MICHIGAN STATE UNIVERSITY

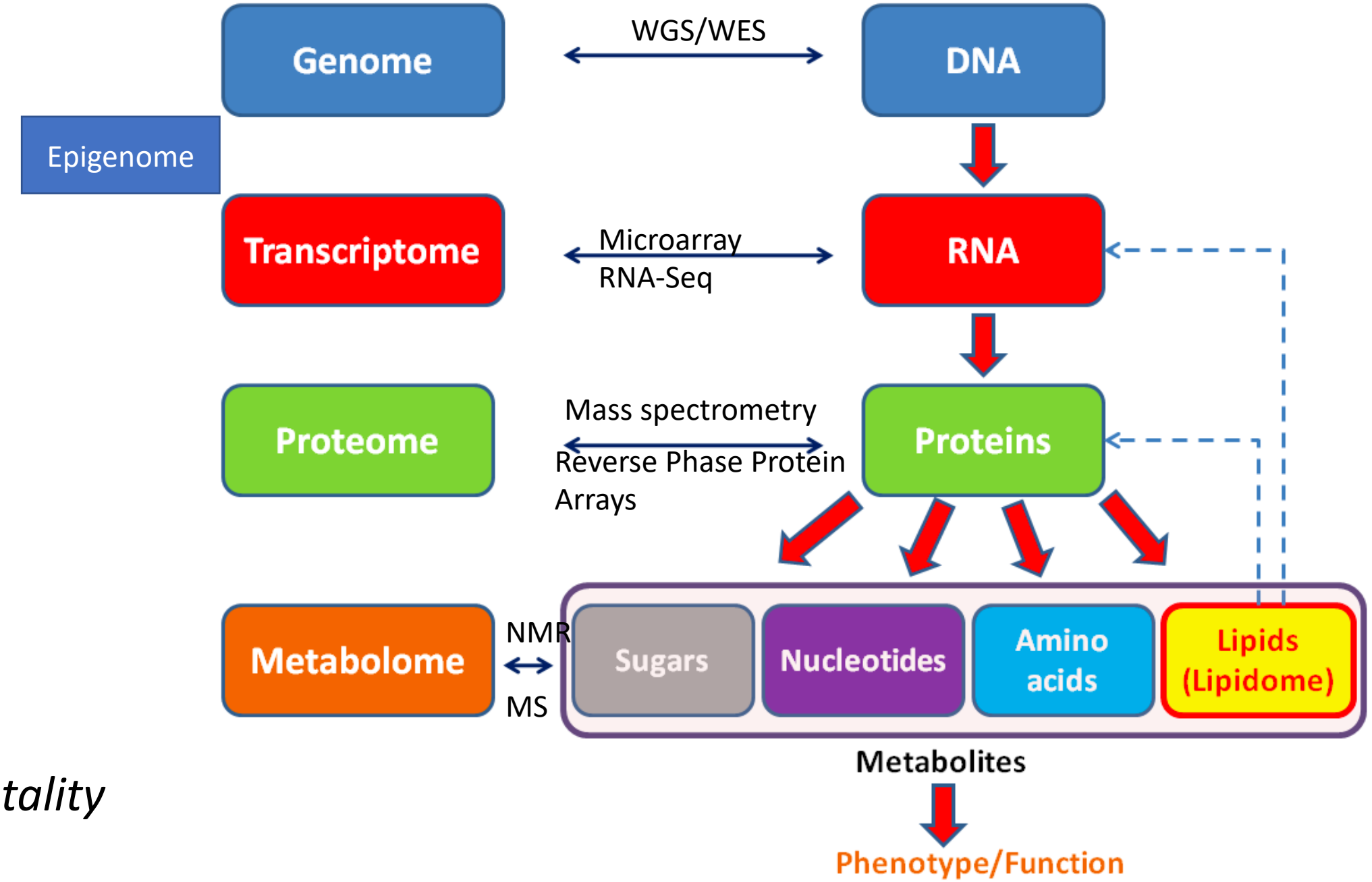
ChenLab
Michigan State University
http://binchenlab.org

# Transcriptomics + drug discovery

**Bin Chen**

Associate Professor
Dept. of Pediatrics and Human Development
Dept. of Pharmacology and Toxicology
College of Human Medicine
Michigan State University
Bin.Chen@hc.msu.edu @DrBinChen
http://binchenlab.org

IPSTP boot camp, Summer, 2024

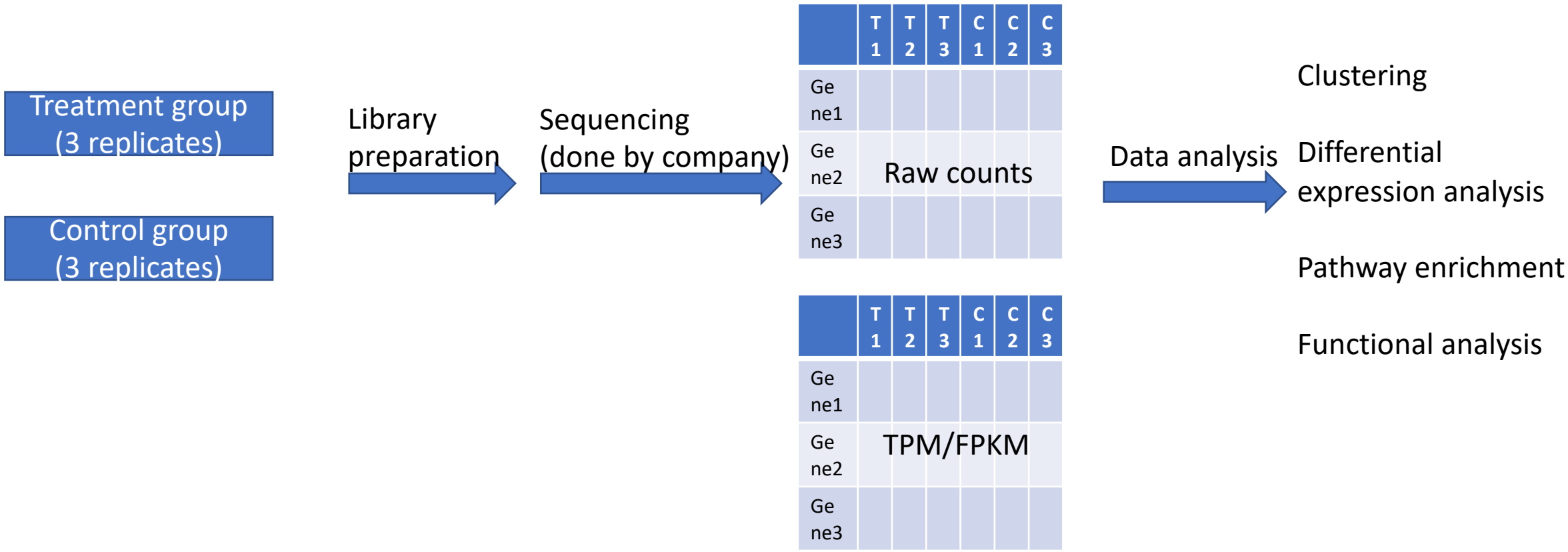



-Ome: *totality*

https://en.wikipedia.org/wiki/Genomics

# Potential questions we can ask using RNASeq

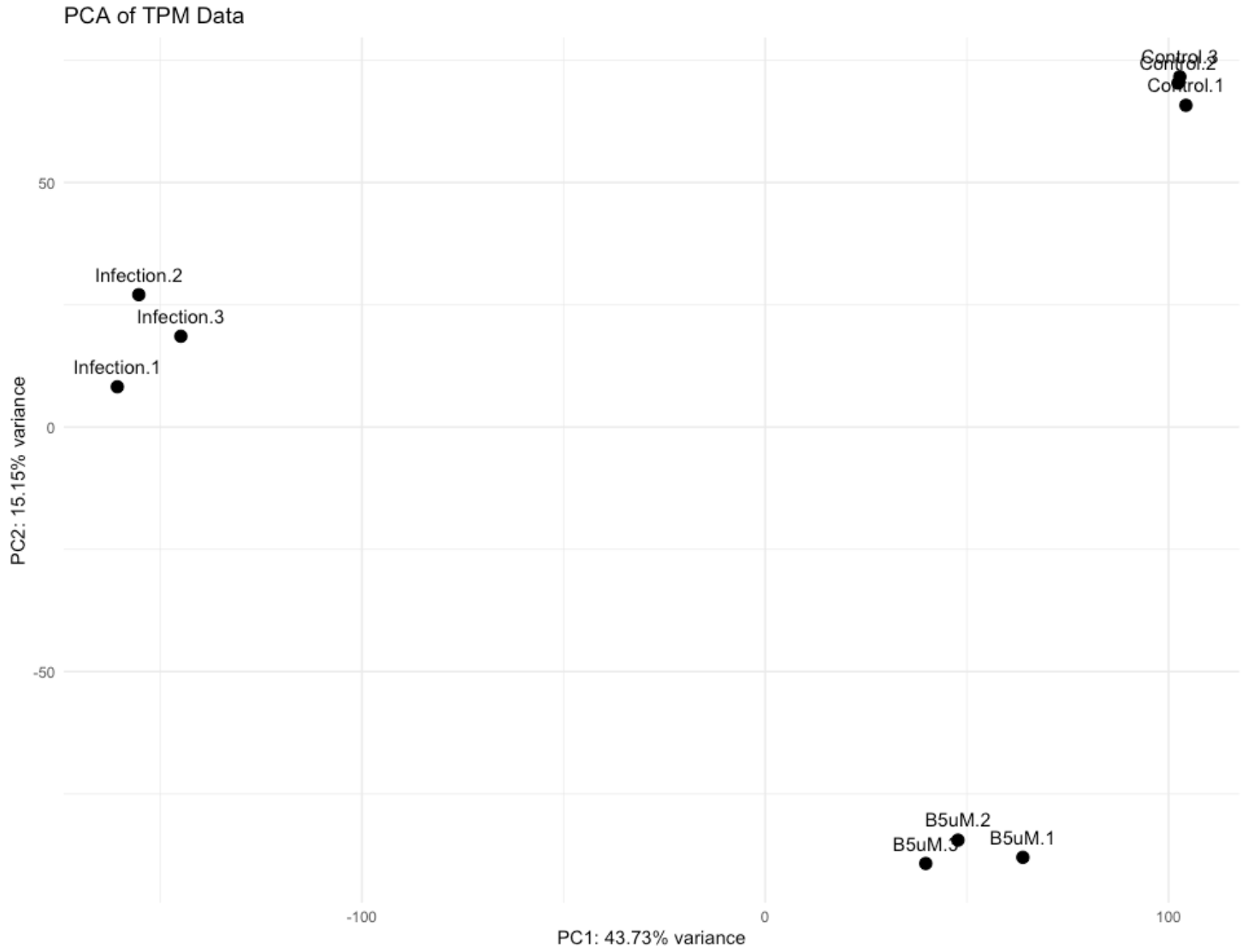
- Identify potential targets/pathways/cellular changes after pharmacological perturbation



```
> count_data <- read.csv(gzfile("~/Downloads/GSE187420_Raw_Read_count.csv.gz"), row.names = 1)
> tpm_data <- read.csv(gzfile("~/Downloads/GSE187420_Normalized_TPM_count.csv.gz"), row.names = 1)
> head(count_data)
                Control.1 Control.2 Control.3 Infection.1 Infection.2 Infection.3    B5uM.1    B5uM.2    B5uM.3
ENSG00000000003 11.627078 11.384244 11.456868    5.727920    5.807355    5.930737 11.535761 11.179287 11.194757
ENSG00000000005  0.000000  0.000000  0.000000    0.000000    0.000000    0.000000  0.000000  0.000000  0.000000
ENSG00000000419 10.817783 10.490851 10.529431    5.930737    6.087463    6.066089 10.303781  9.962896  9.962896
ENSG00000000457  8.926533  8.829786  8.869964    5.891905    6.140779    6.105804  7.676662  7.427439  7.337354
ENSG00000000460  9.336328  9.226364  9.401157    5.479619    5.106013    4.761817  9.288635  9.187006  9.202809
ENSG00000000938  0.000000  0.000000  0.000000    0.000000    1.000000    1.000000  0.000000  0.000000  0.000000
> head(tpm_data)
                Control.1 Control.2 Control.3 Infection.1 Infection.2 Infection.3   B5uM.1   B5uM.2   B5uM.3
ENSG00000000003  5.187847  5.050066  5.122673    2.032101   1.9411063   2.2898345 5.171127 4.934988 4.969473
ENSG00000000005  0.000000  0.000000  0.000000    0.000000   0.0000000   0.0000000 0.000000 0.000000 0.000000
ENSG00000000419  5.602291  5.384741  5.415488    3.144046   3.1921942   3.4235782 5.146900 4.915999 4.960234
ENSG00000000457  2.223423  2.195348  2.150560    1.851999   1.9030383   2.1890338 1.189034 1.220330 1.097611
ENSG00000000460  2.933573  2.967169  3.130931    2.077243   1.2927817   1.4646683 3.152183 3.054848 3.140779
ENSG00000000938  0.000000  0.000000  0.000000    0.000000   0.1634987   0.0976108 0.000000 0.000000 0.000000
```

# Sample visualization

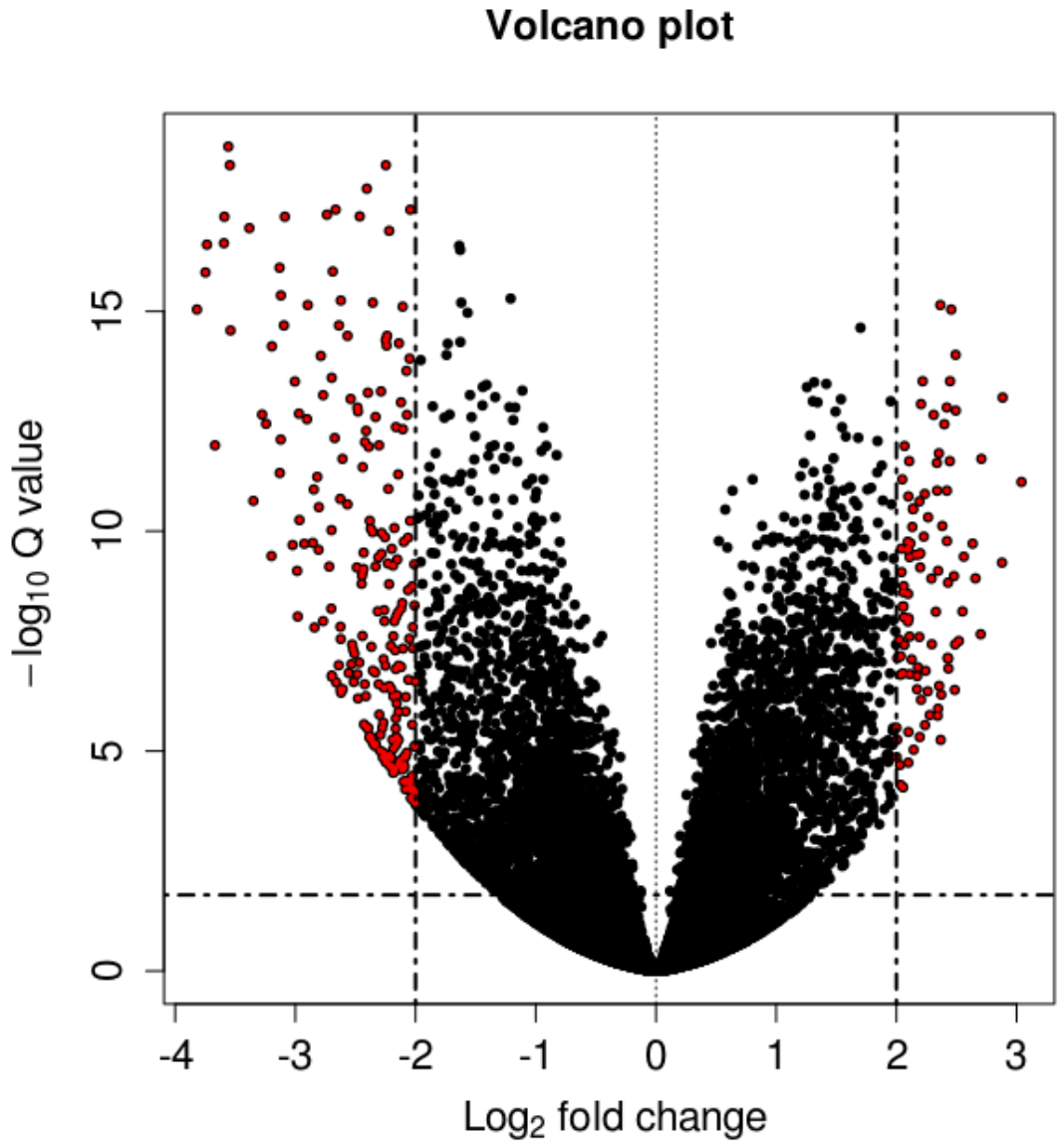
pca_results <- PCA(t(tpm_data))



# Gene expression analysis

| | T 1 | T 2 | T 3 | C 1 | C 2 | C 3 |
|---|---|---|---|---|---|---|
| Gen e1 | | | | | | |
| Gen e2 | | | | | | |
| Gen e3 | | | | | | |
| Gen e4 | | | | | | |

DE analysis
DESeq2, EdgeR, Limma




The goal is to compute fold change and significance!

# A naïve approach

| | T 1 | T 2 | T 3 | C 1 | C 2 | C 3 |
|---|---|---|---|---|---|---|
| Gene1 | | | | | | |
| Gene2 | | | | | | |
| Gene3 | | | | | | |
| Gene4 | | | | | | |

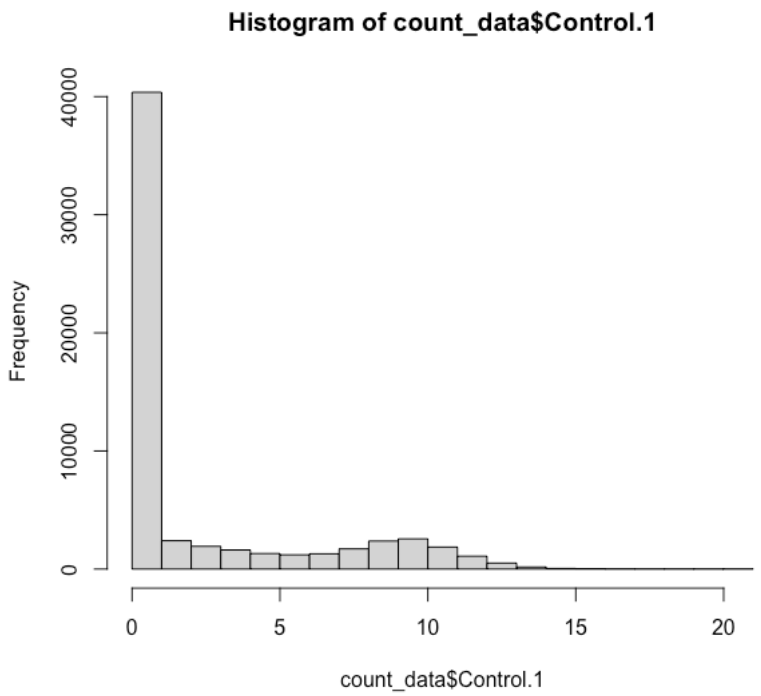
Method 1 $y = X\beta + \varepsilon$ $\beta = (X^T X)^{-1} X^T y$

Count = b0 + b1 * treatment → fold change, p value

Method 2

t. test → fold change, p value

Both linear regression and t test require normal distribution of expression count



# Data normalization

| | T1 | T2 | T3 | C1 | C2 | C3 |
|---|---|---|---|---|---|---|
| Gene1 | 10 | 20 | 25 | 1 | 1 | 1 |
| Gene2 | 24 | 40 | 45 | 1 | 2 | 1 |
| Gene3 | 5 | 10 | 8 | 2 | 1 | 2 |
| Gene4 | 23 | 20 | 20 | 150 | 160 | 90 |

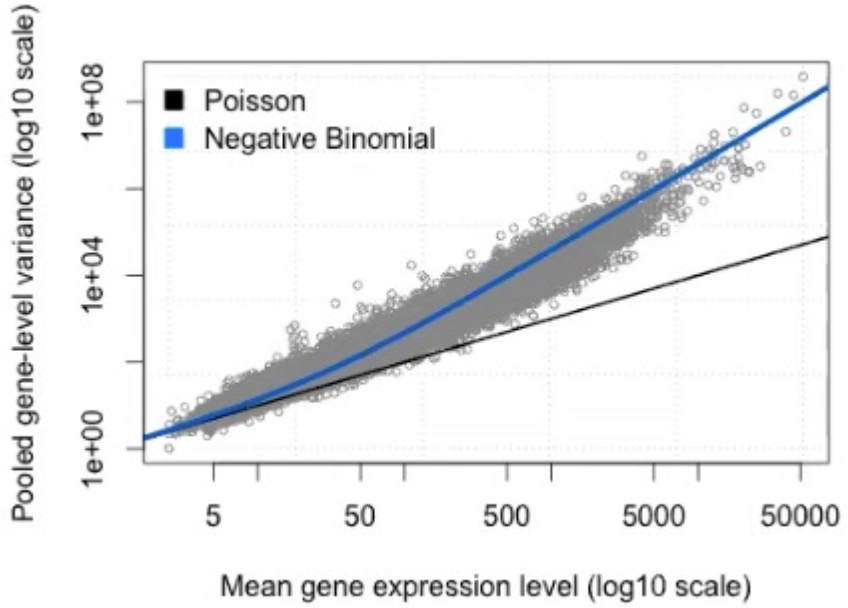
Normalize each gene across samples to adjust for differences in library composition

Normalize total read accounts per sample
Assuming every sample has same transcripts

Regardless of normalization methods, the counts are not normally distributed

# Negative Binomial Distribution

$NB(r, p)$: # of success before the first r failure given success probability p



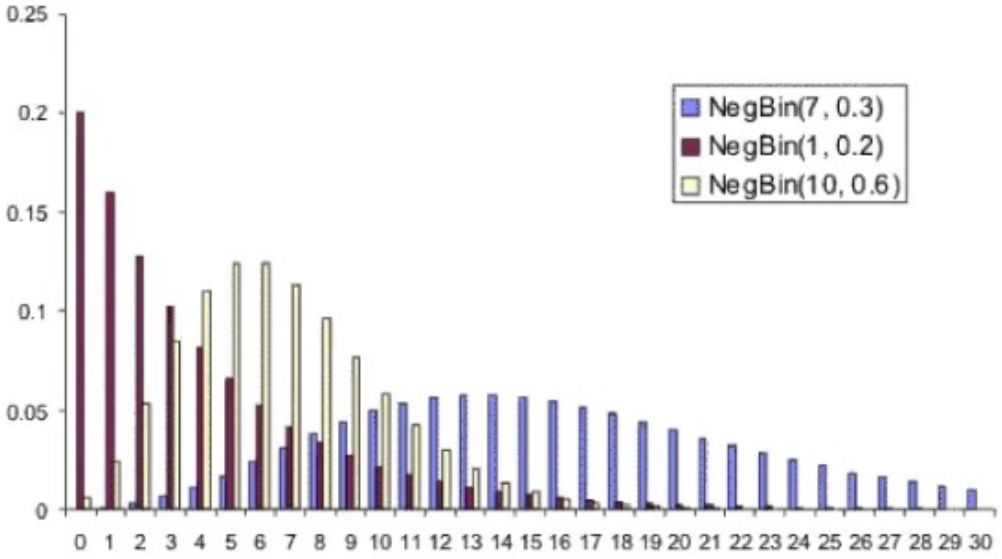

$$K_{ij} \sim NB(s_{ij} q_{ij}, \alpha_i)$$

Raw count for gene i in sample j

Normalization factor (library size)

Expression level

dispersion



# Generalized linear regression

| | T 1 | T 2 | T 3 | C 1 | C 2 | C 3 |
|---|---|---|---|---|---|---|
| Gene1 | | | | | | |
| Gene2 | | | | | | |
| Gene3 | | | | | | |
| Gene4 | | | | | | |

Method 1 $y = X\beta + \varepsilon$ $\beta = (X^T X)^{-1} X^T y$

Count = b0 + b1 * treatment → fold change, p value

$$E(Y|X) = \mu = g^{-1}(X\beta)$$

$$\mu_j = s_j * \exp(X\beta_j)$$

$$\log(\mu_j) = \log(s_j) + X\beta_j$$

Generalized linear model: a unified framework for modeling various types of response variables by specifying a distribution from the exponential family, a linear predictor, and a link function

# DESeq2 approach:

1. **Negative Binomial Distribution:**

Raw counts

dispersion

$$Y_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$$

2. **Mean of Counts:**

$$\mu_{ij} = s_j \cdot \exp(\eta_{ij})$$

where:

- $\mu_{ij}$ is the mean count for gene $i$ in sample $j$.
- $s_j$ is the size factor for sample $j$.
- $\eta_{ij}$ is the linear predictor for gene $i$ in sample $j$.

3. **Variance of Counts:**

$$\text{Var}(Y_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

**Link Function**:

- The GLM in DESeq2 uses a logarithmic link function to relate the mean count $\mu_{gi}$ to the linear predictors:

$$\log(\mu_{gi}) = \eta_{gi}$$

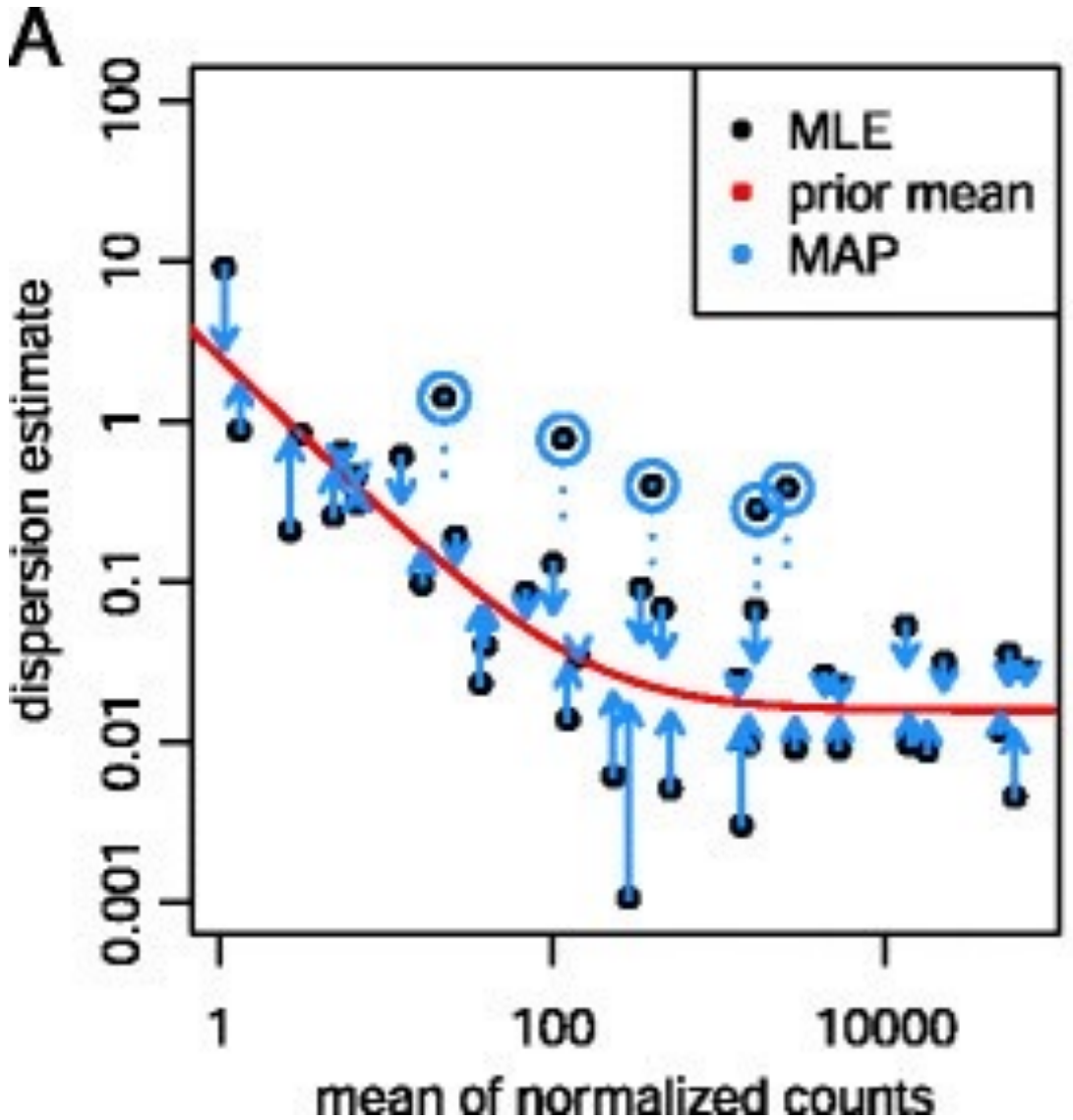
- The linear predictor $\eta_{gi}$ is specified as:

$$\eta_{gi} = \log(s_i) + X_i \beta_g$$

where:

- $s_i$ is the size factor for sample $i$, accounting for differences in sequencing depth.
- $X_i$ is the design matrix row corresponding to sample $i$.
- $\beta_g$ is the vector of coefficients for gene $g$.

# DESeq2: Dispersion estimation



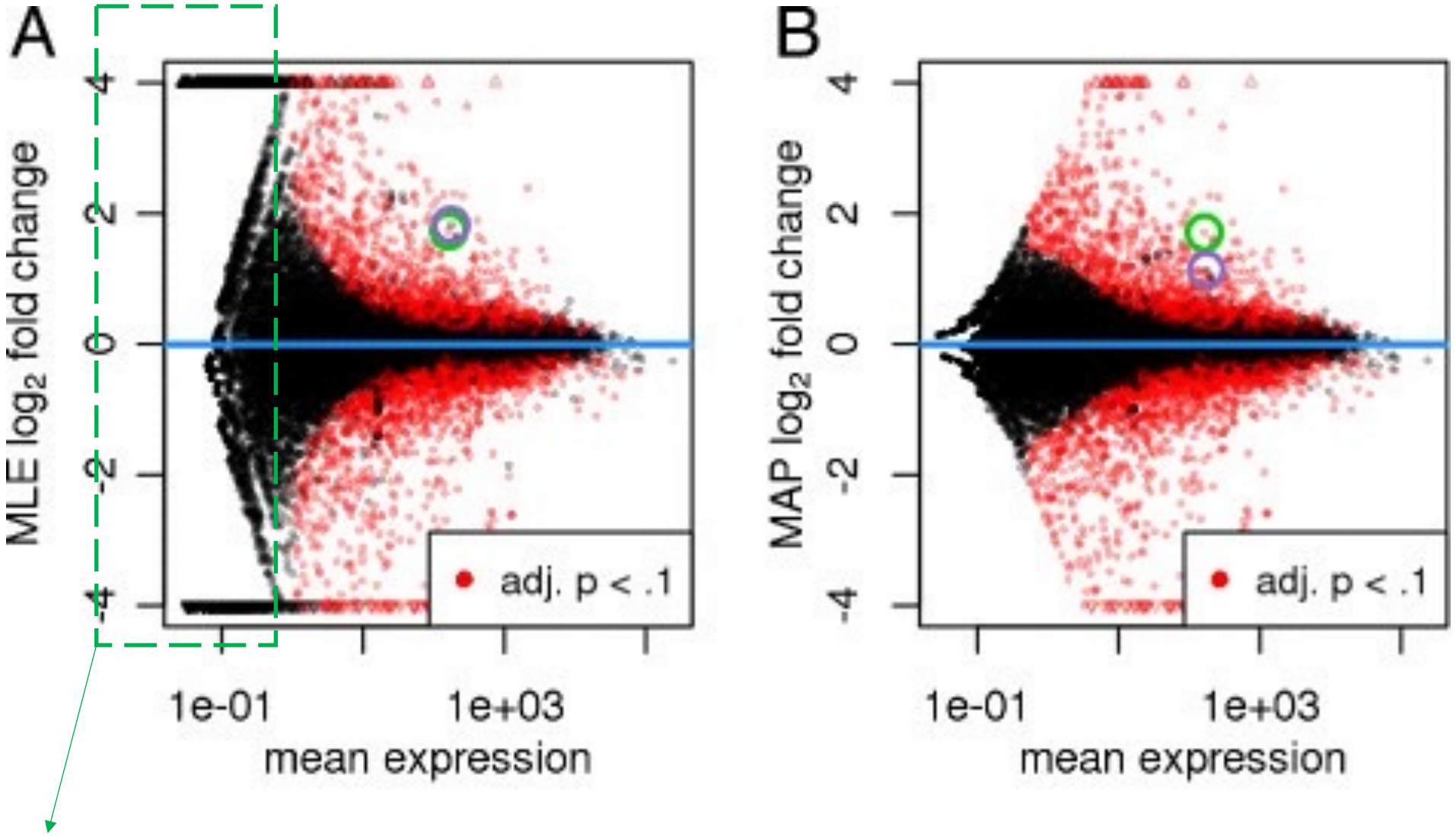
# Design Matrix

```
           (Intercept) conditionInfection conditionTreatment
Control.1            1                  0                  0
Control.2            1                  0                  0
Control.3            1                  0                  0
Infection.1          1                  1                  0
Infection.2          1                  1                  0
Infection.3          1                  1                  0
B5uM.1               1                  0                  1
B5uM.2               1                  0                  1
B5uM.3               1                  0                  1
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$condition
[1] "contr.treatment"
```

# DESeq2:

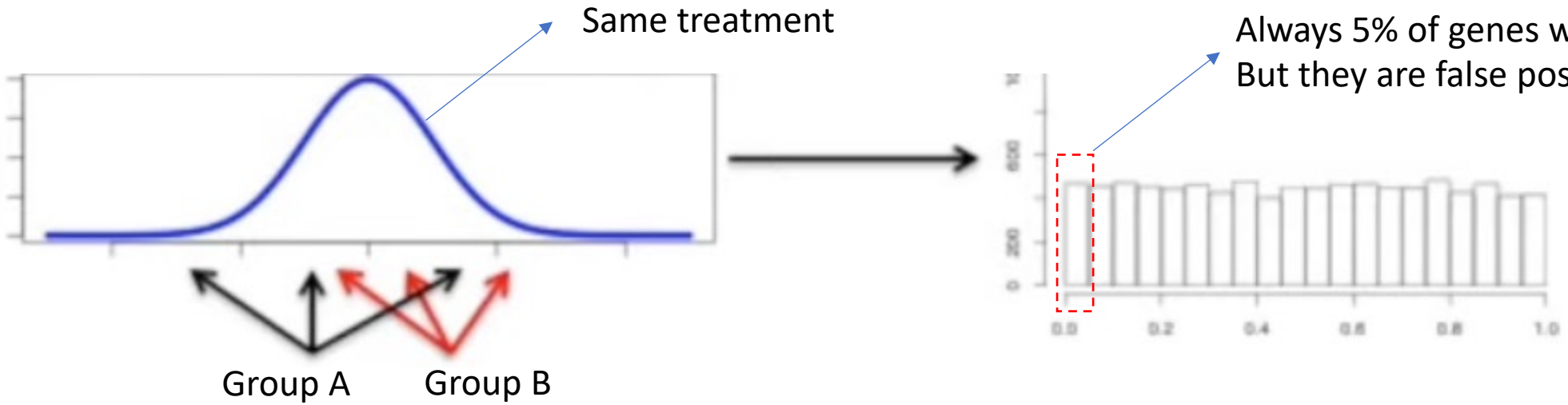
```
dds <- DESeqDataSetFromMatrix(countData = count_data_raw,
                              colData = sample_info,
                              design = ~ condition)

# Run the DESeq pipeline
dds <- DESeq(dds)

# Get the results
res <- results(dds, contrast=c("condition","Infection","Control"))
```

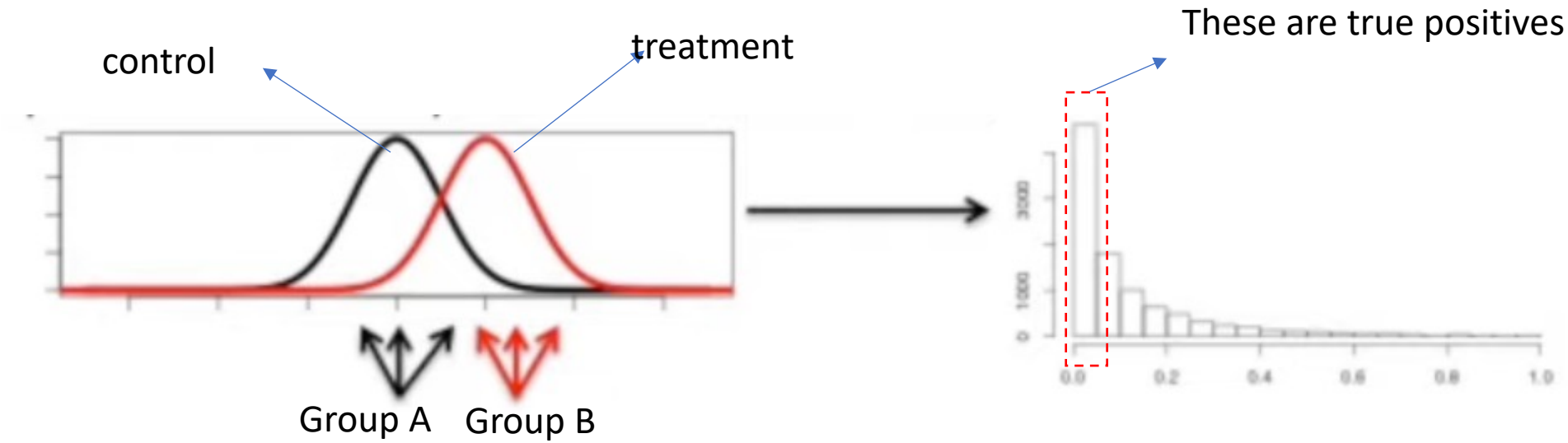
# DESeq2: shrink fold changes



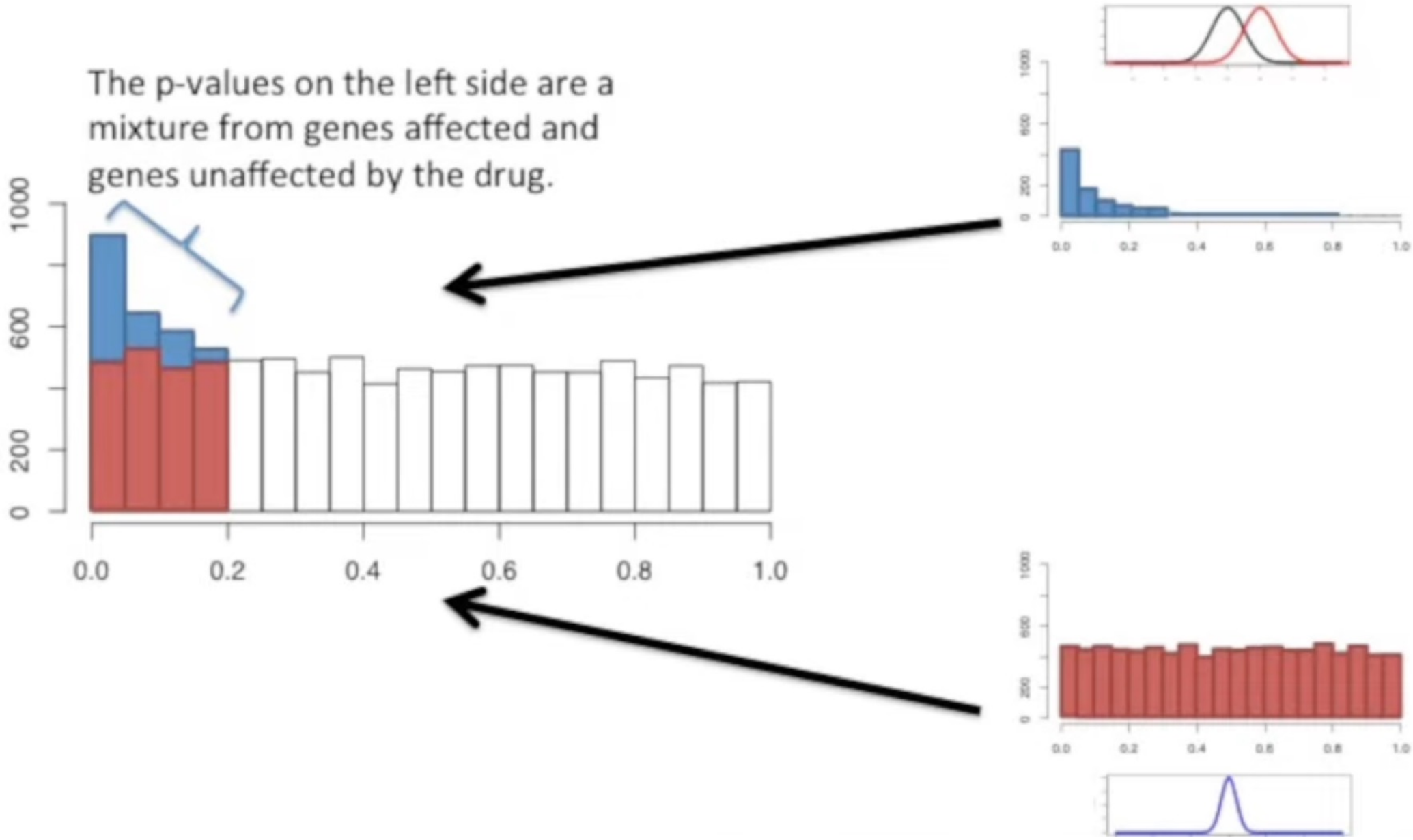

The fold changes of those lowly expressed genes are adjusted

# FDR: why do we need to adjust p value?

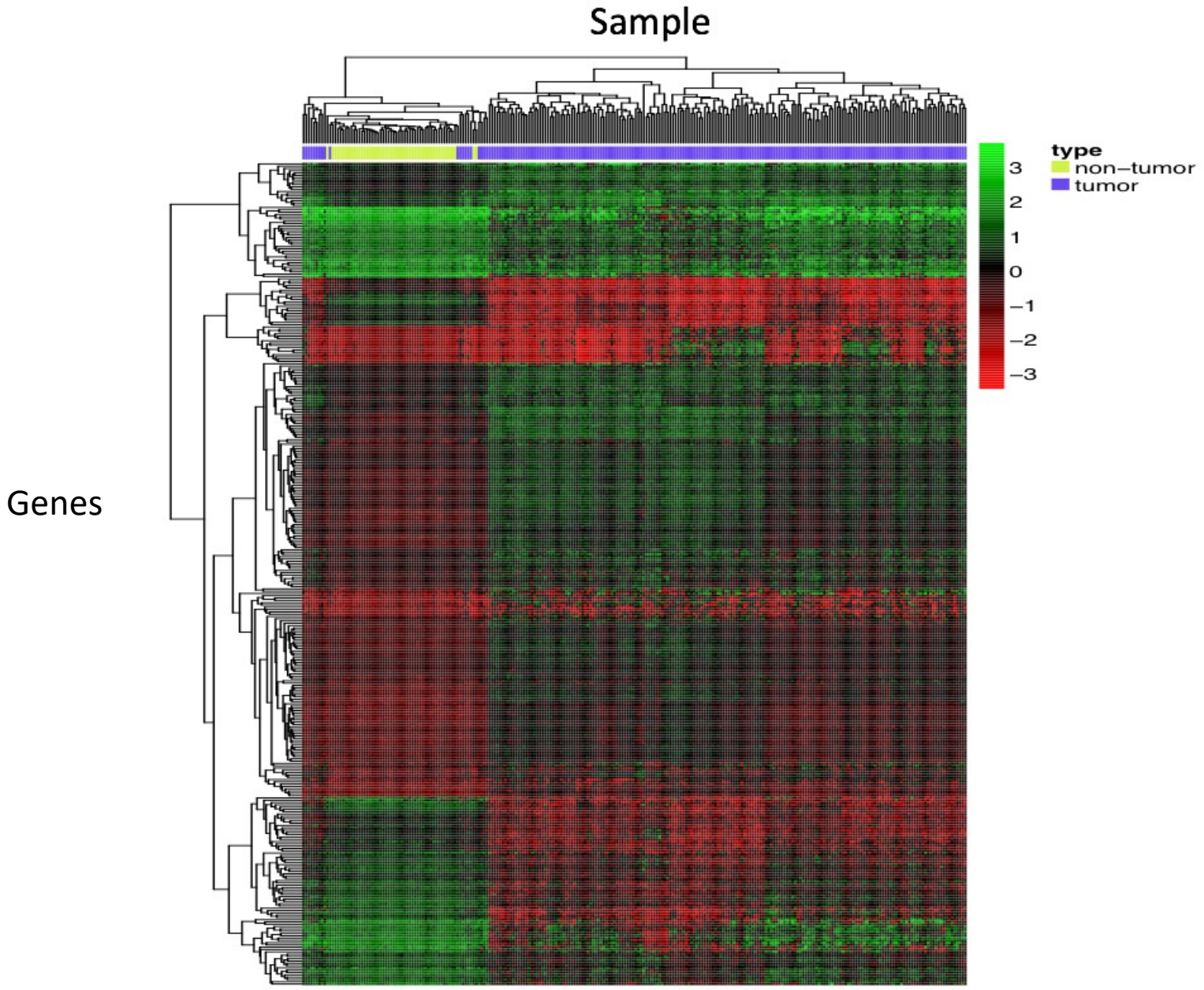


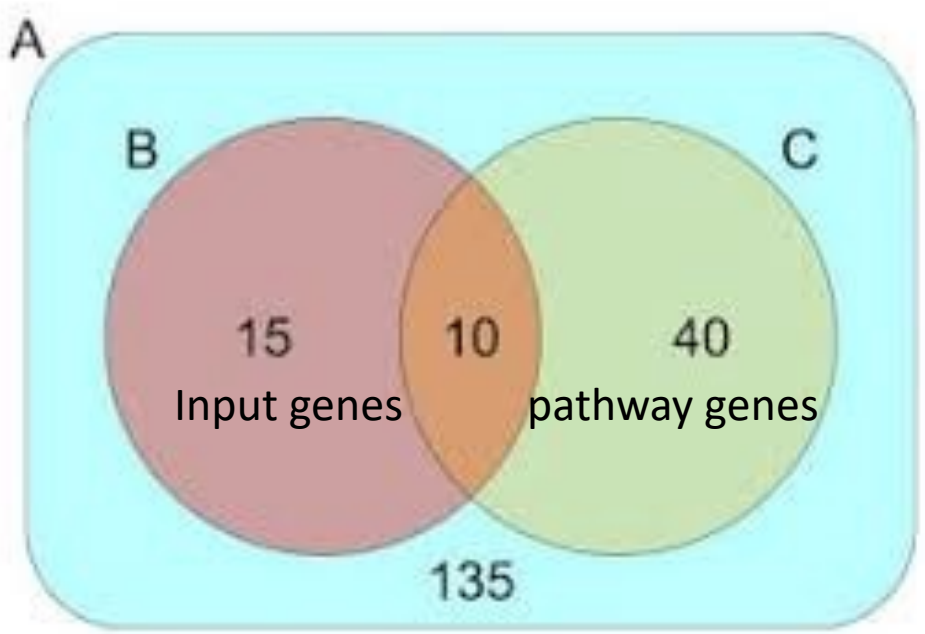
# FDR: how to we correct p value?



# Clustering

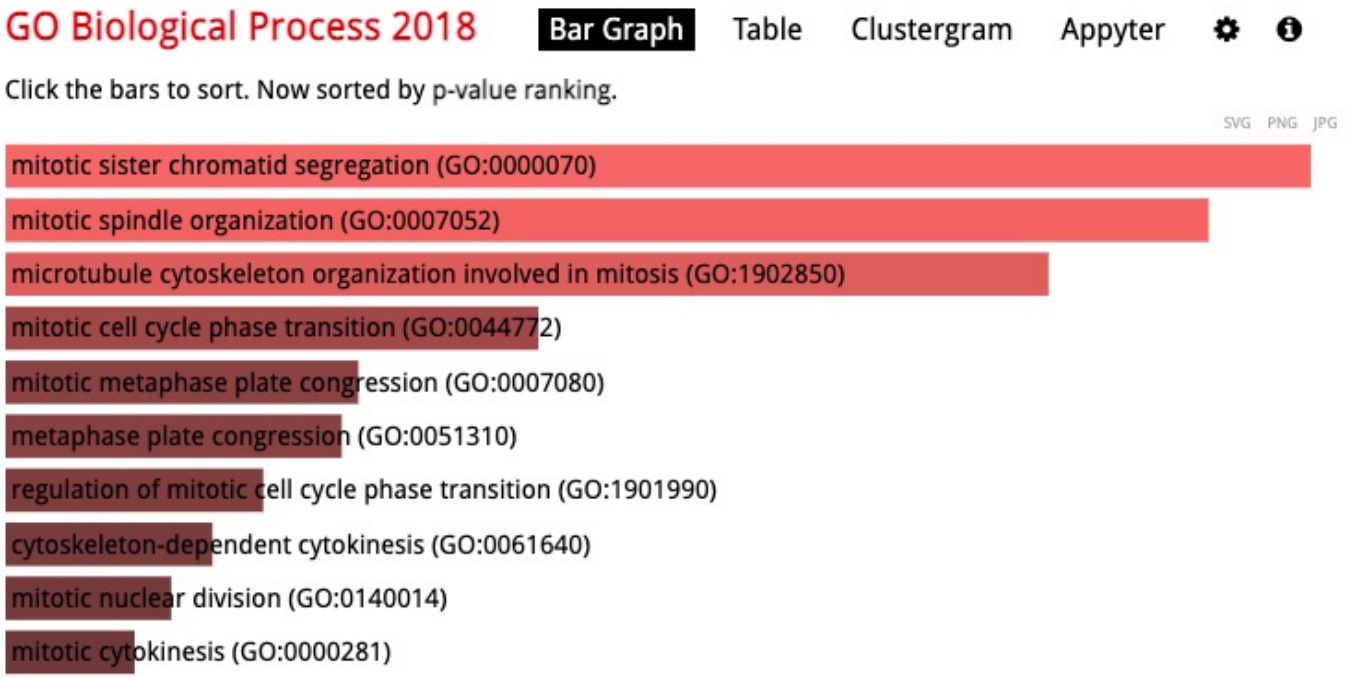


# Gene Set Enrichment Analysis




hypergeometric test

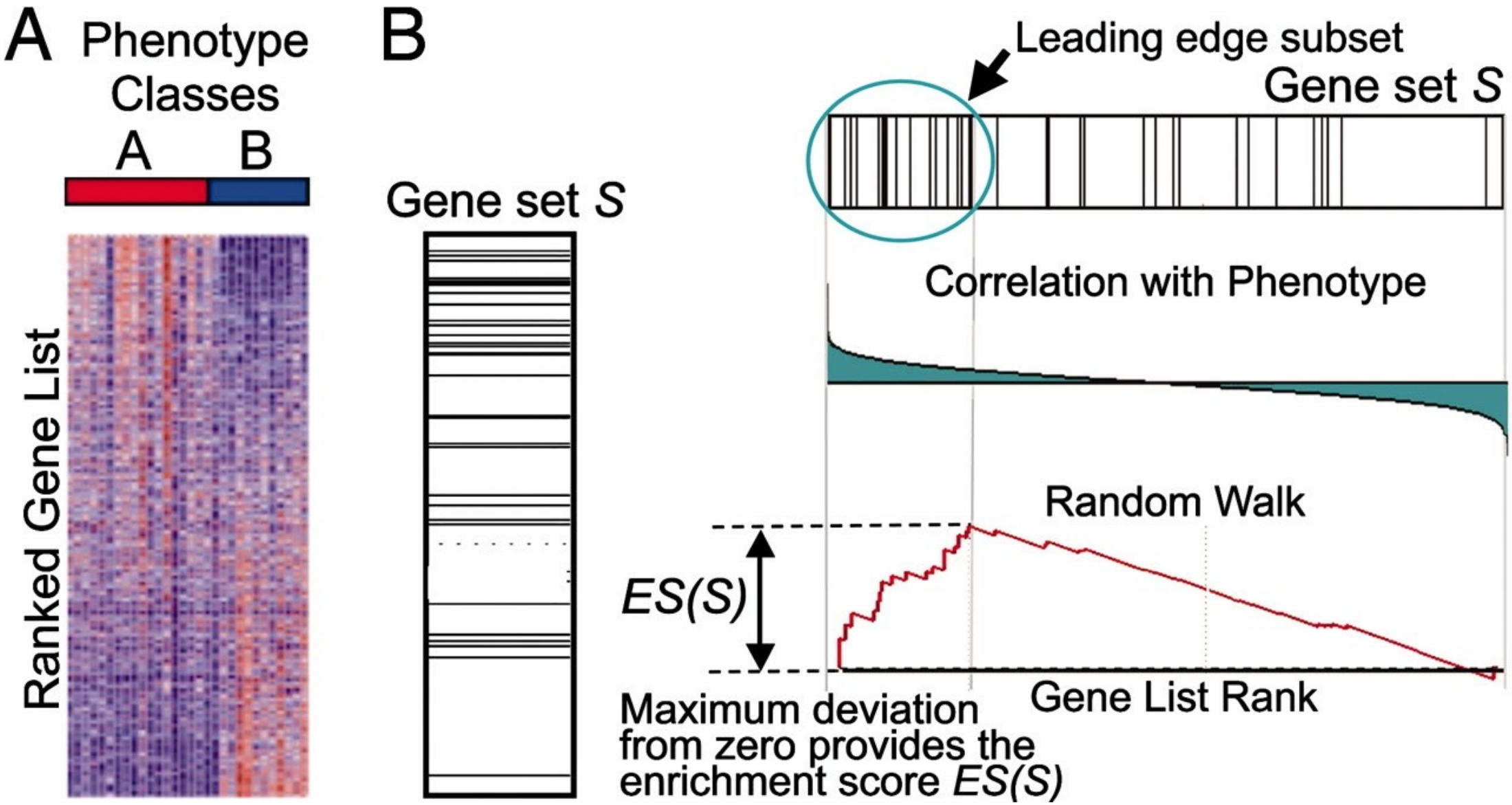
$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$



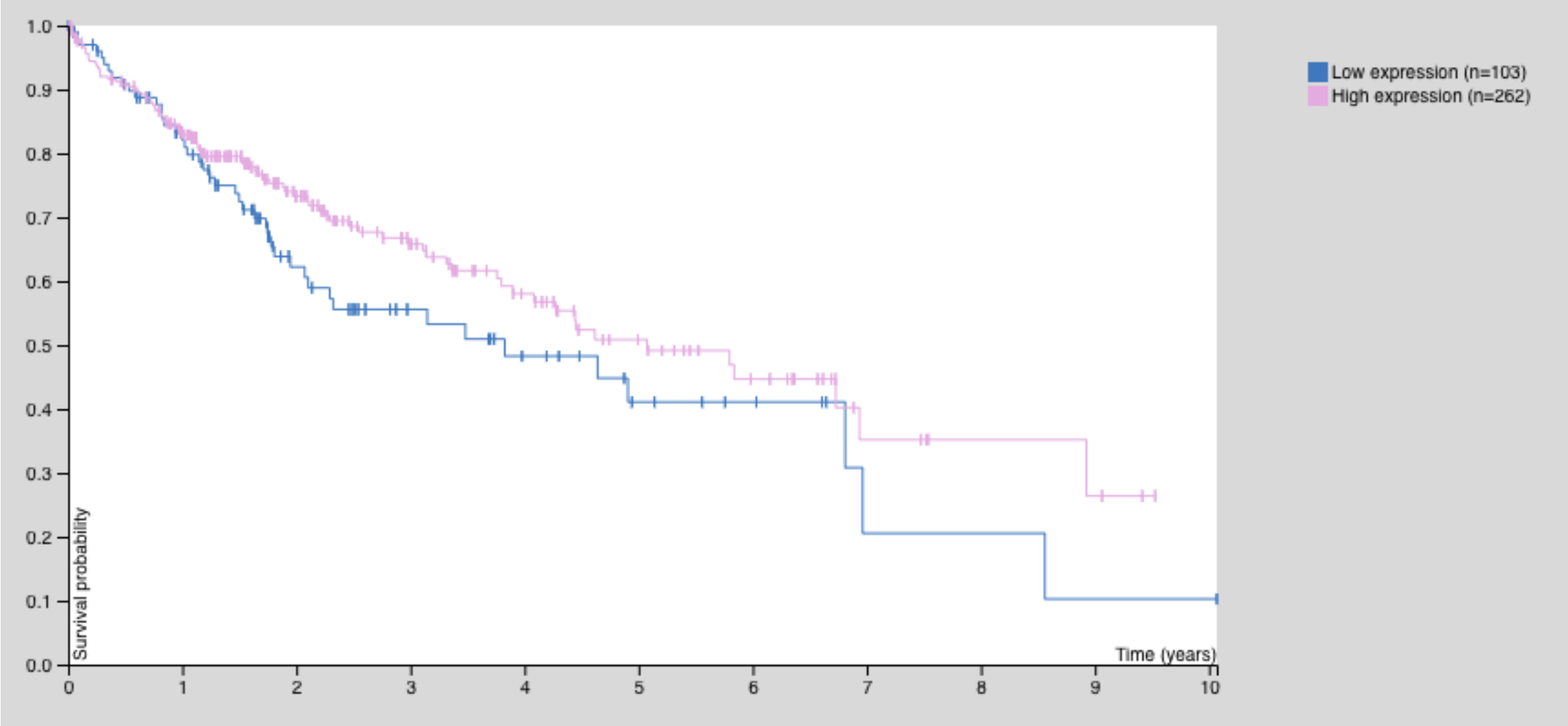

**N**: Total number of genes in the background. **K**: Number of genes in the pathway. **n**: Number of genes in your gene list of interest. **k**: Number of genes in the overlap between your gene list and the pathway.

https://maayanlab.cloud/Enrichr

# Gene Set Enrichment Analysis

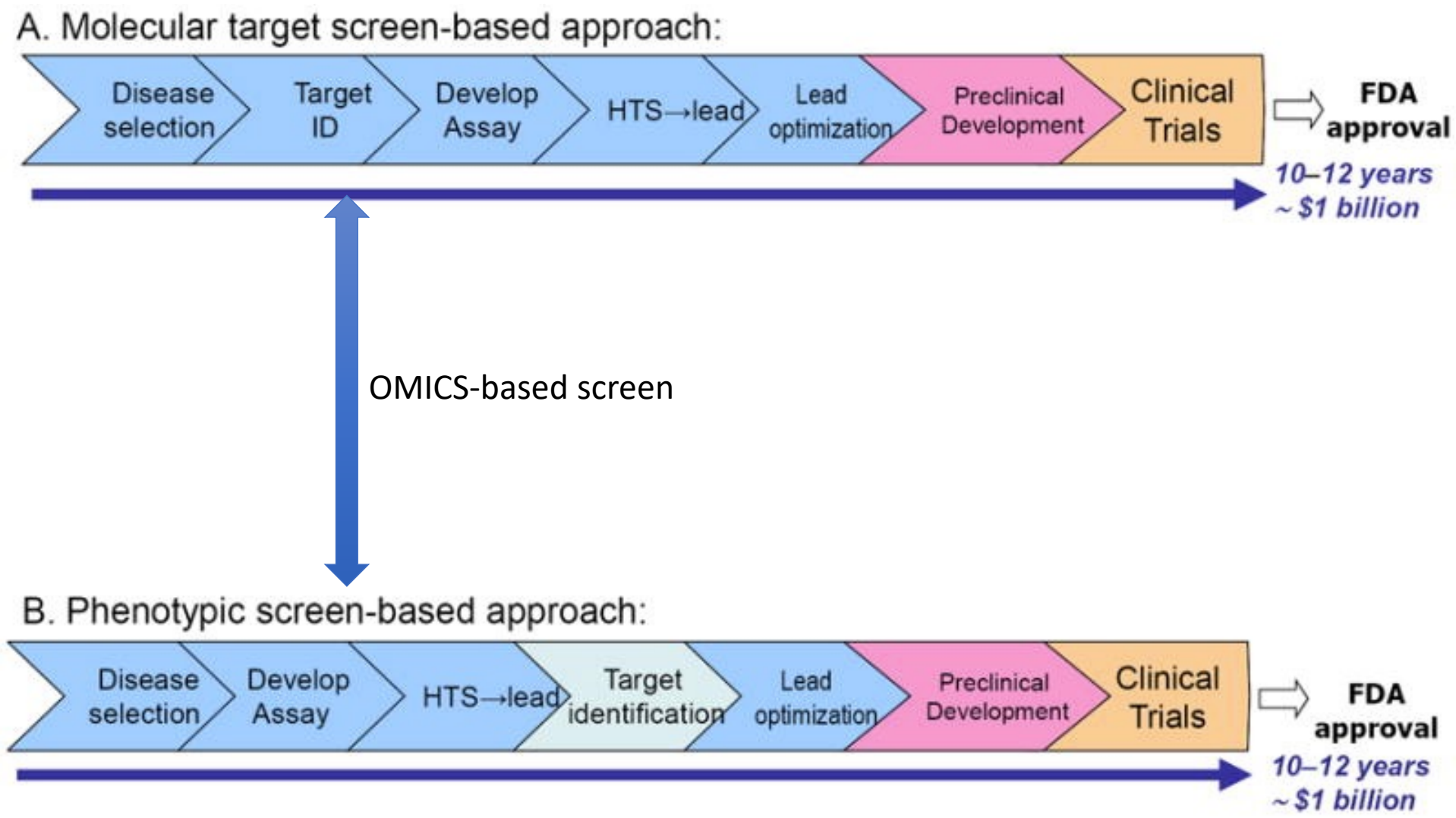


# Survival analysis



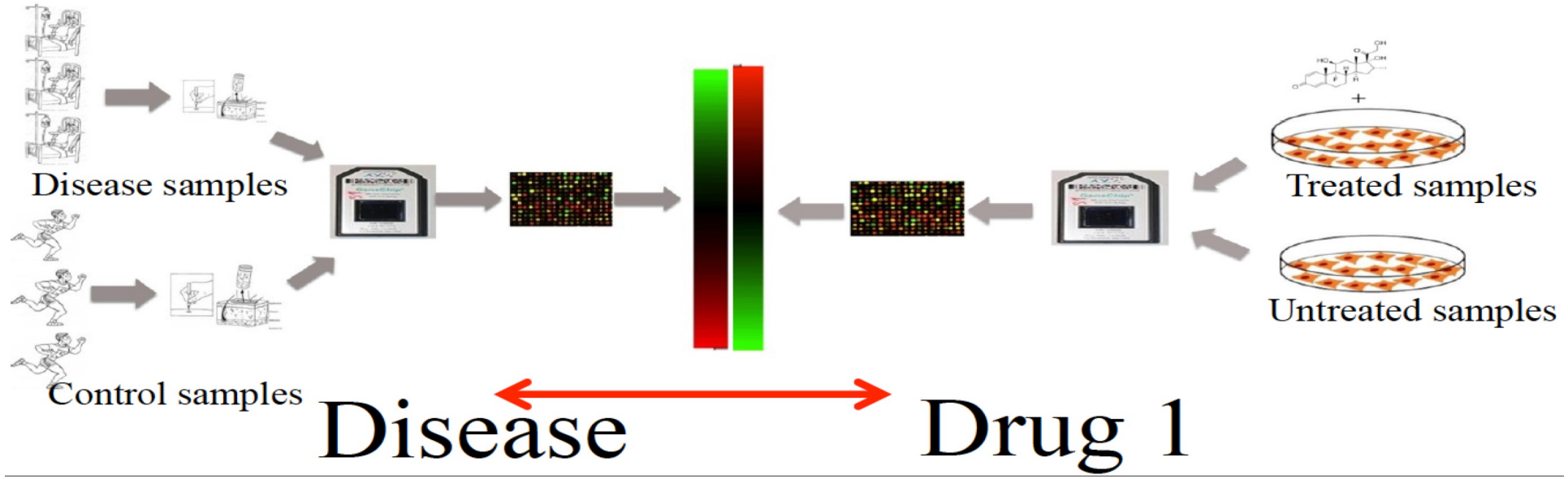

Kaplan–Meier plot

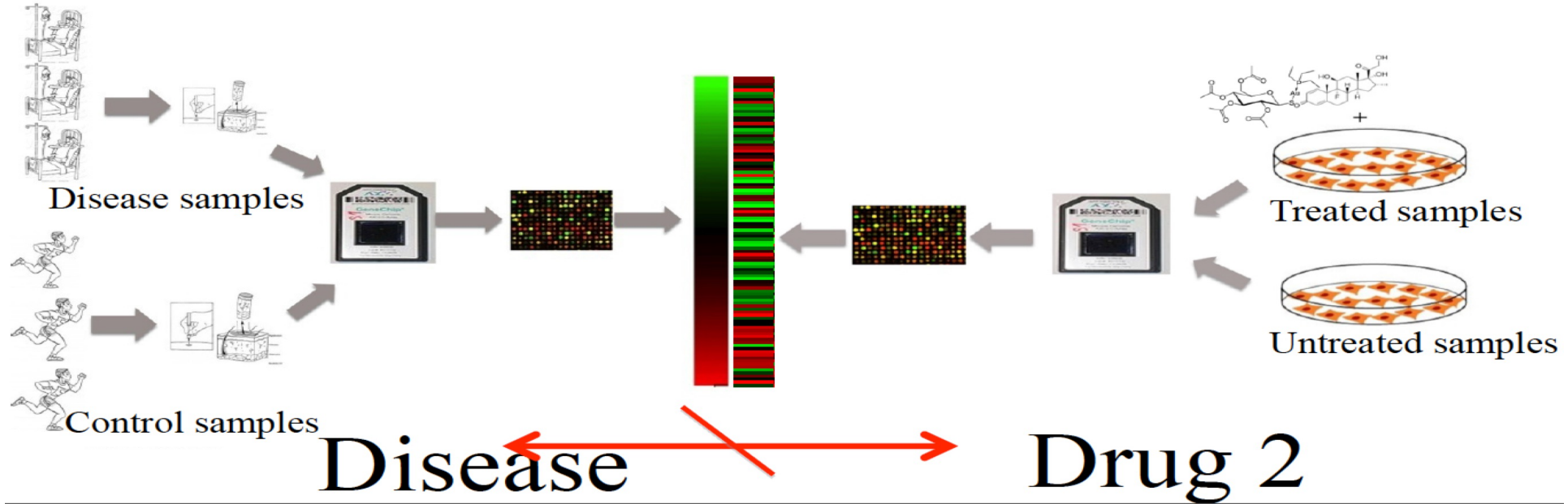
proteinatlas.org/



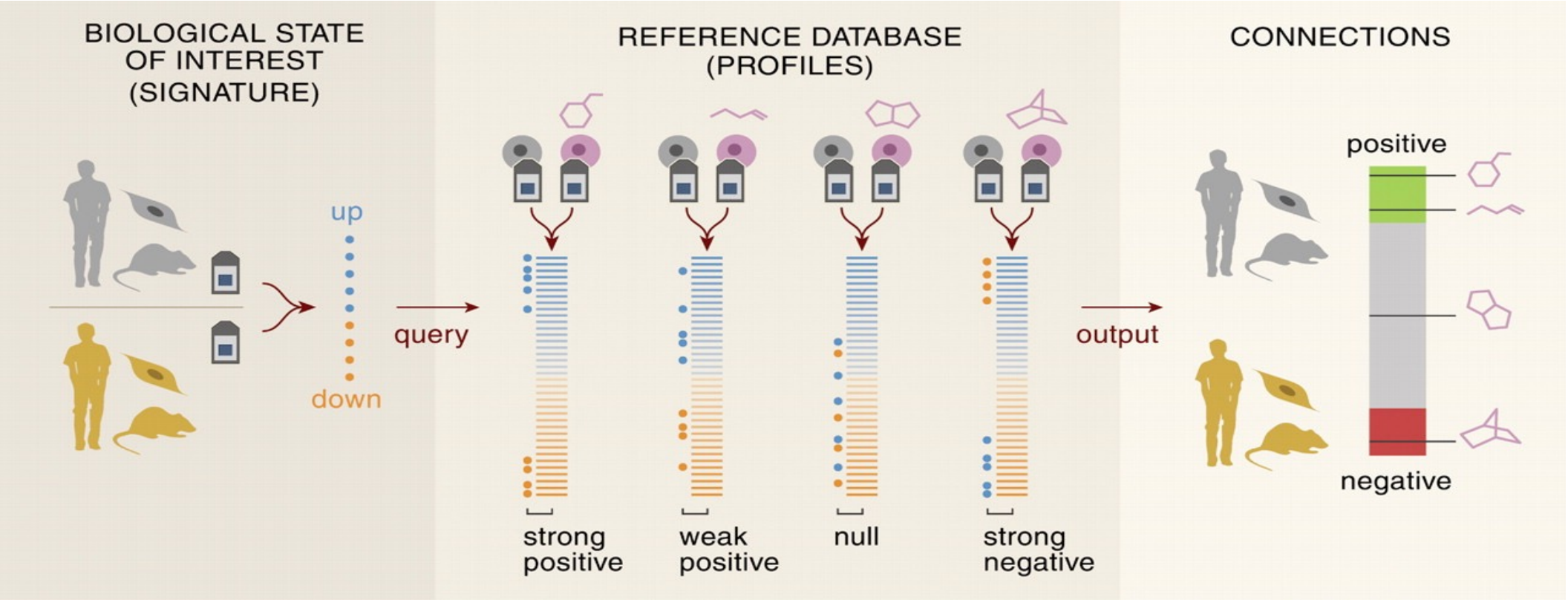

Phenotype = f(omics)

# Systems-based drug discovery approach



# Connectivity Map 1.0

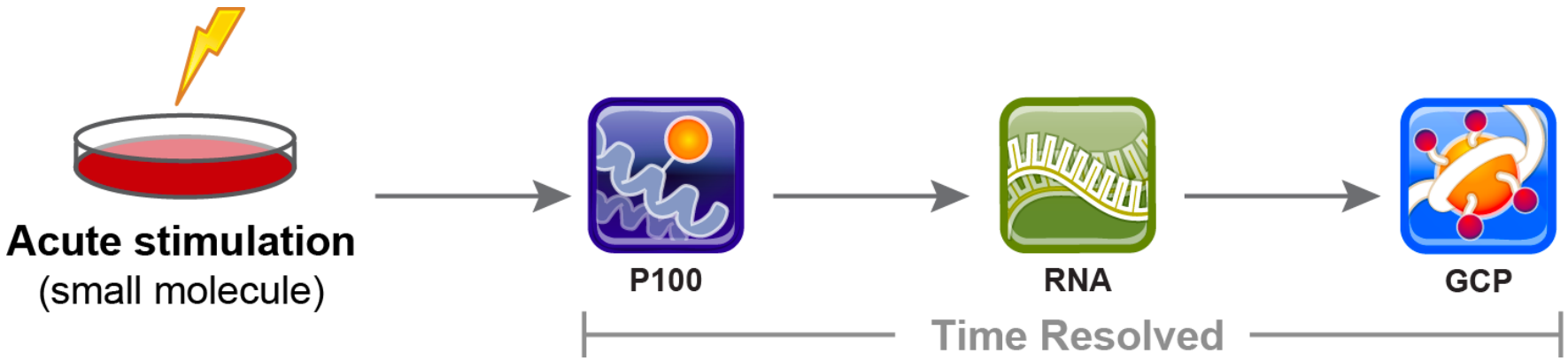



Connectivity Map Score/ reversal score: summary score of enrichment of disease up/down genes on the ranked drug profile

500 gene expression profiles → 6K (CMap) → 1M (LINCS)

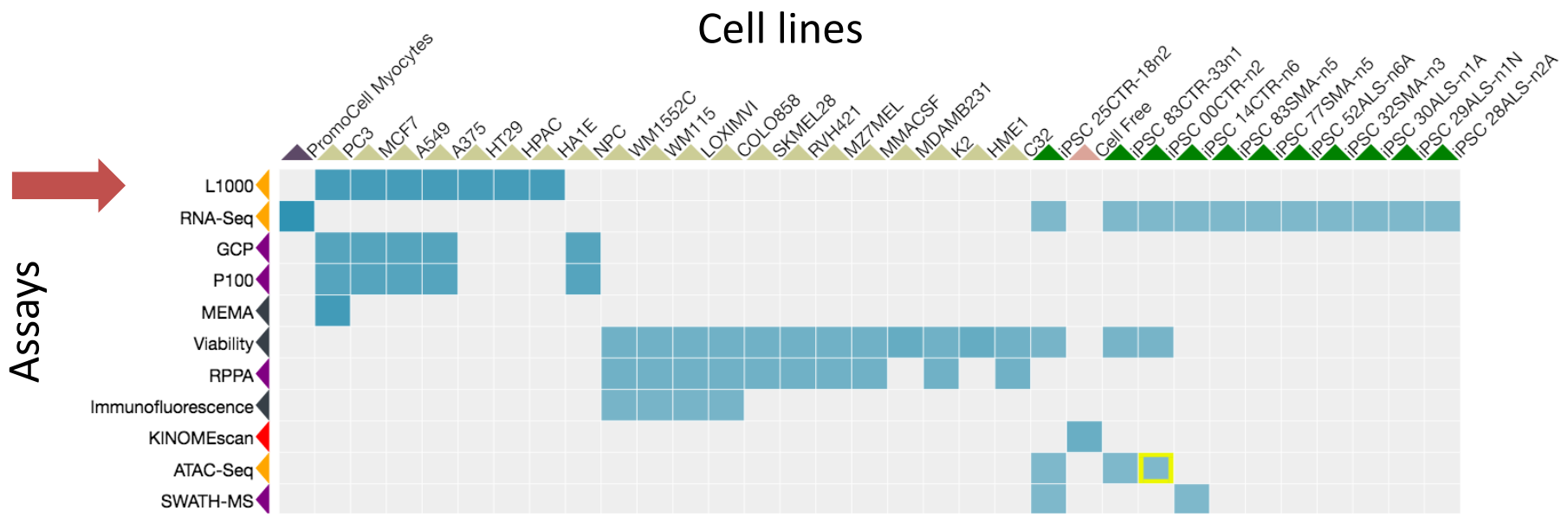
**Justin Lamb, Science 2006**

# Connectivity Map 2.0 (LINCS)




A Subramanian, 2017, Cell

# Library of Network-Based Cellular Signatures (LINCS)




421 Datasets

41847 Small Molecules

1127 Cells

978 Genes

1469 Proteins /155 Peptide Probes

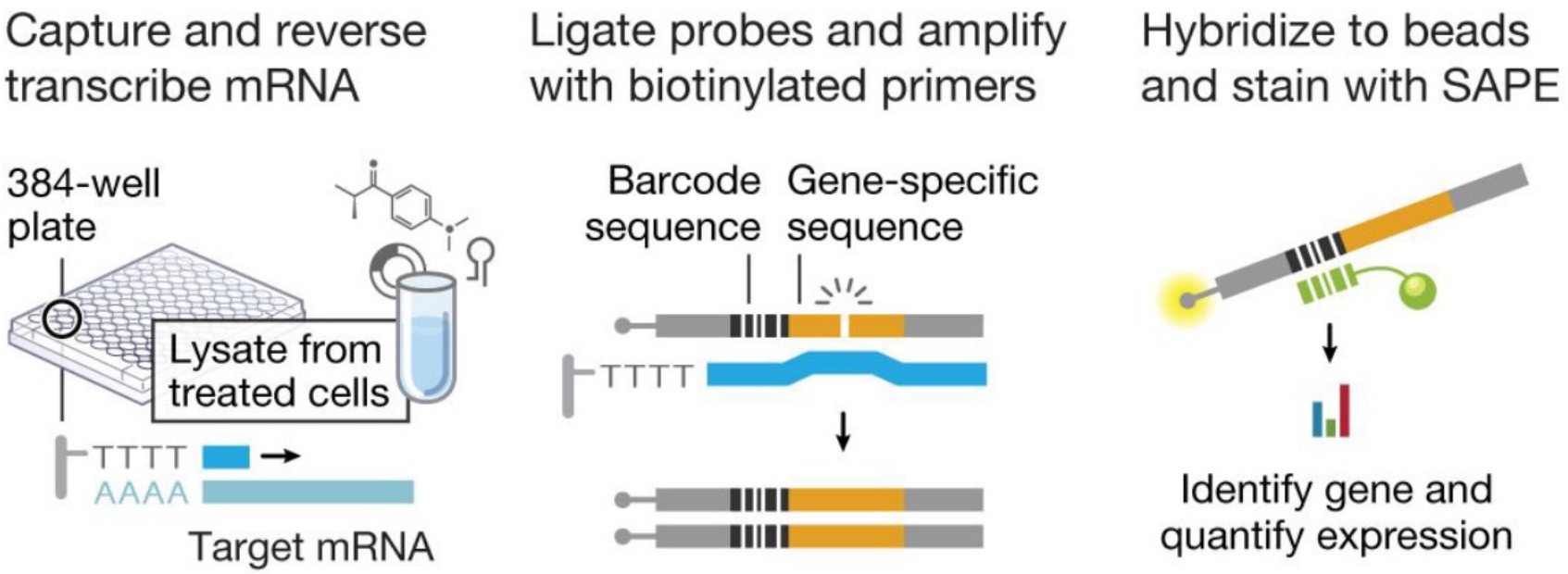
117 Antibodies

http://www.lincsproject.org/data/data-releases/

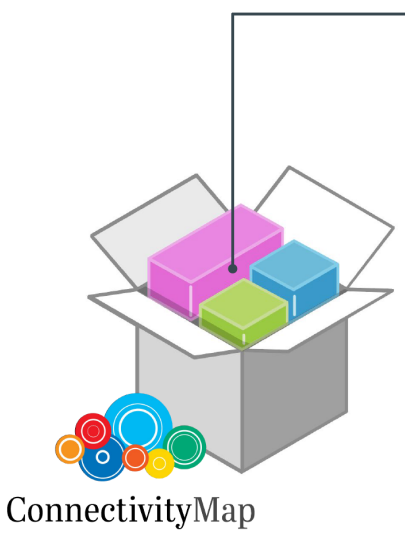
# Cost of profiling 1,000,000 samples

- If using RNAseq
  - 1,000,000 * $200 per sample = $200M
- If using LINCS 1000
  - 1,000,000 * $2 per sample = $2M

# Connectivity Map 2.0 (LINCS L1000)




A Subramanian, 2017, Cell

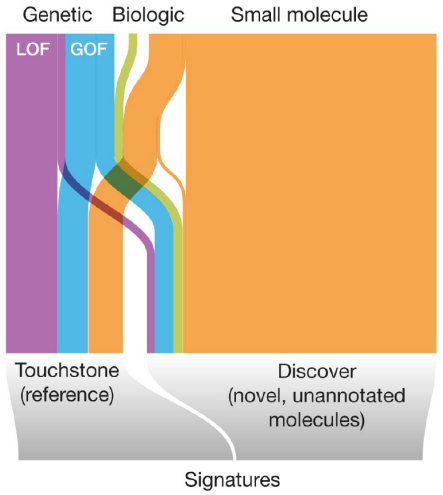



## L1000

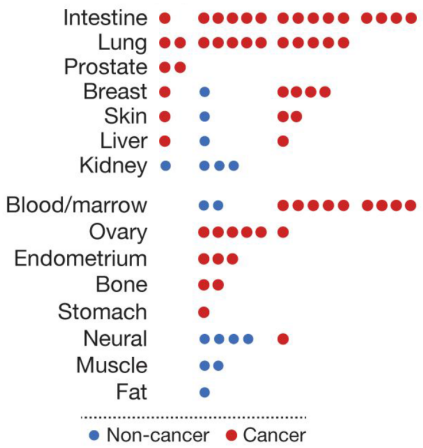
## Pattern-recognition algorithms

## 1.3M profiles

28,000 perturbagens




80 Cell lines




Subramanian et al, *Cell* 2017

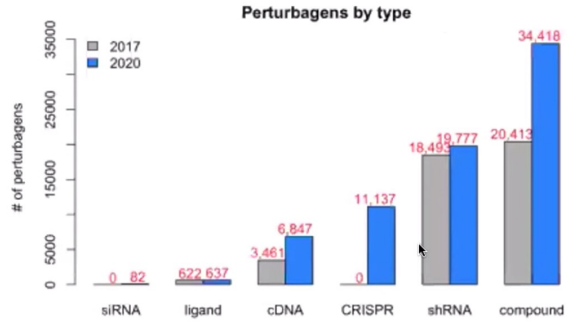
# LINCS L1000 phases

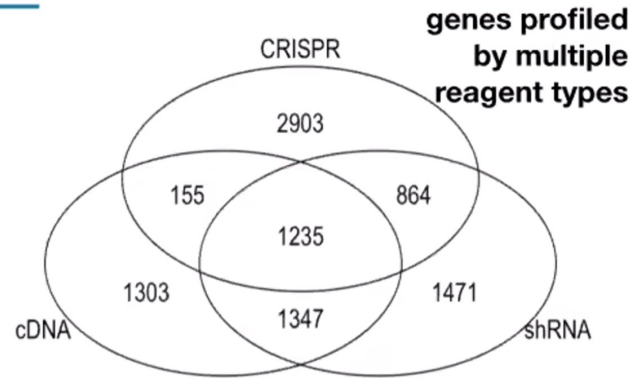
2017: Phase I: 1,319,138 (GSE92742) + Phase II: 354,123 (GSE70138)
2020: 3M (manuscript unpublished)

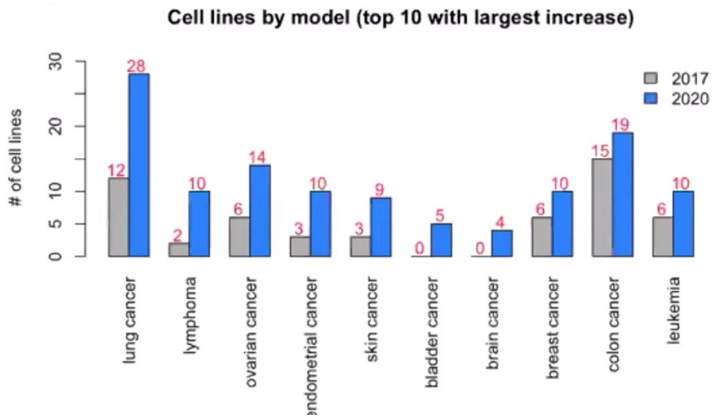
## Overview of updated resource

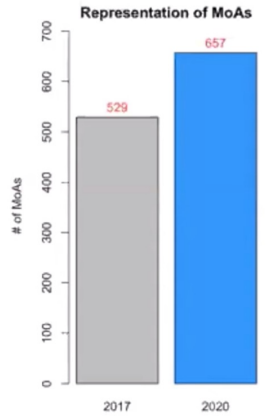
### expansion upon previous release

| | |
|---|---|
| **# perturbagens** | 81,979 |
| **# compounds** | 33,609 |
| **# MoAs** | 657 |
| **# genes** | 9,288 |
| **# cell contexts** | 240 (12 primary) |
| **# profiles** | 3.02M |
| **# signatures** | 1.16M |

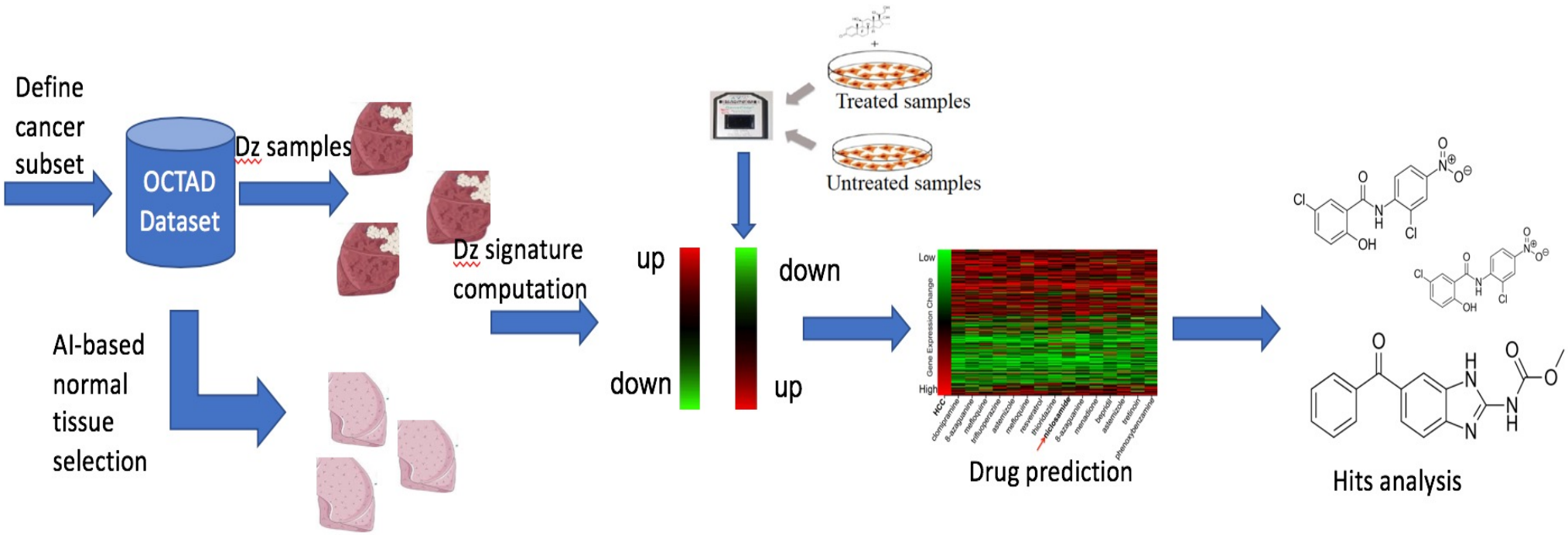


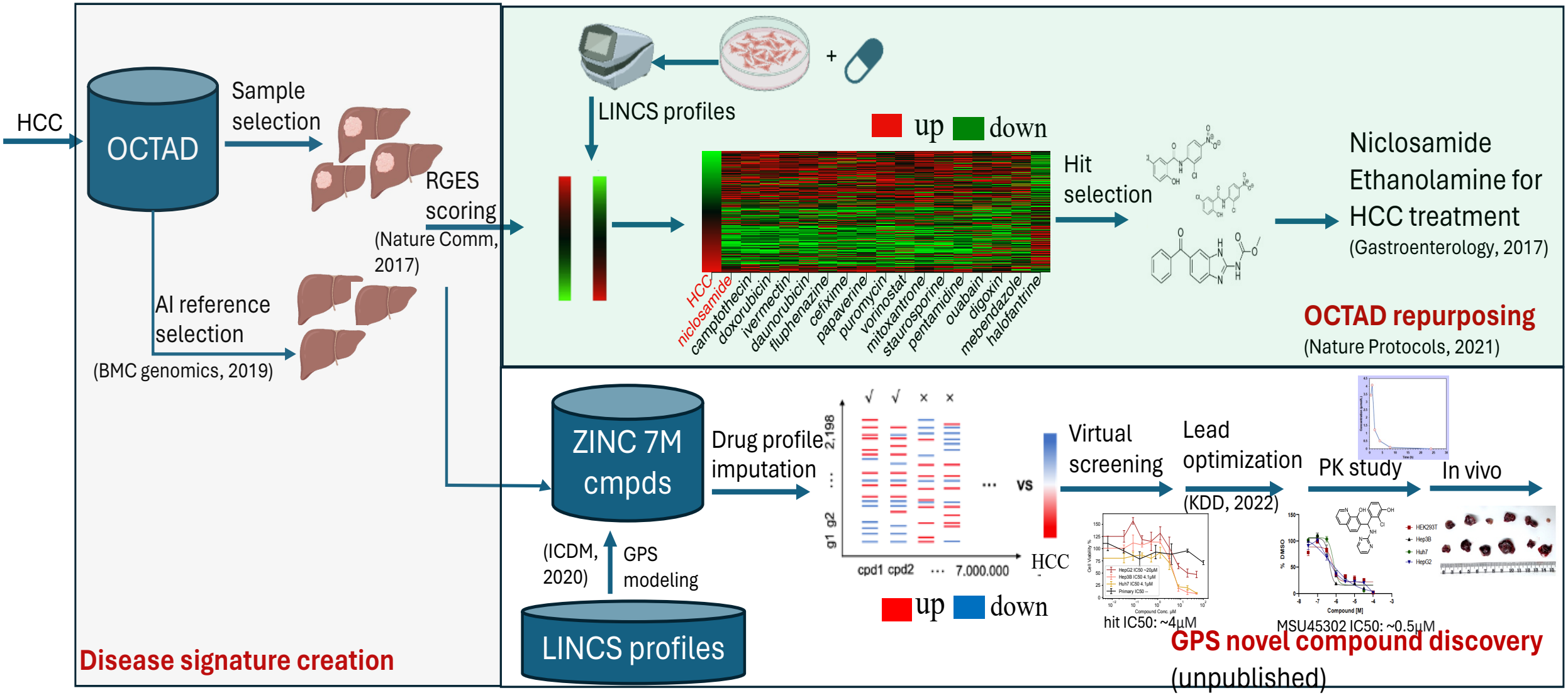



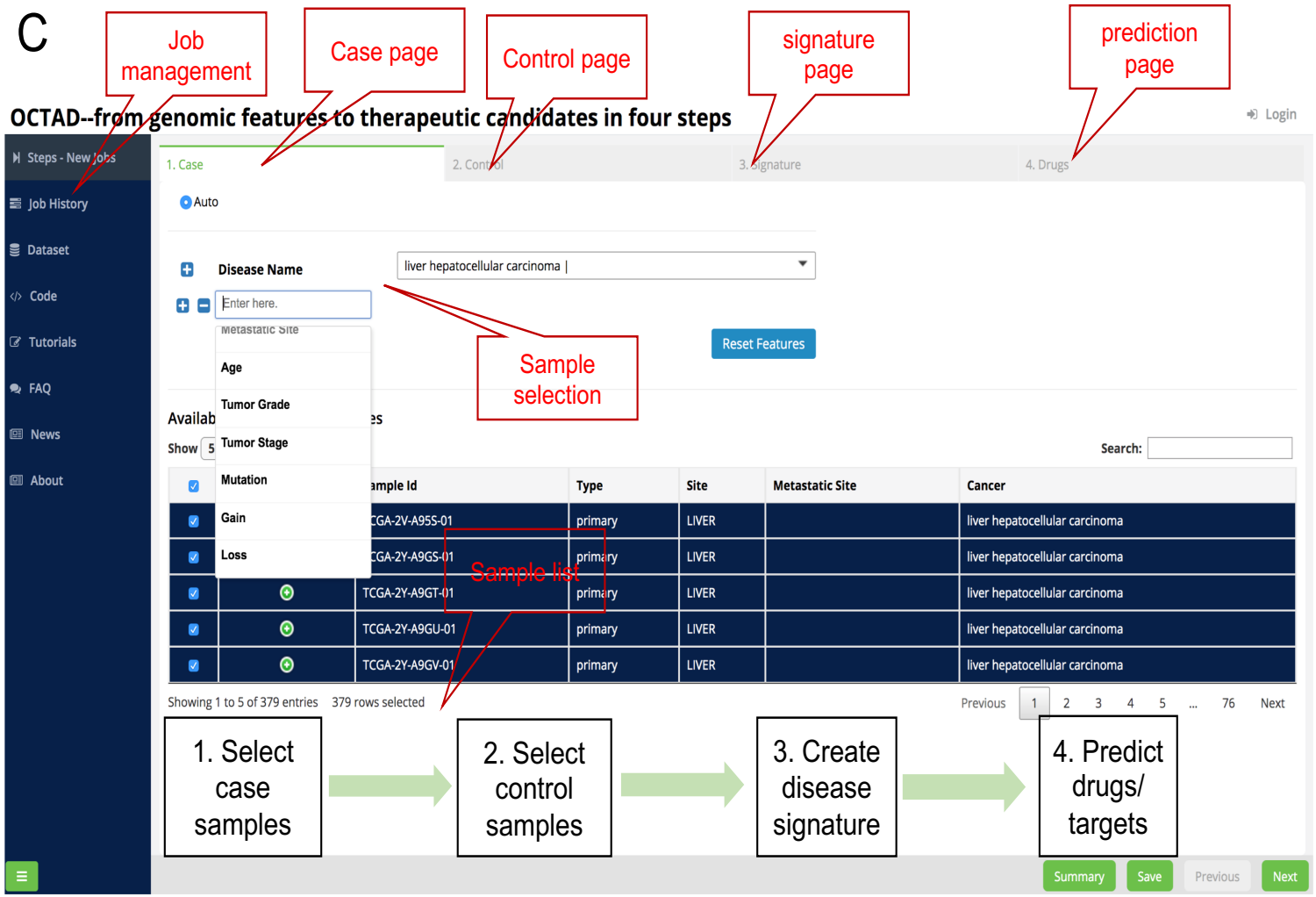
# Transcriptomics based drug repurposing pipeline



# Transcriptomics-based drug discovery platform



# OCTAD Web portal




#Billy Zeng, #Benjamin S. Glicksberg, #Patrick Newbury, #Evgenii Chekalin, Jing Xing, Ke Liu, Anita Wen, Caven Chow, Bin Chen, OCTAD: an open workplace for virtually screening therapeutics targeting precise cancer patient groups using gene expression features accepted, Nature Protocols

# OCTAD R package

```
case_id=subset(phenoDF,cancer=='liver hepatocellular carcinoma'&sample.type == 'primary’, select = c(“sample.id”)

HCC_adjacent=subset(phenoDF,cancer=='liver hepatocellular carcinoma'&sample.type == 'adjacent'&data.source == 'TCGA’, select = c(“sample.id”)

res=diffExp(case_id,control_id,source='octad.whole',output=T,n_topGenes=10000,file='octad.counts.and.tpm.h5')

sRGES=runsRGES(res,max_gene_size=500,permutations=10000)

head(sRGES)
```