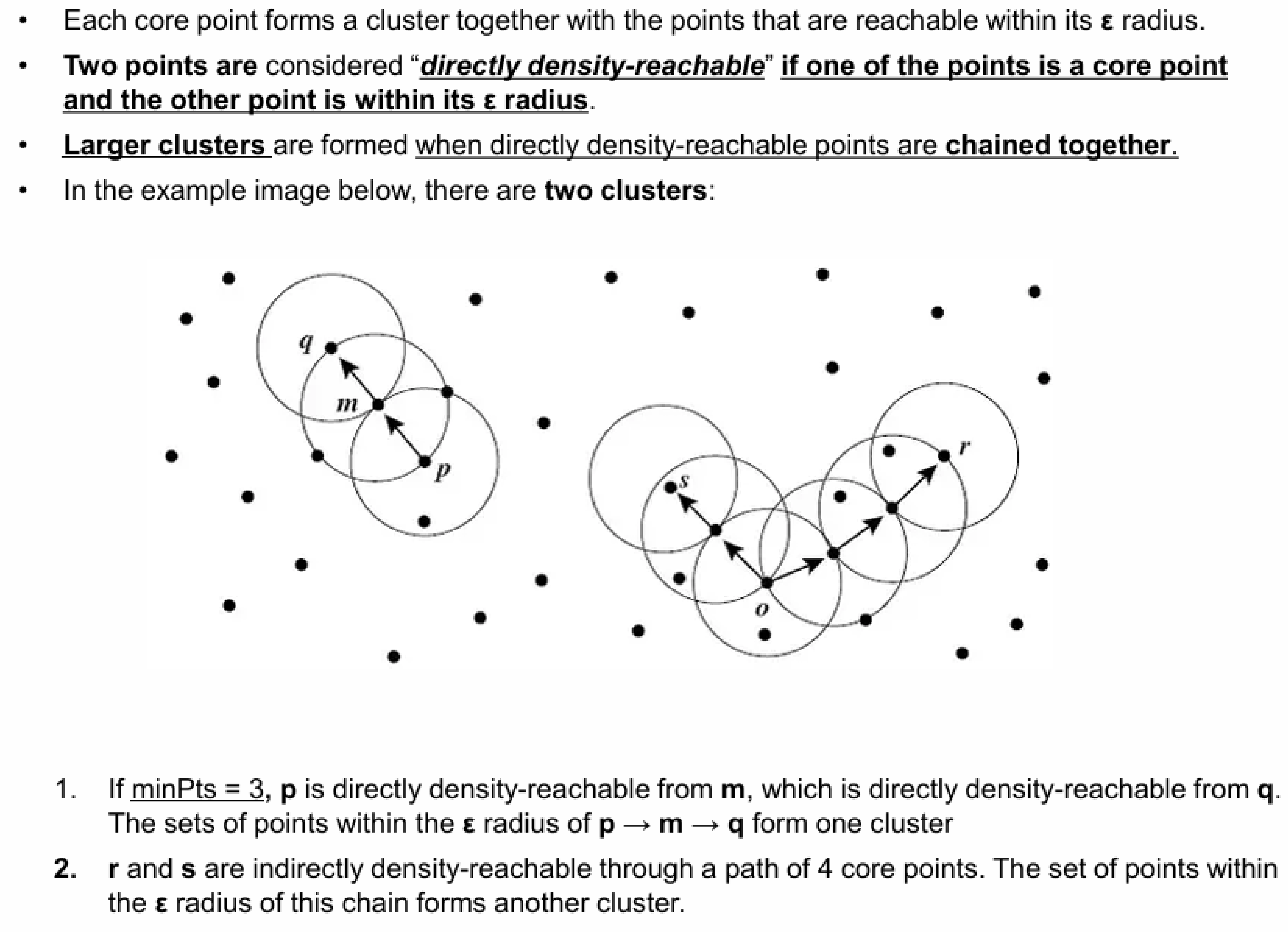
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)



- Clustering algorithm that divides a dataset into **subgroups of high density regions**.
- Two parameters required for DBSCAN:
  - **Epsilon (ε)**: a distance parameter that defines the radius to search for nearby neighbors
  - **MinPts:** minimum number of other points required to form a cluster
- **Core point** – a point that **has at least the minPts of other points within its ε radius**.
- **Border point** – a point within the **ε** radius of a *core point* BUT has less than the **minPts** within its own **ε** radius
- **Noise point** – a point that is neither a *core point* or a *border point*



- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

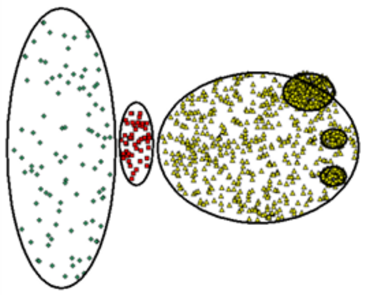


- Each core point forms a cluster together with the points that are reachable within its **ε** radius.
- **Two points are** considered “***directly density-reachable***” **if one of the points is a core point and the other point is within its ε radius**.
- **Larger clusters** are formed when directly density-reachable points are **chained together**.
- In the example image below, there are **two clusters**:

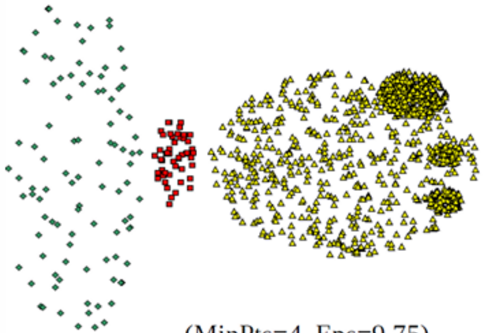
1. If minPts = 3**, p** is directly density-reachable from **m**, which is directly density-reachable from **q**. The sets of points within the **ε** radius of **p** → **m** → **q** form one cluster
2. **r** and **s** are indirectly density-reachable through a path of 4 core points. The set of points within the **ε** radius of this chain forms another cluster.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

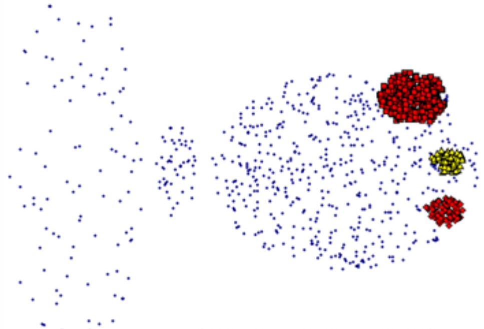
- The DBSCAN algorithm **repeats the following process until all points have been assigned to a cluster** or are labeled as visited:
  1. **Arbitrarily select a point P**.
  2. **Retrieve all points** directly density-reachable from P **with respect to ε**.
  3. **If P is a core point**, a cluster is formed. Find recursively all its density connected points and **assign them to the same cluster as P**.
  4. **If P is not a core point**, DBSCAN **iterates through the remaining unvisited points** in the dataset.
- DBSCAN does not require us to specify the number of clusters.
- It can handle clusters of arbitrarily shapes and sizes.
- It is robust to noise.



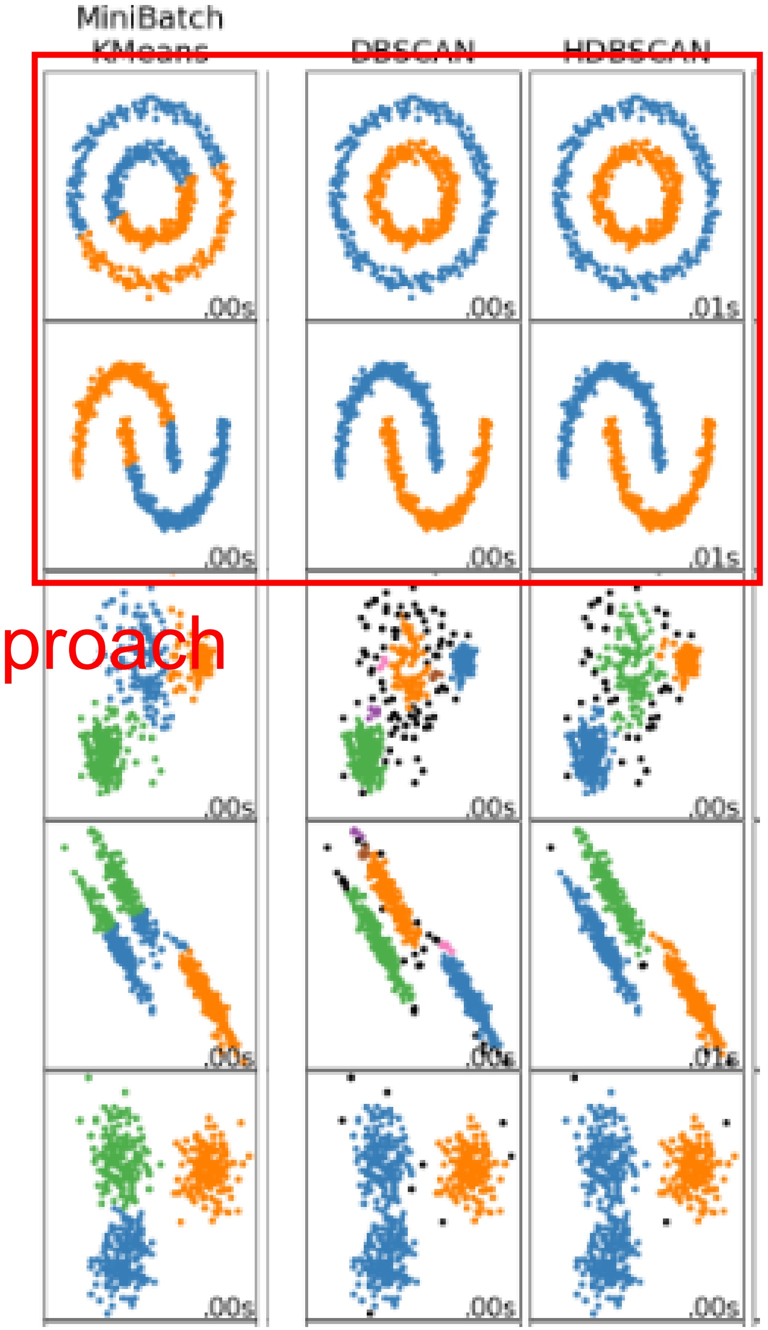
Original Points



(MinPts=4, Eps=9.75).



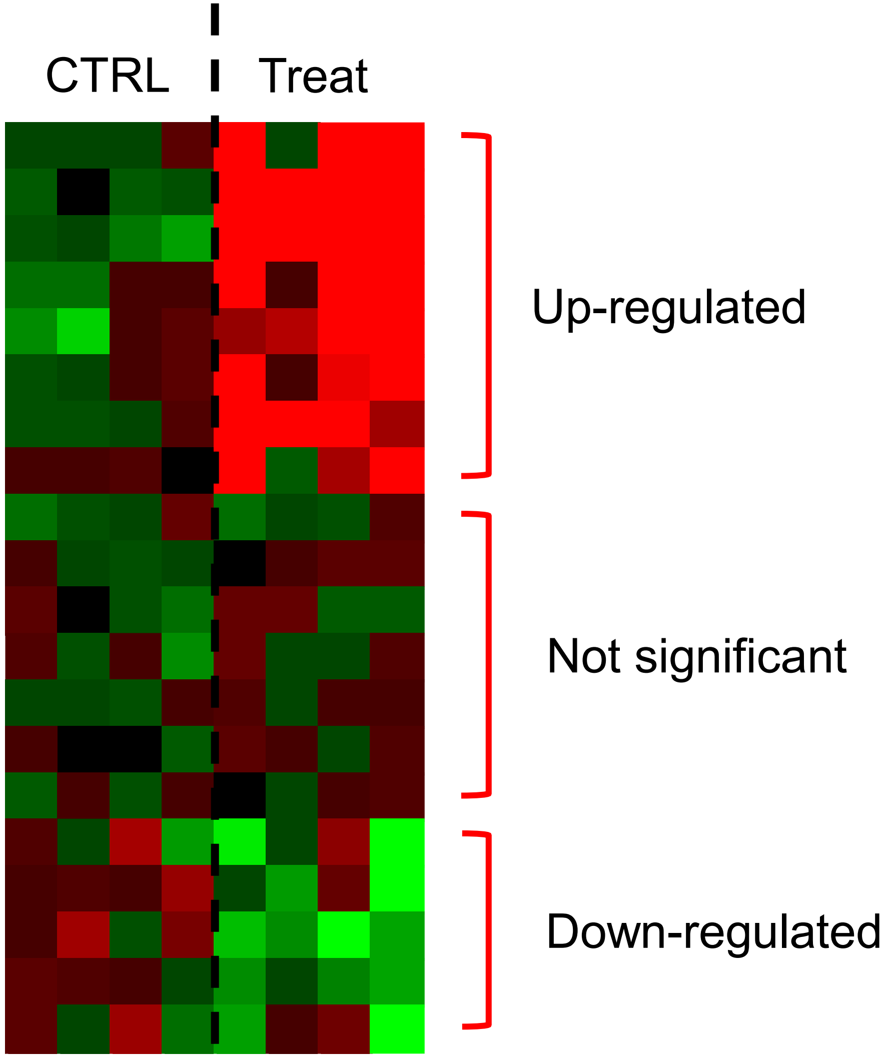
(MinPts=4, Eps=9.92)



Nonlinear approach

- DEG (Differentially expressed gene) analysis

-Which gene is expressing differently between two groups



CTRL

Treat

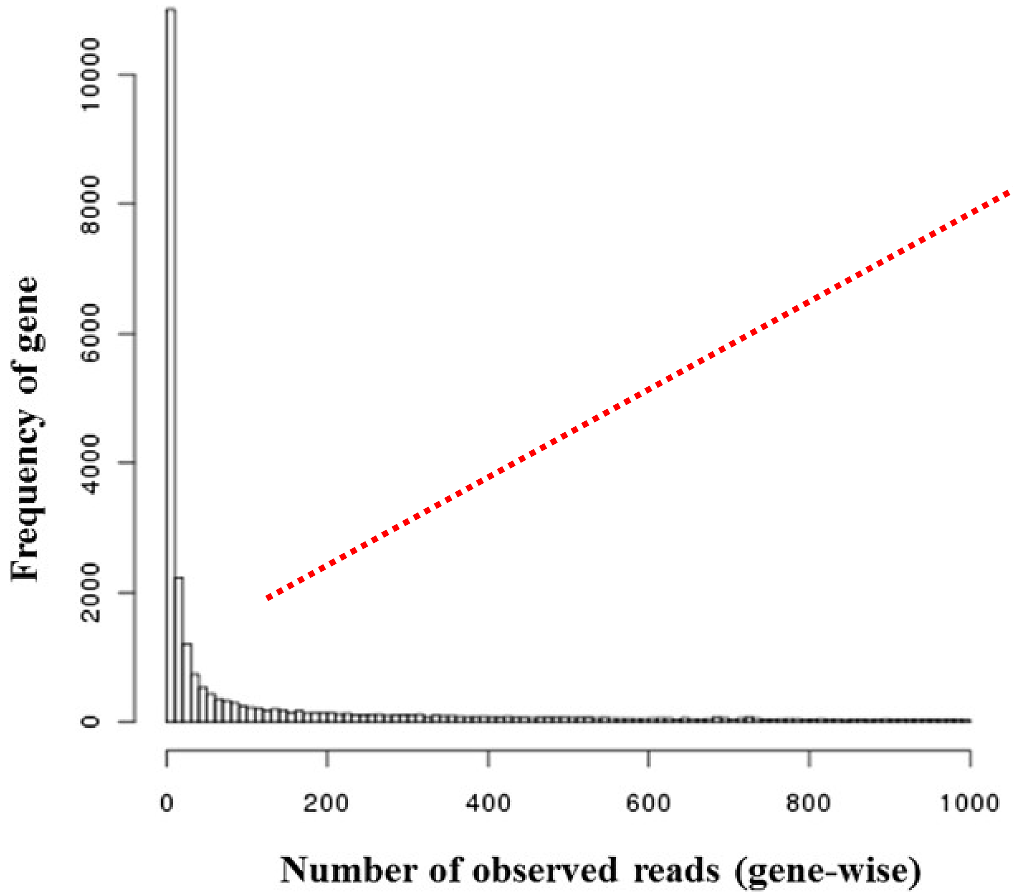
Up-regulated

Not significant

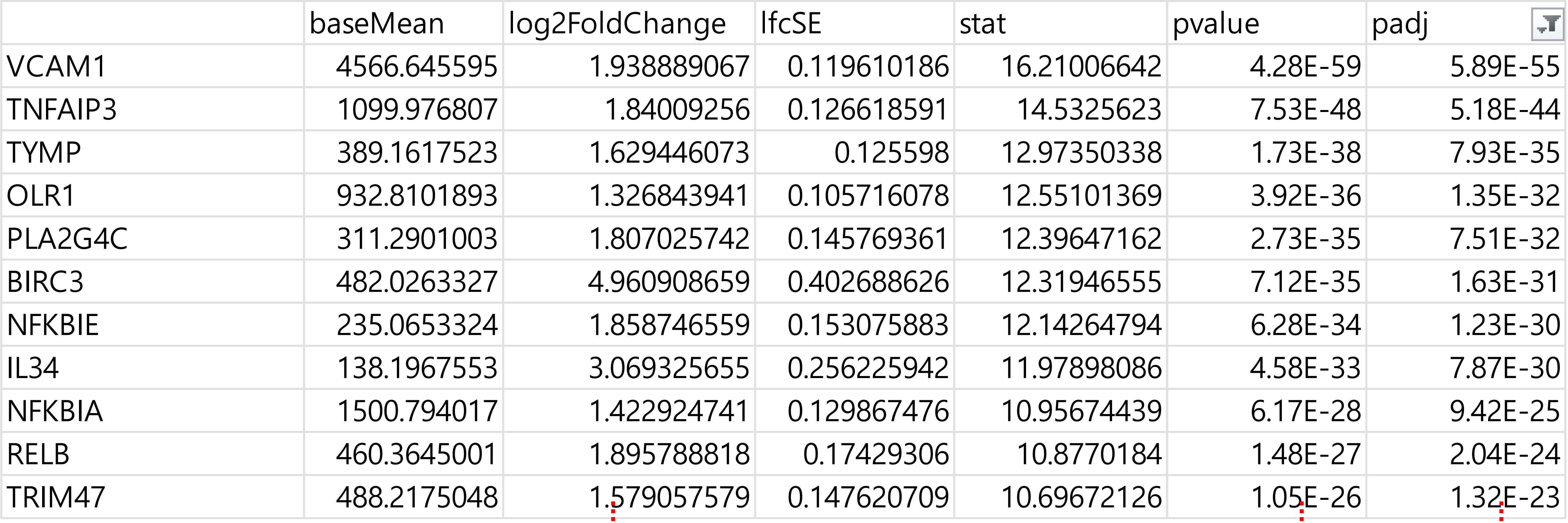
Down-regulated

- DEG (Differentially expressed gene) analysis

-DESeq2: Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution



- DEG (Differentially expressed gene) analysis



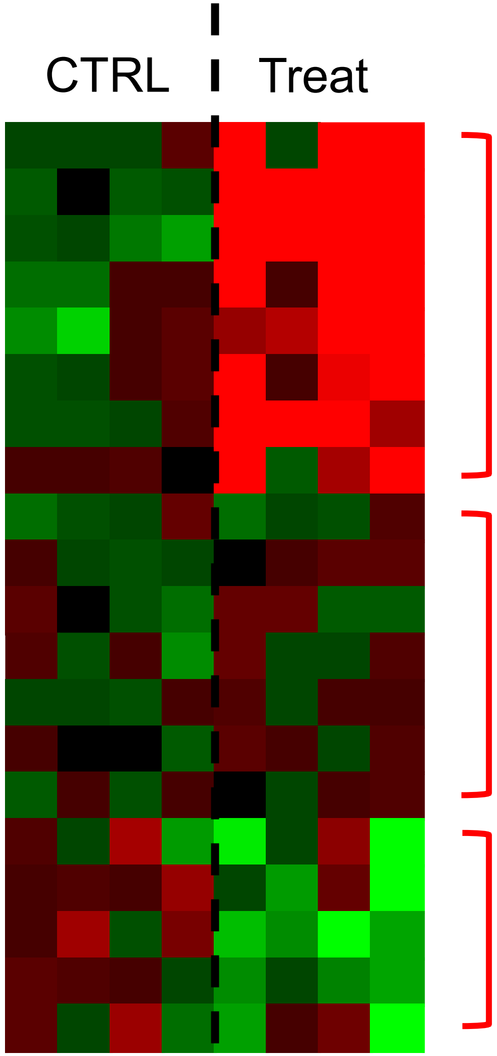
	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
VCAM1	4566.645595	1.938889067	0.119610186	16.21006642	4.28E-59	5.89E-55
TNFAIP3	1099.976807	1.84009256	0.126618591	14.5325623	7.53E-48	5.18E-44
TYMP	389.1617523	1.629446073	0.125598	12.97350338	1.73E-38	7.93E-35
OLR1	932.8101893	1.326843941	0.105716078	12.55101369	3.92E-36	1.35E-32
PLA2G4C	311.2901003	1.807025742	0.145769361	12.39647162	2.73E-35	7.51E-32
BIRC3	482.0263327	4.960908659	0.402688626	12.31946555	7.12E-35	1.63E-31
NFKBIE	235.0653324	1.858746559	0.153075883	12.14264794	6.28E-34	1.23E-30
IL34	138.1967553	3.069325655	0.256225942	11.97898086	4.58E-33	7.87E-30
NFKBIA	1500.794017	1.422924741	0.129867476	10.95674439	6.17E-28	9.42E-25
RELB	460.3645001	1.895788818	0.17429306	10.8770184	1.48E-27	2.04E-24
TRIM47	488.2175048	1.579057579	0.147620709	10.69672126	1.05E-26	1.32E-23

Log2(treat / ctrl)
→ 0: no change
→ +: increase, -: decrease
→ Value 1: 2 fold

P value → Adjusted p value
- Multiple hypothesis correction
- Avoid lucky hits my multiple testing

- Gene set analysis

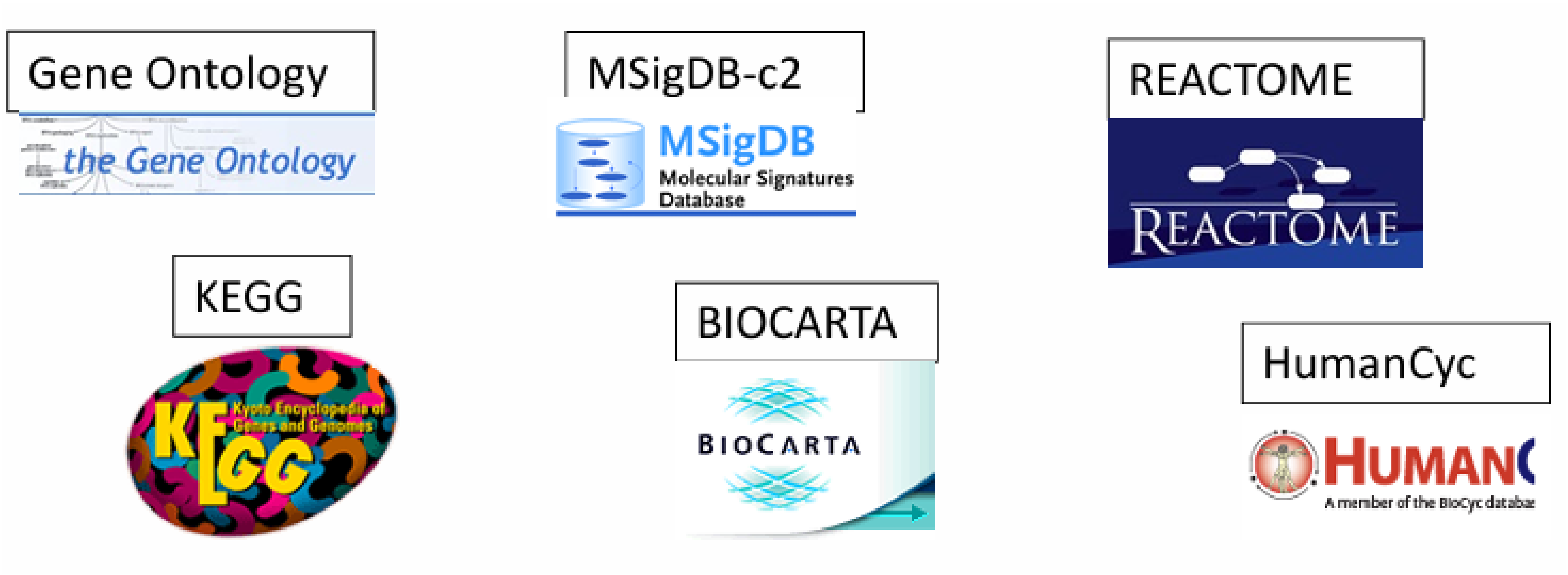
We can perform gene-centric analysi
But! Too many! (20k genes)
Let’s see whether there is a coherent pathway between genes



Up-regulated genes: ex) Immune activation, chemotaxis, …

- Geneset Database

Researchers already studied a lot !
We don’t need to start from the scratch



Gene Ontology

MSigDB-c2

REACTOME

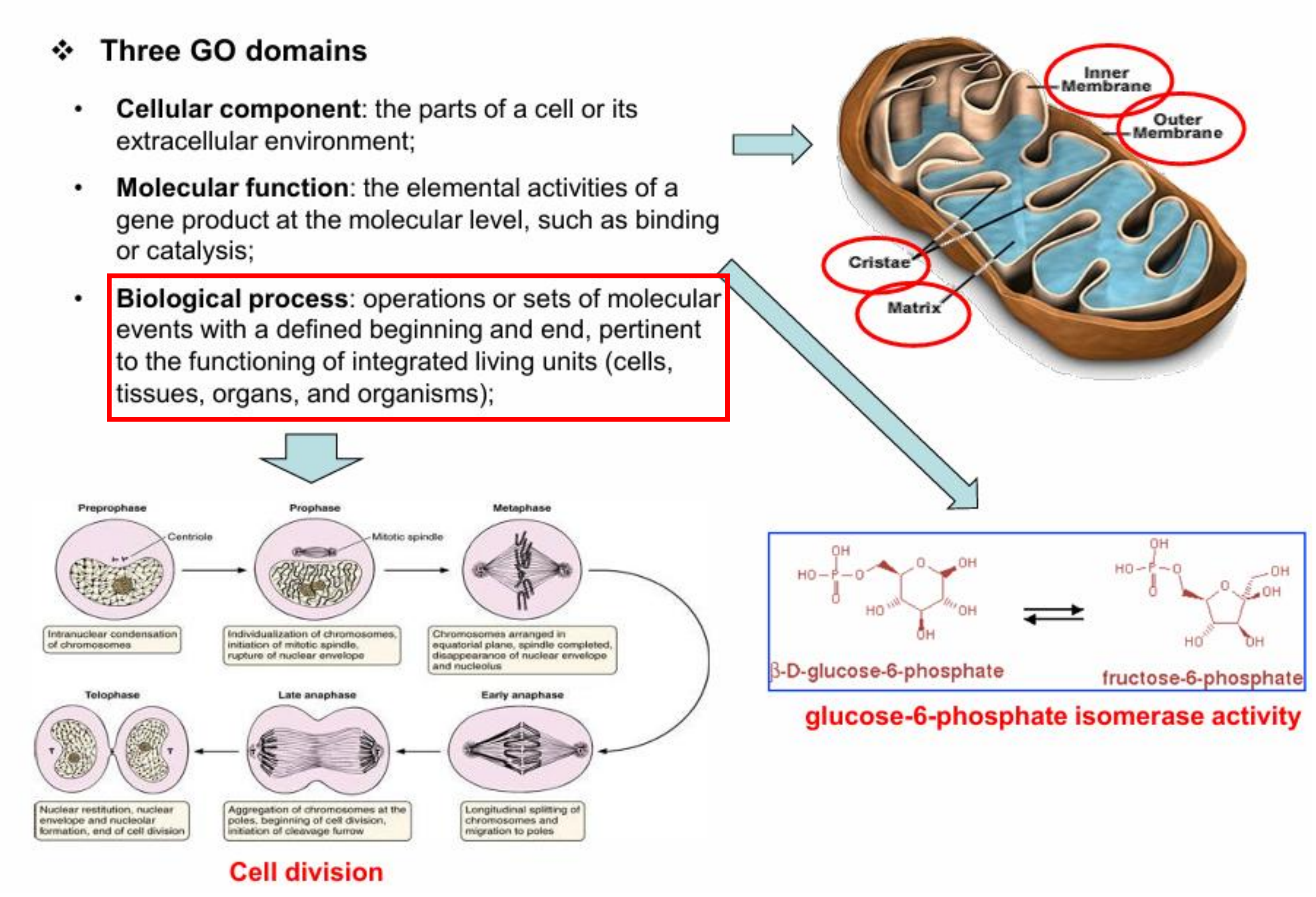
KEGG

BIOCARTA

HumanCyc

- GO (Gene ontology)

-Description about a given gene



Three GO domains

- **Cellular component**: the parts of a cell or its extracellular environment;
- **Molecular function**: the elemental activities of a gene product at the molecular level, such as binding or catalysis;
- **Biological process**: operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units (cells, tissues, organs, and organisms);

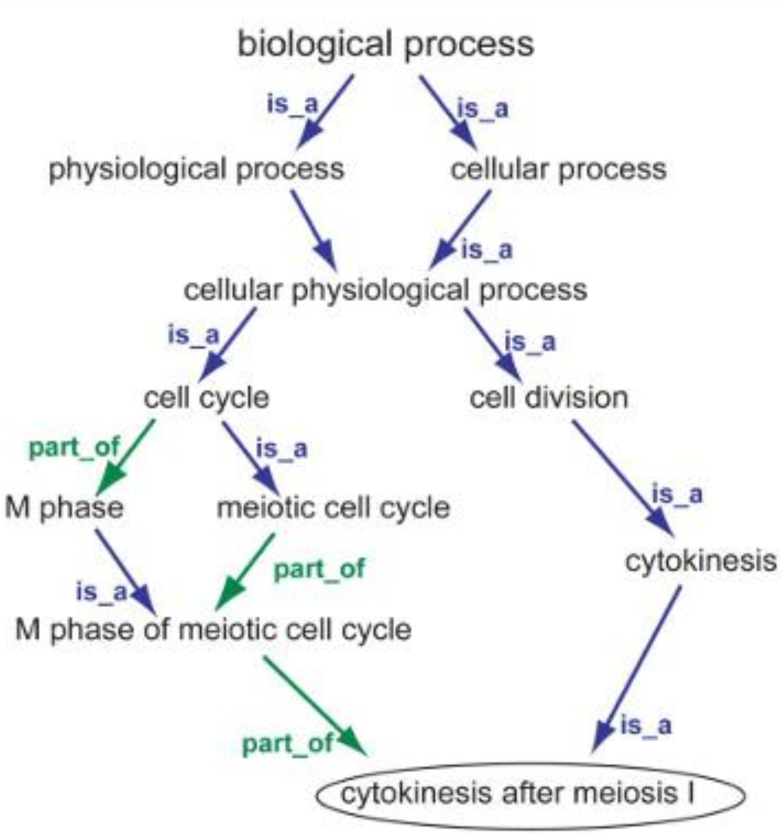
Inner Membrane
Outer Membrane
Cristae
Matrix
Preprophase
Centriole
Prophase
Mitotic spindle
Metaphase
Intranuclear condensation of chromosomes
Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope
Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus
Telophase
Late anaphase
Early anaphase
Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division
Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow
Longitudinal splitting of chromosomes and migration to poles
β-D-glucose-6-phosphate
fructose-6-phosphate

Cell division

glucose-6-phosphate isomerase activity

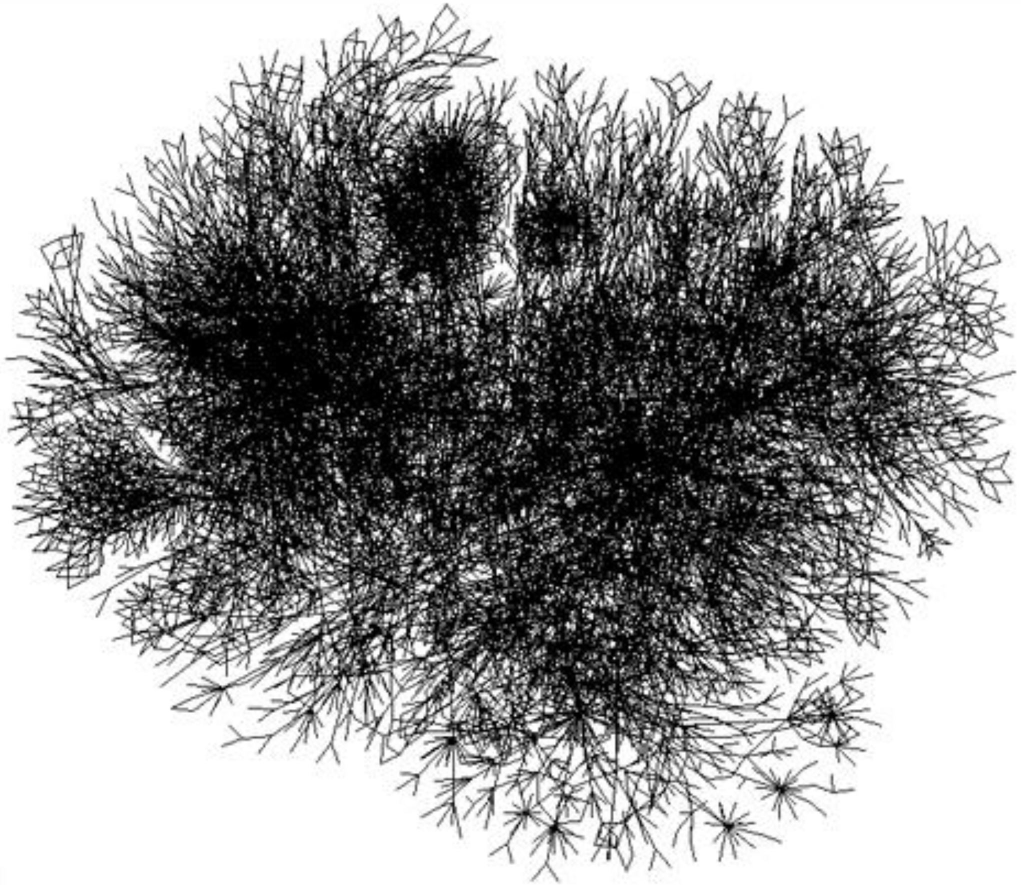
- GO (Gene ontology)

➢ **Example of GO tree**



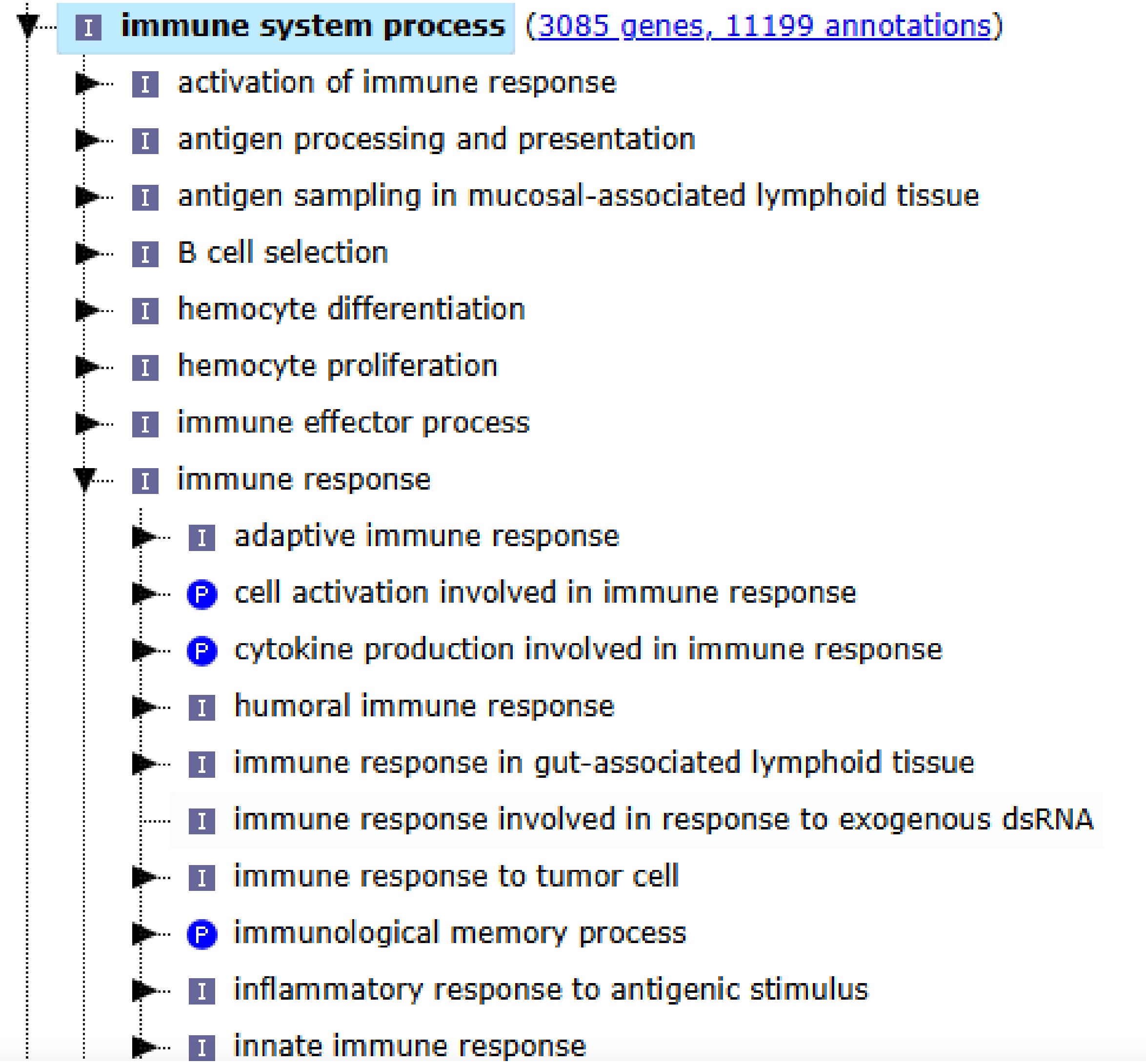
"cellular physiological process", "M phase of meiotic cell cycle" and "cytokinesis after meiosis I" have two parents.

➢ **Layout of whole GO structure**



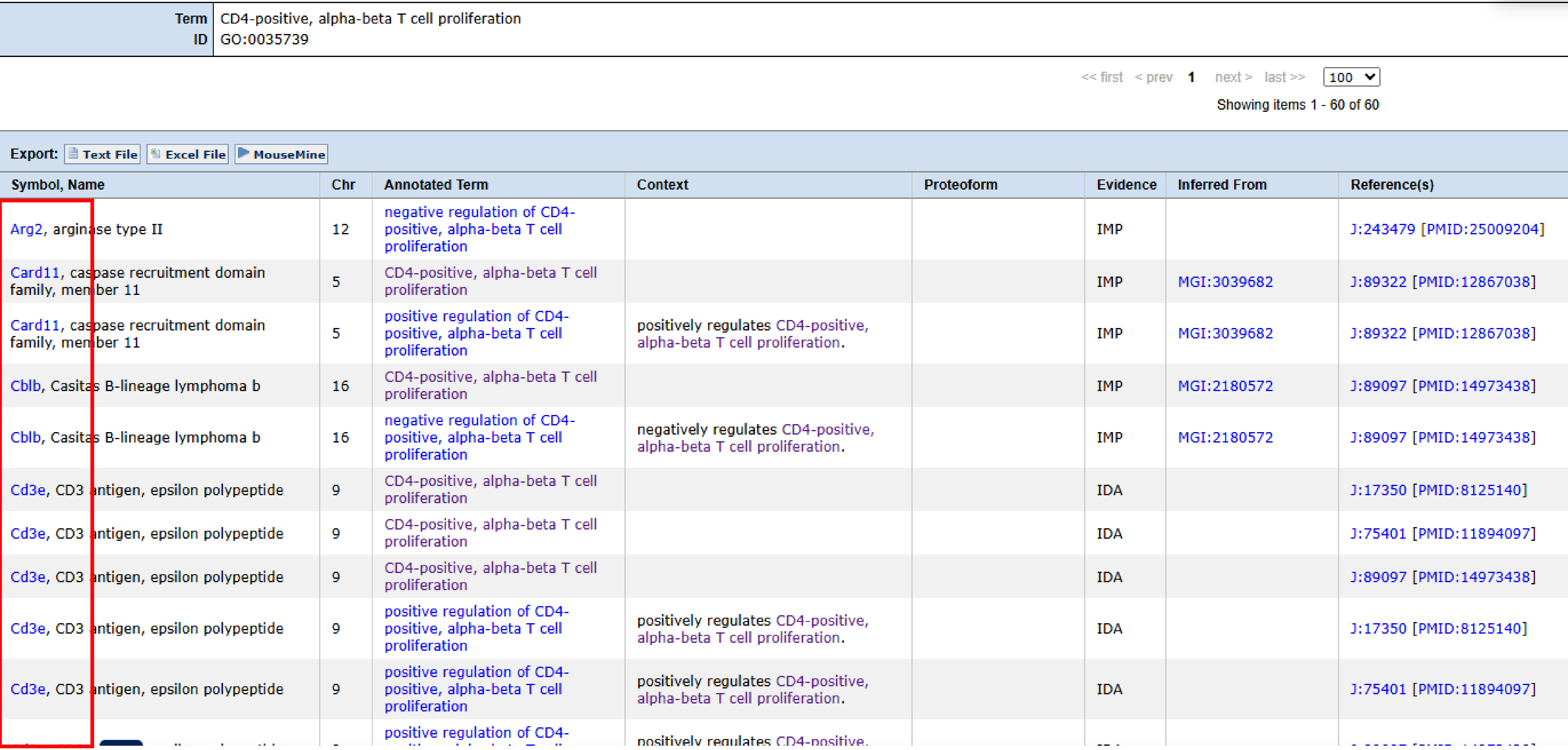
15,335 Is_a or part_of relationships between 9,199 GO *biological process* terms (as of March 2005, by Insuk Lee)

- GO (Gene ontology)



- ▼ I **immune system process** (3085 genes, 11199 annotations)
  - ▶ I activation of immune response
  - ▶ I antigen processing and presentation
  - ▶ I antigen sampling in mucosal-associated lymphoid tissue
  - ▶ I B cell selection
  - ▶ I hemocyte differentiation
  - ▶ I hemocyte proliferation
  - ▶ I immune effector process
  - ▼ I immune response
    - ▶ I adaptive immune response
    - ▶ P cell activation involved in immune response
    - ▶ P cytokine production involved in immune response
    - ▶ I humoral immune response
    - ▶ I immune response in gut-associated lymphoid tissue
    - I immune response involved in response to exogenous dsRNA
    - ▶ I immune response to tumor cell
    - ▶ P immunological memory process
    - ▶ I inflammatory response to antigenic stimulus
    - ▶ I innate immune response

- GO (Gene ontology); CD4-positive, alpha-beta T cell proliferation



Term	CD4-positive, alpha-beta T cell proliferation
ID	GO:0035739

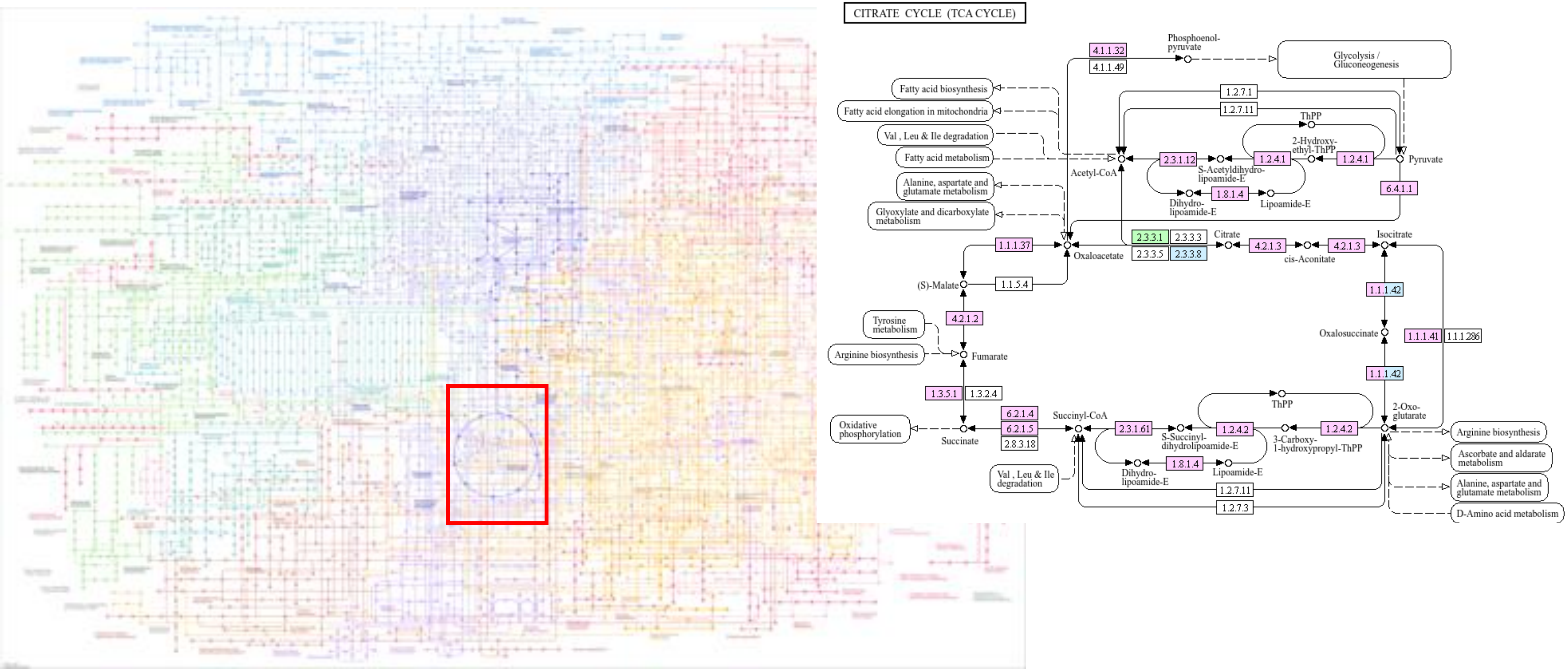
<< first < prev **1** next > last >> 100

Showing items 1 - 60 of 60

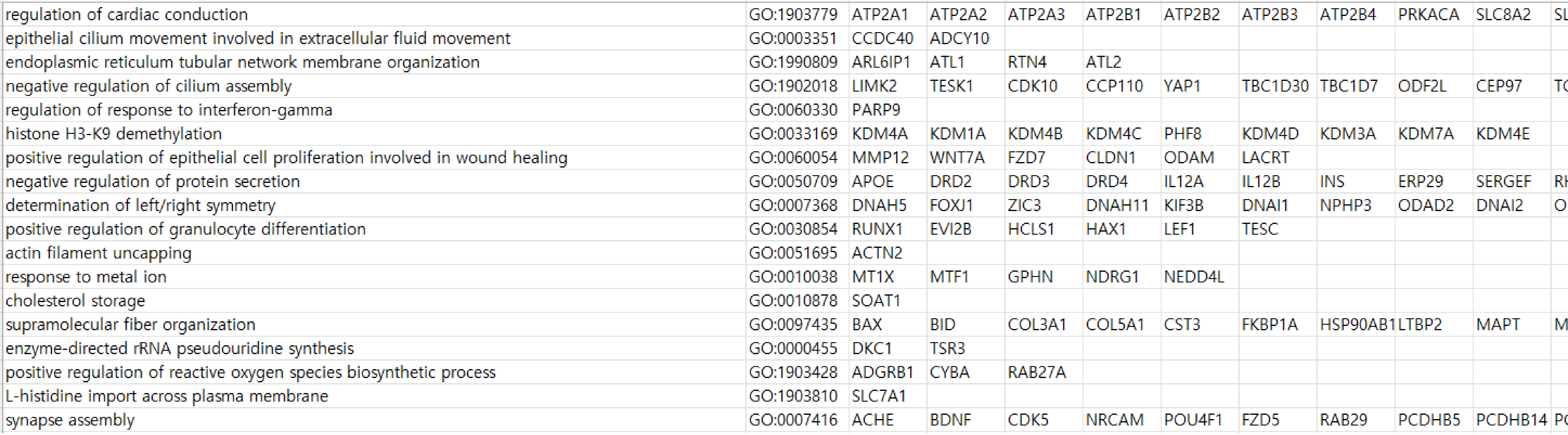
Export: Text File Excel File MouseMine

Symbol, Name	Chr	Annotated Term	Context	Proteoform	Evidence	Inferred From	Reference(s)
Arg2, arginase type II	12	negative regulation of CD4-positive, alpha-beta T cell proliferation			IMP		J:243479 [PMID:25009204]
Card11, caspase recruitment domain family, member 11	5	CD4-positive, alpha-beta T cell proliferation			IMP	MGI:3039682	J:89322 [PMID:12867038]
Card11, caspase recruitment domain family, member 11	5	positive regulation of CD4-positive, alpha-beta T cell proliferation	positively regulates CD4-positive, alpha-beta T cell proliferation.		IMP	MGI:3039682	J:89322 [PMID:12867038]
Cblb, Casitas B-lineage lymphoma b	16	CD4-positive, alpha-beta T cell proliferation			IMP	MGI:2180572	J:89097 [PMID:14973438]
Cblb, Casitas B-lineage lymphoma b	16	negative regulation of CD4-positive, alpha-beta T cell proliferation	negatively regulates CD4-positive, alpha-beta T cell proliferation.		IMP	MGI:2180572	J:89097 [PMID:14973438]
Cd3e, CD3 antigen, epsilon polypeptide	9	CD4-positive, alpha-beta T cell proliferation			IDA		J:17350 [PMID:8125140]
Cd3e, CD3 antigen, epsilon polypeptide	9	CD4-positive, alpha-beta T cell proliferation			IDA		J:75401 [PMID:11894097]
Cd3e, CD3 antigen, epsilon polypeptide	9	CD4-positive, alpha-beta T cell proliferation			IDA		J:89097 [PMID:14973438]
Cd3e, CD3 antigen, epsilon polypeptide	9	positive regulation of CD4-positive, alpha-beta T cell proliferation	positively regulates CD4-positive, alpha-beta T cell proliferation.		IDA		J:17350 [PMID:8125140]
Cd3e, CD3 antigen, epsilon polypeptide	9	positive regulation of CD4-positive, alpha-beta T cell proliferation	positively regulates CD4-positive, alpha-beta T cell proliferation.		IDA		J:75401 [PMID:11894097]
		positive regulation of CD4-	positively regulates CD4-positive,				

- KEGG (Kyoto Encyclopedia of Genes and Genomes)

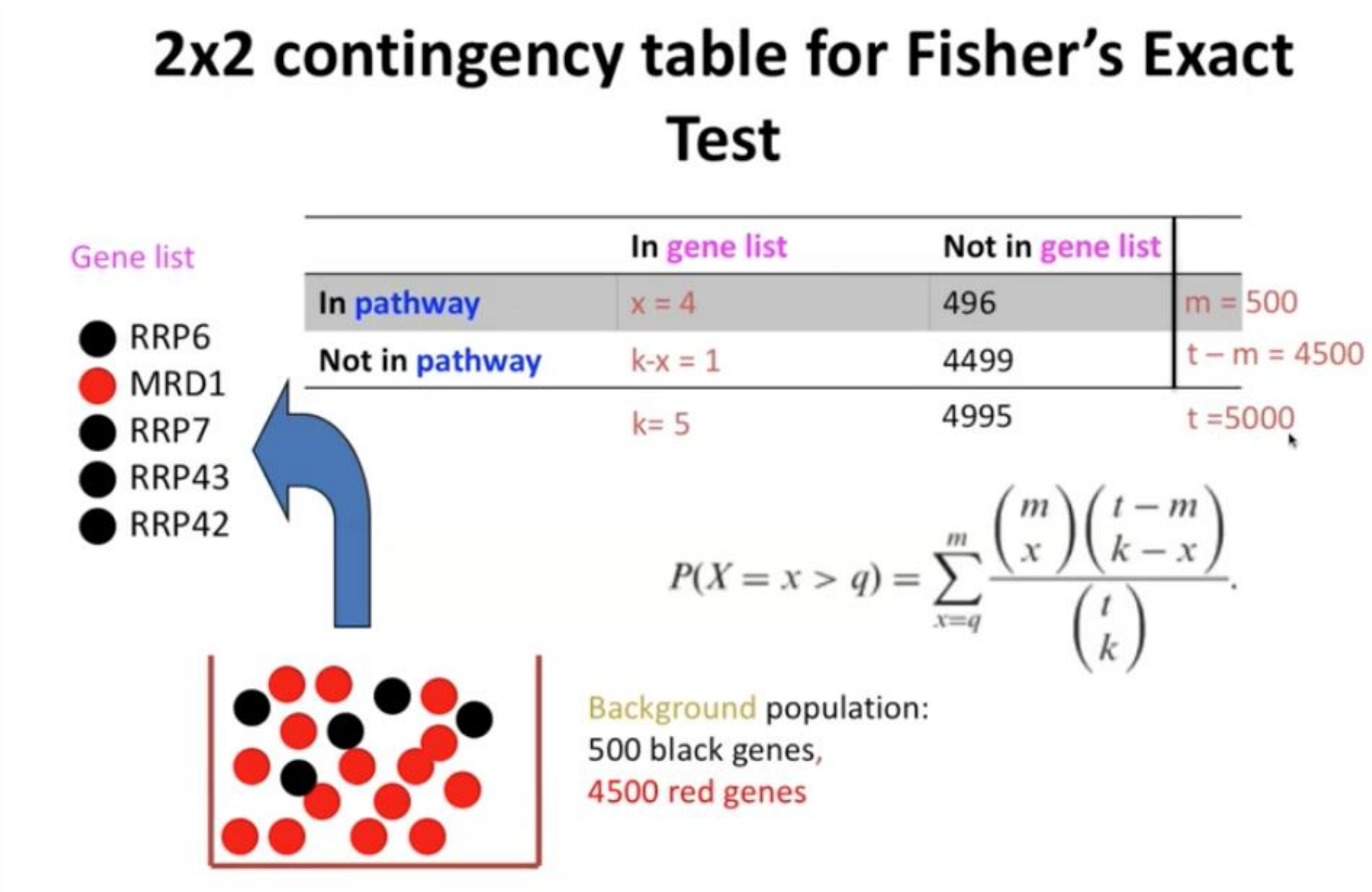


- GMT format



regulation of cardiac conduction	GO:1903779	ATP2A1	ATP2A2	ATP2A3	ATP2B1	ATP2B2	ATP2B3	ATP2B4	PRKACA	SLC8A2	SL
epithelial cilium movement involved in extracellular fluid movement	GO:0003351	CCDC40	ADCY10								
endoplasmic reticulum tubular network membrane organization	GO:1990809	ARL6IP1	ATL1	RTN4	ATL2						
negative regulation of cilium assembly	GO:1902018	LIMK2	TESK1	CDK10	CCP110	YAP1	TBC1D30	TBC1D7	ODF2L	CEP97	TC
regulation of response to interferon-gamma	GO:0060330	PARP9									
histone H3-K9 demethylation	GO:0033169	KDM4A	KDM1A	KDM4B	KDM4C	PHF8	KDM4D	KDM3A	KDM7A	KDM4E	
positive regulation of epithelial cell proliferation involved in wound healing	GO:0060054	MMP12	WNT7A	FZD7	CLDN1	ODAM	LACRT				
negative regulation of protein secretion	GO:0050709	APOE	DRD2	DRD3	DRD4	IL12A	IL12B	INS	ERP29	SERGEF	RH
determination of left/right symmetry	GO:0007368	DNAH5	FOXJ1	ZIC3	DNAH11	KIF3B	DNAI1	NPHP3	ODAD2	DNAI2	O
positive regulation of granulocyte differentiation	GO:0030854	RUNX1	EVI2B	HCLS1	HAX1	LEF1	TESC				
actin filament uncapping	GO:0051695	ACTN2									
response to metal ion	GO:0010038	MT1X	MTF1	GPHN	NDRG1	NEDD4L					
cholesterol storage	GO:0010878	SOAT1									
supramolecular fiber organization	GO:0097435	BAX	BID	COL3A1	COL5A1	CST3	FKBP1A	HSP90AB1	LTBP2	MAPT	M
enzyme-directed rRNA pseudouridine synthesis	GO:0000455	DKC1	TSR3								
positive regulation of reactive oxygen species biosynthetic process	GO:1903428	ADGRB1	CYBA	RAB27A							
L-histidine import across plasma membrane	GO:1903810	SLC7A1									
synapse assembly	GO:0007416	ACHE	BDNF	CDK5	NRCAM	POU4F1	FZD5	RAB29	PCDHB5	PCDHB14	PC

- Geneset analysis
  - ❖ ***Hypergeometric test*** (also known as ***Fisher's exact test***)
    - **Null hypothesis:** Observed list is a random sample from population.
    - **Alternative hypothesis:** More black genes than expected in my list.



2x2 contingency table for Fisher's Exact Test

	In gene list	Not in gene list	
In pathway	x = 4	496	m = 500
Not in pathway	k-x = 1	4499	t – m = 4500
	k= 5	4995	t =5000

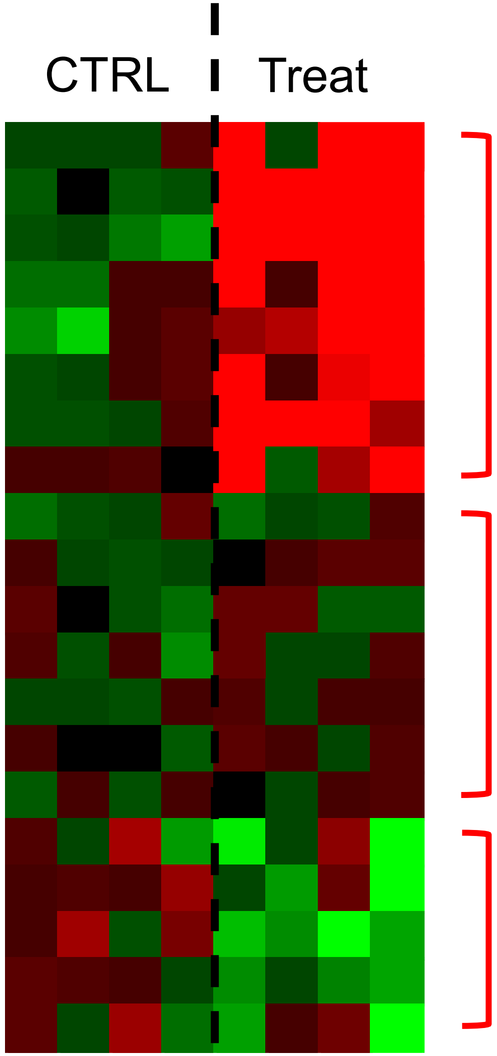
$$P(X = x > q) = \sum_{x=q}^{m} \frac{\binom{m}{x}\binom{t-m}{k-x}}{\binom{t}{k}}.$$

Gene list

- ● RRP6
- 🔴 MRD1
- ● RRP7
- ● RRP43
- ● RRP42

Background population:
500 black genes,
4500 red genes

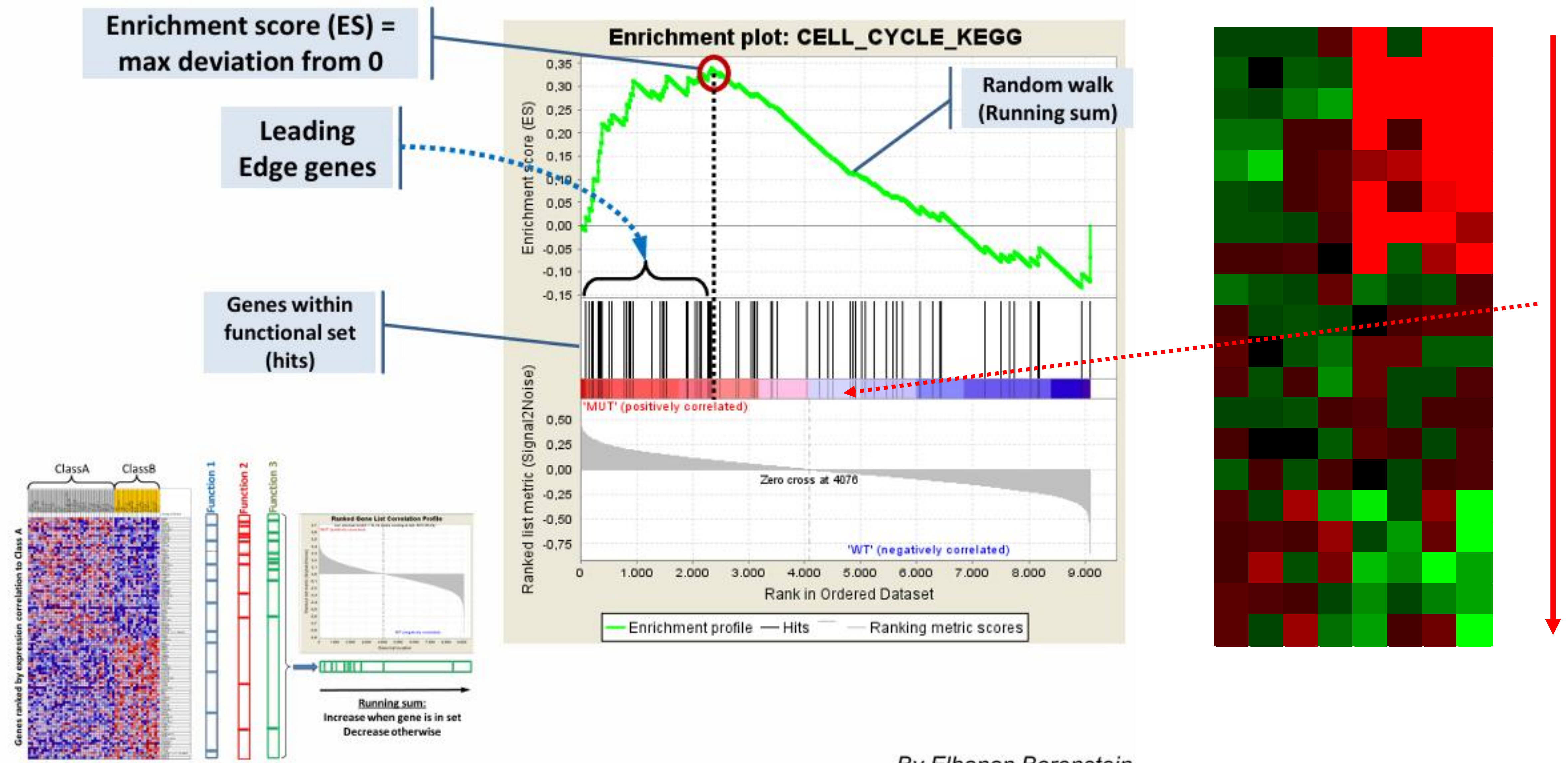
- Geneset analysis



Up-regulated genes → Fisher's exact test
→ More genes detected from "Immune activation"

- GSEA (Geneset enrichment analysis)

Adjusting gene weight (Fold change, Gene expression, etc)



Enrichment score (ES) = max deviation from 0

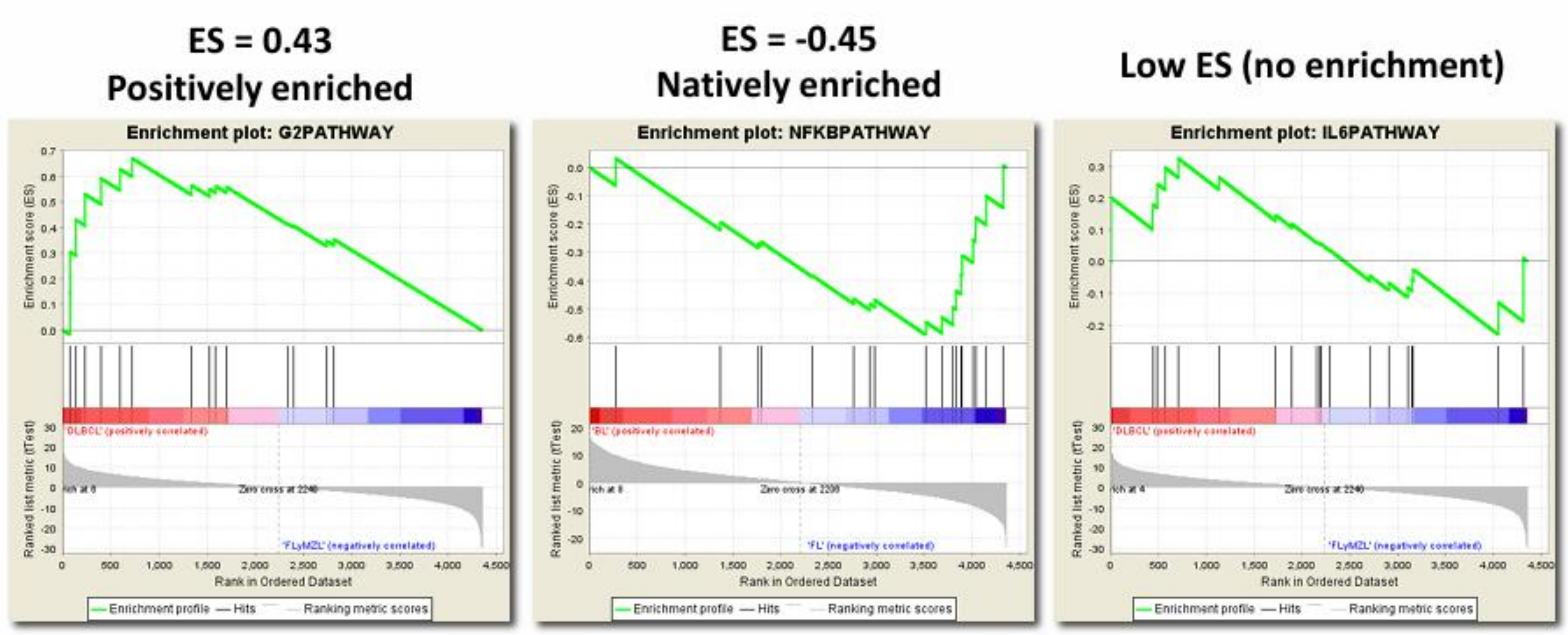
Leading Edge genes

Genes within functional set (hits)

Random walk (Running sum)

By Elhanan Borenstein

- GSEA (Geneset enrichment analysis)



ES = 0.43
Positively enriched

ES = -0.45
Natively enriched

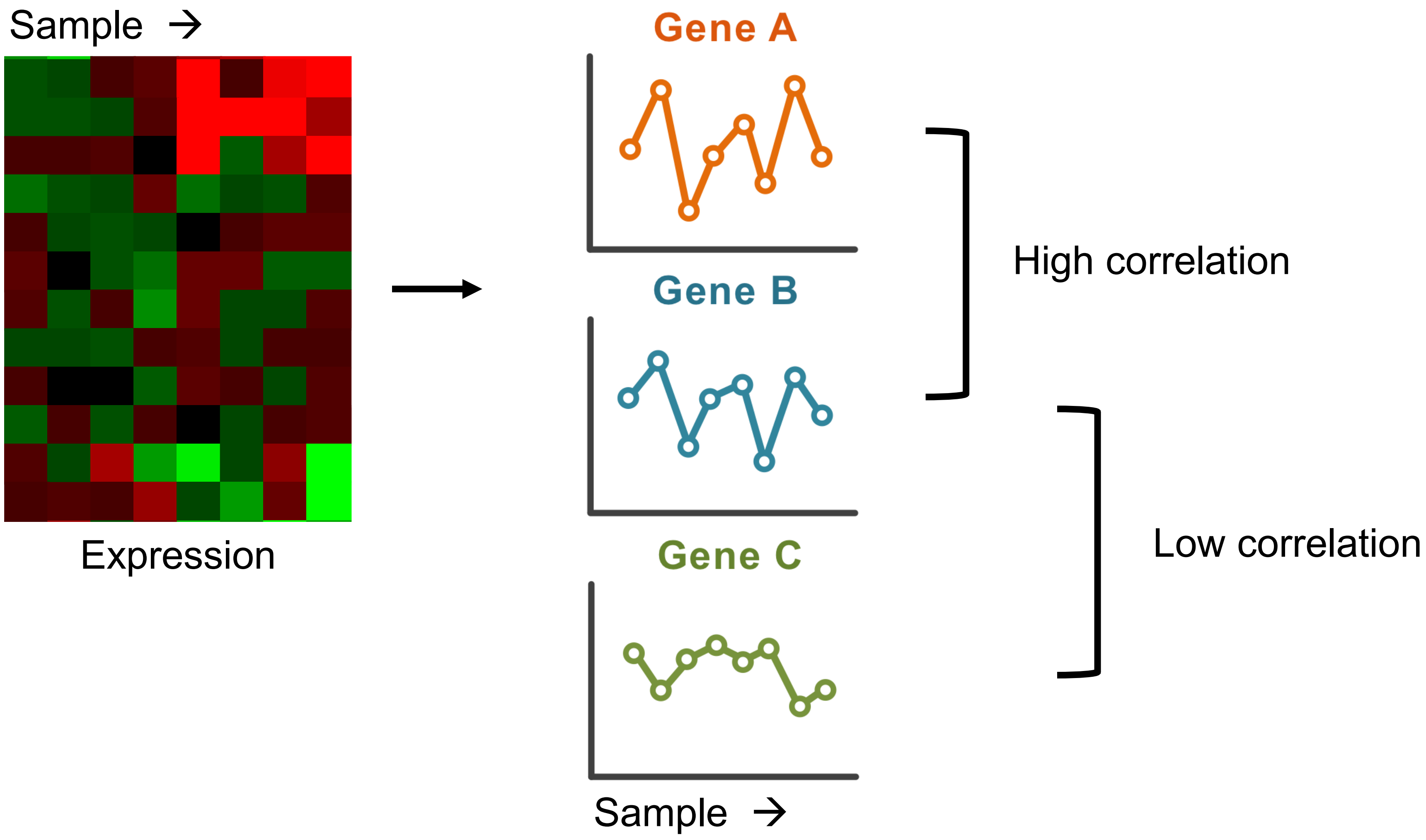
Low ES (no enrichment)

Enrichment plot: G2PATHWAY

Enrichment plot: NFKBPATHWAY

Enrichment plot: IL6PATHWAY

- Network analysis

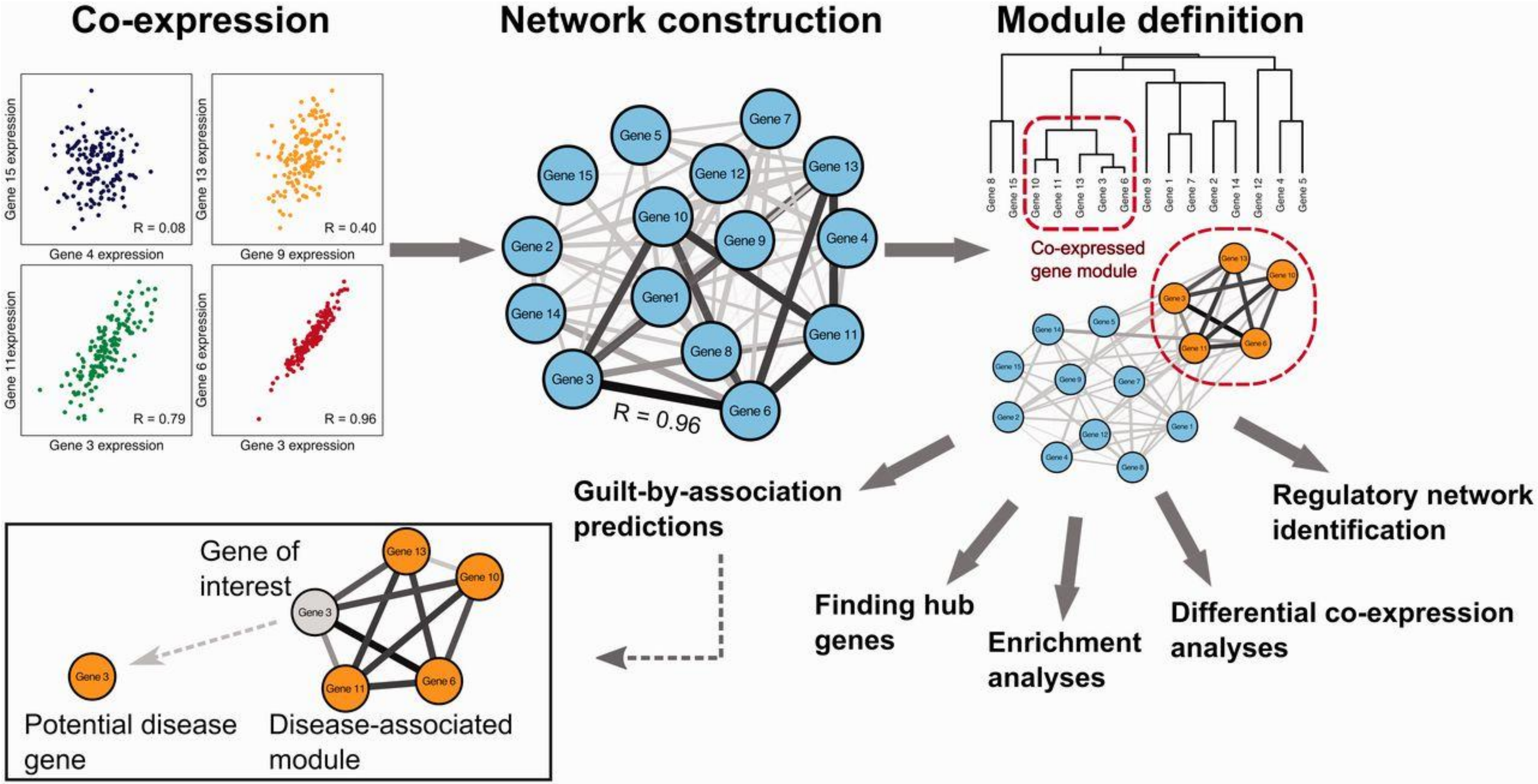


Gene pair: correlation (Pearson correlation coefficient: PCC / Spearman correlation coefficient (SCC)

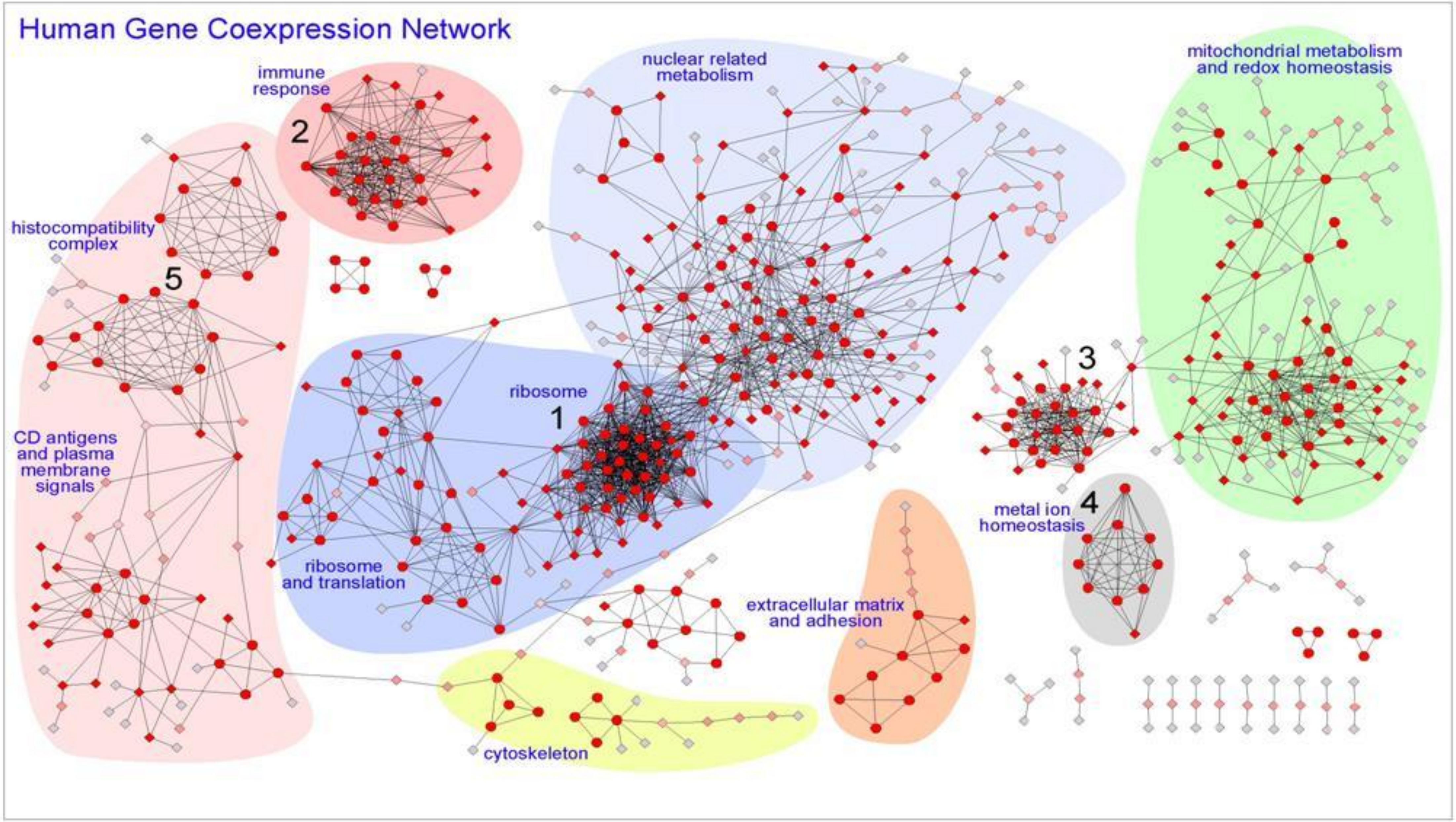
$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

SCC: rank-based (non-parametric)

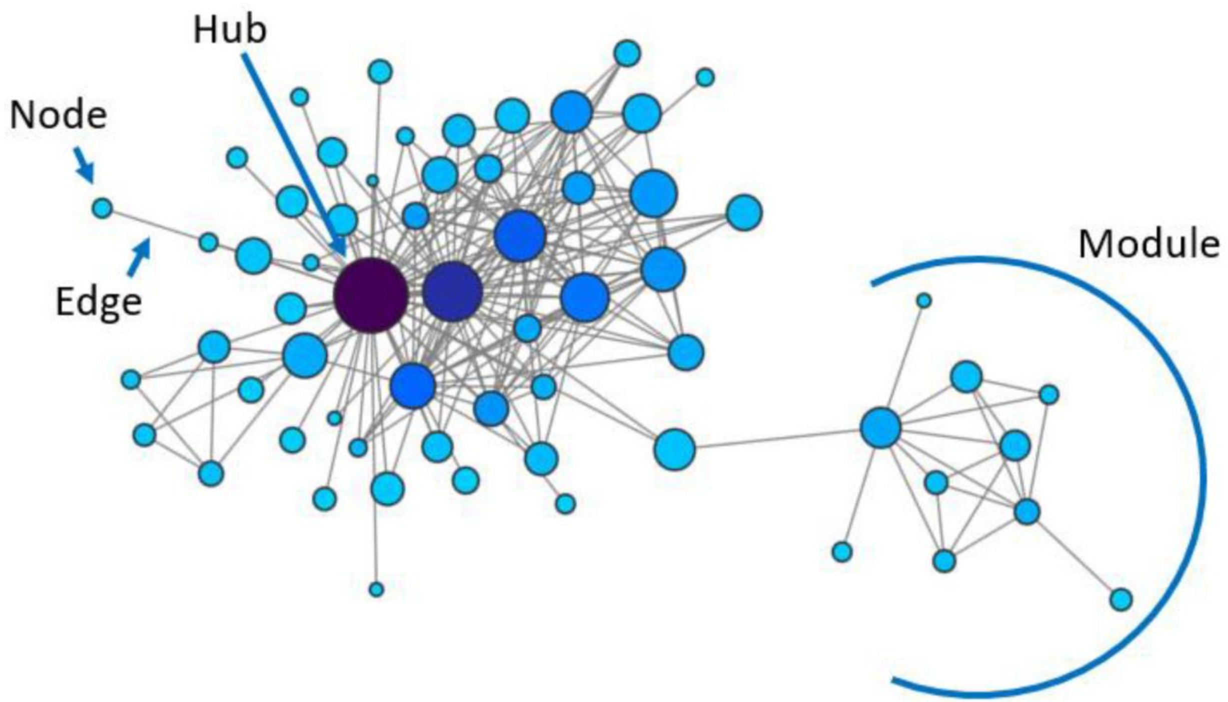
- Network analysis



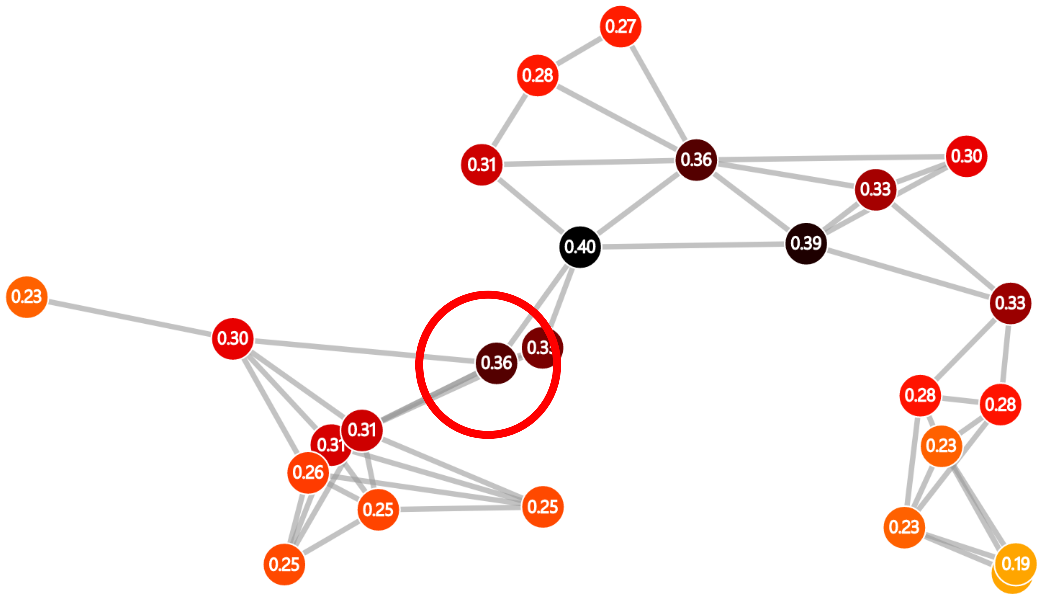
- Network analysis


Human Gene Coexpression Network
immune response
2
nuclear related metabolism
mitochondrial metabolism and redox homeostasis
histocompatibility complex
5
ribosome
1
3
CD antigens and plasma membrane signals
metal ion homeostasis
4
ribosome and translation
extracellular matrix and adhesion
cytoskeleton

- Network analysis



Centrality by node degree



Centrality by betweenness

-Node degree: number of edges for each node → highest
-Betweenness: find shortest path for each node pair → sort by how many shortest paths pass each node → highest

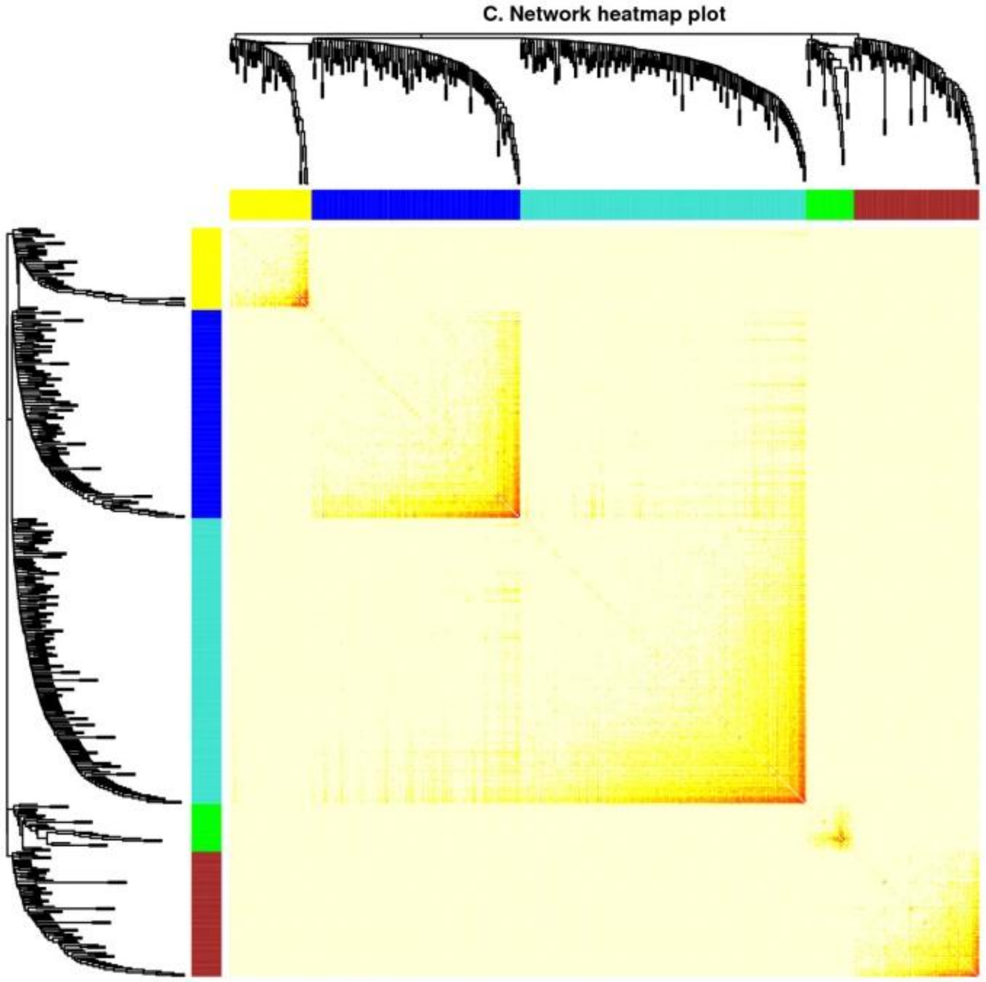
- WGCNA: Weighted Gene Co-expression Network Analysis

Gene pair → correlation
→ Weight * corr → thresholding
→ Hierarchical clustering, tree cutting

Gene expression program / module

Weight: until it maintains scale-free topology

Scale-free topology
-only some of nodes have most of the edges
-edge follows power law


C. Network heatmap plot