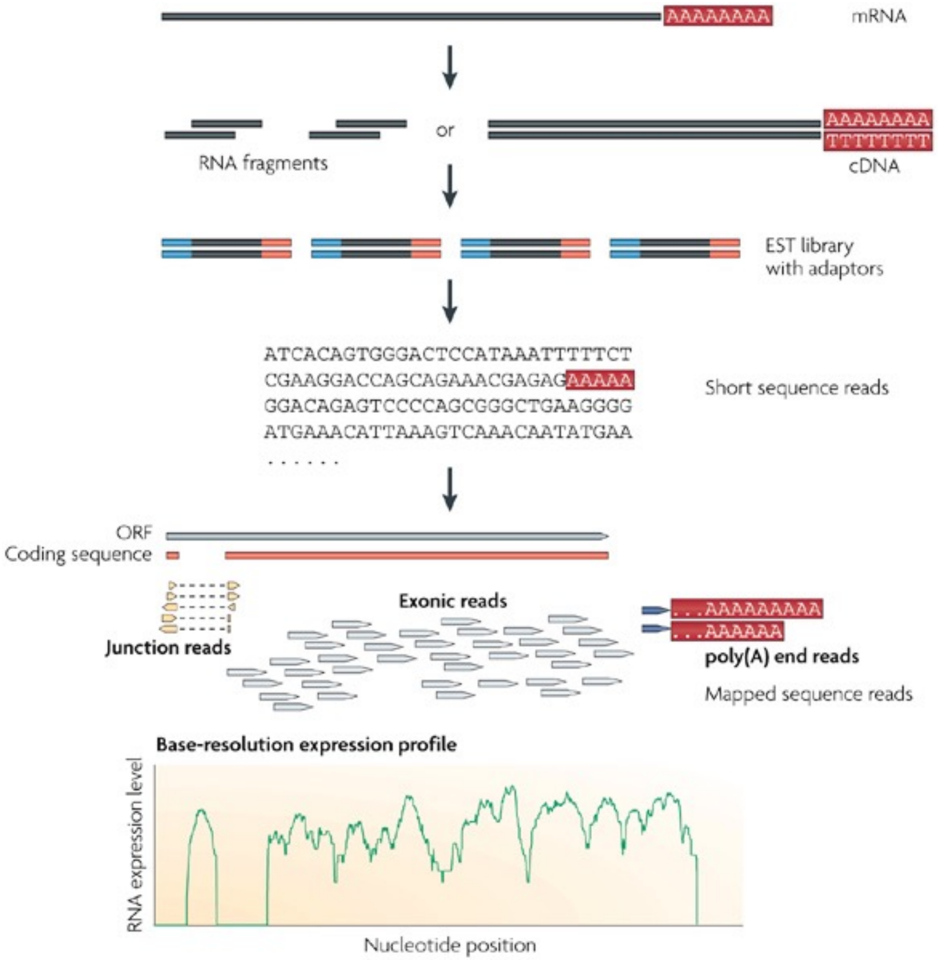
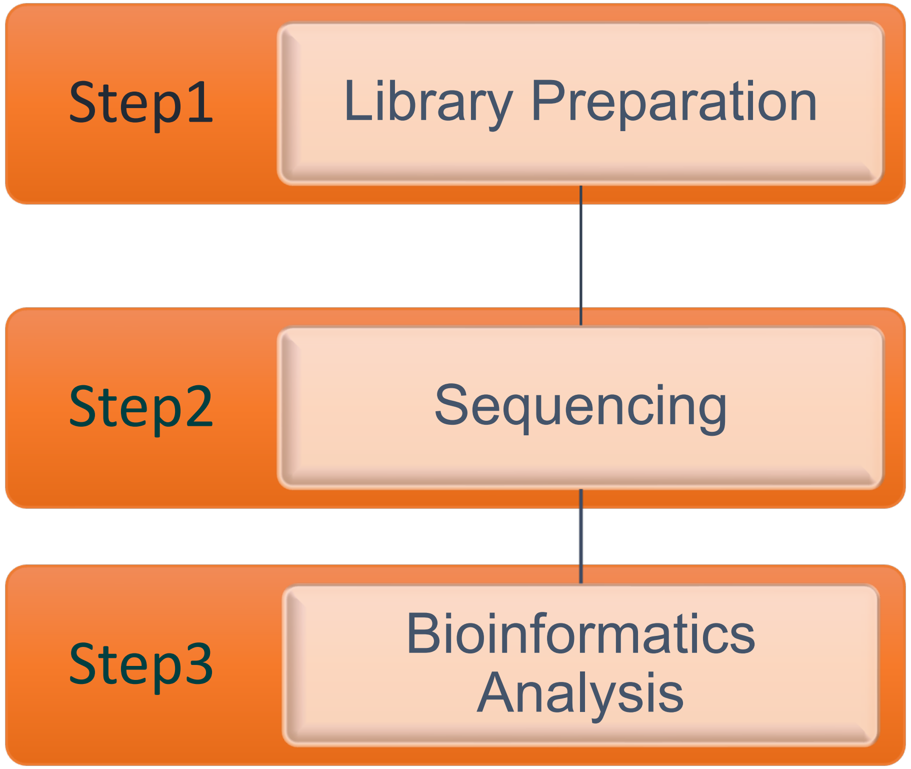
Bulk RNA-seq analysis and pathway analysis

Chia-Ni Hsiung

National Tsing Hua University

RNA-seq workflow





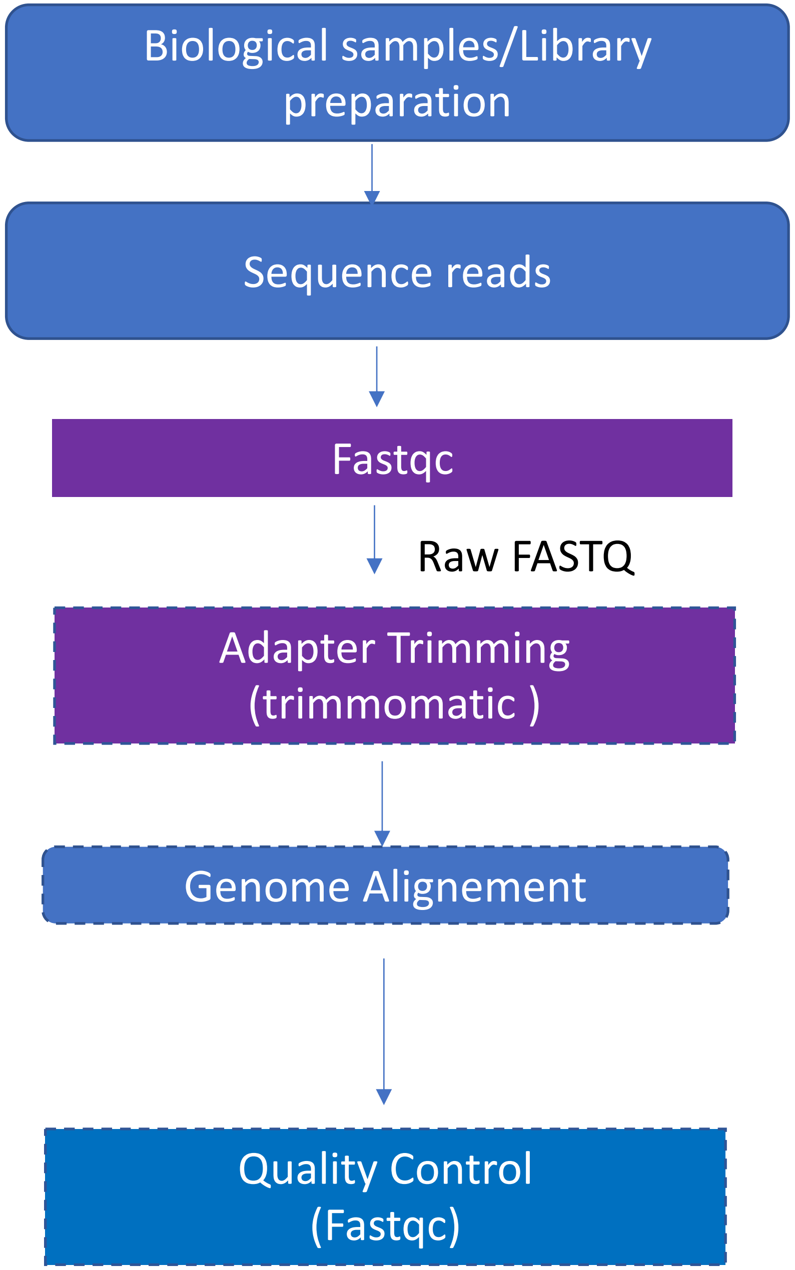
Step1 Library Preparation

Step2 Sequencing

Step3 Bioinformatics Analysis

Nature Reviews | Genetics

Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.



Biological samples/Library preparation

Sequence reads

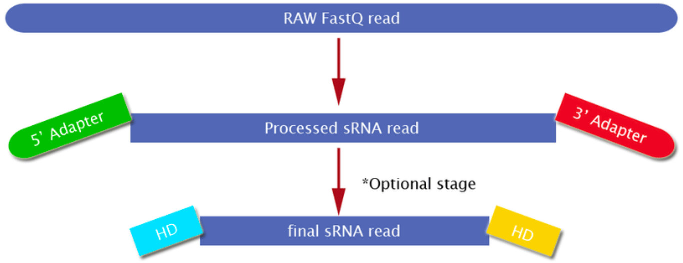
Fastqc

Raw FASTQ

Adapter Trimming (trimmomatic)

Genome Alignement

Quality Control (Fastqc)



Benefits and opportunities of RNA-seq

- Whole transcriptome sequencing
  - Annotation of new exons, transcribed regions, genes or non-coding RNAs
  - The ability to look at alternative splicing
  - Allele specific expression
  - RNA editing
  - Differential expression

Fastqc

- Provide not only the problem from the sequencer, but the sample library
- Main functions
  - Import of data from BAM, SAM or FastQ files (any variant)
  - Providing a quick overview to tell you in which areas there may be problems
  - Summary graphs and tables to quickly assess your data
  - Export of results to an HTML based permanent report
  - Offline operation to allow automated generation of reports without running the interactive application

Quality score

- Q = -10 x log10(P)

where P is the probability that a base call is erroneous

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

We expected quality score is less than 20

Fastqc example

Expect

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

Bad trimmed adapter

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/RNA-Seq_fastqc.html

RNAseq Software

1. Short Read Alignment

➢	STAR	https://github.com/alexdobin/STAR/releases
➢	HISAT2	https://ccb.jhu.edu/software/hisat2/index.shtml

2. Read counting

➢	HTseq	http://www-huber.embl.de/HTSeq/doc/overview.html
➢	SAMtools	http://www.htslib.org/

3. Differential Expression

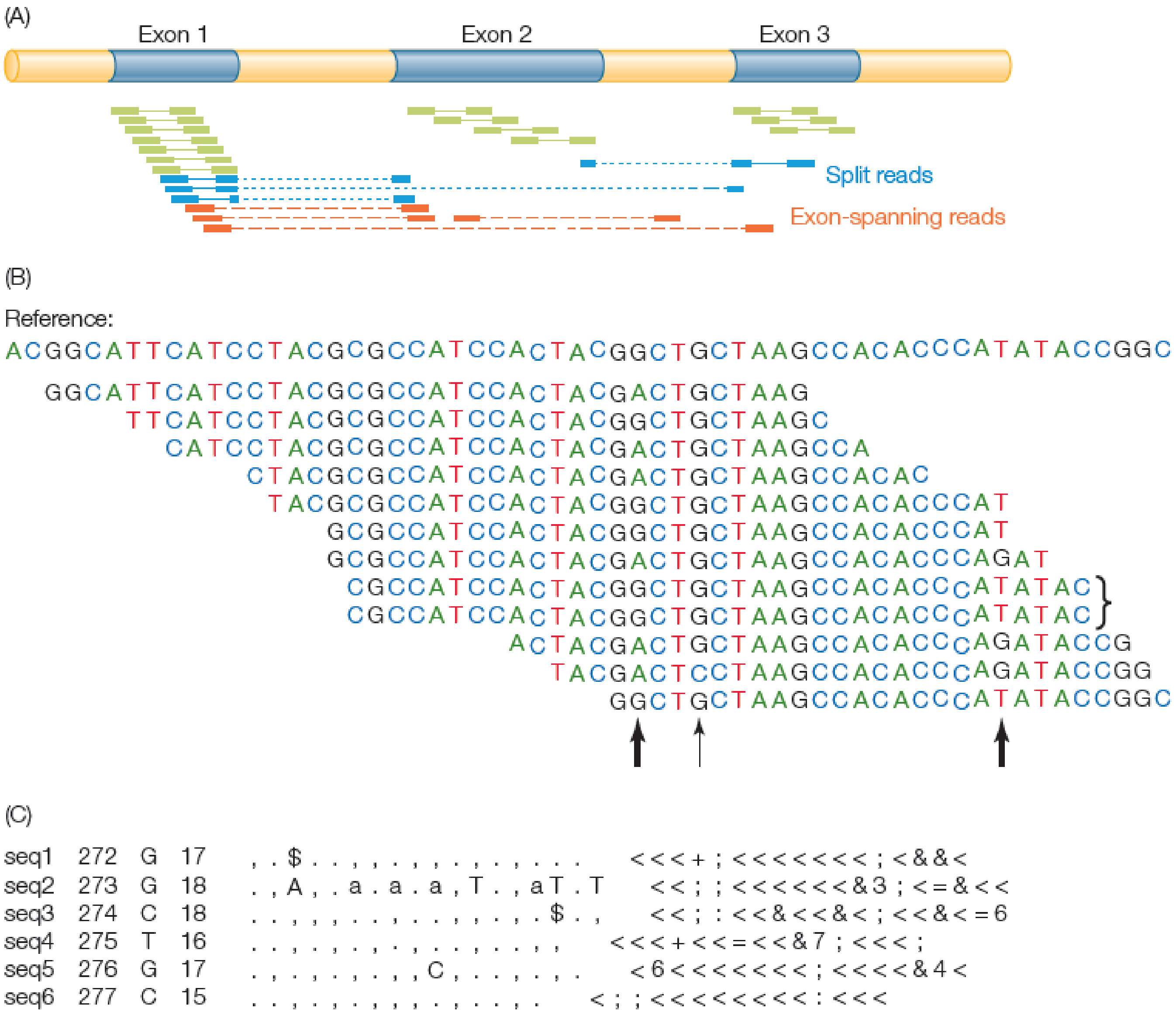
➢	DESeq	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
➢	DExSeq	https://www.bioconductor.org/packages/release/bioc/html/DEXSeq.html
➢	edgeR	https://bioconductor.org/packages/release/bioc/html/edgeR.html
➢	Voom	http://web.mit.edu/~r/current/arch/i386_linux26/lib/R/library/limma/html/voom.html

4. Data Normalization

➢	SVASeq	https://www.bioconductor.org/packages/release/bioc/html/sva.html
➢	Combat	https://www.rdocumentation.org/packages/sva/versions/3.20.0/topics/ComBat
➢	PEER	http://www.sanger.ac.uk/science/tools/peer
➢	SNM	https://www.bioconductor.org/packages/release/bioc/html/snm.html

Another option is the Tuxedo protocol (Bowtie, Tophat, Cufflinks, Cuffdiff,
https://ugene.net/wiki/display/WDD31/RNA-seq+Analysis+with+Tuxedo+Tools

Read Alignment



Reference:

```
ACGGCATTCATCCTACGCGCCATCCACTACGGCTGCTAAGCCACACCCATATACCGGC

  GGCATTCATCCTACGCGCCATCCACTACGACTGCTAAG
      TTCATCCTACGCGCCATCCACTACGGCTGCTAAGC
        CATCCTACGCGCCATCCACTACGACTGCTAAGCCA
            CTACGCGCCATCCACTACGACTGCTAAGCCACAC
             TACGCGCCATCCACTACGGCTGCTAAGCCACACCCAT
                GCGCCATCCACTACGGCTGCTAAGCCACACCCAT
                GCGCCATCCACTACGACTGCTAAGCCACACCCAGAT
                 CGCCATCCACTACGGCTGCTAAGCCACACCCATATAC}
                 CGCCATCCACTACGGCTGCTAAGCCACACCCATATAC}
                         ACTACGACTGCTAAGCCACACCCAGATACCG
                           TACGACTCCTAAGCCACACCCAGATACCGG
                              GGCTGCTAAGCCACACCCATATACCGGC
```

```
seq1   272   G   17    , . $ . . , , . , , . , , . . . .    < < < + ; < < < < < < < ; < & & <
seq2   273   G   18    . , A , . a . a . a , T . , a T . T    < < ; ; < < < < < < & 3 ; < = & < <
seq3   274   C   18    . . , , , , . . , . . . , , . , . $ . ,    < < ; : < < & < < & < ; < < & < = 6
seq4   275   T   16    . . , . . , . . , . . , . . , ,    < < < + < < = < < & 7 ; < < < ;
seq5   276   G   17    . , . . , . . , . , C , . . . , . .    < 6 < < < < < < < ; < < < < & 4 <
seq6   277   C   15    . . , . , . , , . , , . , . . , .    < ; ; < < < < < < < < : < < <
```

Basics of Experimental Design: Levels of Replication

Often you will have a fixed budget that constrains how many arrays can be processed. So your first task is to determine what levels of replication you can afford, and how they will impact statistical power.

Technical Replication:

- RNA preparation (eg. from adjacent biopsies)
- cDNA synthesis (pooling minimizes outlier effects)
- library preparation
- sequencing lane or array hybridization (usually a minimal effect)

Biological Replication:

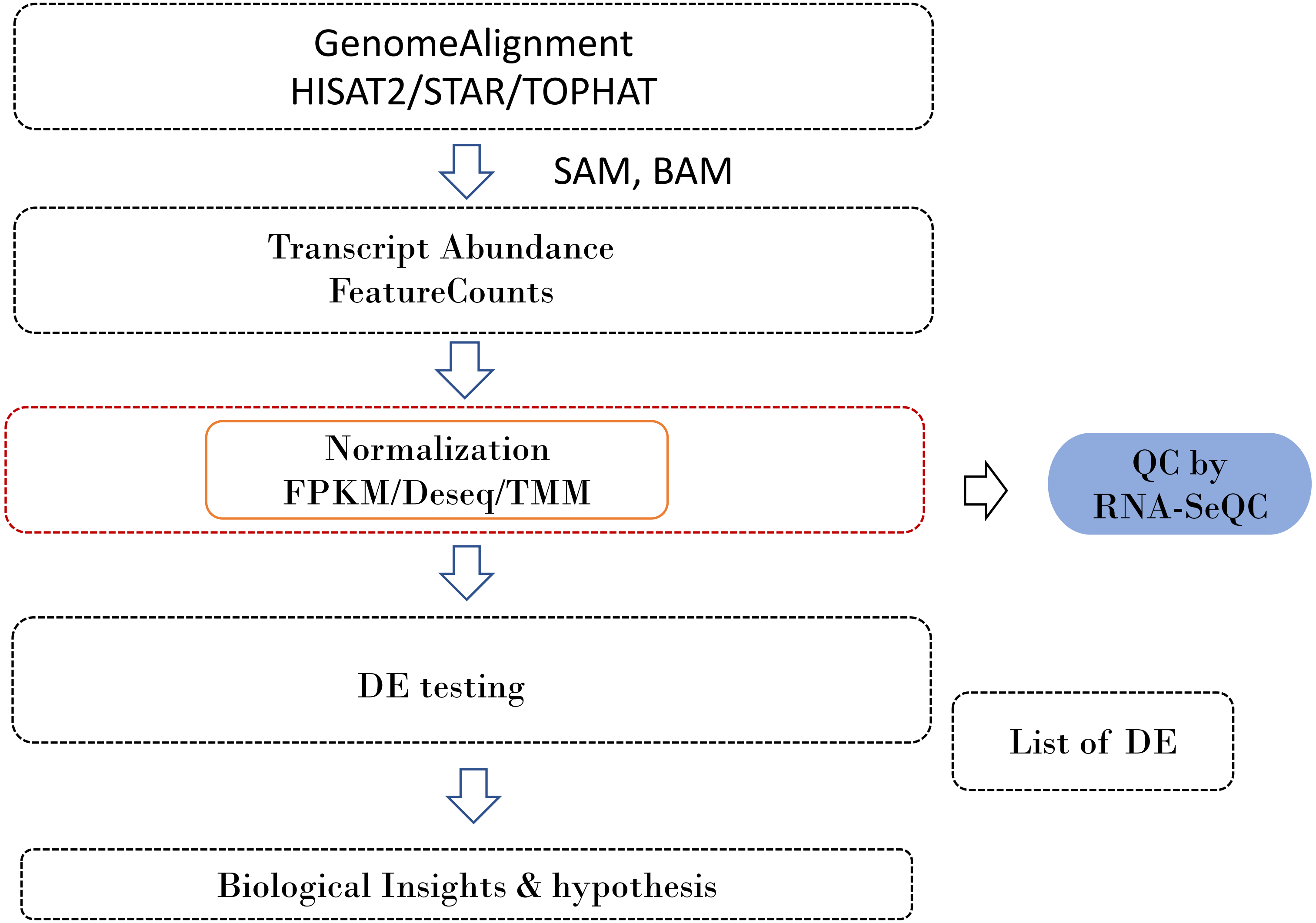
Fixed effects:

- sex
- treatment (drug, growth regimen, tissue)
- time of sampling (repeated measures in some cases)
- genotype (IF specifically chosen and resampled)

Random effects

- individual from a population
- field plot

From reads to differential expression



GenomeAlignment
HISAT2/STAR/TOPHAT

SAM, BAM

Transcript Abundance
FeatureCounts

Normalization
FPKM/Deseq/TMM

QC by
RNA-SeQC

DE testing

List of DE

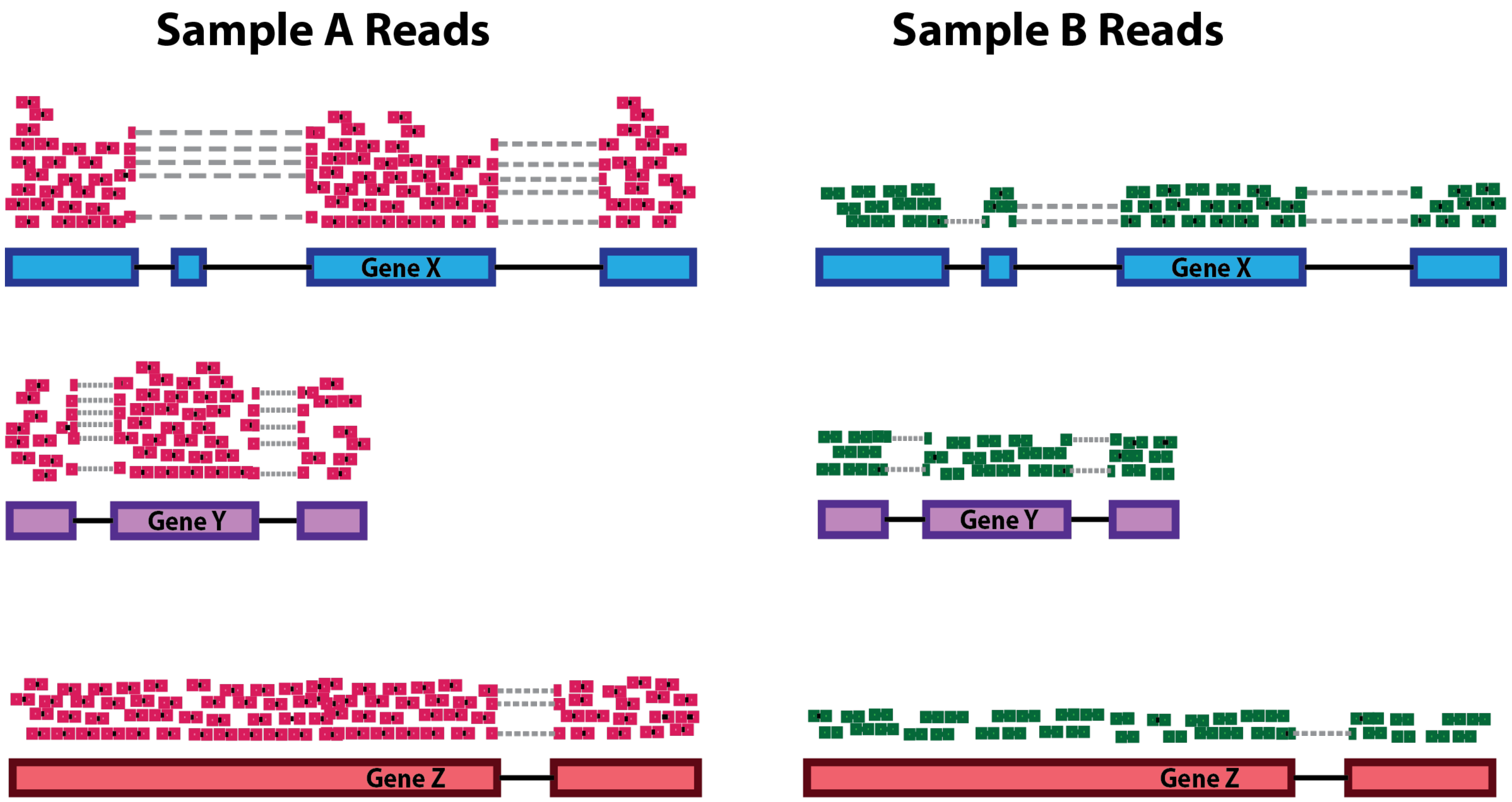
Biological Insights & hypothesis

Aim of normalization

- Normalization aims to ensure our expression estimates are:
    - comparable across features (genes, isoforms, etc)
    - comparable across libraries (different samples)
    - on a human-friendly scale (interpretable magnitude)

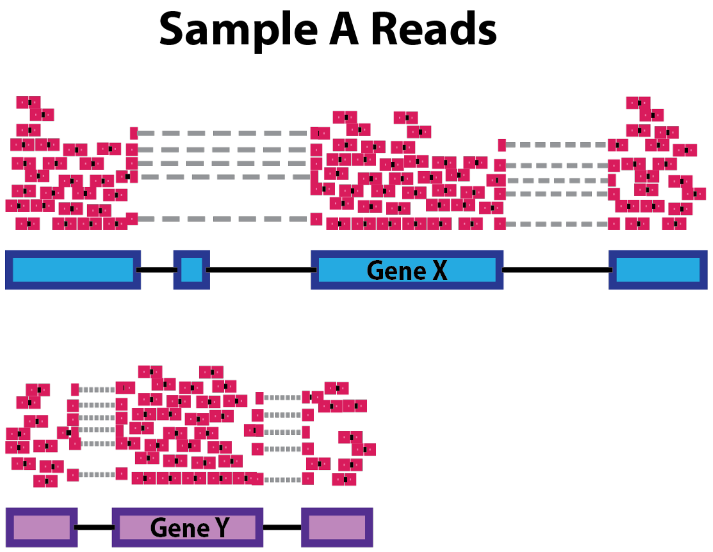
The main factors considered during normalization

- **Sequencing depth**

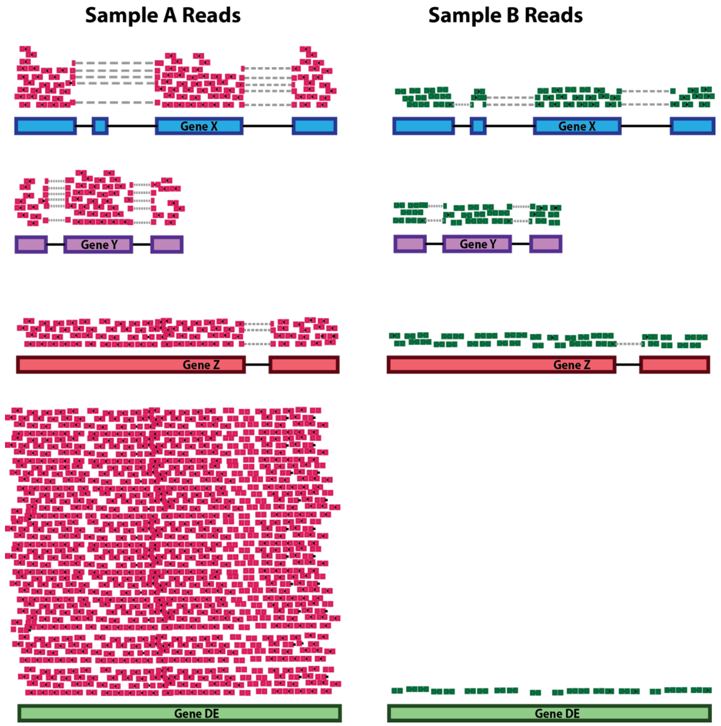


__NOTE:__ In the figure above, each pink and green rectangle represents a read aligned to a gene. Reads connected by dashed lines connect a read spanning an intron.

- **Gene length**



- RNA composition



Normalization method

Normalization method	Description	Accounted factors	Recommendations for use	Between/ within
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same samplegroup; **NOT for within sample comparisons or DE analysis**	
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**	Within
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**	Within
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**	Between
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for **DE analysis**	Between

RPKM normalization
(Reads Per Kilobase per Million)

- $RPKM = \frac{r_g}{fl_g \times R} \times 10^9$

r_g :No. of gene reads

R : Total number of reads

Flg : gene length

FPKM

- FPKM= $\frac{f_g}{fl_g \times R} \times 10^9$

f_g :No. of gene fragments

flg : Length of gene

R: total reads counts

Example

	Replicate 1	Replicate 2	Replicate 3
Gene A (2kb)	10,000,000	12,000,000	30,000,000
Gene B (4kb)	20,000,000	25,000,000	60,000,000
Gene C (1kb)	5,000,000	8,000,000	15,000,000
Gene D (10kb)	0	0	1,000,000
Sum	35,000,000	45,000,000	106,000,000

$$\text{RPKM} = \frac{\text{total exone reads}}{mapped\ reads\ (millions) * exon\ length\ (KB)} = \frac{10{,}000{,}000}{(10+20+5)*2} = 142857$$

RPKM/FPKM limitation

- Limitation

Using RPKM/FPKM normalization, the total number of RPKM/FPKM normalized counts for each sample will be different. Therefore, you cannot compare the normalized counts for each gene equally between samples

RPKM-normalized counts table

gene	sampleA	sampleB
XCR1	5.5	5.5
WASHC1	73.4	21.8
…	…	…
Total RPKM-normalized counts	1,000,000	1,500,000

Between sample normalization

- To improvement the samples compare

- Methods
  - TMM (Trimmed mean of M-values)
  - DeSeq

Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

Assumptions behind the method

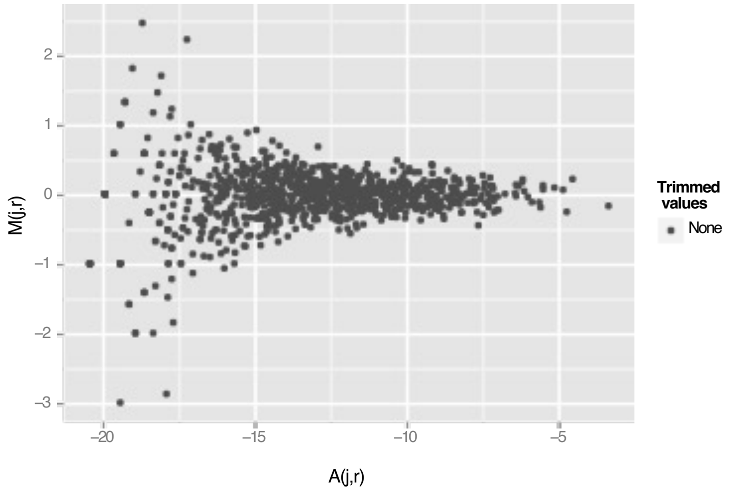
- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

⇒ remove extreme data for fold-changed (M) and average intensity (A)

$$M_g(j,r) = \log_2 \frac{K_{gj}}{D_j} - \log_2 \frac{K_{gr}}{D_r} \qquad A_g(j,r) = \frac{1}{2} \log_2 \frac{K_{gj}}{D_j} + \log_2 \frac{K_{gr}}{D_r}$$

select as a reference sample, the sample r with the upper quartile closest to the average upper quartile

M- vs A-values



Trimmed Mean of M-values (TMM)

[Robinson and Oshlack, 2010], **edgeR**

Assumptions behind the method

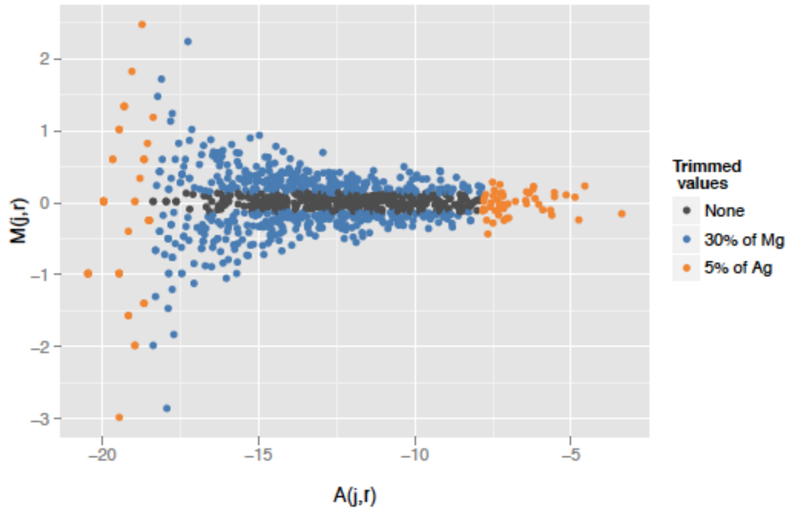
- the total read count strongly depends on a few highly expressed genes
- most genes are not differentially expressed

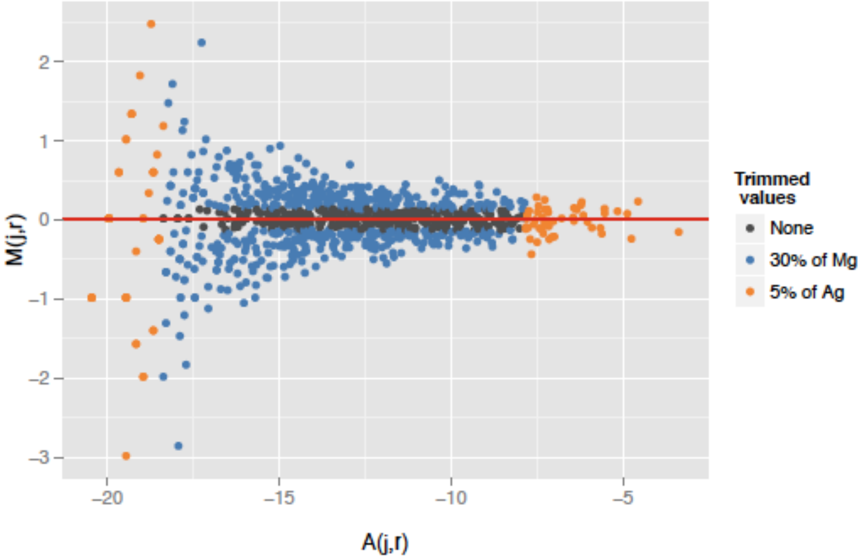
$\Rightarrow$ remove extreme data for fold-changed (M) and average intensity (A)

$$M_g(j,r) = \log_2\left(\frac{K_{gj}}{D_j}\right) - \log_2\left(\frac{K_{gr}}{D_r}\right) \qquad A_g(j,r) = \frac{1}{2}\left[\log_2\left(\frac{K_{gj}}{D_j}\right) + \log_2\left(\frac{K_{gr}}{D_r}\right)\right]$$

Trim 30% on M-values

Trim 5% on A-values





On remaining data, calculate the weighted mean of M-values:

$$\mathrm{TMM}(j,r) = \frac{\sum_{g:\text{not trimmed}} w_g(j,r) M_g(j,r)}{\sum_{g:\text{not trimmed}} w_g(j,r)}$$

$$\text{with } w_g(j,r) = \left(\frac{D_j - K_{gj}}{D_j K_{gj}} + \frac{D_r - K_{gr}}{D_r K_{gr}} \right).$$

Robinson and Oshlack, 2010

```
calcNormFactors(..., method="TMM")
```

Relative Log Expression (RLE) Deseq2

- RLE uses the median of ratios method

Step 1: creates a pseudo-reference sample (row-wise geometric mean)
For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
…	…	…	…

Anders and Huber 2010

RLE

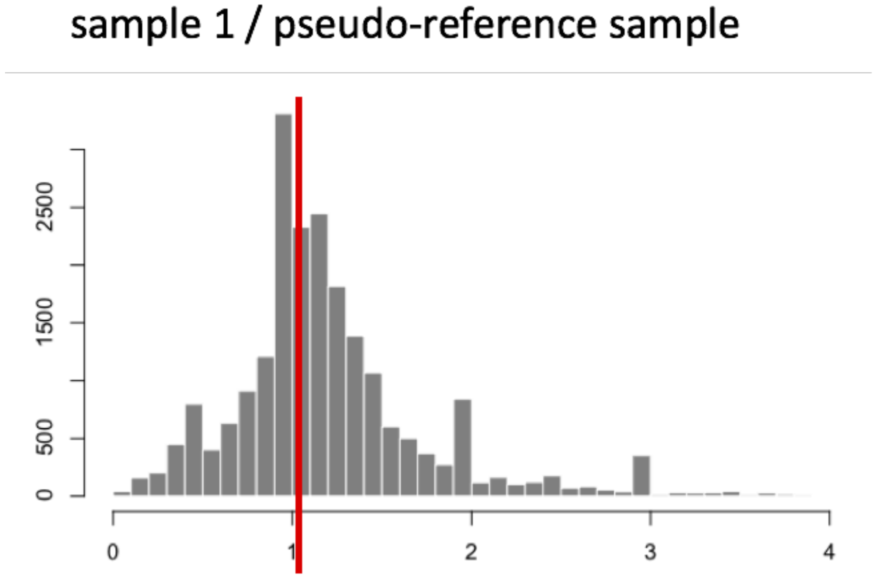
Step 2: calculates ratio of each sample to the reference

For every gene in a sample, the ratios (sample/ref) are calculated (as shown below). This is performed for each sample in the dataset. Since the majority of genes are not differentially expressed, the majority of genes in each sample should have similar ratios within the sample.

gene	**sampleA**	**sampleB**	**pseudo-reference sample**	**ratio of sampleA/ref**	**ratio of sampleB/ref**
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
…	…	…	…		

RLE

- **Step 3: calculate the normalization factor for each sample (size factor)**
- The median value from all genes of all ratios for a given sample is taken as the normalization factor (size factor) for that sample, as calculated below. Notice that the differentially expressed genes should not affect the median value:



sample 1 / pseudo-reference sample

- **Step 4: calculate the normalized count values using the normalization factor**

For example, if the median ratio for SampleA was 1.3 and the median ratio for SampleB was 0.77, you could calculate normalized counts as follows:
SampleA median ratio = 1.3
SampleB median ratio = 0.77

Raw Counts

gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
…	…	…

Normalized Counts

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
…	…	…

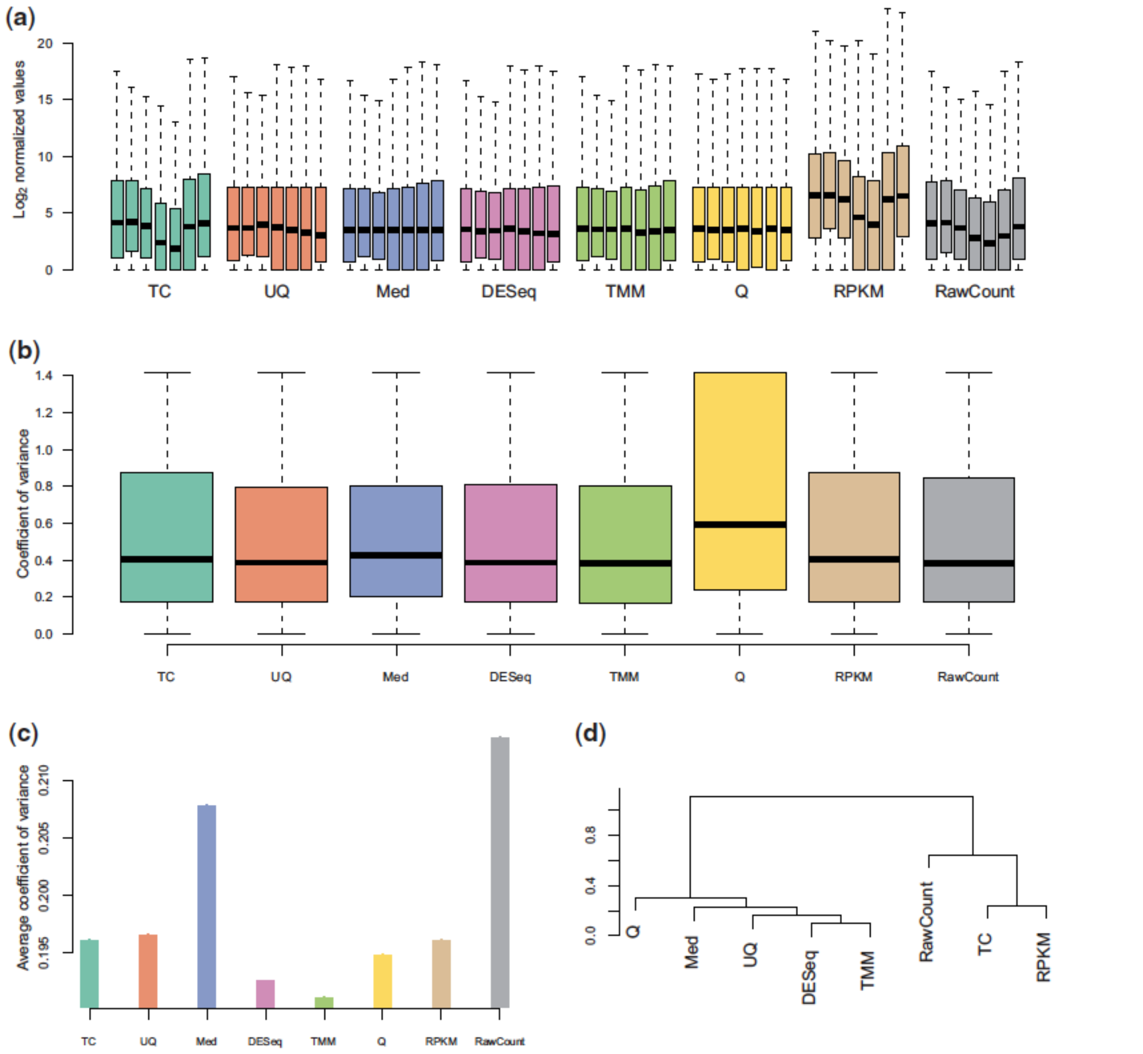
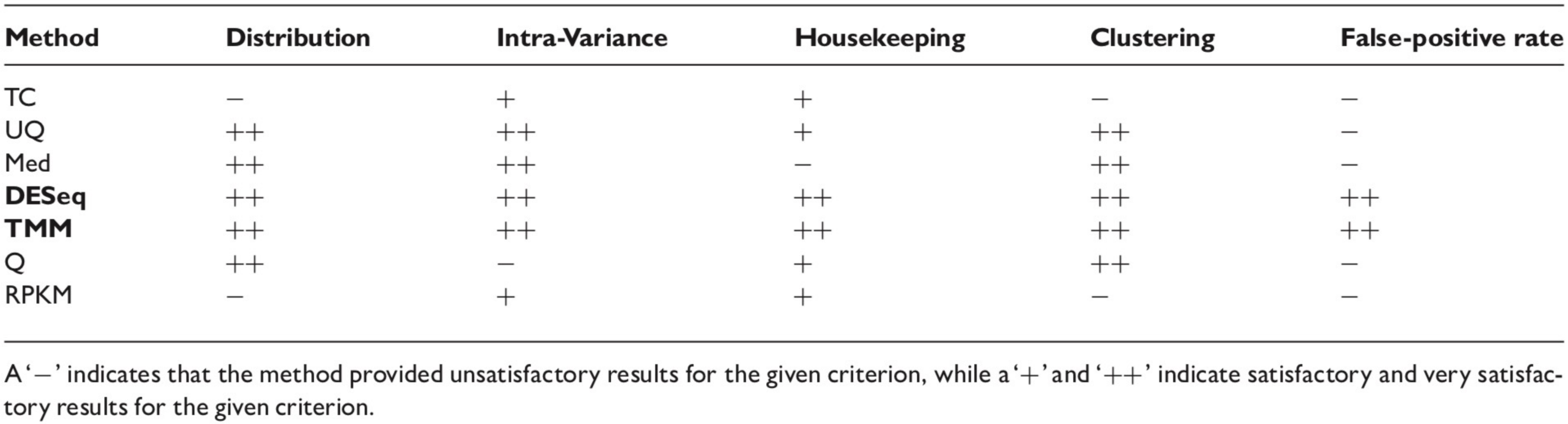


Figure 1: Comparison of normalization methods for real data. (**A**) Boxplots of log2(counts + 1) for all conditions and replicates in the *M. musculus* data, by normalization method. (**B**) Boxplots of intra-group variance for one of the conditions (labeled 'B' in the corresponding data found in Supplementary Data) in the *M. musculus* data, by normalization method. (**C**) Analysis of housekeeping genes for the *H. sapiens* data. (**D**) Consensus dendrogram of differential analysis results, using the **DESeq** Bioconductor package, for all normalization methods across the four datasets under consideration.

Dillies et al. 2012

Normalization result

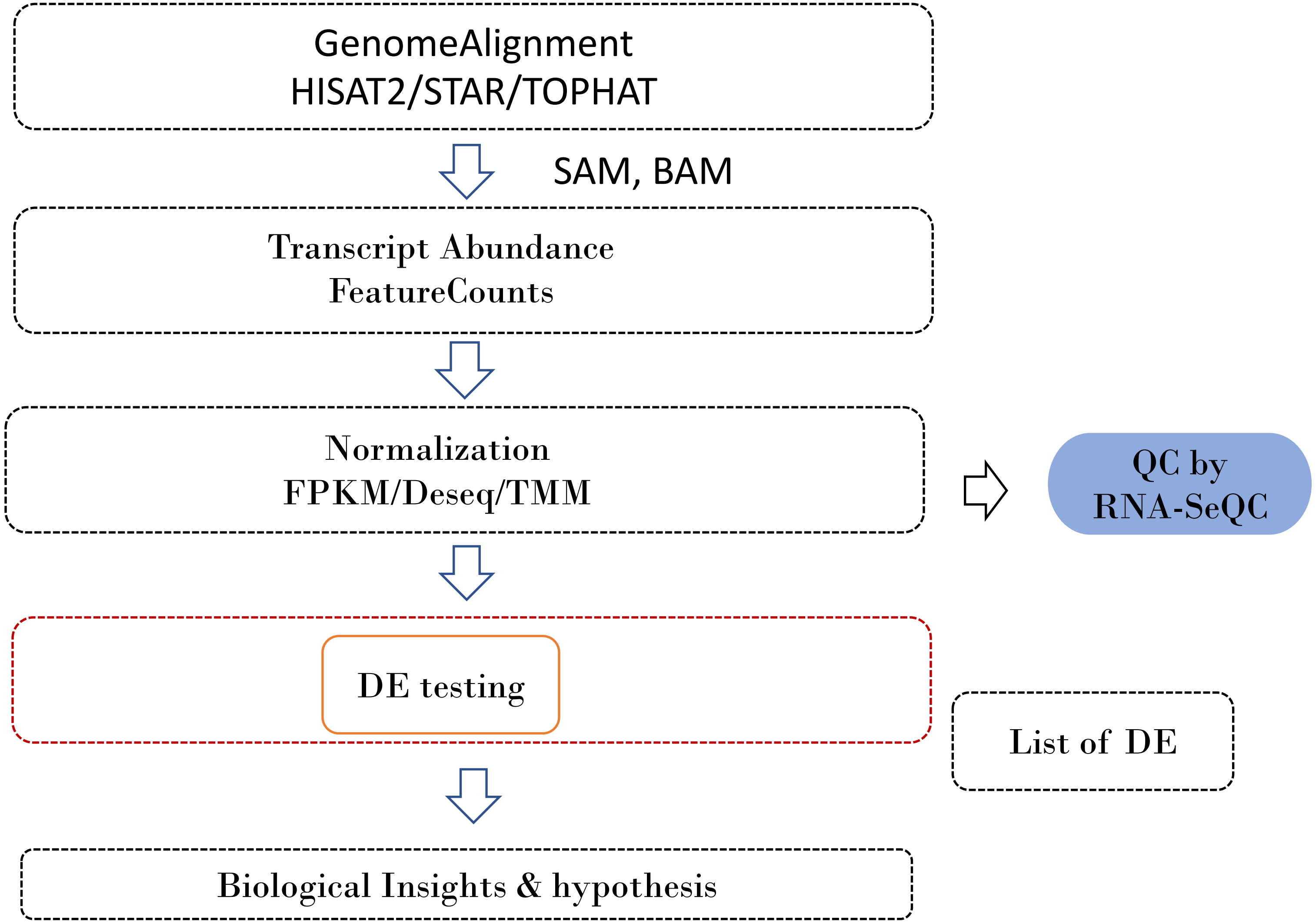


Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	−	+	+	−	−
UQ	++	++	+	++	−
Med	++	++	−	++	−
DESeq	++	++	++	++	++
TMM	++	++	++	++	++
Q	++	−	+	++	−
RPKM	−	+	+	−	−

A '−' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nature Communications

From reads to differential expression



GenomeAlignment
HISAT2/STAR/TOPHAT

SAM, BAM

Transcript Abundance
FeatureCounts

Normalization
FPKM/Deseq/TMM

QC by
RNA-SeQC

DE testing

List of DE

Biological Insights & hypothesis

Differential Expression Analysis

How do the expression levels differ across several conditions?

Challenges:

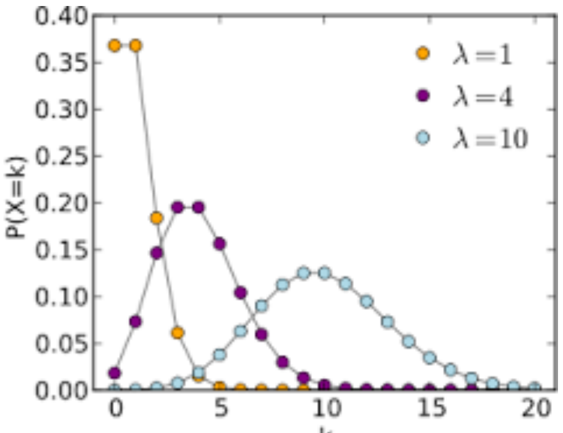
1. Count data is discrete – no normal distribution. Cannot perform t-test.
2. Small number of replicates – cannot use permutation methods.
3. Account for variability in measurements across biological replicates of an experiment.

Adopted from Soumya Luthra's presentation ("RNA-Seq analysis in R (Bioconductor)"*)*

Poisson Distribution?

In probability theory and statistics, the Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume, e.g. the number of phone calls received by a call center per hour.

- Mean = Variance
  - **Mean** is the average of the numbers
  - **Variance** (σ^2) in statistics is a measurement of the spread between numbers in a data set. That is, *it measures how far each number in the set is from the mean and therefore from every other number in the set.*
- Is read count data Poisson Distributed?
- **Over-dispersion** - variance in RNA-Seq measurements of gene expression are larger than the theoretical values



- In statistics, **overdispersion** is the presence of greater variability in a data set than would be expected based on a given statistical model.

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Adopted from Soumya Luthra's presentation ("RNA-Seq analysis in R (Bioconductor)")

Negative Binomial Distribution

In probability theory and statistics, the **negative binomial distribution** is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of failures (denoted *r*) occurs. For example, if we define a 1 as failure, all non-1s as successes, and we throw a dice repeatedly until 1 appears the third time (*r* = three failures), then the probability distribution of the number of non-1s that appeared will be a negative binomial distribution.

- NB has been shown to be a good fit to RNA-Seq data
- It is flexible enough to account for biological variability

Model:

- Makes the assumption that an observation say $Y_{gj\,(\text{observed number})}$ of reads for gene g sample j, has a mean μ_{gj} and a variance of $\mu_{gj} + \Phi_g \mu^2$, where Φ_g represents over-dispersion relative to poisson distribution.
- The mean parameter depends on the sequencing depth as well as on the amount of RNA from gene in the sample
- Obtaining good estimates of each gene's dispersion is critical for statistical testing.

Tools:

- EdgeR and DESeq count data using a Negative Binomial Distribution and perform statistical tests for differential expression.

Adopted from Soumya Luthra's presentation ("RNA-Seq analysis in R (Bioconductor)"*)*

edgeR

EdgeR treats the Poisson variance as simple sampling variance, and refers to the dispersion estimate as the "biological coefficient of variation."

Estimating dispersion:

- EdgeR shares information across genes to determine a common dispersion. It then calculates a dispersion estimate per gene and shrinks it towards the common dispersion. The gene-specific (referred to in edgeR as tagwise) dispersion estimates are used in the test for differential expression.

Statistical Test:

- **Simple design** - Fischer's exact test

(statistical significance test that is one of a class of exact tests, so called because the significance of the deviation from a null hypothesis (e.g., P-value) can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity, as with many statistical tests).

- **Complex design** - Generalized linear model (GLM) framework

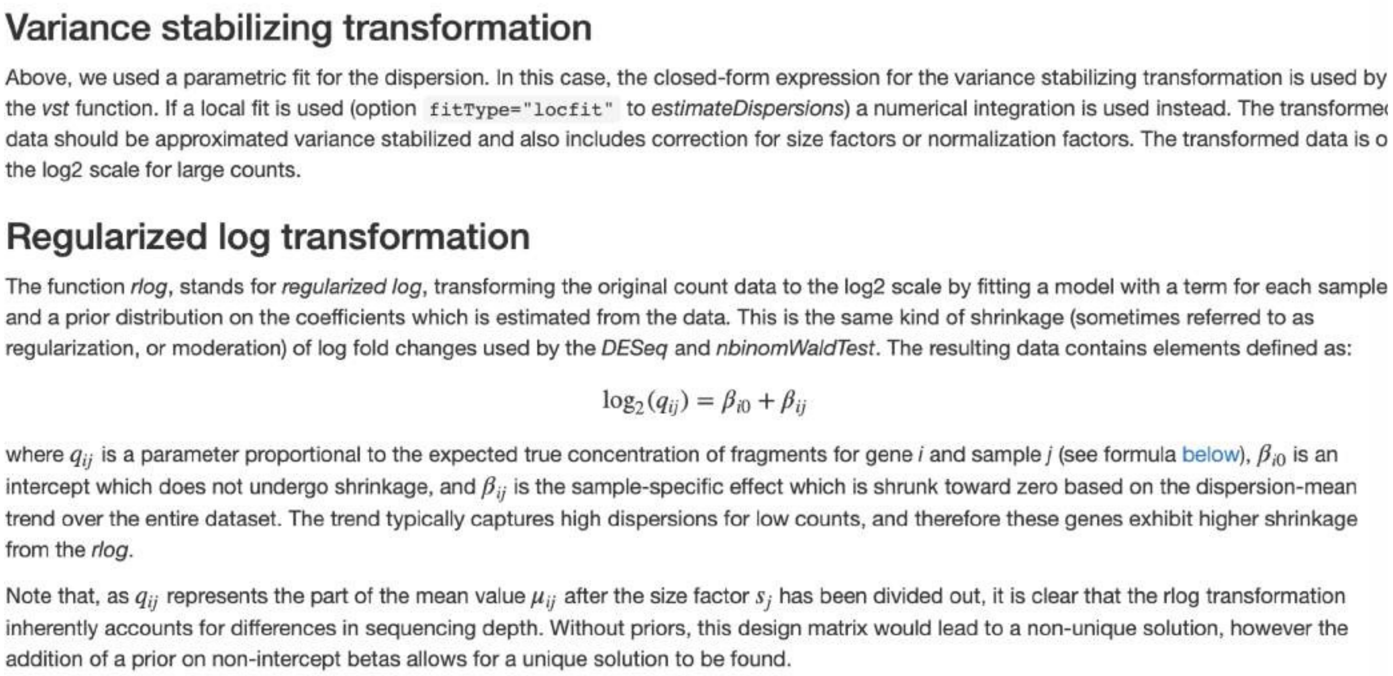
(In statistics, the generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.)

Adopted from Soumya Luthra's presentation ("RNA-Seq analysis in R (Bioconductor)"*)*

DESeq

- Differential gene expression from count data based on negative binomial distribution.
- Offers two transformations for stabilizing the variance of count data:
  - **VST** – Variance stabilizing transformation
  - **Regularized log transformation** (**rlog**)

http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html



Variance stabilizing transformation

Above, we used a parametric fit for the dispersion. In this case, the closed-form expression for the variance stabilizing transformation is used by the *vst* function. If a local fit is used (option `fitType="locfit"` to *estimateDispersions*) a numerical integration is used instead. The transforme data should be approximated variance stabilized and also includes correction for size factors or normalization factors. The transformed data is o the log2 scale for large counts.

Regularized log transformation

The function *rlog*, stands for *regularized log*, transforming the original count data to the log2 scale by fitting a model with a term for each sample and a prior distribution on the coefficients which is estimated from the data. This is the same kind of shrinkage (sometimes referred to as regularization, or moderation) of log fold changes used by the *DESeq* and *nbinomWaldTest*. The resulting data contains elements defined as:

$$\log_2(q_{ij}) = \beta_{i0} + \beta_{ij}$$

where q_{ij} is a parameter proportional to the expected true concentration of fragments for gene *i* and sample *j* (see formula below), β_{i0} is an intercept which does not undergo shrinkage, and β_{ij} is the sample-specific effect which is shrunk toward zero based on the dispersion-mean trend over the entire dataset. The trend typically captures high dispersions for low counts, and therefore these genes exhibit higher shrinkage from the *rlog*.

Note that, as q_{ij} represents the part of the mean value μ_{ij} after the size factor s_j has been divided out, it is clear that the rlog transformation inherently accounts for differences in sequencing depth. Without priors, this design matrix would lead to a non-unique solution, however the addition of a prior on non-intercept betas allows for a unique solution to be found.

Adopted from Soumya Luthra's presentation ("RNA-Seq analysis in R (Bioconductor)"*)*

How do I use VST or rlog data for differential testing?

The variance stabilizing and rlog transformations are provided for applications other than differential testing, for example clustering of samples or other machine learning applications. For differential testing we recommend the *DESeq* function applied to raw counts.

mRNAs

DE mRNA expression, p <0.05 (Top 1000 mRNAs)

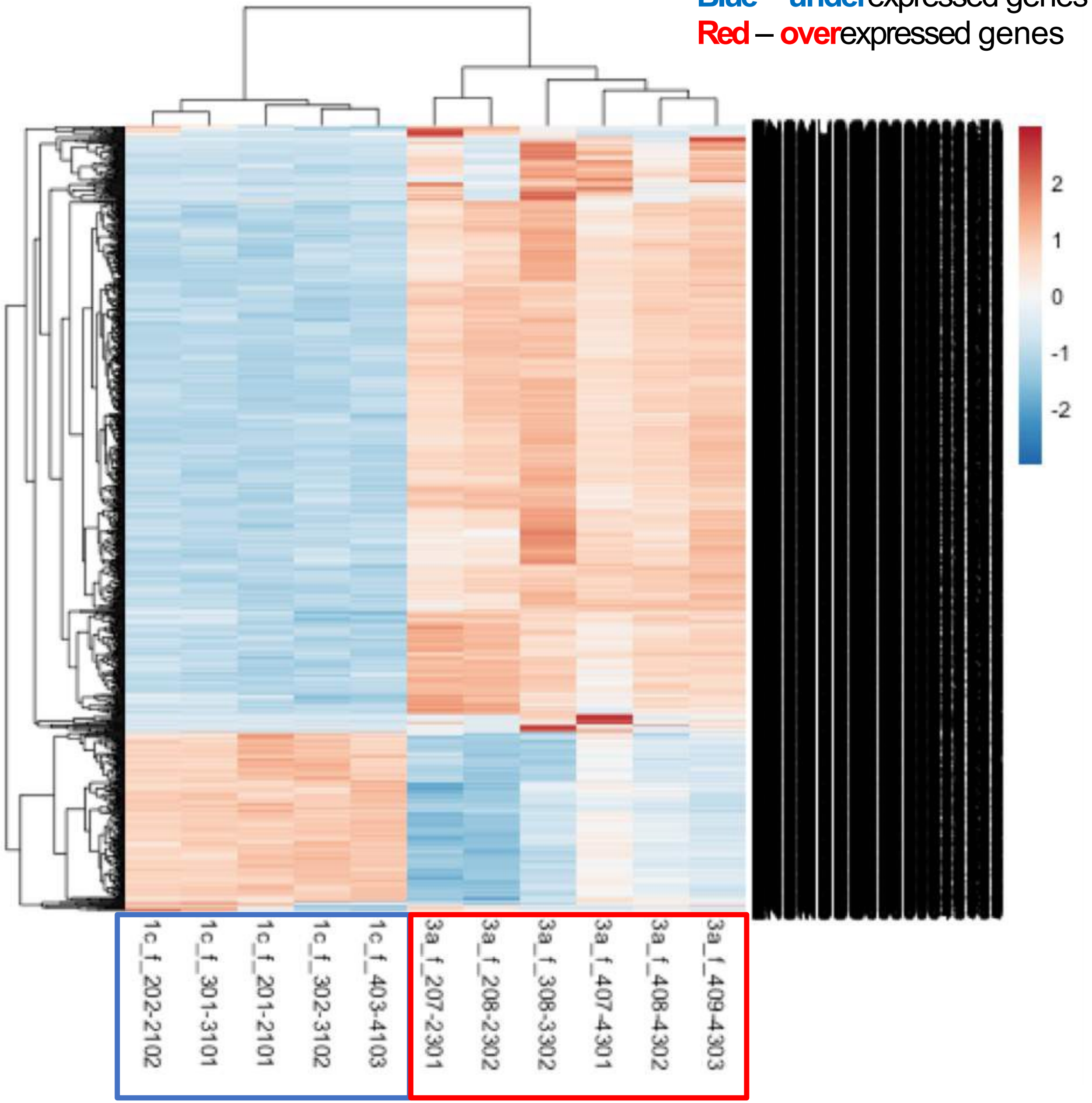
adenine (3a) vs. control (1c)

Blue – underexpressed genes
Red – overexpressed genes

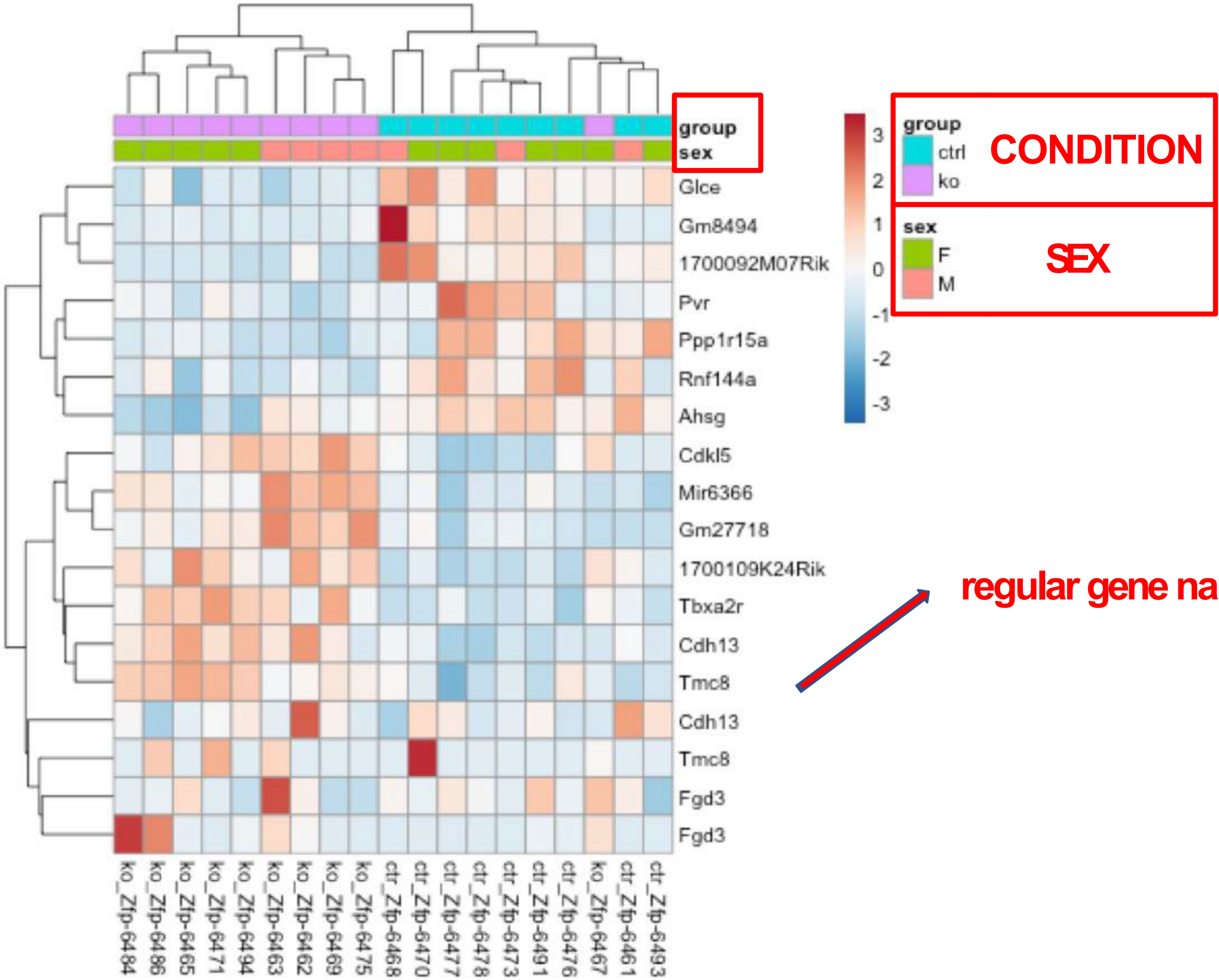
Dendrogram at the side shows us a hierarchical clustering for the genes.
Since the clustering is only relevant for genes that actually carry signal, one usually carries it out only for a subset of most highly variable genes (genes with the highest variance across samples)

The heatmap becomes more interesting if we do not look at absolute expression strength but rather at **the amount by which each gene deviates in a specific sample from the gene's average across all samples.** Hence, we center and scale each genes' values across samples, and plot a heatmap.

Heatmap is a graphical representation of data where individual values contained in a matrix are represented as colors. It allows to visualize expression of many genes in many samples.



Adding other parameters for the heatmaps….



CONDITION

SEX

regular gene names

mRNAs

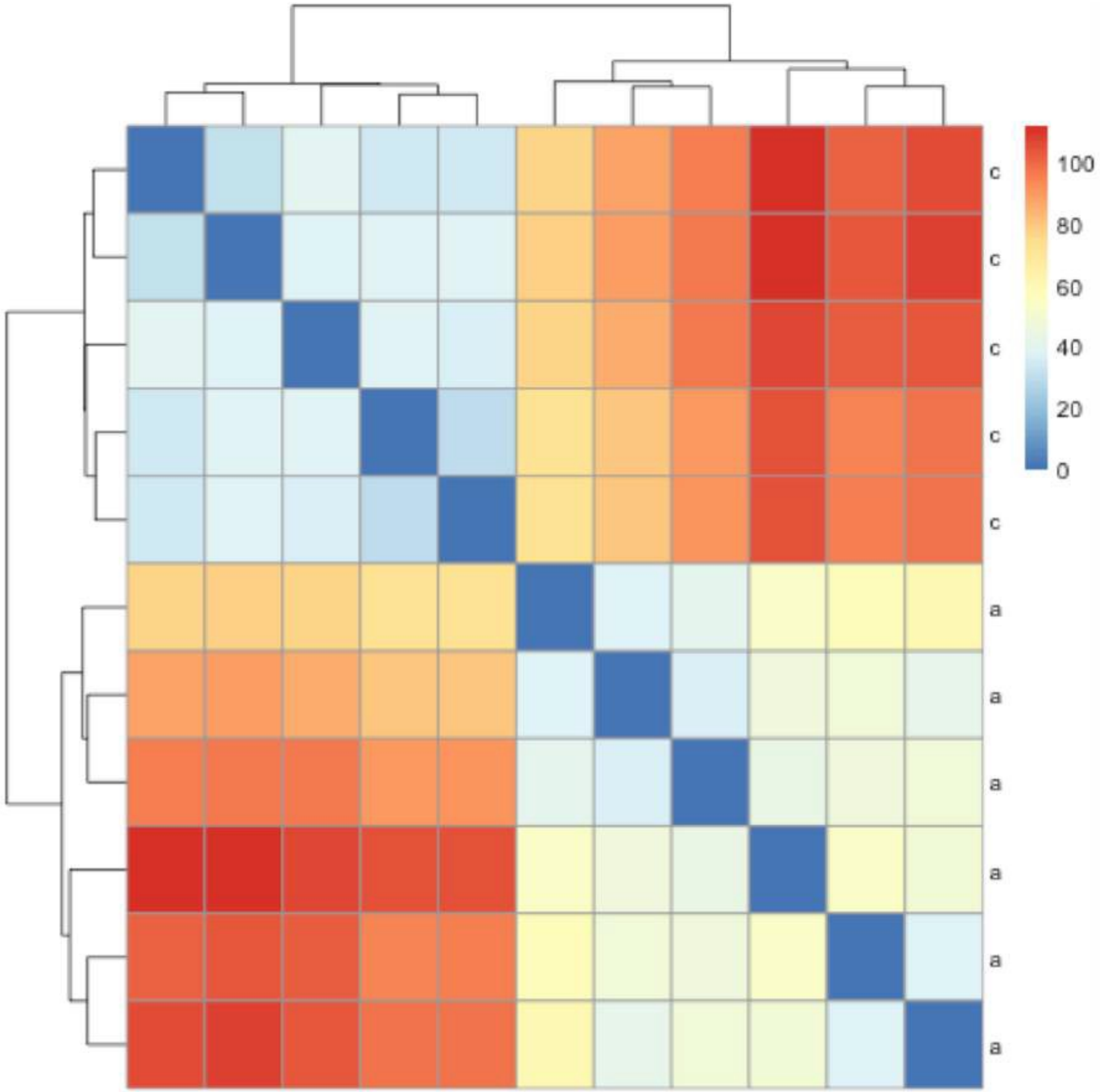
Sample-To-Sample distance (Euclidian)

adenine (a) vs. control (c)

Goal:
to assess overall similarity between samples

A heatmap of this distance matrix gives us an **overview over similarities and dissimilarities between samples.**

We have to provide a hierarchical clustering (hc) to the heatmap function based on the sample distances, or else the heatmap function would calculate a clustering based on the distances between the rows/columns of the distance matrix.



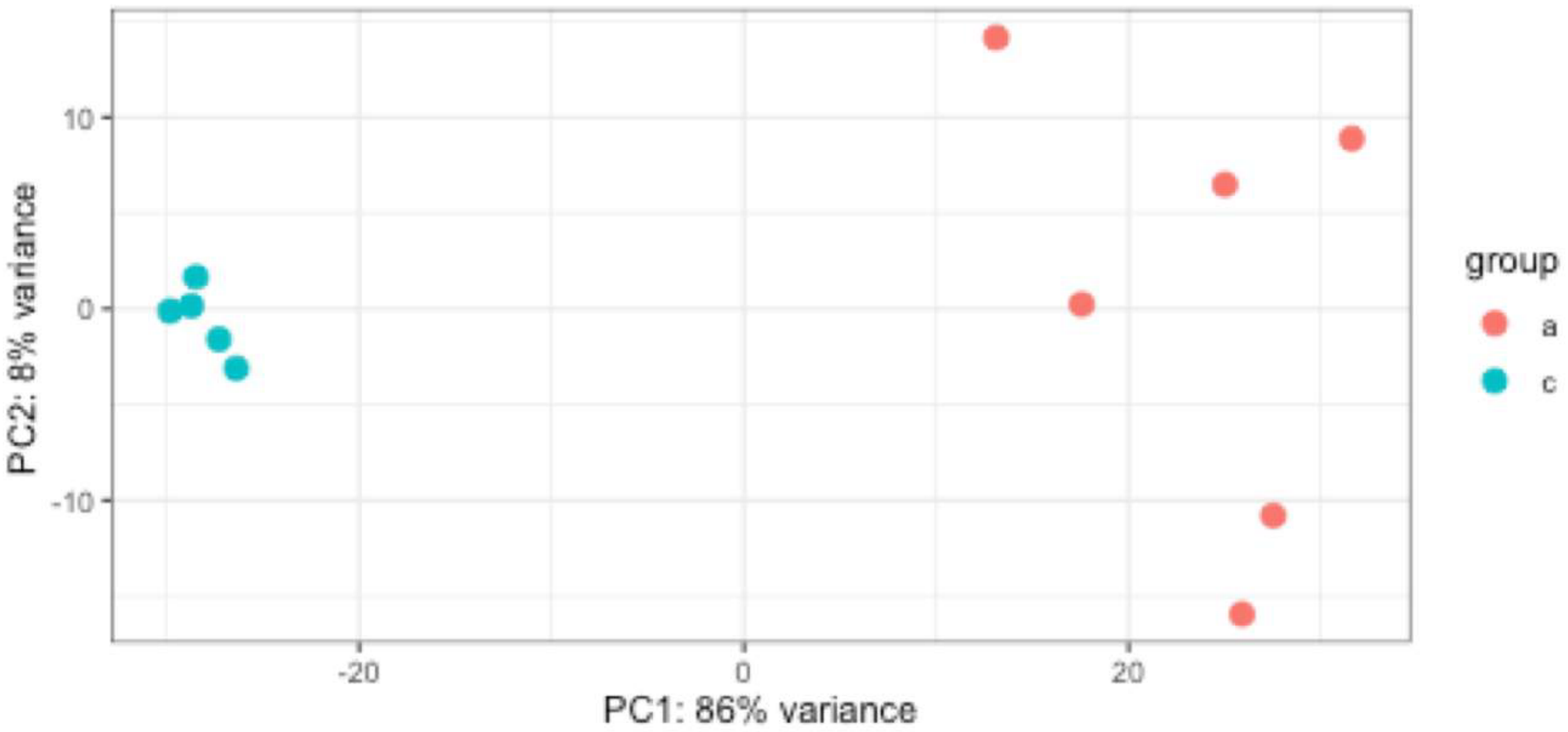
mRNAs

adenine (a) vs. control (c)

PCA plot

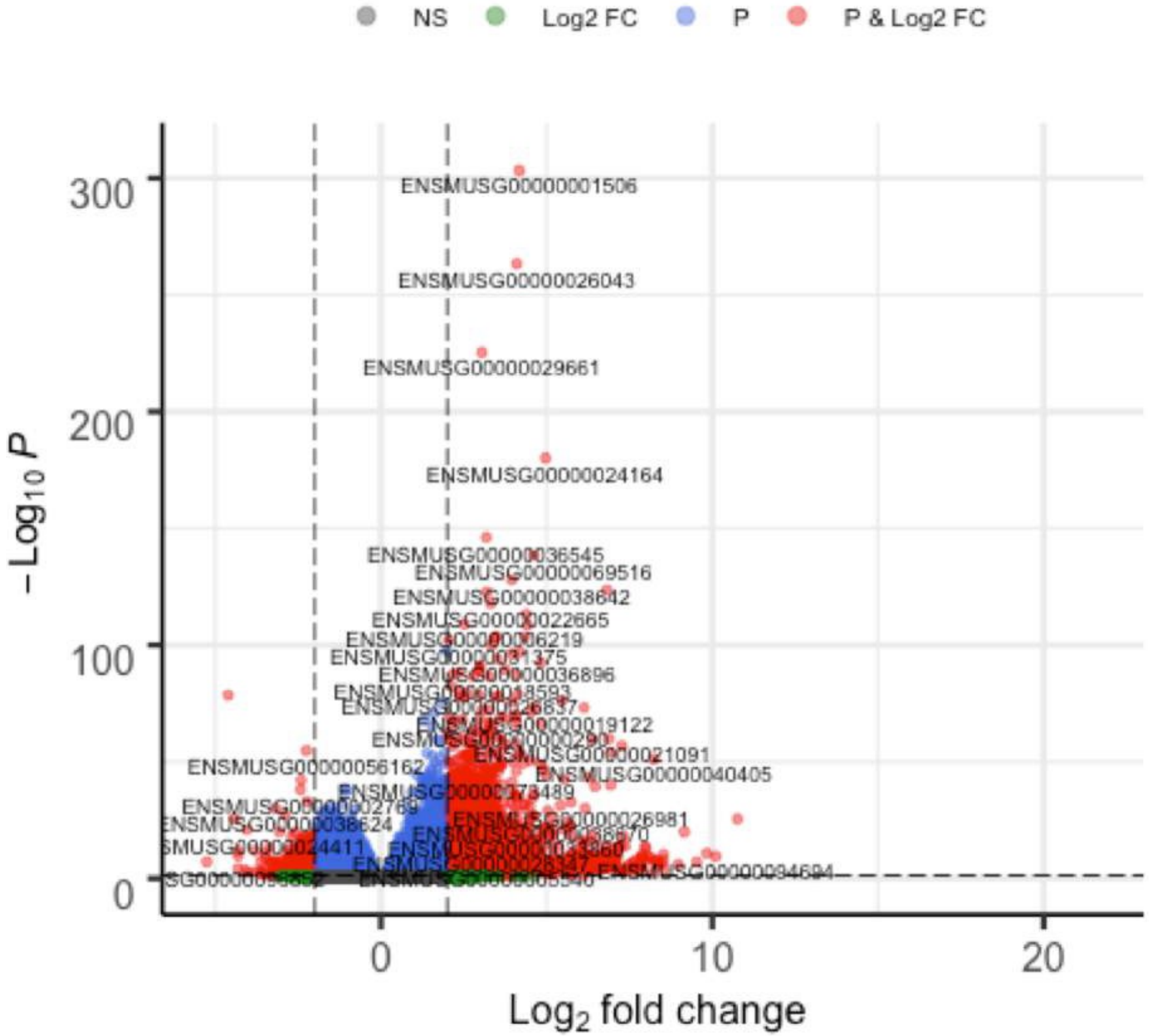
Principal component plot of the samples

Related to the distance matrix is the PCA plot, which shows the samples in the 2D plane spanned by their first two principal components. This type of plot is useful for **visualizing the overall effect of experimental covariates** and **batch effects.**



mRNAs

Volcano plot - adenine (3a) vs. control (1c)



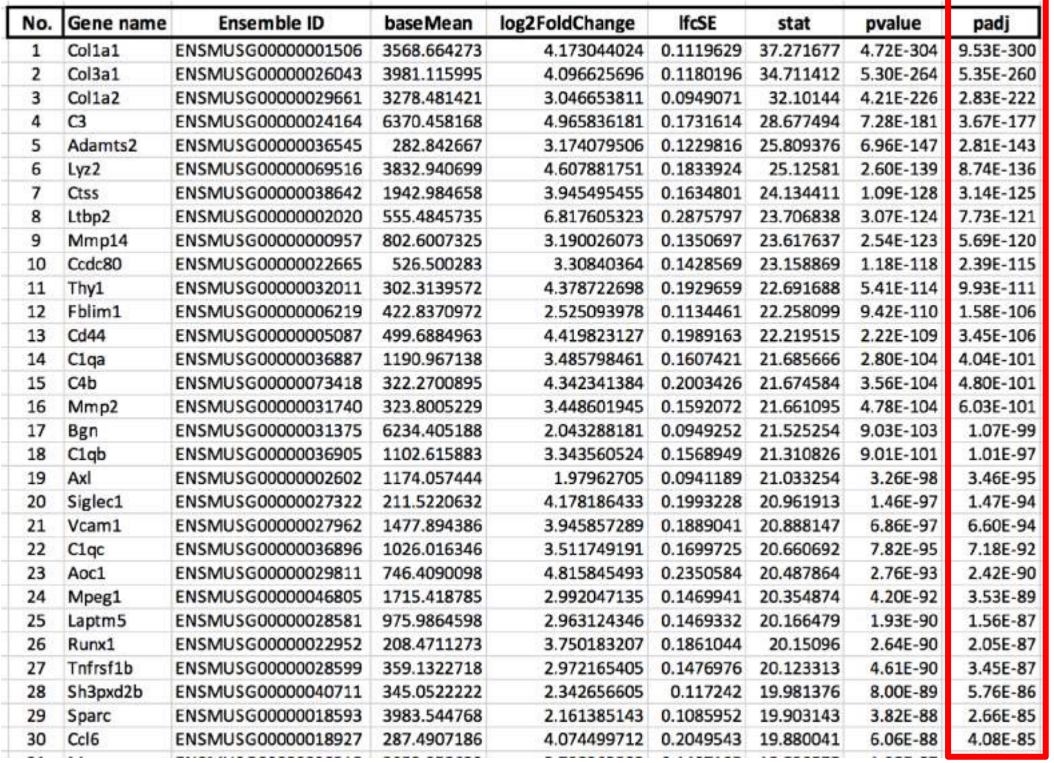
mRNAs

adenine (3a) vs. control (1c) vs. – first top 30 mRNAs

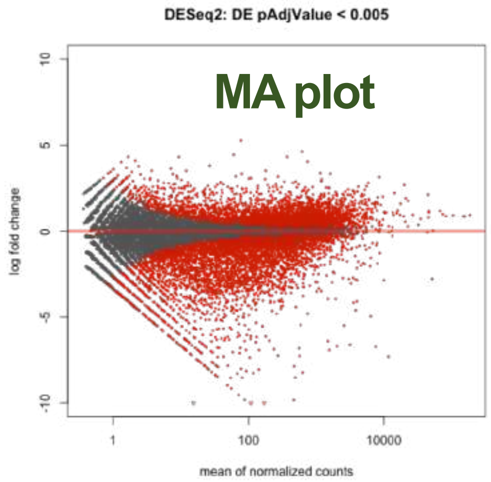
Exporting results to **CSV** files

sorted by padj

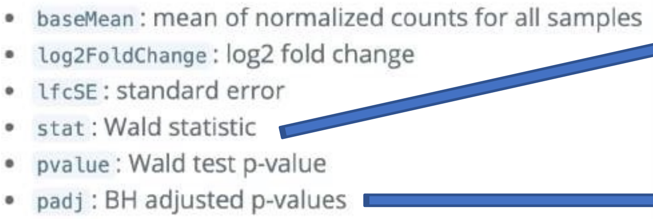
(from the smallest to the largest & expand selection)



No.	Gene name	Ensemble ID	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
1	Col1a1	ENSMUSG00000001506	3568.664273	4.173044024	0.1119629	37.271677	4.72E-304	9.53E-300
2	Col3a1	ENSMUSG00000026043	3981.115995	4.096625696	0.1180196	34.711412	5.30E-264	5.35E-260
3	Col1a2	ENSMUSG00000029661	3278.481421	3.046653811	0.0949071	32.10144	4.21E-226	2.83E-222
4	C3	ENSMUSG00000024164	6370.458168	4.965836181	0.1731614	28.677494	7.28E-181	3.67E-177
5	Adamts2	ENSMUSG00000036545	282.842667	3.174079506	0.1229816	25.809376	6.96E-147	2.81E-143
6	Lyz2	ENSMUSG00000069516	3832.940699	4.607881751	0.1833924	25.12581	2.60E-139	8.74E-136
7	Ctss	ENSMUSG00000038642	1942.984658	3.945495455	0.1634801	24.134411	1.09E-128	3.14E-125
8	Ltbp2	ENSMUSG00000002020	555.4845735	6.817605323	0.2875797	23.706838	3.07E-124	7.73E-121
9	Mmp14	ENSMUSG00000000957	802.6007325	3.190026073	0.1350697	23.617637	2.54E-123	5.69E-120
10	Ccdc80	ENSMUSG00000022665	526.500283	3.30840364	0.1428569	23.158869	1.18E-118	2.39E-115
11	Thy1	ENSMUSG00000032011	302.3139572	4.378722698	0.1929659	22.691688	5.41E-114	9.93E-111
12	Fblim1	ENSMUSG00000006219	422.8370972	2.525093978	0.1134461	22.258099	9.42E-110	1.58E-106
13	Cd44	ENSMUSG00000005087	499.6884963	4.419823127	0.1989163	22.219515	2.22E-109	3.45E-106
14	C1qa	ENSMUSG00000036887	1190.967138	3.485798461	0.1607421	21.685666	2.80E-104	4.04E-101
15	C4b	ENSMUSG00000073418	322.2700895	4.342341384	0.2003426	21.674584	3.56E-104	4.80E-101
16	Mmp2	ENSMUSG00000031740	323.8005229	3.448601945	0.1592072	21.661095	4.78E-104	6.03E-101
17	Bgn	ENSMUSG00000031375	6234.405188	2.043288181	0.0949252	21.525254	9.03E-103	1.07E-99
18	C1qb	ENSMUSG00000036905	1102.615883	3.343560524	0.1568949	21.310826	9.01E-101	1.01E-97
19	Axl	ENSMUSG00000002602	1174.057444	1.97962705	0.0941189	21.033254	3.26E-98	3.46E-95
20	Siglec1	ENSMUSG00000027322	211.5220632	4.178186433	0.1993228	20.961913	1.46E-97	1.47E-94
21	Vcam1	ENSMUSG00000027962	1477.894386	3.945857289	0.1889041	20.888147	6.86E-97	6.60E-94
22	C1qc	ENSMUSG00000036896	1026.016346	3.511749191	0.1699725	20.660692	7.82E-95	7.18E-92
23	Aoc1	ENSMUSG00000029811	746.4090098	4.815845493	0.2350584	20.487864	2.76E-93	2.42E-90
24	Mpeg1	ENSMUSG00000046805	1715.418785	2.992047135	0.1469941	20.354874	4.20E-92	3.53E-89
25	Laptm5	ENSMUSG00000028581	975.9864598	2.963124346	0.1469332	20.166479	1.93E-90	1.56E-87
26	Runx1	ENSMUSG00000022952	208.4711273	3.750183207	0.1861044	20.15096	2.64E-90	2.05E-87
27	Tnfrsf1b	ENSMUSG00000028599	359.1322718	2.972165405	0.1476976	20.123313	4.61E-90	3.45E-87
28	Sh3pxd2b	ENSMUSG00000040711	345.0522222	2.342656605	0.117242	19.981376	8.00E-89	5.76E-86
29	Sparc	ENSMUSG00000018593	3983.544768	2.161385143	0.1085952	19.903143	3.82E-88	2.66E-85
30	Ccl6	ENSMUSG00000018927	287.4907186	4.074499712	0.2049543	19.880041	6.06E-88	4.08E-85



- The function *plotMA* shows **the log2 fold changes** attributable to a given **variable** over the mean of normalized counts for all the samples in the *DESeqDataSet*.
- Points will be colored **red** if the **adjusted *p* value is < 0.1**.



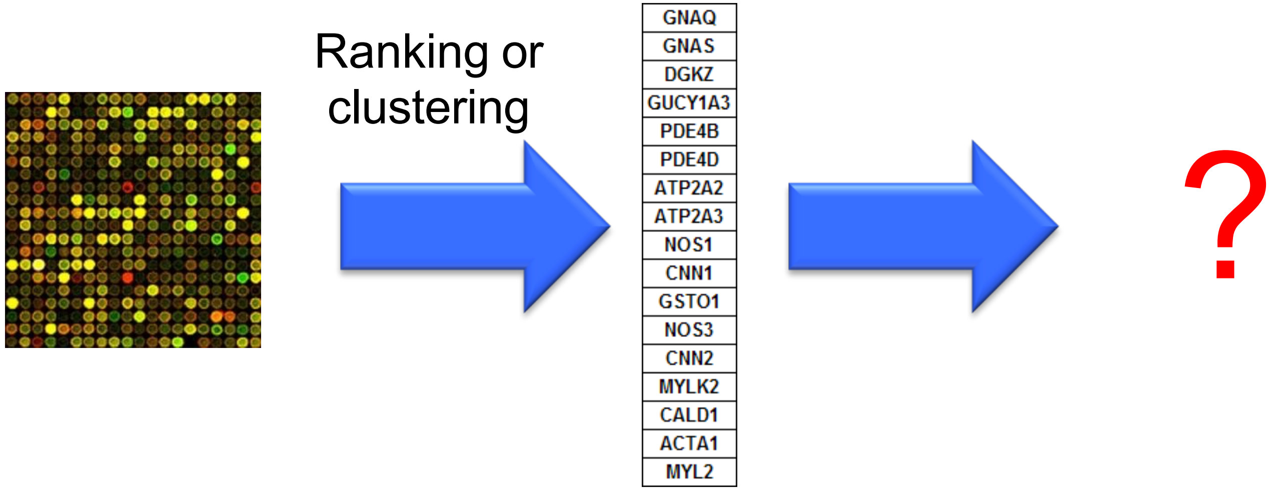
- `baseMean` : mean of normalized counts for all samples
- `log2FoldChange` : log2 fold change
- `lfcSE` : standard error
- `stat` : Wald statistic
- `pvalue` : Wald test p-value
- `padj` : BH adjusted p-values

The **Wald statistic** is the logfoldchange (LFC) divided by its standard error (lfcSE) . This Wald statistic is used to calculate p-values (it is compared to a standard normal distribution) . So it's the ratio of LFC and SE which determines significance.

The Benjamini-Hochberg (BH) procedure is a powerful tool that decreases the false discovery rate. Adjusting the rate helps to control for the fact that sometimes small p-values (less than 5%) happen by chance, which could lead you to incorrectly reject the true null hypotheses. In other words, the BH Procedure helps you to avoid Type I errors (false positives).

Interpreting gene lists

- Genome-Scale Analysis (Omics)
  - Genomics, Proteomics
- Tell me what's interesting about these genes

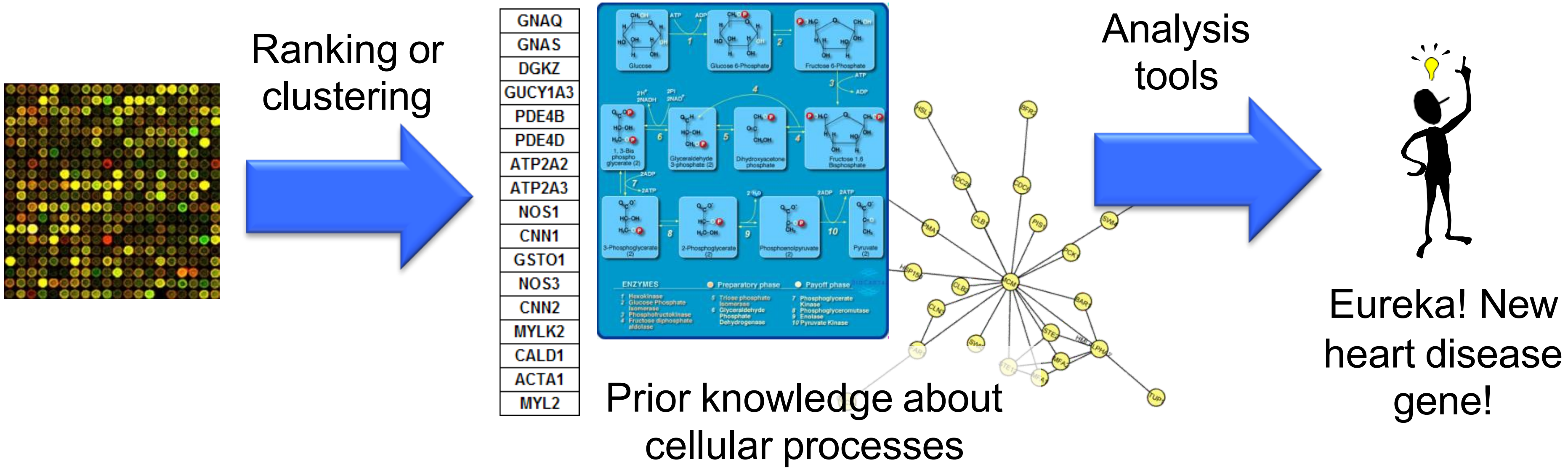


Ranking or clustering

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

Interpreting gene lists

- Genome-Scale Analysis (Omics)
  - Genomics, Proteomics
- Tell me what's interesting about these genes
  - Are they enriched in known pathways, complexes, functions



Ranking or clustering

GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

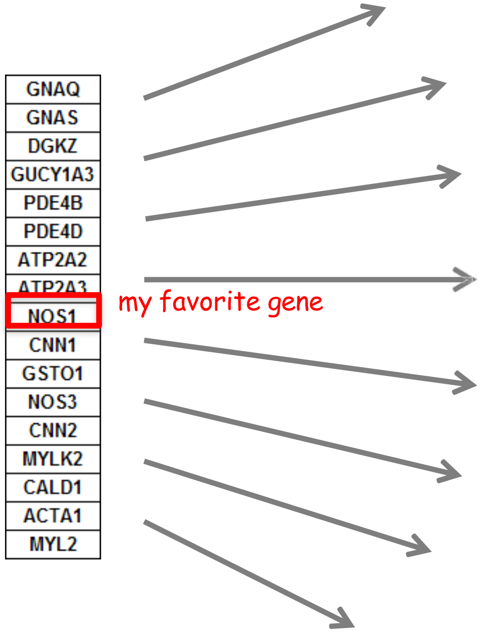
Prior knowledge about cellular processes

Analysis tools

Eureka! New heart disease gene!

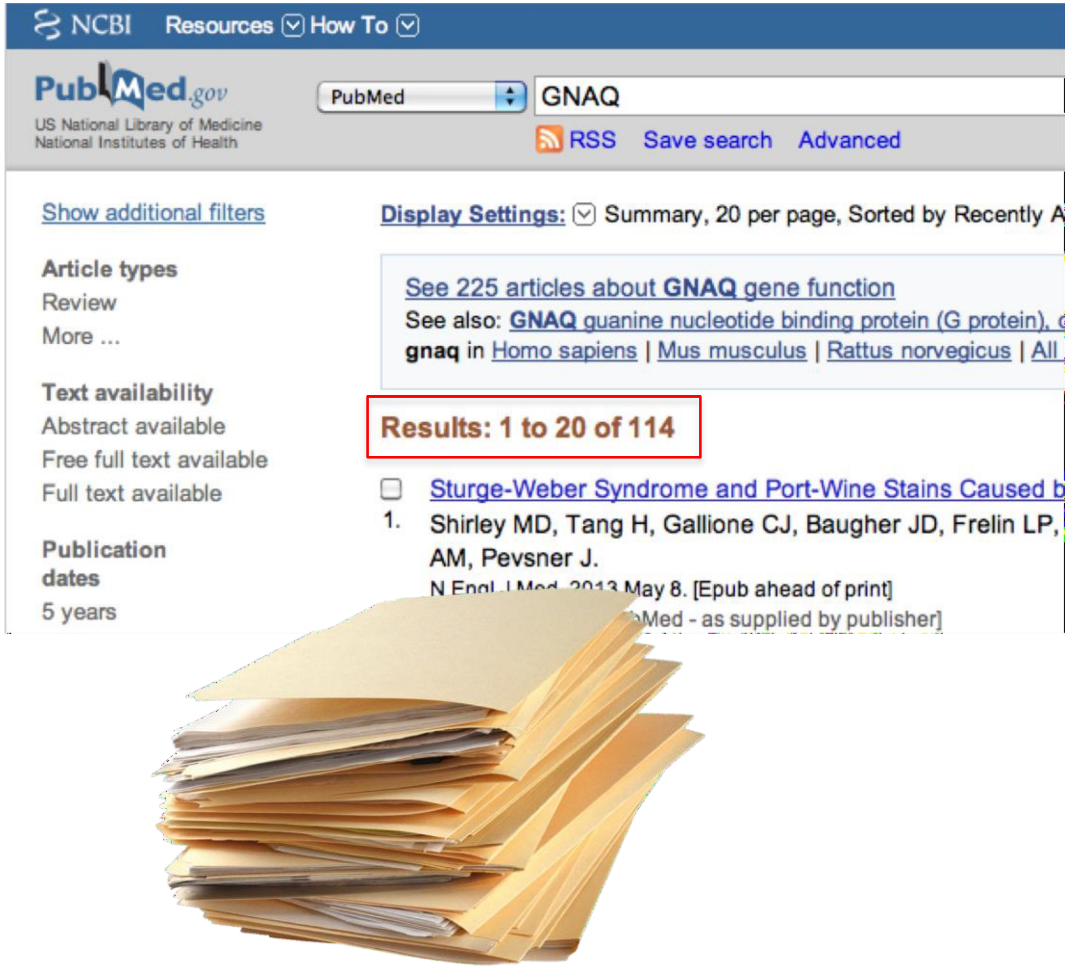
Pathway and network analysis

- Save time compared to traditional approach



GNAQ
GNAS
DGKZ
GUCY1A3
PDE4B
PDE4D
ATP2A2
ATP2A3
NOS1
CNN1
GSTO1
NOS3
CNN2
MYLK2
CALD1
ACTA1
MYL2

my favorite gene



NCBI
Resources
How To
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
RSS
Save search
Advanced

GNAQ

Show additional filters

Article types
Review
More ...

Text availability
Abstract available
Free full text available
Full text available

Publication dates
5 years

Display Settings: Summary, 20 per page, Sorted by Recently A

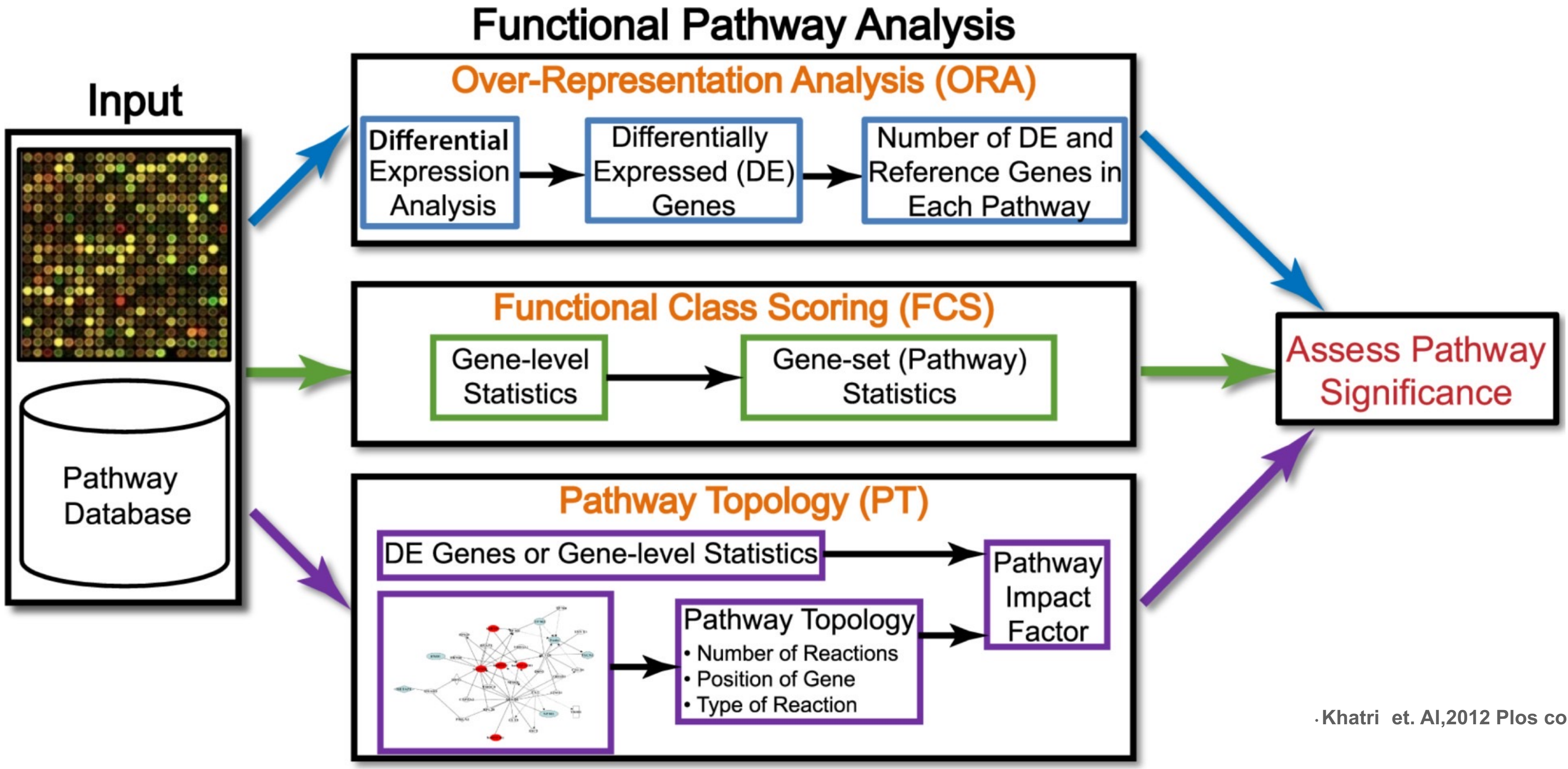
See 225 articles about GNAQ gene function
See also: GNAQ guanine nucleotide binding protein (G protein), c
gnaq in Homo sapiens | Mus musculus | Rattus norvegicus | All

Results: 1 to 20 of 114

□ 1. Sturge-Weber Syndrome and Port-Wine Stains Caused b
Shirley MD, Tang H, Gallione CJ, Baugher JD, Frelin LP, AM, Pevsner J.
N Engl J Med. 2013 May 8. [Epub ahead of print]
Med - as supplied by publisher]

Pathway and network analysis

- Helps gain mechanistic insight into ‘omics data
  - Identifying a master regulator, drug targets, characterizing pathways active in a sample
- Any type of analysis that involves pathway or network information
- Most commonly applied to help interpret lists of genes
- Most popular type is pathway enrichment analysis, but many others are useful



Functional Pathway Analysis

Input

Pathway Database

Over-Representation Analysis (ORA)

Differential Expression Analysis → Differentially Expressed (DE) Genes → Number of DE and Reference Genes in Each Pathway

Functional Class Scoring (FCS)

Gene-level Statistics → Gene-set (Pathway) Statistics

Pathway Topology (PT)

DE Genes or Gene-level Statistics → Pathway Impact Factor

Pathway Topology

- Number of Reactions
- Position of Gene
- Type of Reaction

Assess Pathway Significance

- Khatri et. Al,2012 Plos computational biology

- The data generated by an experiment using a high-throughput technology (e.g., microarray, proteomics, metabolomics), along with functional annotations (pathway database) of the corresponding genome, are input to virtually all pathway analysis methods.
- ORA methods require that the input is a list of differentially expressed genes
- FCS methods use the entire data matrix as input
- PT-based methods additionally utilize the number and type of interactions between gene products, which may or may not be a part of a pathway database.
- The result of every pathway analysis method is a list of significant pathways in the condition under study.

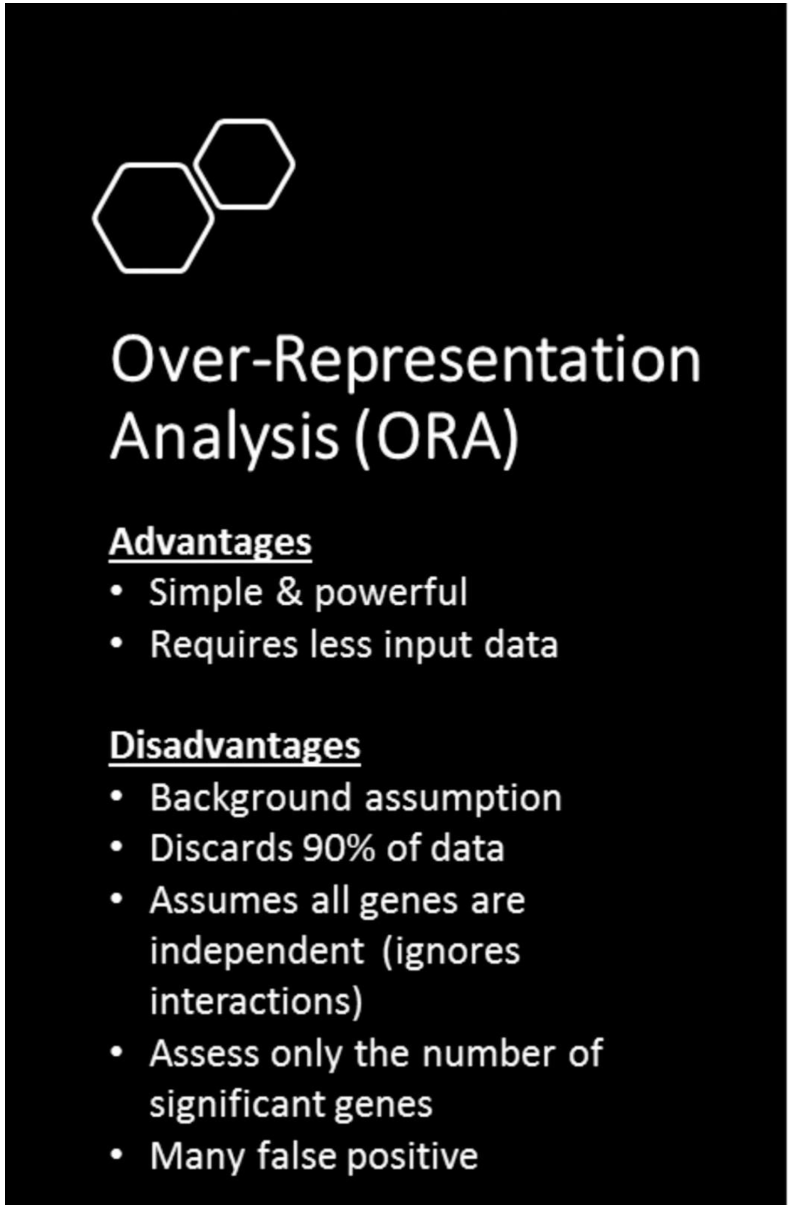
Over-Representation Analysis (ORA) Approaches

- Earliest methods ➔ over-representation analysis (ORA)

- Statistically evaluates the fraction of genes in a particular pathway found among the set of genes showing changes in expression

- “2×2 table methods”

Over-representation Analysis (ORA)



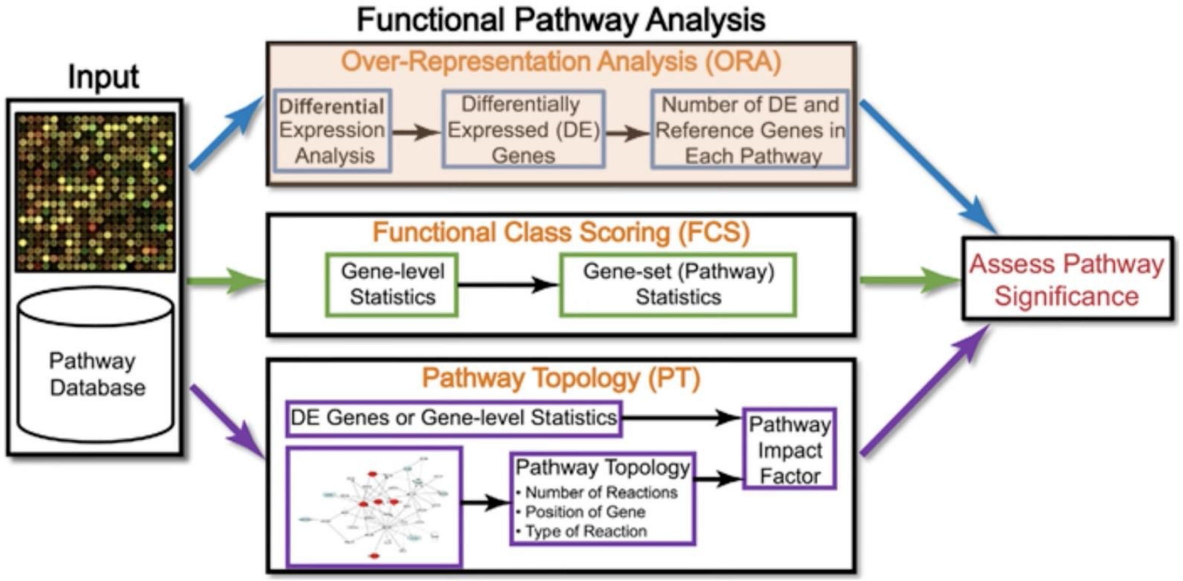
Over-Representation Analysis (ORA)

Advantages

- Simple & powerful
- Requires less input data

Disadvantages

- Background assumption
- Discards 90% of data
- Assumes all genes are independent (ignores interactions)
- Assess only the number of significant genes
- Many false positive

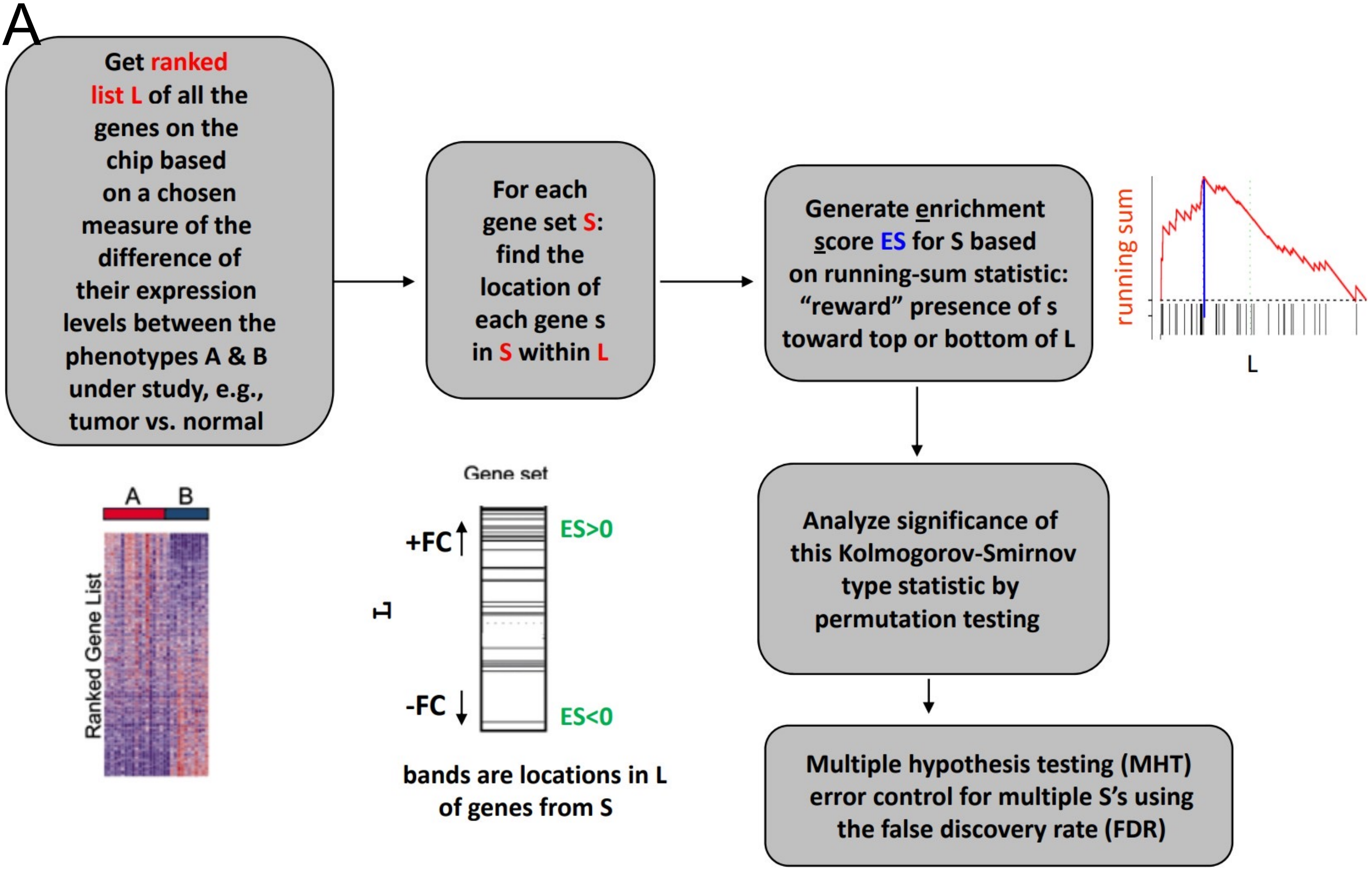


Functional Pathway Analysis

Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Impact Factor
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Assess Pathway Significance

Functional Class Scoring

Example --- GSEA



Get ranked list L of all the genes on the chip based on a chosen measure of the difference of their expression levels between the phenotypes A & B under study, e.g., tumor vs. normal

For each gene set S: find the location of each gene s in S within L

Generate enrichment score ES for S based on running-sum statistic: "reward" presence of s toward top or bottom of L

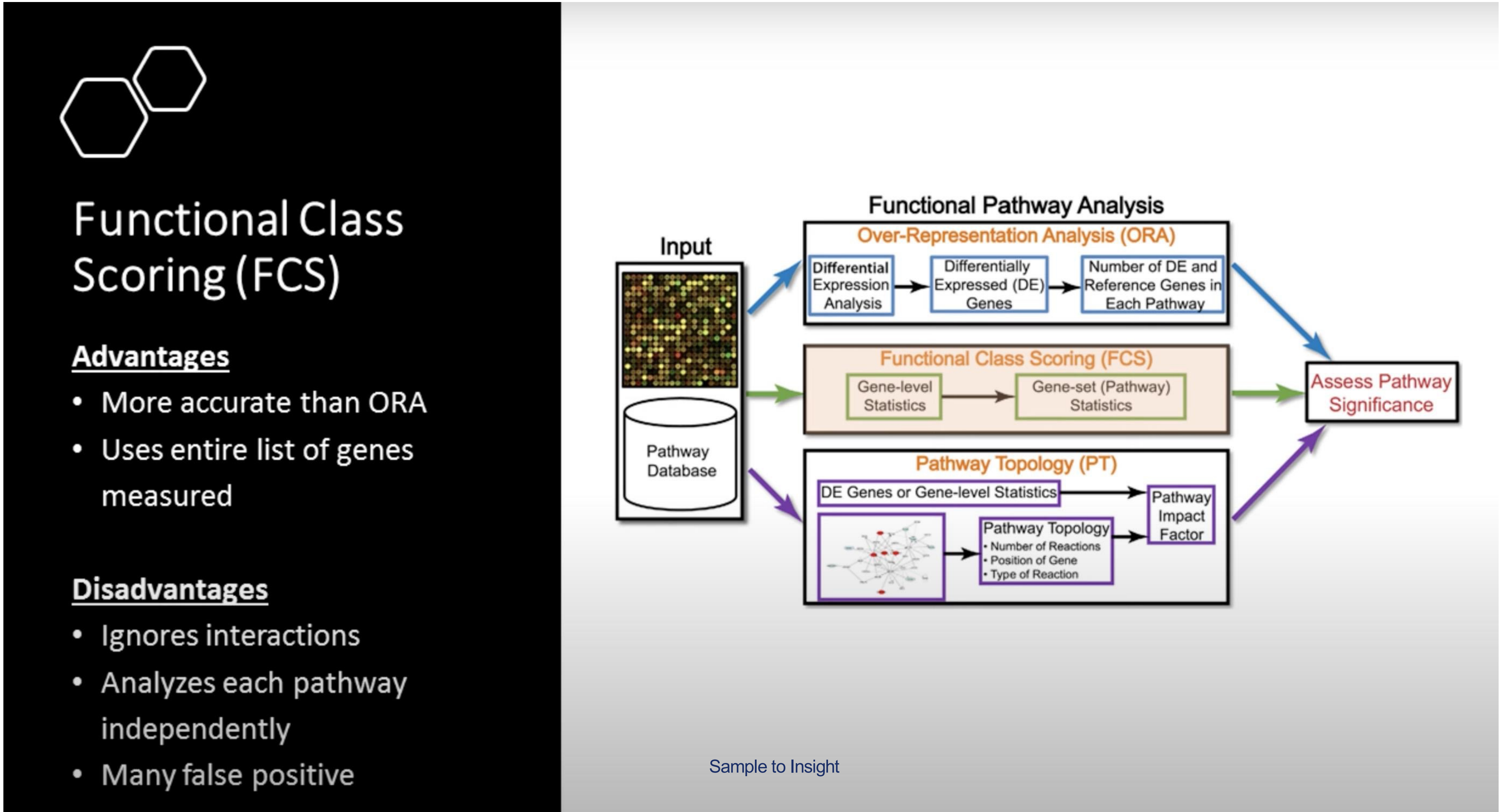
Analyze significance of this Kolmogorov-Smirnov type statistic by permutation testing

Multiple hypothesis testing (MHT) error control for multiple S's using the false discovery rate (FDR)

bands are locations in L of genes from S

1. Identified differential expression gene
2. gene-level statics combined to pathway-level statistics, statistics methods Kolmogorov-Smimov statistic, sum, mean, median of gene-level statistics。
3. Test pathway-level statistics

Functional Class Scoring



Functional Class Scoring (FCS)

Advantages

- More accurate than ORA
- Uses entire list of genes measured

Disadvantages

- Ignores interactions
- Analyzes each pathway independently
- Many false positive

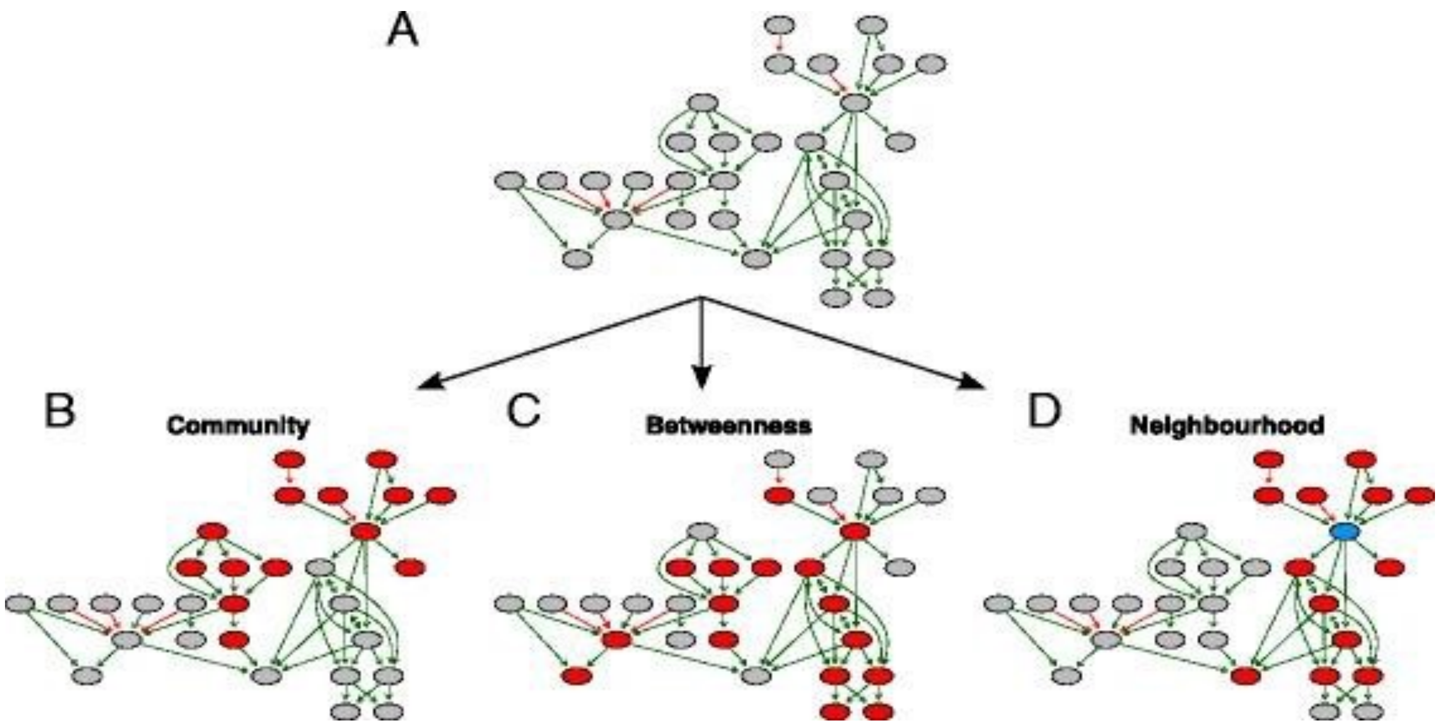
Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Pathway Topology (PT)-Based Approaches

- A large number of publicly available pathway knowledge bases provide information beyond simple lists of genes for each pathway
  - KEGG
  - MetaCyc
  - Reactome
  - RegulonDB
  - STKE
  - BioCarta
  - PantherDB
  - …

- These knowledge bases also provide information about gene products that interact with each other in a given pathway, how they interact (e.g., activation, inhibition, etc.), and where they interact (e.g., cytoplasm, nucleus, etc.)

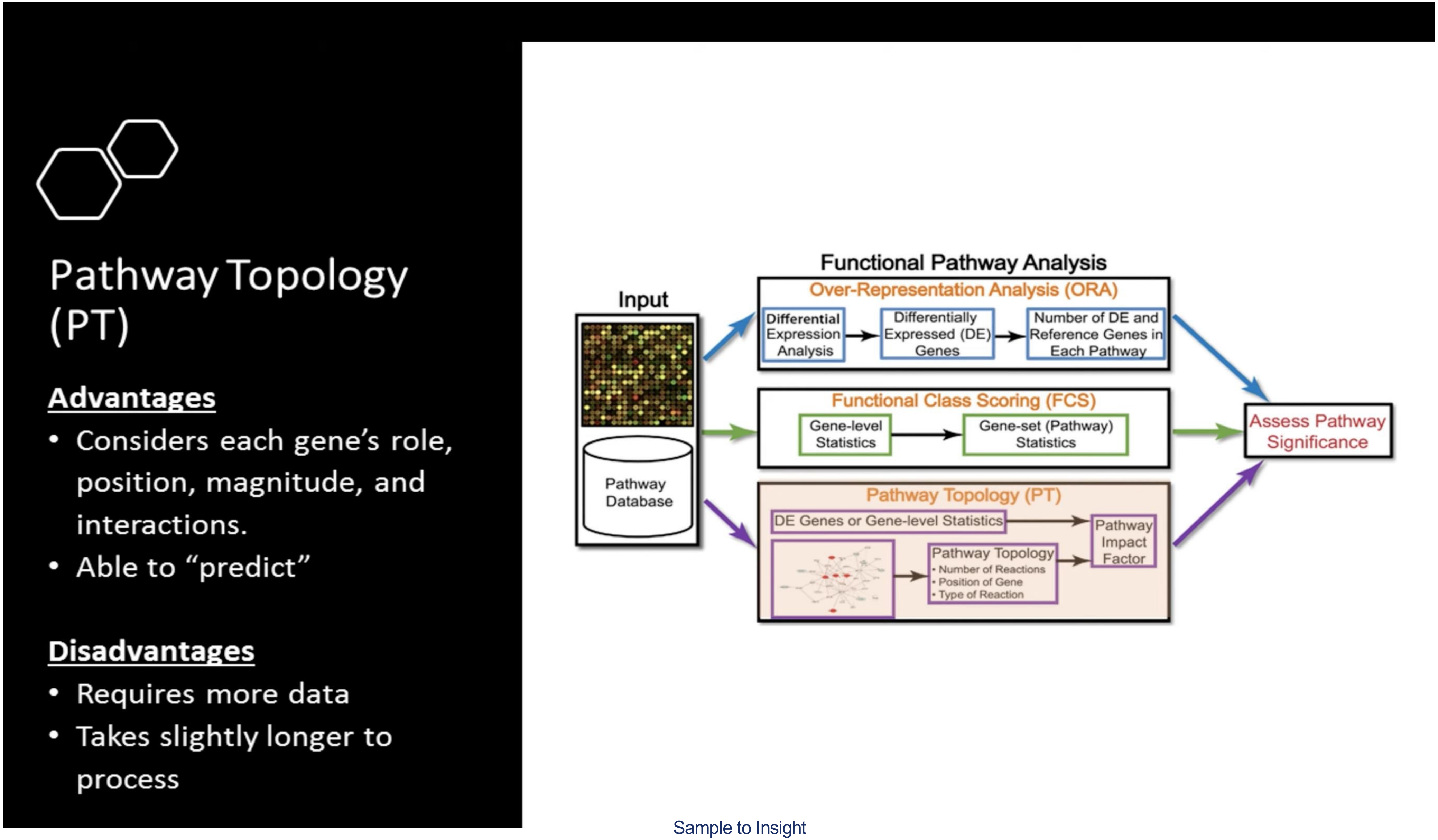
Topology method



Topology designs for pathway deregulation. **a** Example of a particular pathway with 30 genes. In order to deregulate this pathway on detection call level e.g. *DC* = 50 % (+/– 5 %) we needed to assign 14–16 affected gene to this pathway and allocate them on the pathway graph according to 3 topology approaches. **b** In the community design two gene communities were selected to be affected (depicted in red). **c** Top scored betweenness genes were depicted in red. **d** Gene neighbourhood of order 2 of the blue gene was affected (in red). The colour coding of graph edges represents activation (green) and inhibition (red) interactions between the nodes

Bayerlová, M., Jung, K., Kramer, F. *et al.*BMC Bioinformatics **12(335) 2015**

Pathway Topology



Pathway Topology (PT)

Advantages

- Considers each gene's role, position, magnitude, and interactions.
- Able to "predict"

Disadvantages

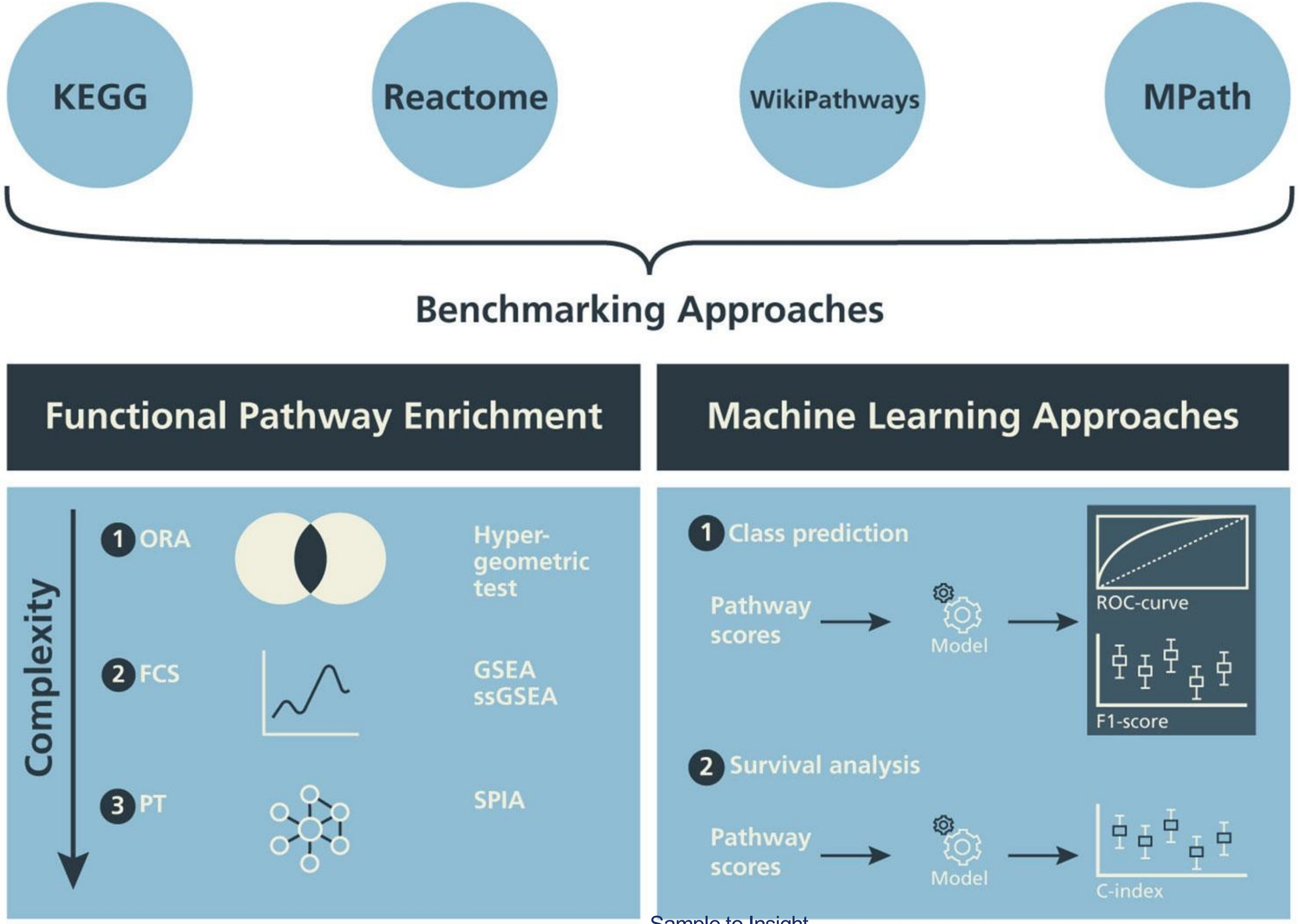
- Requires more data
- Takes slightly longer to process

Functional Pathway Analysis
Input
Pathway Database
Over-Representation Analysis (ORA)
Differential Expression Analysis
Differentially Expressed (DE) Genes
Number of DE and Reference Genes in Each Pathway
Functional Class Scoring (FCS)
Gene-level Statistics
Gene-set (Pathway) Statistics
Pathway Topology (PT)
DE Genes or Gene-level Statistics
Pathway Topology
• Number of Reactions
• Position of Gene
• Type of Reaction
Pathway Impact Factor
Assess Pathway Significance

Benefits of pathway analysis

- Easier to interpret
  - Familiar concepts e.g. cell cycle
- Identifies possible causal mechanisms
- Predicts new roles for genes
- Improves statistical power
  - Fewer tests, aggregates data from multiple genes into one pathway
- More reproducible
  - E.g. gene expression signatures
- Facilitates integration of multiple data types

combine all of them


KEGG
Reactome
WikiPathways
MPath
Benchmarking Approaches
Functional Pathway Enrichment
Machine Learning Approaches
Complexity
1 ORA
Hyper-geometric test
2 FCS
GSEA
ssGSEA
3 PT
SPIA
1 Class prediction
Pathway scores
Model
ROC-curve
F1-score
2 Survival analysis
Pathway scores
Model
C-index