

Statistical Methods for High-dimensional Biology



Gene set enrichment analysis

Yongjin Park, UBC Path&Lab, STAT, BC Cancer

Today's lecture: Enrichment Analysis

- **Motivations: What's next after genomics analysis?**
 - What have we learned?
 - How do we know that our discovery is meaningful?
- **Gene set enrichment analysis**
 - Set-based approach: Hypergeometric test
 - Rank-based approach: GSEA by KS statistic

edgeR in action

```
1 des <- model.matrix(~ Sex + Group + DPC, data = colData(sumexp))
2 dge <- DGEList(counts = assays(sumexp)$counts,
3                  samples = colData(sumexp))
4 dge <- calcNormFactors(dge)
5 dge <- estimateDisp(dge)
6 QLfit <- glmQLFit(dge, design = des)
7 QLtest_Group <- glmQLFTest(QLfit, coef = "GroupMu")
8
9 topTags(QLtest_Group)
```

Coefficient: GroupMu

		logFC	logCPM	F	PValue	FDR
Chd8	-0.5910169	7.168410	136.93350	1.060551e-14	1.289418e-10	
Dnajc4	0.3318225	3.483504	45.44089	3.682335e-08	2.238491e-04	
Vrk3	0.2278501	5.018357	40.30631	1.334186e-07	5.407010e-04	
Hmgcll1	-0.2725199	4.679259	33.17377	9.253342e-07	2.463745e-03	
Gh	-1.0930426	4.327084	35.11644	1.013220e-06	2.463745e-03	
Xrcc4	0.3206540	3.800252	31.97145	1.307488e-06	2.649407e-03	
Myef2	-0.2631618	6.465549	29.11564	3.045538e-06	4.068467e-03	
Lrrc48	0.4081822	3.078380	28.94894	3.204690e-06	4.068467e-03	
Usp11	-0.2680673	7.332344	28.89415	3.256932e-06	4.068467e-03	
Anxa11	0.4735432	2.957357	28.47633	3.716860e-06	4.068467e-03	

DESeq2 in action

```
1 des <- model.matrix(~ Sex + Group + DPC, data = colData(sumexp))
2 dds <- DESeqDataSet(sumexp, design = des)
3 dds <- estimateSizeFactors(dds)
4 dds <- DESeq(dds)
5
6 res <- results(dds, name = "GroupMu")
7 res[order(res$padj), ]
```

log2 fold change (MLE): GroupMu

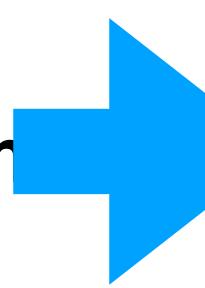
Wald test p-value: GroupMu

DataFrame with 12158 rows and 6 columns

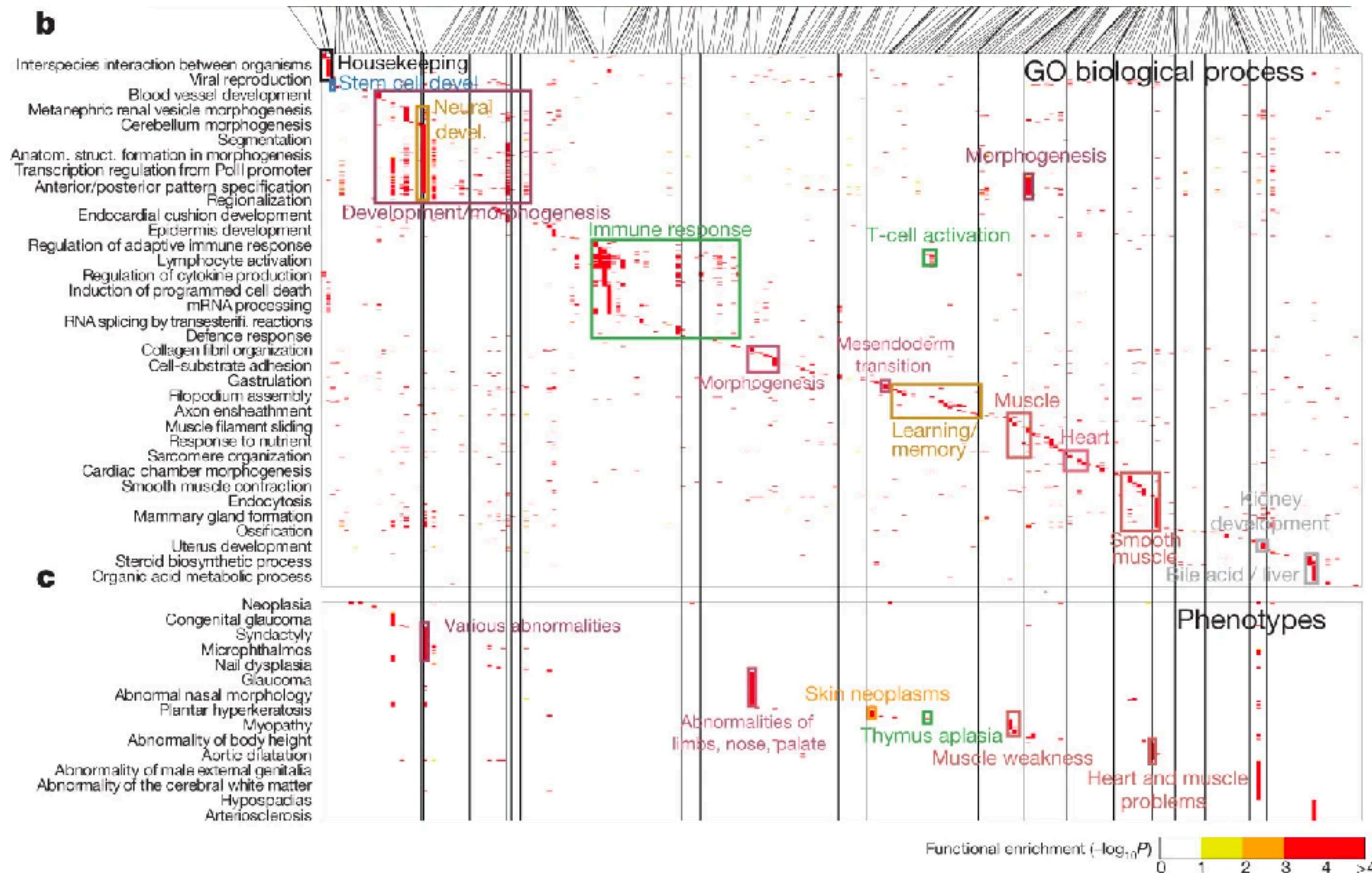
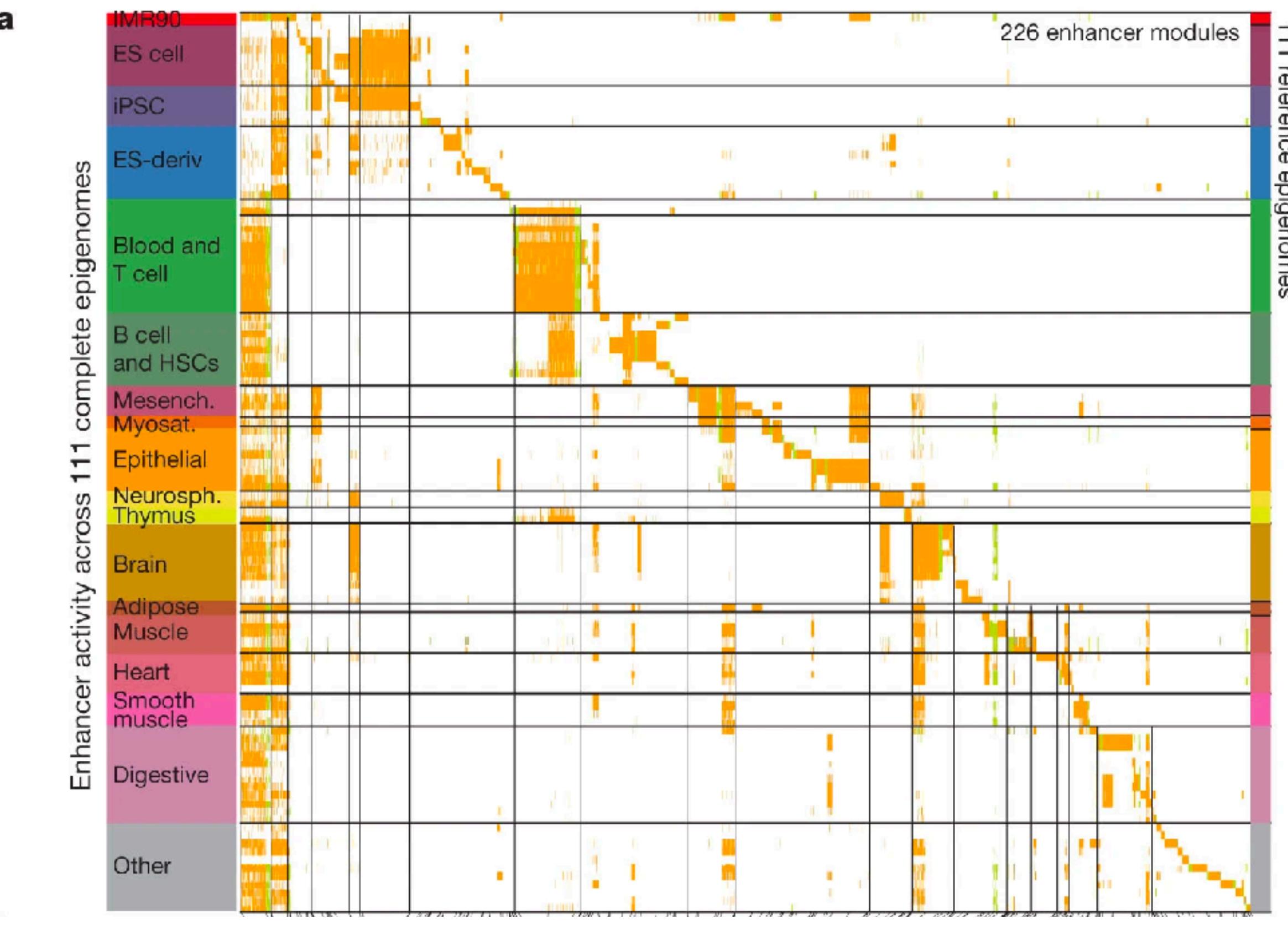
	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
Chd8	3559.907	-0.595924	0.0531048	-11.22165	3.19249e-29	3.80513e-25
Dnajc4	276.096	0.334895	0.0512012	6.54076	6.12065e-11	3.64760e-07
Vrk3	799.700	0.217710	0.0345255	6.30578	2.86745e-10	1.13924e-06
Hmgcll1	633.353	-0.279578	0.0494532	-5.65339	1.57318e-08	4.68767e-05
Anxa11	194.990	0.514626	0.0930202	5.53241	3.15858e-08	7.52943e-05
...
Wdr52	113.6225	-0.3634976	0.1600699	-2.270868	0.0231550	NA
Wisp1	88.5170	0.3151848	0.1661647	1.896822	0.0578514	NA
Wnt10a	85.9890	0.2485353	0.2584913	0.961484	0.3363089	NA
Wnt5b	110.0166	0.0966134	0.0960656	1.005703	0.3145585	NA
Xdh	46.9601	0.5700222	0.2805920	1.071520	0.0496622	NA

Biological mechanisms/processes are modular

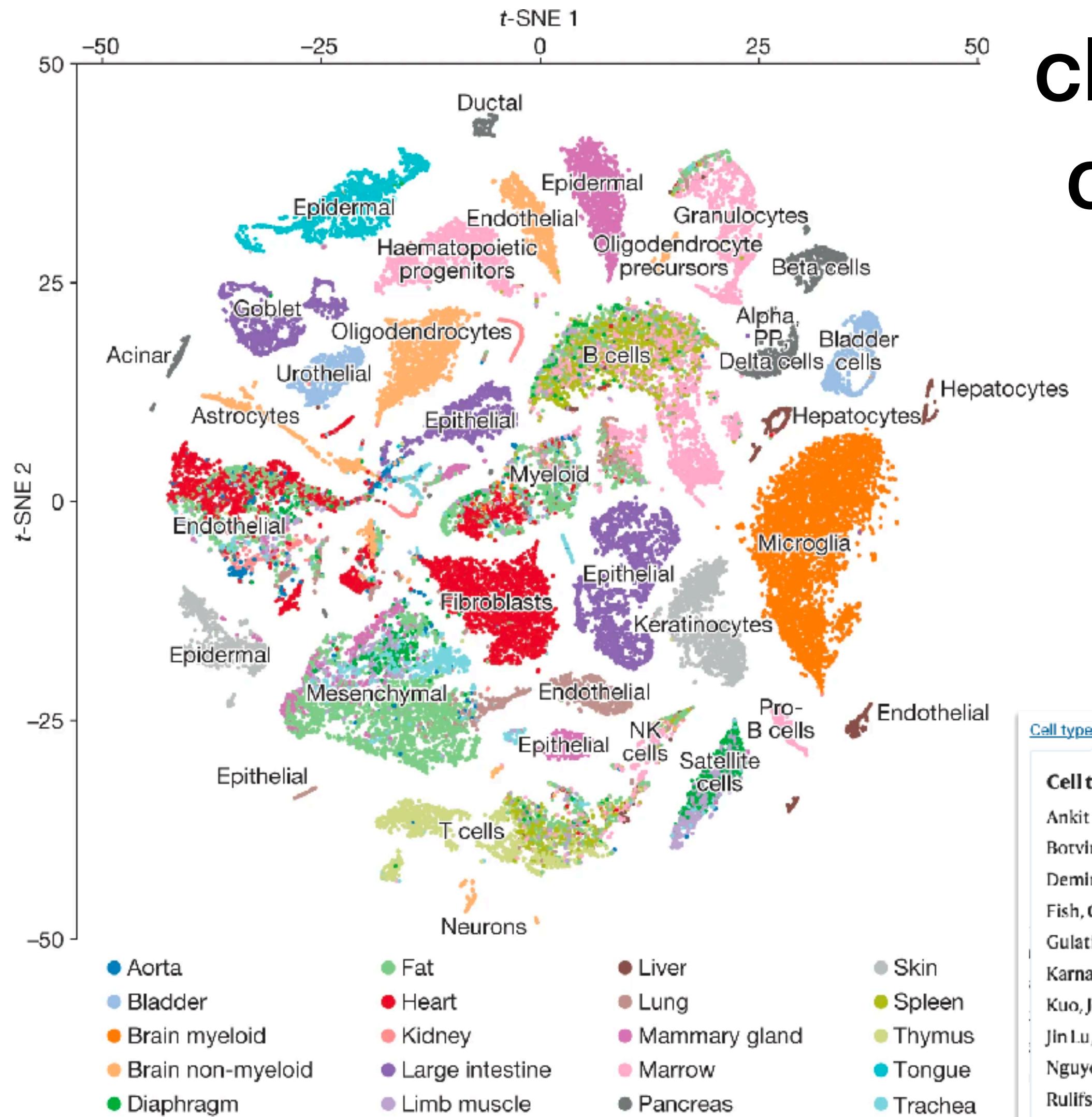
Epigenomic (enhancer element) modules across 111 tissues



Gene ontology information can be used to predict or tissue-specific biological pathways



How do we know clusters of cells correspond to cell types?



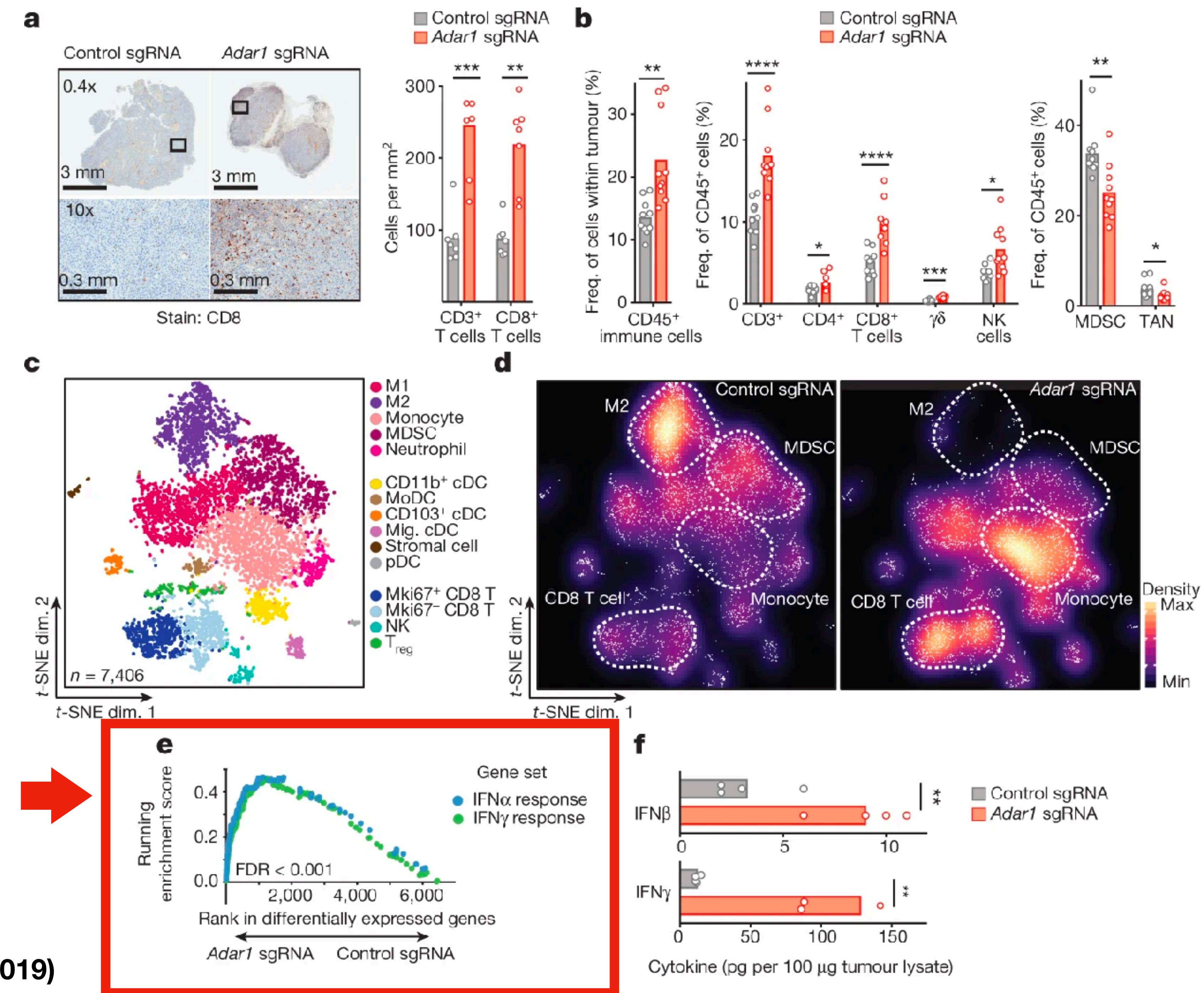
Tabular Muris, Nature (2018)

Cell type annotation		
Section	Criteria	References
Cell type annotation		
Ankit S. Baghel, Isaac Bakerman, Joshua Batson, Biter Bilen, Olga Botvinnik, Douglas Brownfield, Michelle B. Chen, Spyros Darmanis, Kubilay Demir, Antoine de Morree, Hamid Ebadi, F. Hernán Espinoza, Matt Fish, Qiang Gan, Benson M. George, Astrid Gillich, Xueying Gu, Gunsagar S. Gulati, Yan Hang, Albin Huang, Tal Iram, Taichi Isobe, Guruswamy Karnam, Aaron M. Kershner, Bernhard M. Kiss, William Kong, Christin S. Kuo, Jonathan Y. Lam, Renot Lehallier, Guang Li, Qingyun Li, Ling Liu, Wan-Jin Lu, Dullel Min, Ahmad N. Nabhan, Katharine M. Ng, Patricia K. Nguyen, Rasika Patkar, Weng Chuan Peng, Lolita Penland, Eric J. Rulifson, Nicholas Schaum, Shaheen S. Sikandar, Rahul Sinha, Krzysztof Szade, Serena Y. Tan, Krissie Tellez, Kyle J. Travaglini, Carolina Tropini, Linda J. van Weele, Bruce M. Wang, Michael N. Woscyna, Jinyi Xiang, Hanadie Yousef & Lu Zhou		

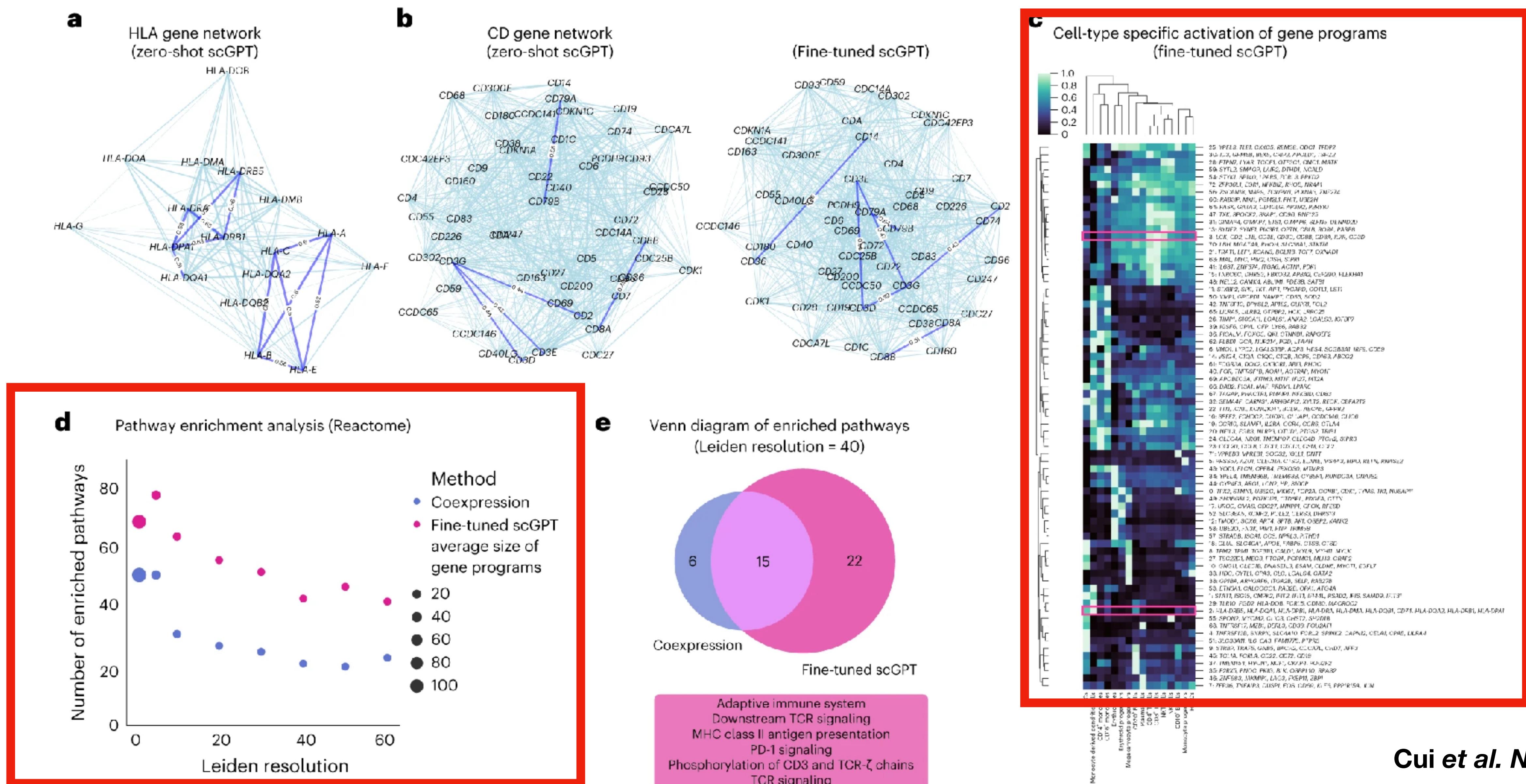


Enrichment analysis is also embedded a larger analysis

Gene set enrichment analysis checks with previous knowledge



Even in scGPT era...



Today's lecture: Enrichment Analysis

- **Motivations: What's next after genomics analysis?**
 - What have we learned?
 - How do we know that our discovery is meaningful?
- **Gene set enrichment analysis**
 - Set-based approach: Hypergeometric test
 - Rank-based approach: GSEA by KS statistic

What is Gene Set Analysis?

(Discrete) Gene Set Analysis

Input:

1. A dictionary of gene sets that map genes to sets (gene-to-set mapping)
2. A list of **top** genes identified in our own study (after FDR control)

Output:

A table of scores for all the gene sets in the dictionary.

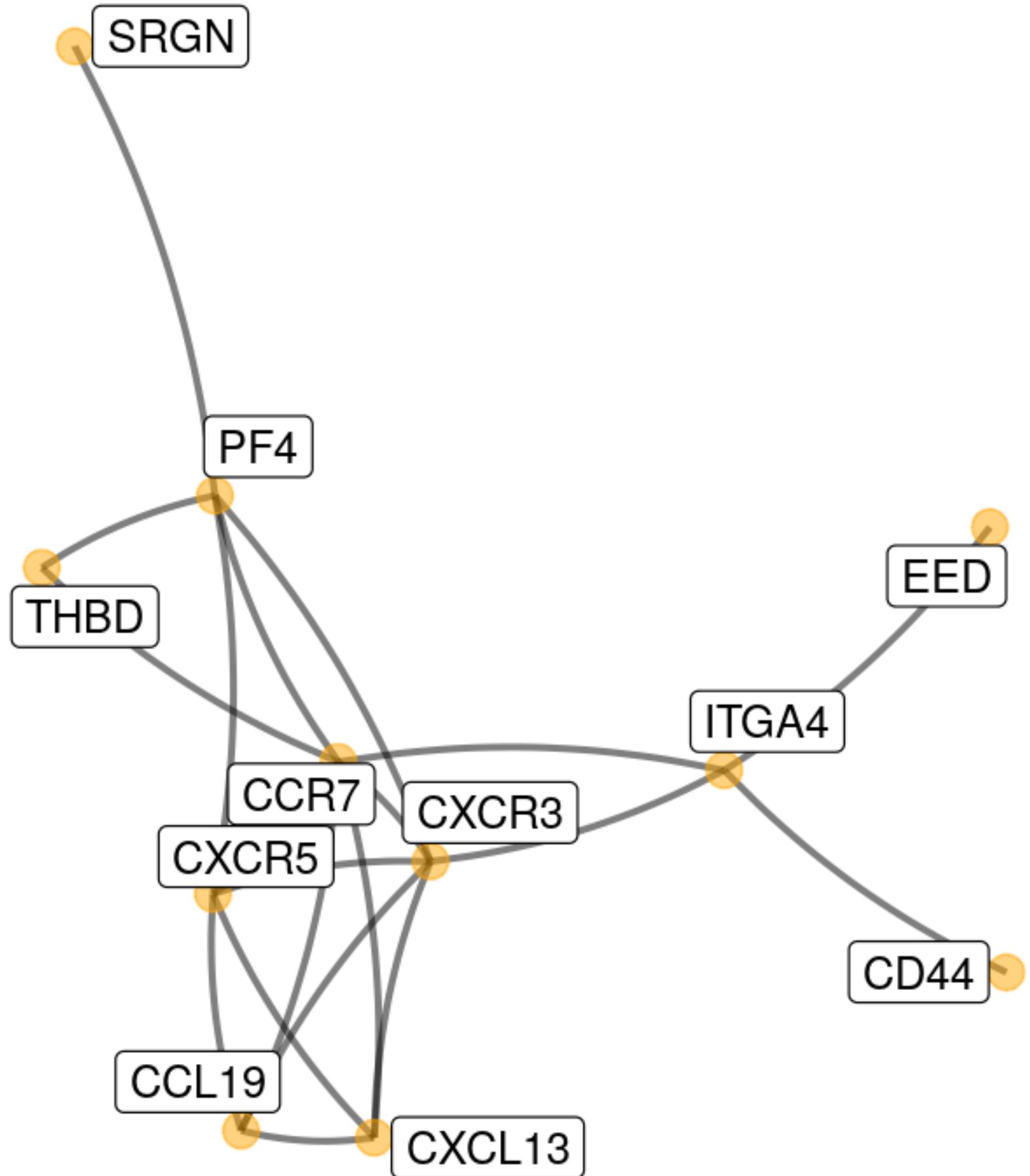
(Rank-based) Gene Set Enrichment

Input:

1. A dictionary of gene sets that map genes to sets
2. A **full** list of gene-level **scores** (e.g., p-values)

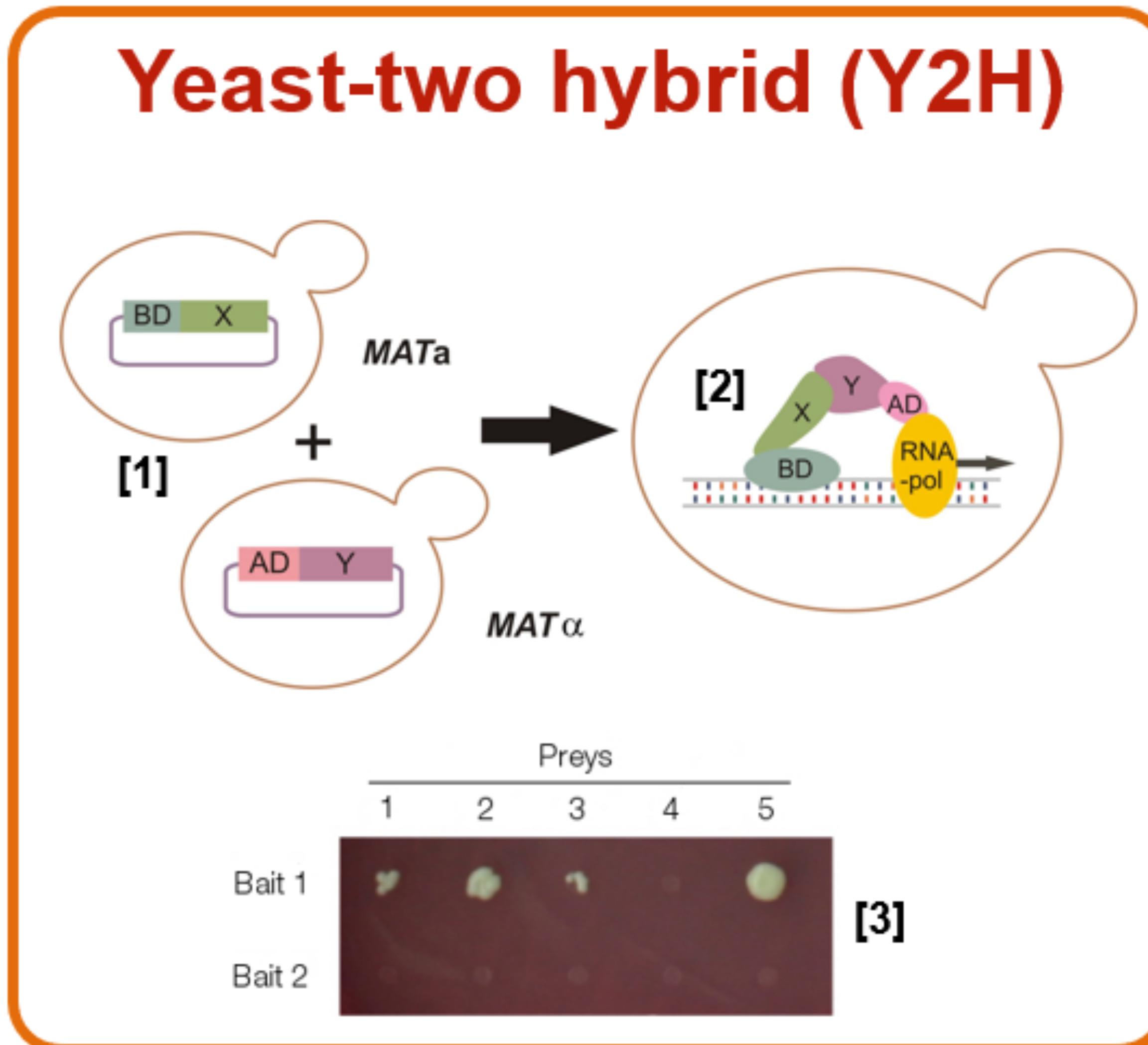
What are gene sets?

Original idea:



- Gene-gene (protein-protein) interaction network
- Functional modules
- Protein complex
- Pathways

Y2H: Where do we get gene sets?

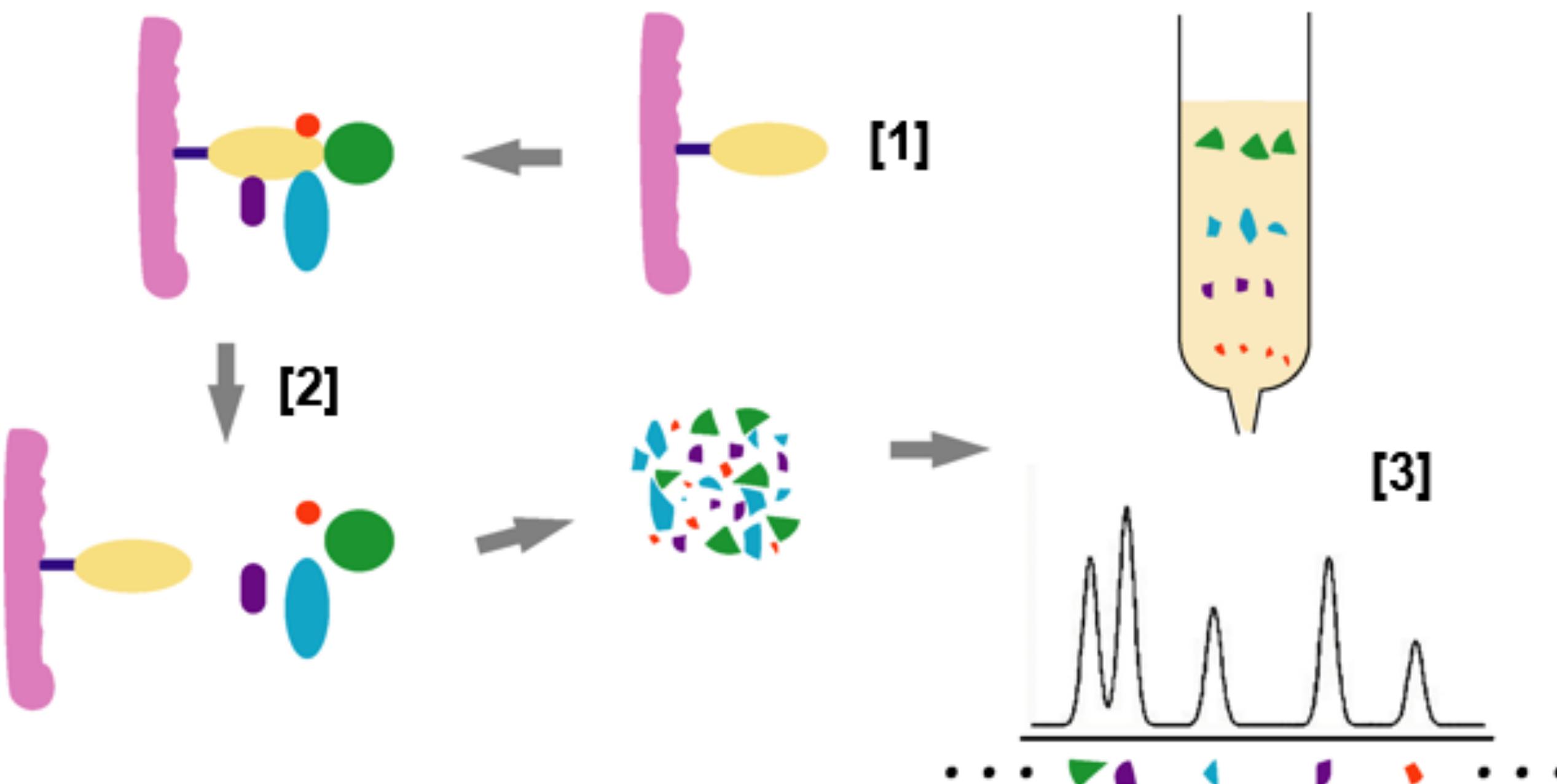


Original idea:

- The interaction of two proteins can be seen *in vivo* (yeast)
- Pros? Cons?

APMS: Where do we get gene sets?

Affinity purification+ mass spectrometry (AP-MS)



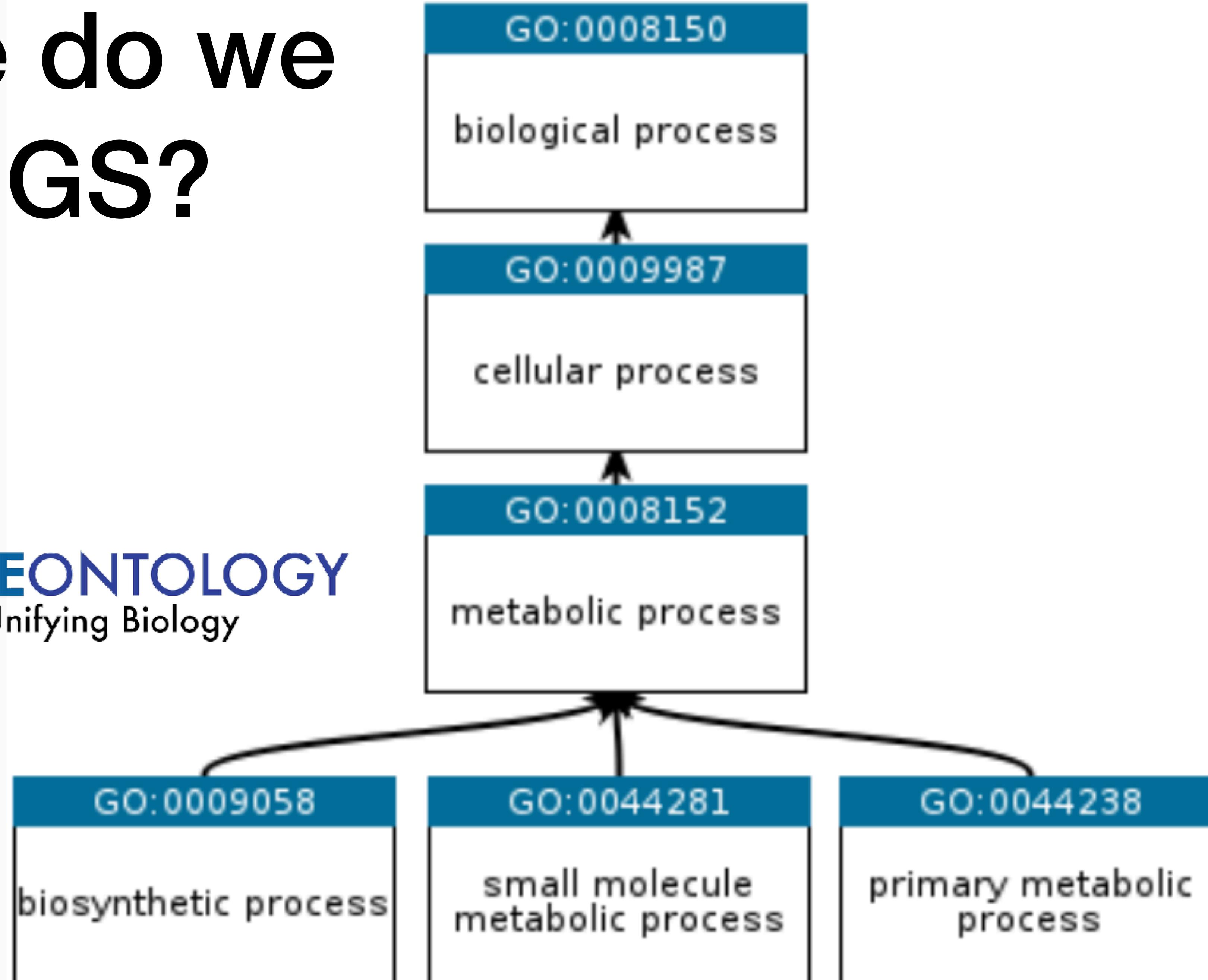
Original idea:

- Direct observation of many proteins combined *in vitro*
- Pros? Cons?

Where do we get GS?



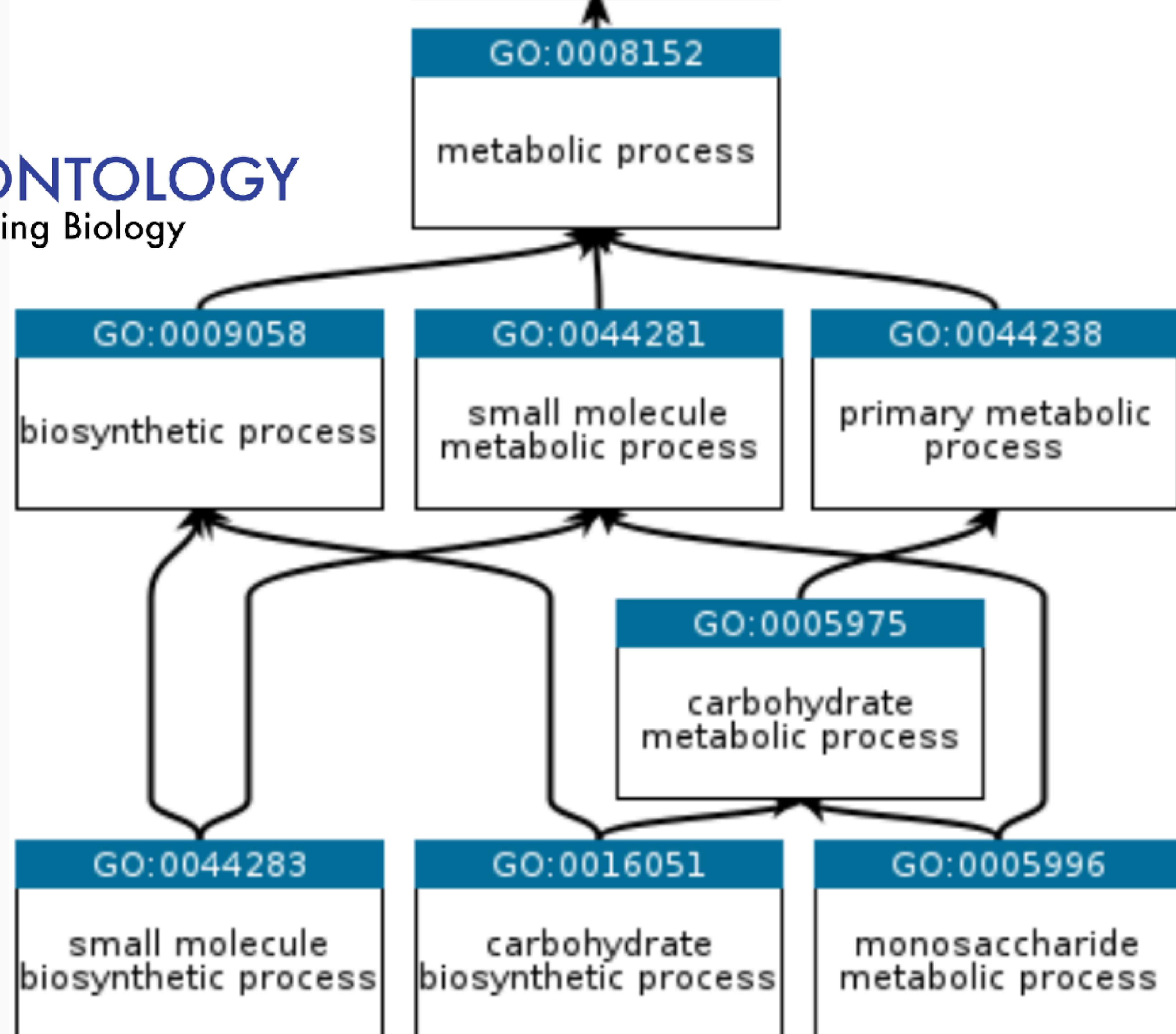
GENEONTOLOGY
Unifying Biology





GENEONTOLOGY

Unifying Biology



Who started Gene Set Enrichment Analysis?

Table 1 • Enrichment of clusters for ORFs within functional categories

Cluster	Periodicity index	Number of ORFs (<i>n</i>)	MIPS functional category (total ORFs)	ORFs within functional category (<i>k</i>)	P value $-\log_{10}$
1	0.07	164	ribosomal proteins (206)	64	54
			organization of cytoplasm (555)	79	39
			organization of chromosome structure (41)	7	4
2	0.38	186	DNA synthesis and replication (82)	23	16
			cell-cycle control and mitosis (312)	30	8
			recombination and DNA repair (84)	11	5
			nuclear organization (720)	40	4

Tavazoie .. Church, Nature Genetics

Who started GSEA?

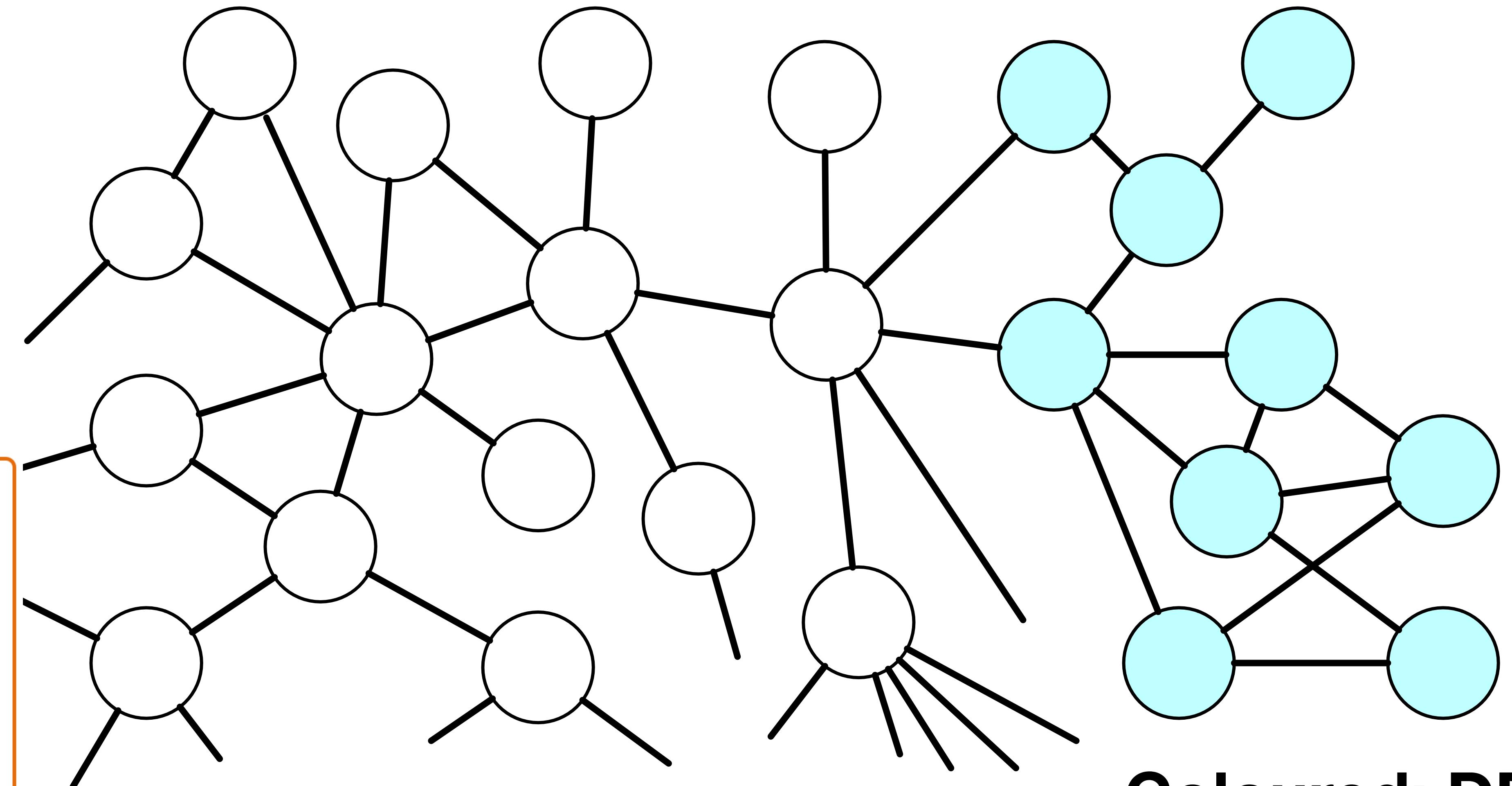
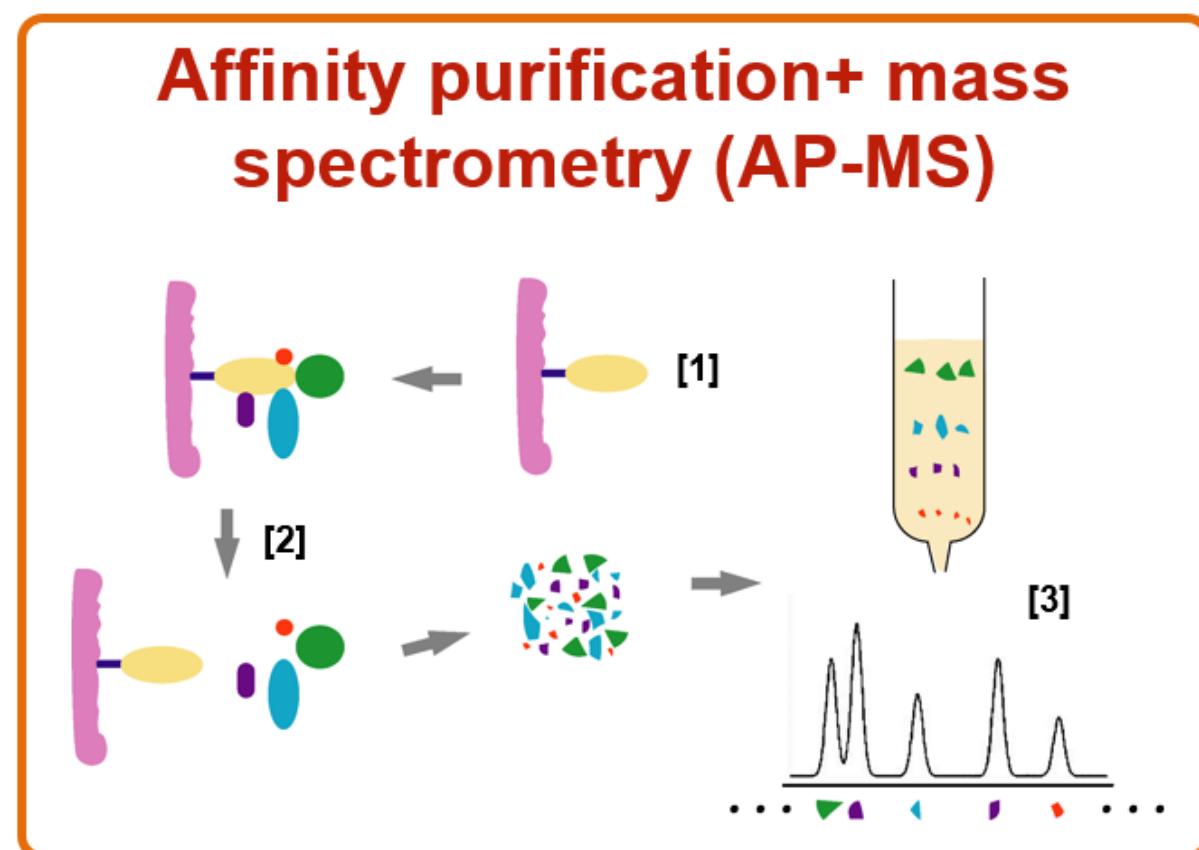
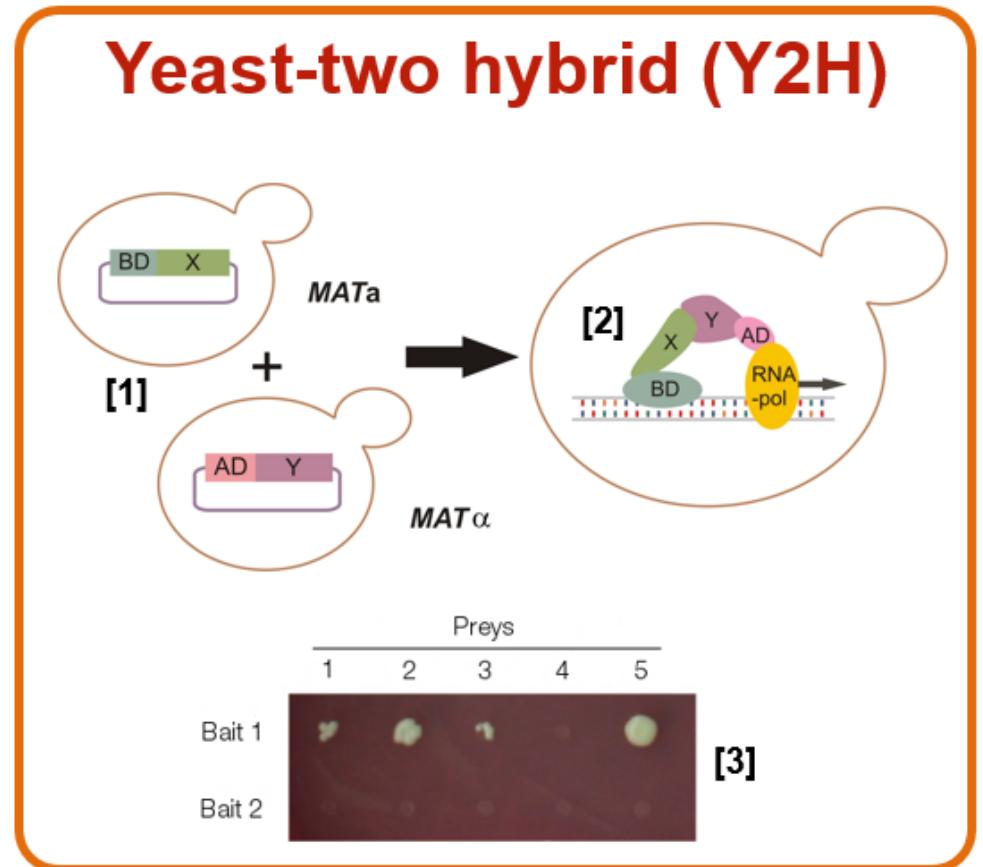
Determination of statistical significance for functional category enrichment.
The hypergeometric distribution was used to obtain the chance probability of observing the number of genes from a particular MIPS functional category within each cluster. More specifically, the probability of observing at least (k) ORFs from a functional category within a cluster of size (n) is given by:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i} \binom{g-f}{n-i}}{\binom{g}{n}},$$



Tavazoie, Mootha, Church, Nature Genetics (1999)

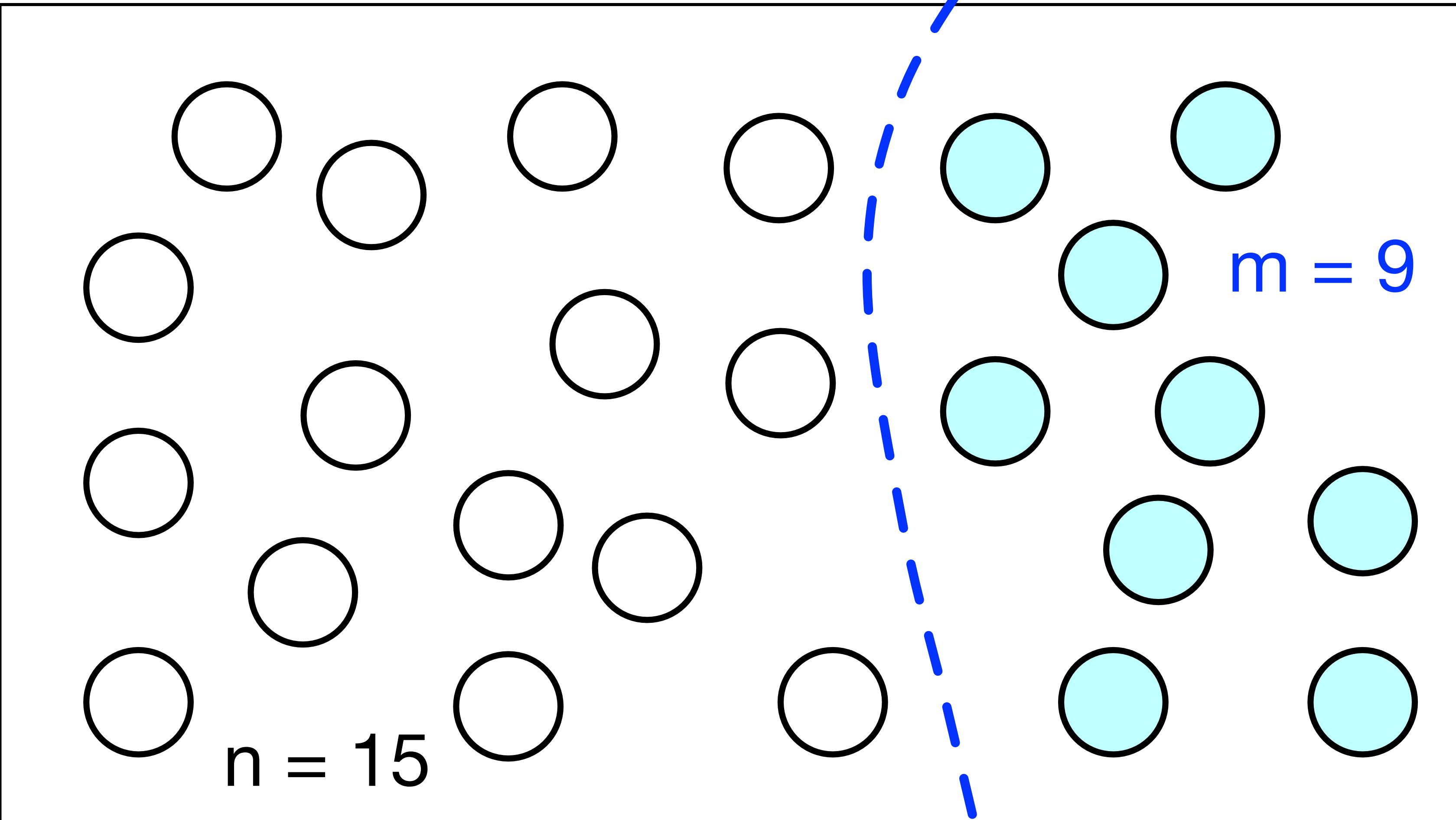
A list of DEGs cover nodes/vertices in a network



Coloured: DEG

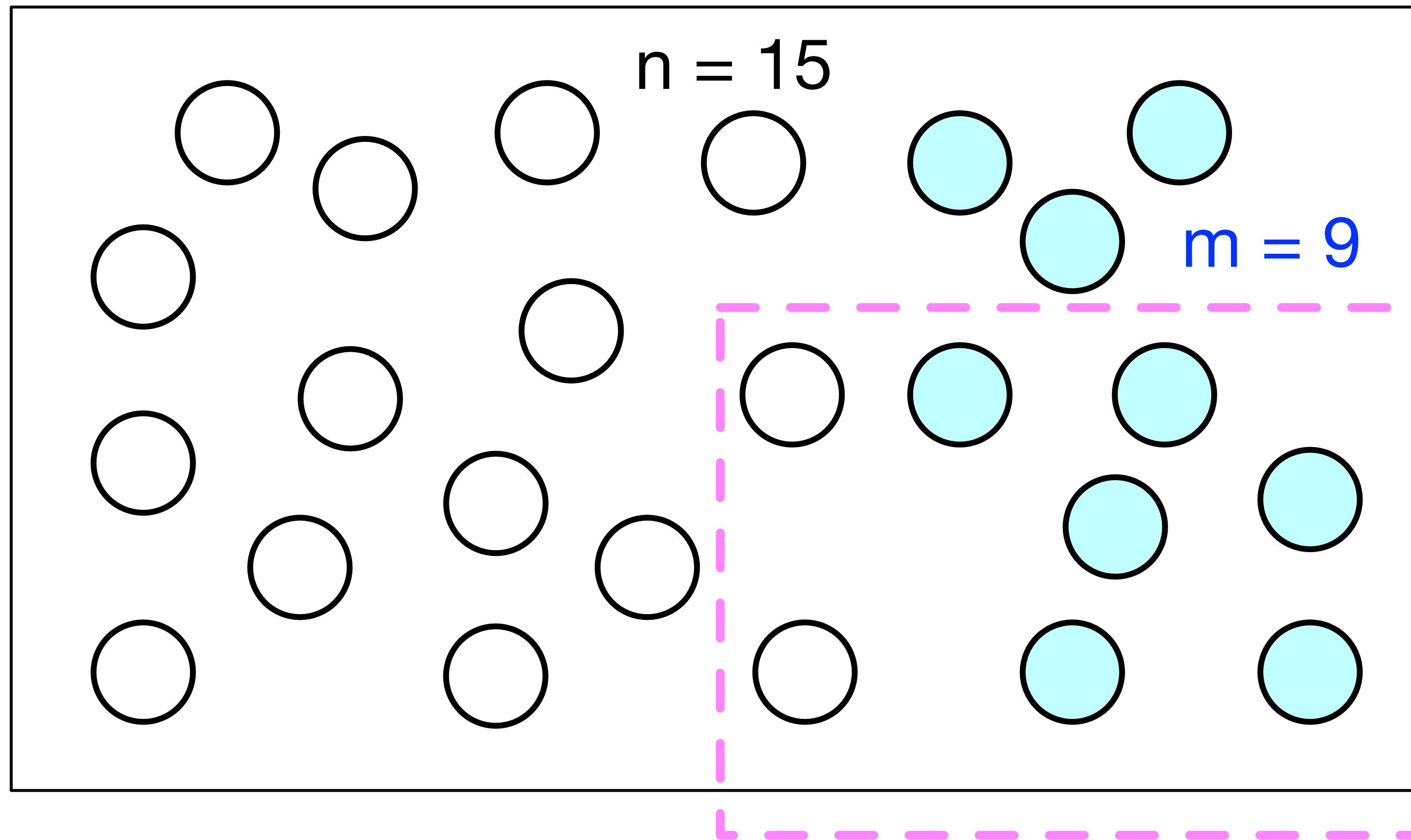
Let's drop the edges

$N = 24$



Does this pathway overlap with our DEG list?

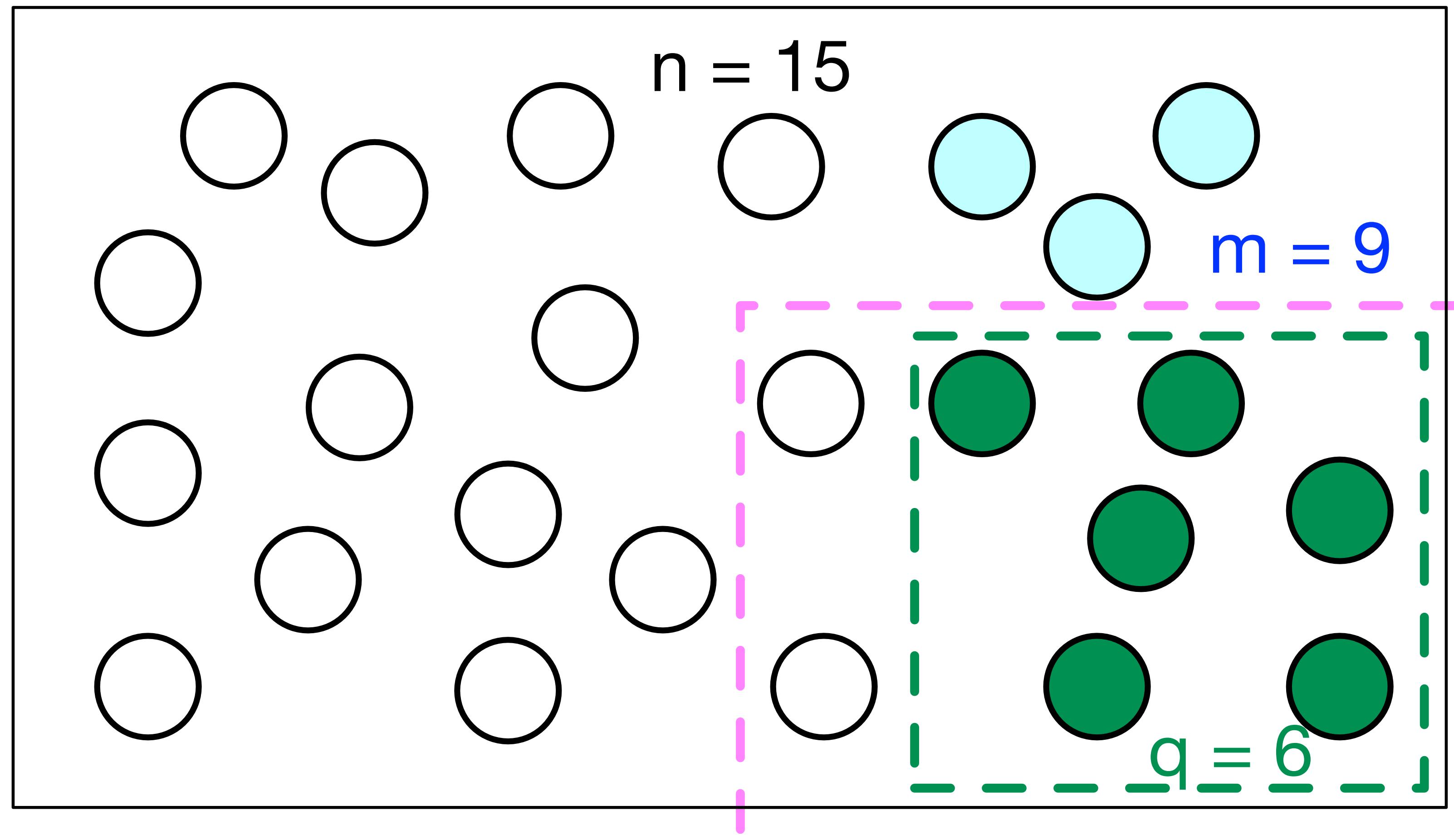
$N = 24$



$k = 8$

Does this pathway overlap with our DEG list?

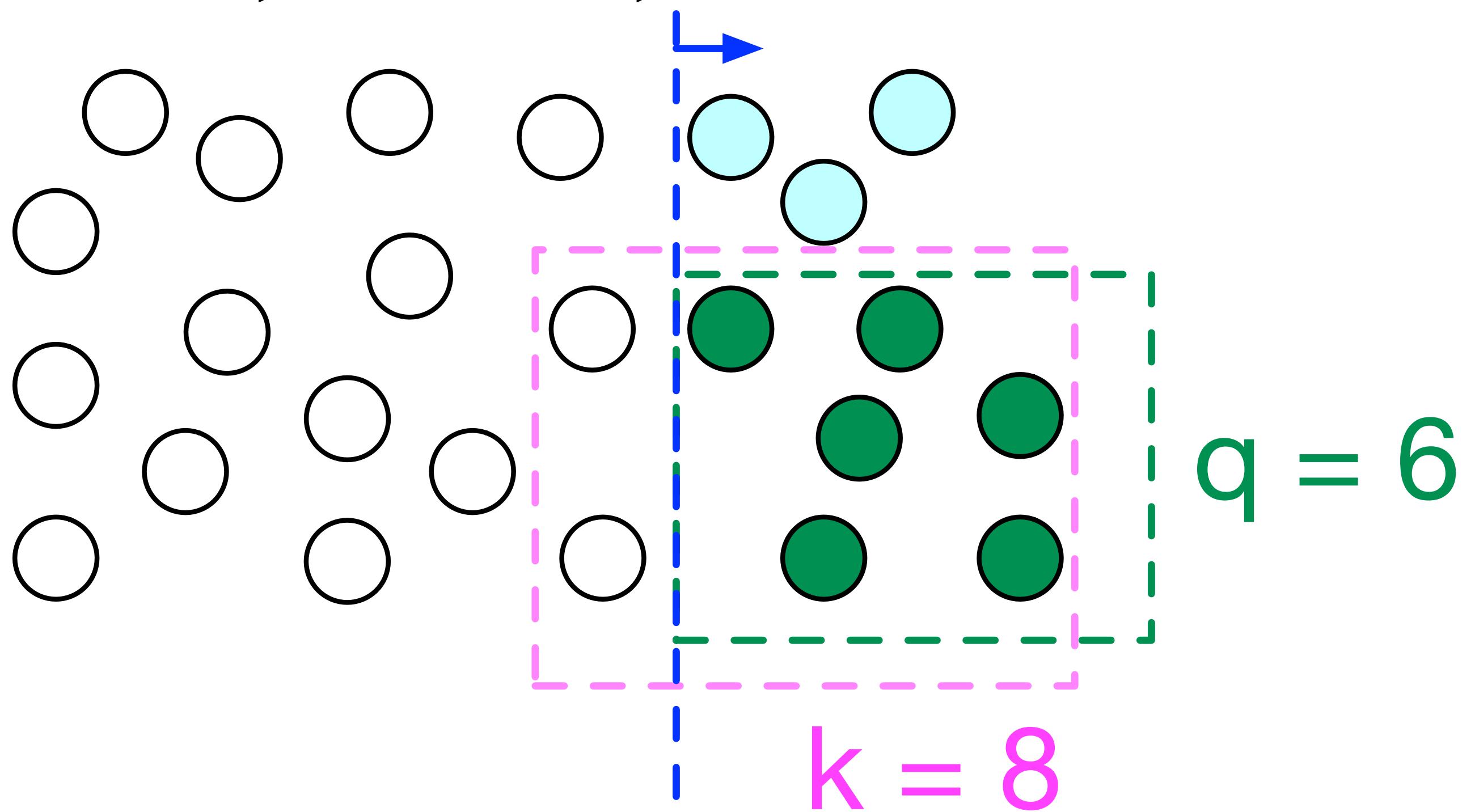
$N = 24$



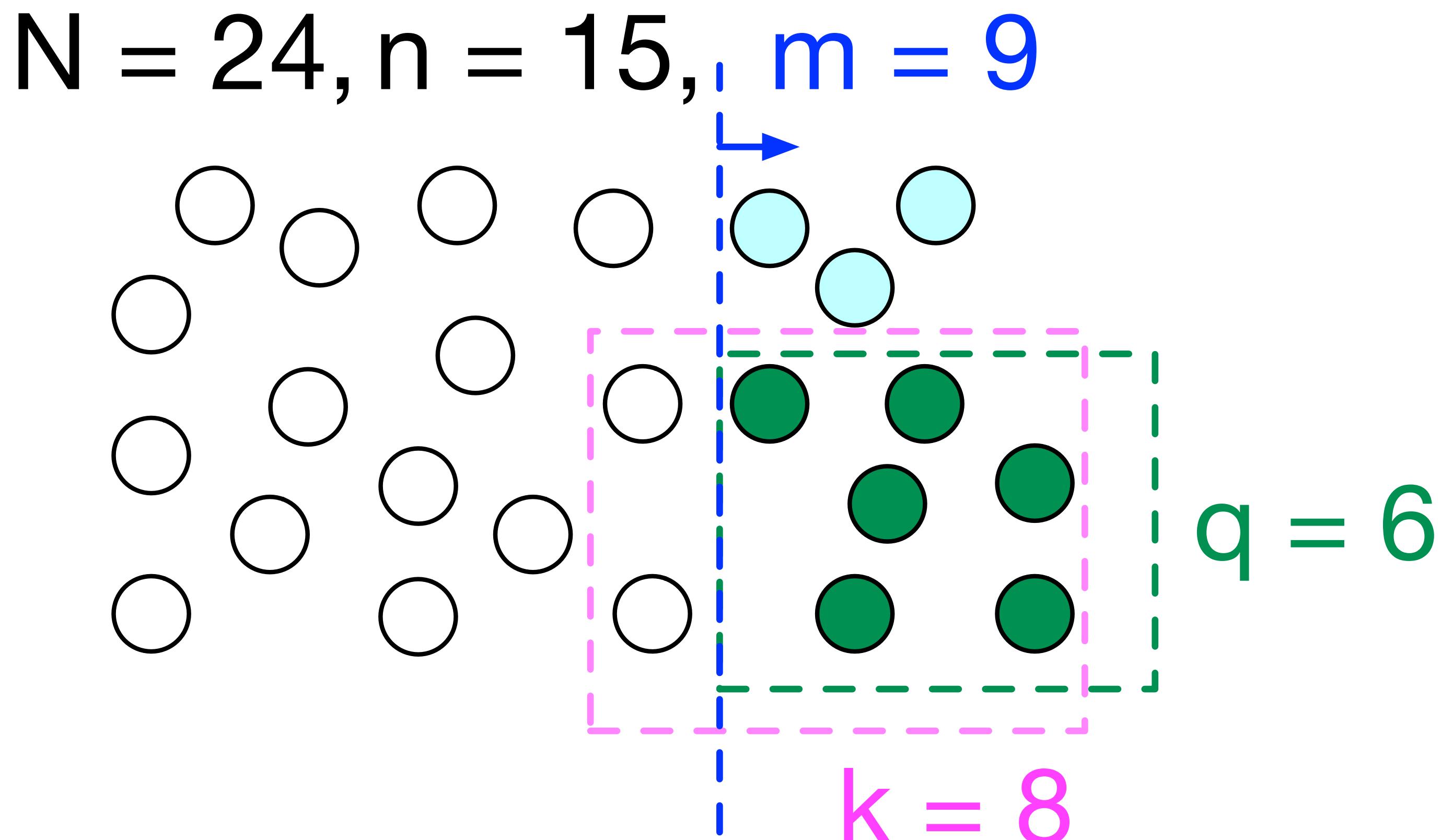
$k = 8$

Let's sort out the numbers

$N = 24, n = 15, m = 9$



GSA testing the over-representation of a discovery

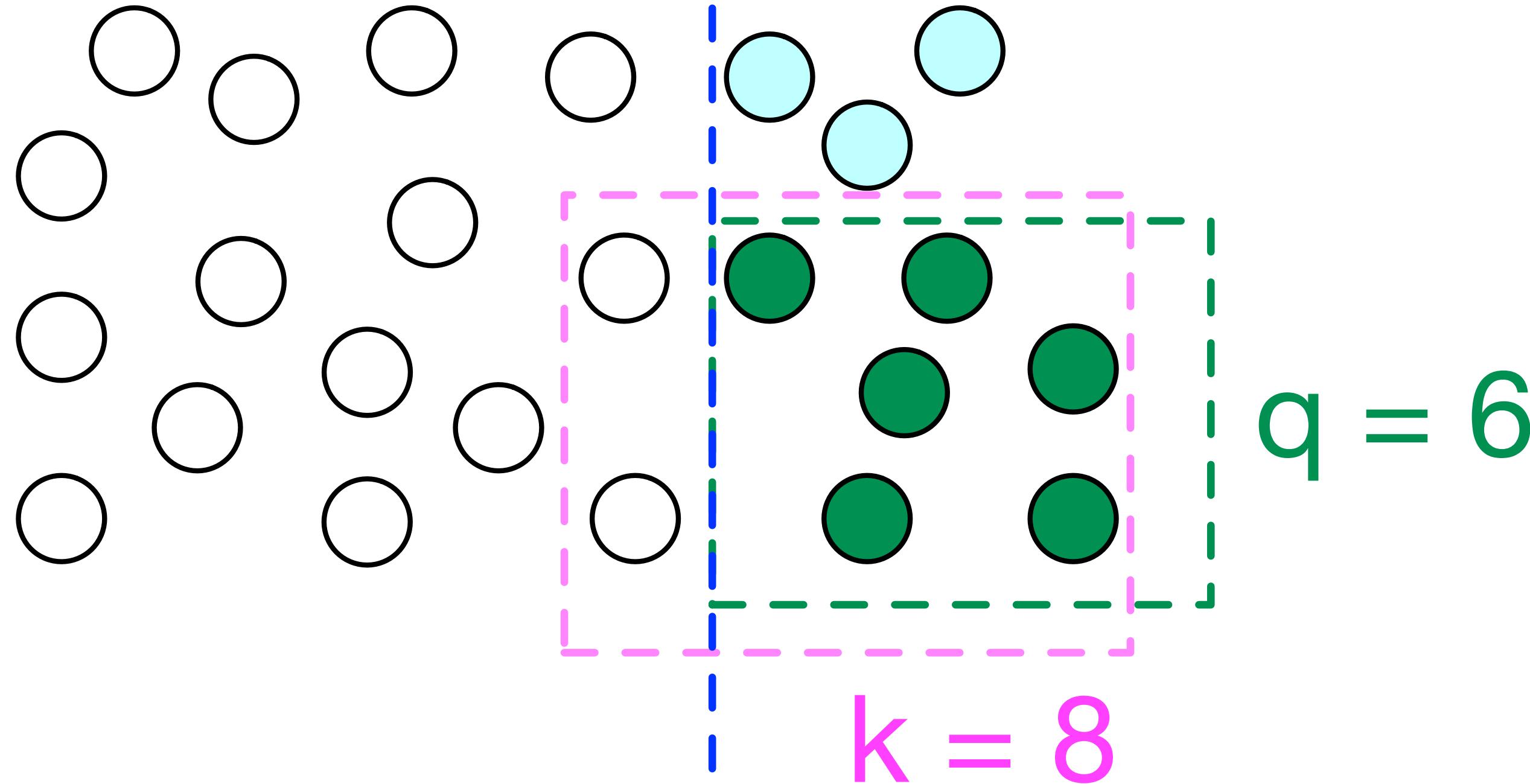


What are the numbers to count?

- ▶ N : # genes in this universe
- ▶ m : # genes in this set
- ▶ n : # genes *not* in this gene set
- ▶ k : # DEGs in our analysis
- ▶ q : # DEGs (of k) overlapping with the set of m genes

How much overlap?

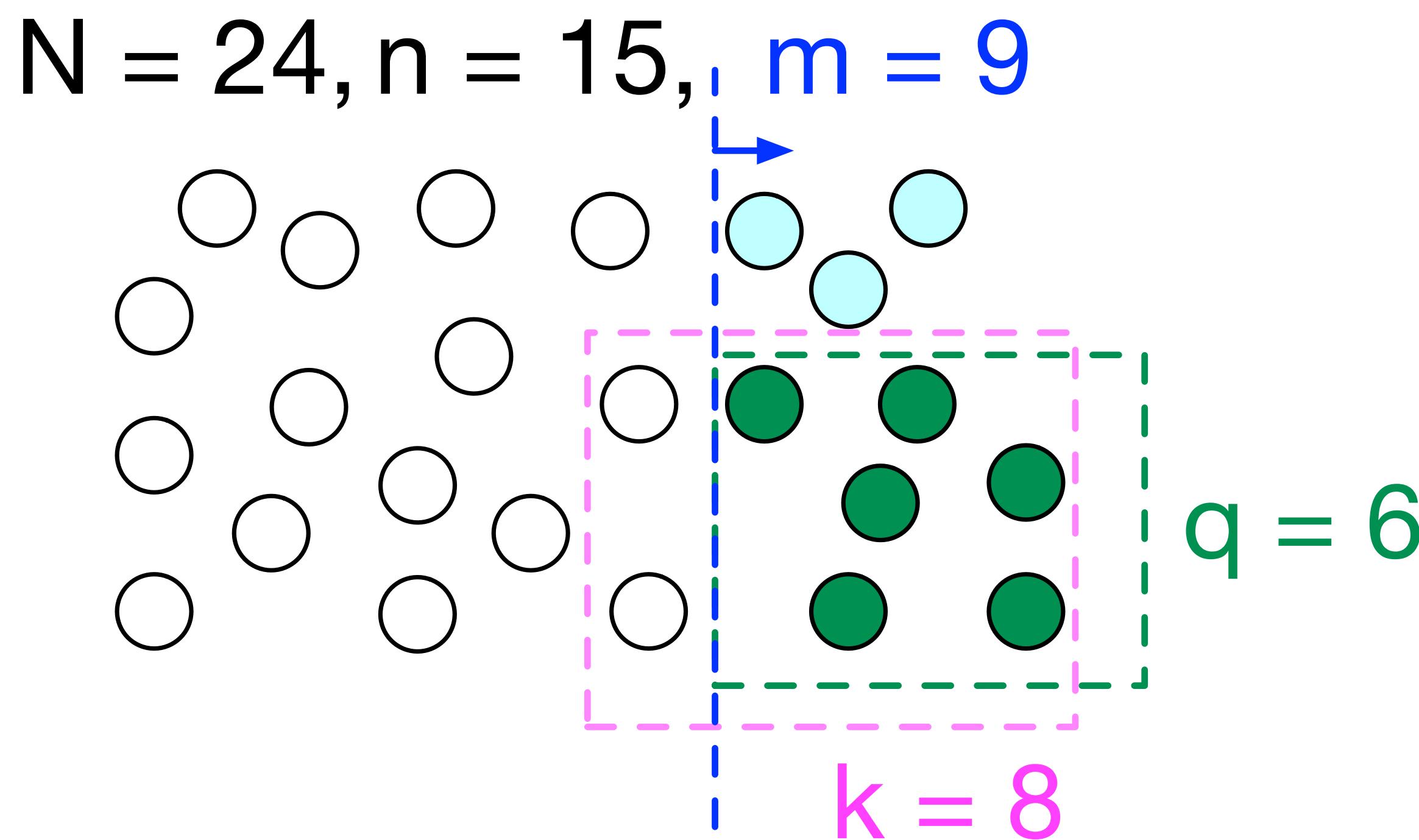
$N = 24, n = 15, m = 9$



Questions:

- ▶ Is it meaningful enough to report?
- ▶ Is it surprising enough that we recapitulated 6/9 (~67 %)?
- ▶ What is the null distribution?
- ▶ What is the generative/simulation scheme?

Can we count all possible cases under the null?



Under the null of hypergeometric distribution

1. Sample k DEGs out of N genes
2. Of these k genes, q overlap with a gene set consisting of m genes
3. The rest $k - q$ genes overlap with genes outside of the gene set $N - m$

Let's count them all under the null

- ▶ How many all possible ways to select $k = 8$ out of $N = 24$ genes, ignoring the order of k selected genes and $N - k$ *not* selected genes?
- ▶ We can think of this as three steps: (1) enumerating N genes, (2) partition them into the first k genes and the rest, (3) ignore the order within each partition.

Binomial coefficient

$$\binom{24}{8} = \frac{\{\text{all possible ways to enumerate 24}\}}{\{\text{enumerate 8}\}\{\text{enumerate 16}\}}$$
$$= \frac{24!}{8!16!}$$

Hypergeometric distribution

- ▶ What is the probability of uniformly selecting k DEGs out of N genes? $\binom{N}{k}^{-1}$

Hypergeometric distribution

- ▶ What is the probability of uniformly selecting k DEGs out of N genes? $\binom{N}{k}^{-1}$
- ▶ How many possible ways of finding q DEGs overlapping with m genes in the gene set? $\binom{m}{q}$

Hypergeometric distribution

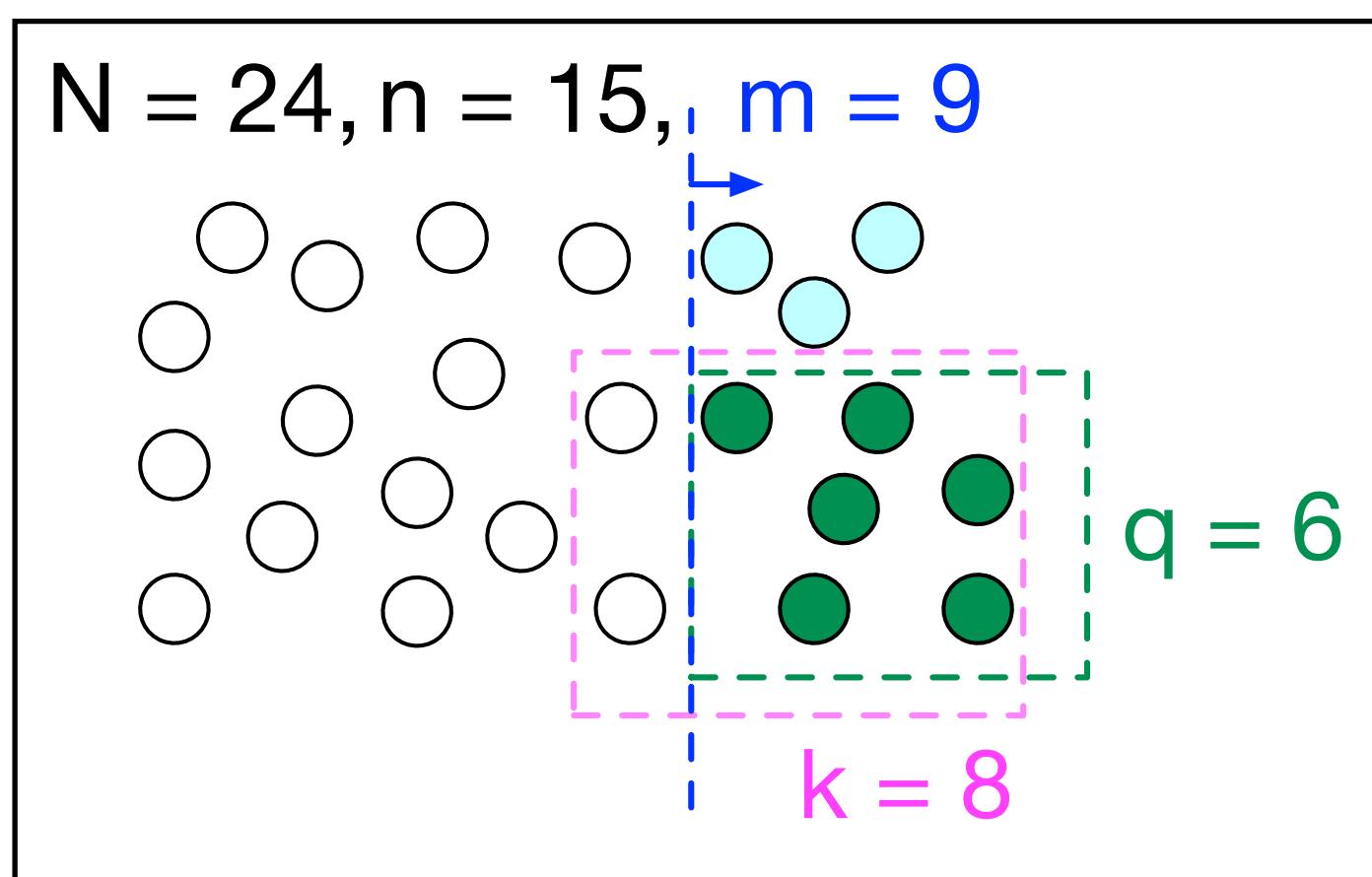
- ▶ What is the probability of uniformly selecting k DEGs out of N genes? $\binom{N}{k}^{-1}$
- ▶ How many possible ways of finding q DEGs overlapping with m genes in the gene set? $\binom{m}{q}$
- ▶ How many possible ways of finding the rest $k - q$ DEGs overlapping with $(N - m = n)$ genes in the gene set? $\binom{N-m}{k-q}$

Hypergeometric distribution

$$P_0(q|N, m, k) = \sum \frac{\# \text{ ways to select } q \text{ out of } m}{\# \text{ ways to select } (k - q) \text{ out of } N - m}$$

the probability
of choosing a set size k
out of total N

=



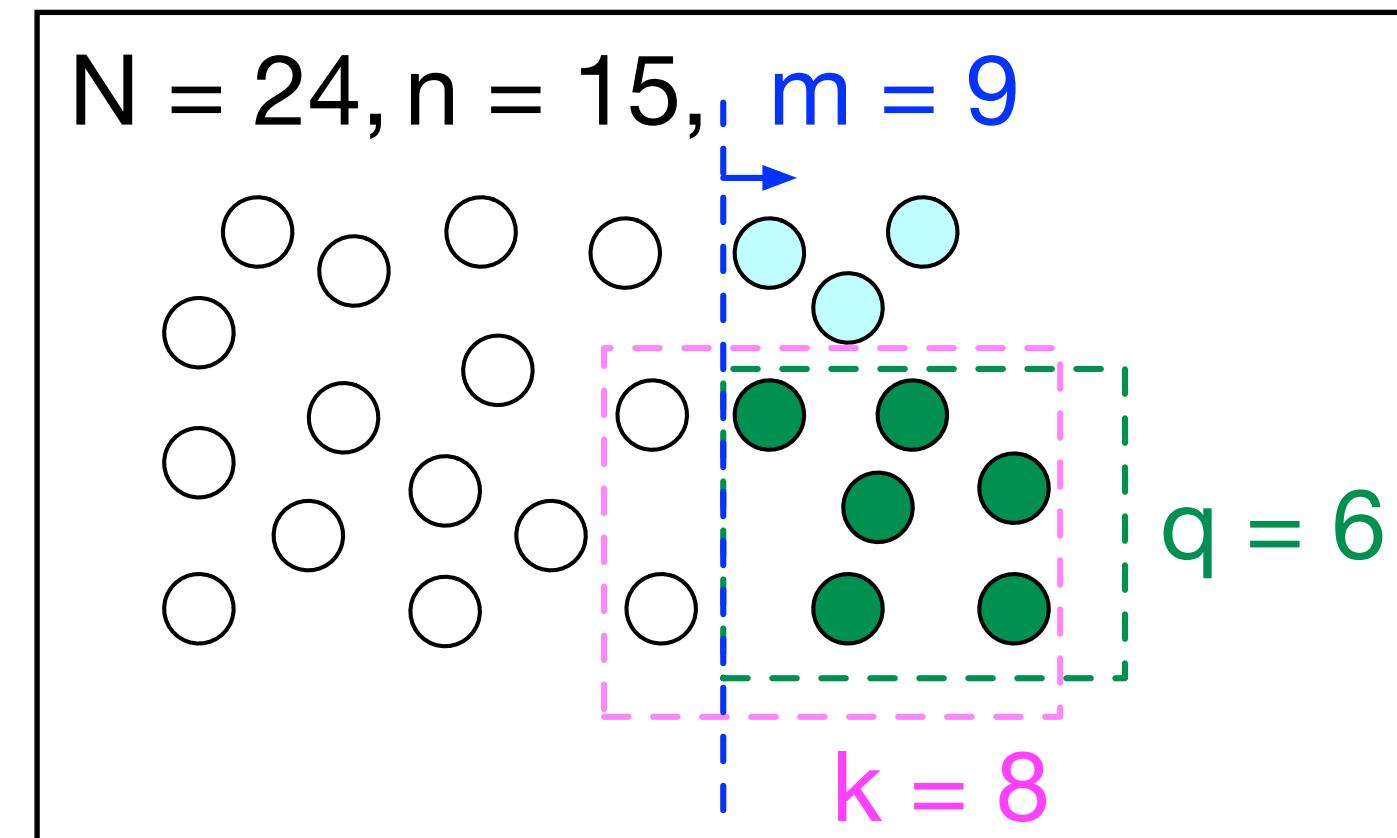
Hypergeometric distribution

$$P_0(q|N, m, k) = \frac{\sum \begin{array}{l} \# \text{ ways to select } q \text{ out of } m \\ \# \text{ ways to select } (k - q) \text{ out of } N - m \end{array}}{\binom{m}{q}}$$

= $\binom{m}{q}$

ways to choose
q overlap out of m

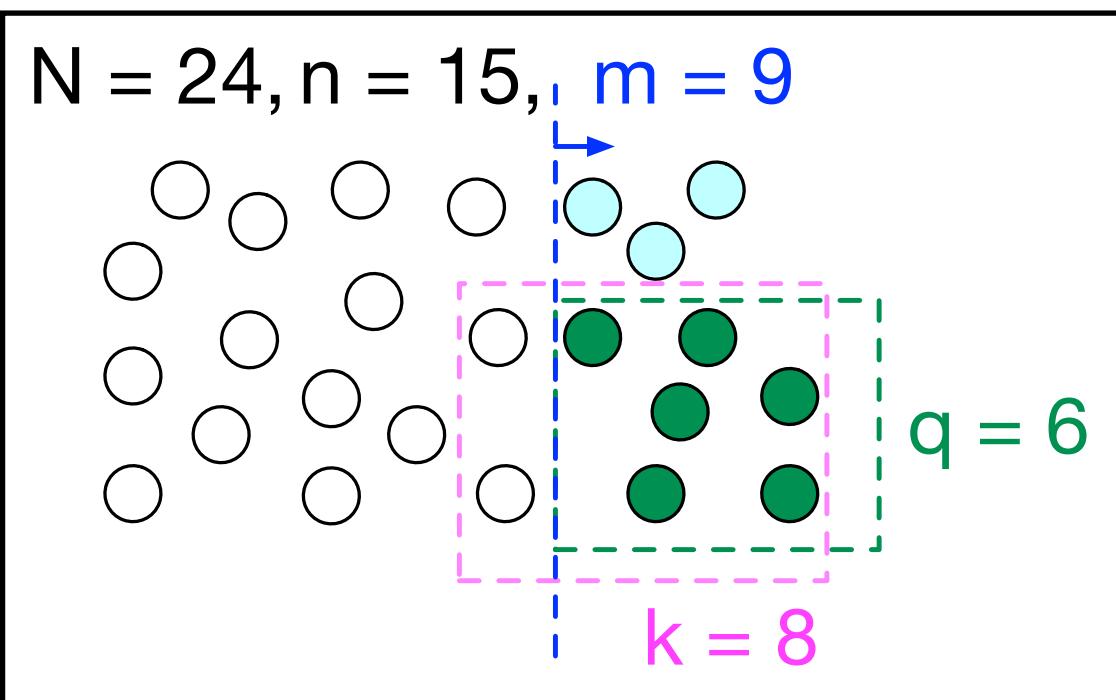
the probability
of choosing a set size k
out of total N



Hypergeometric distribution

$$P_0(q|N, m, k) = \frac{\sum \text{# ways to select } q \text{ out of } m}{\text{# ways to select } (k - q) \text{ out of } N - m}$$
$$= \frac{\binom{m}{q}}{\binom{N - m}{k - q}} \times \frac{\text{# ways to choose } q \text{ overlap out of } m}{\text{# ways to choose } (k - q) \text{ out of } N - m}$$

the probability
of choosing a set size k
out of total N

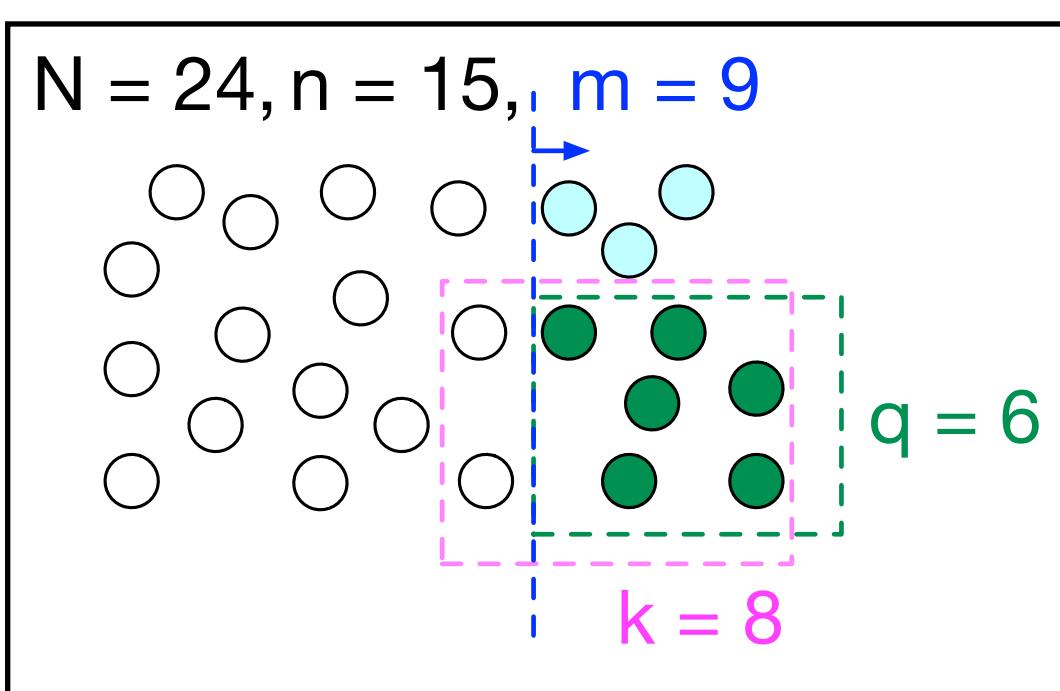


Hypergeometric distribution

$$P_0(q|N, m, k) = \sum_{\substack{\# \text{ ways to select } q \text{ out of } m \\ \# \text{ ways to select } (k - q) \text{ out of } N - m}} \binom{m}{q} \times \binom{N - m}{k - q} \times \binom{N}{k}^{-1}$$

ways to choose
q overlap out of m # ways to choose
($k - q$) out of $N - m$

the probability
of choosing a set size k
out of total N



What is the probability of k overlapping DEGs?

Hypergeometric PMF

$$p(x|N, m, k) = \frac{\binom{m}{x} \binom{N-m}{k-x}}{\binom{N}{k}}$$

Hypergeometric CDF

$$p(q|N, m, k) = \sum_{x=0}^q \frac{\binom{m}{x} \binom{N-m}{k-x}}{\binom{N}{k}}$$

Hypergeometric test for testing significant overlap

$$H_0 : x \leq q \quad \text{vs.} \quad H_1 : x > q$$

We may observe overlap q genes by random sampling of k genes **without** replacement.

Therefore, we can calculate the p-value:

What is the definition of p-value?

```
phyper(q=6, m=9, n=15, k=8, lower.tail=FALSE)
```

```
## [1] 0.0007464604
```

Hypergeometric test for testing significant overlap

$$H_0 : x \leq q \quad \text{vs.} \quad H_1 : x > q$$

We may observe overlap q genes by random sampling of k genes **without** replacement.

Therefore, we can calculate the p-value:

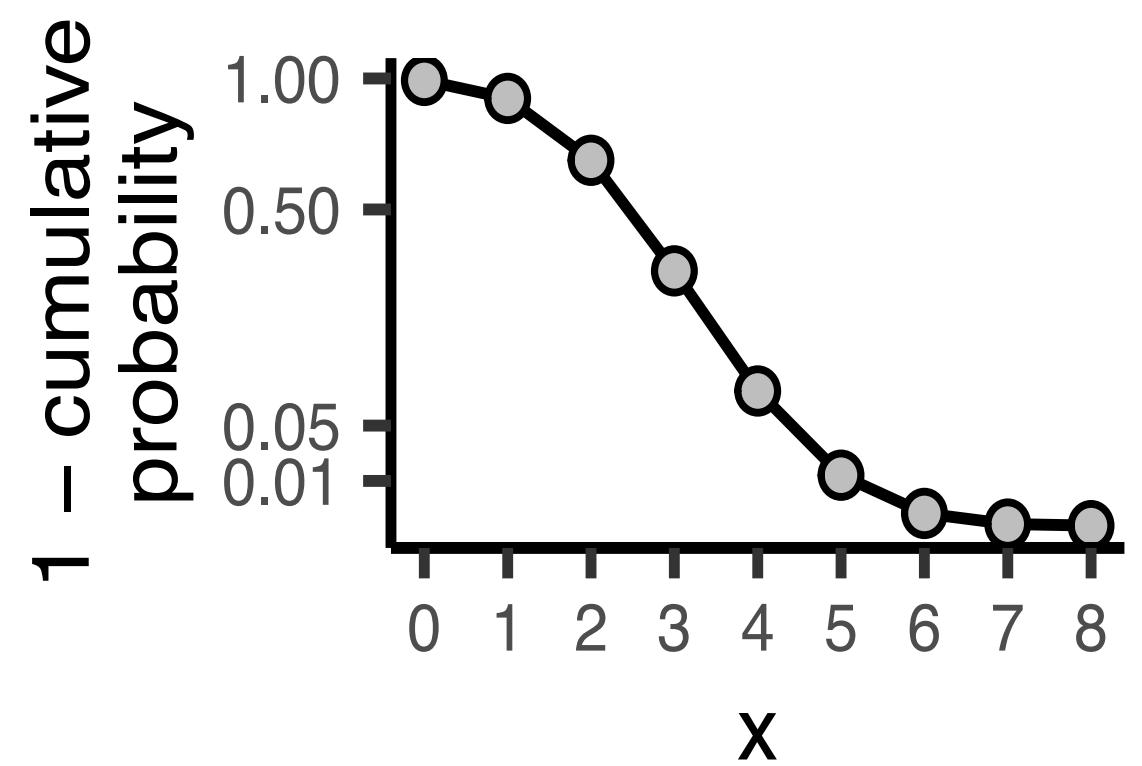
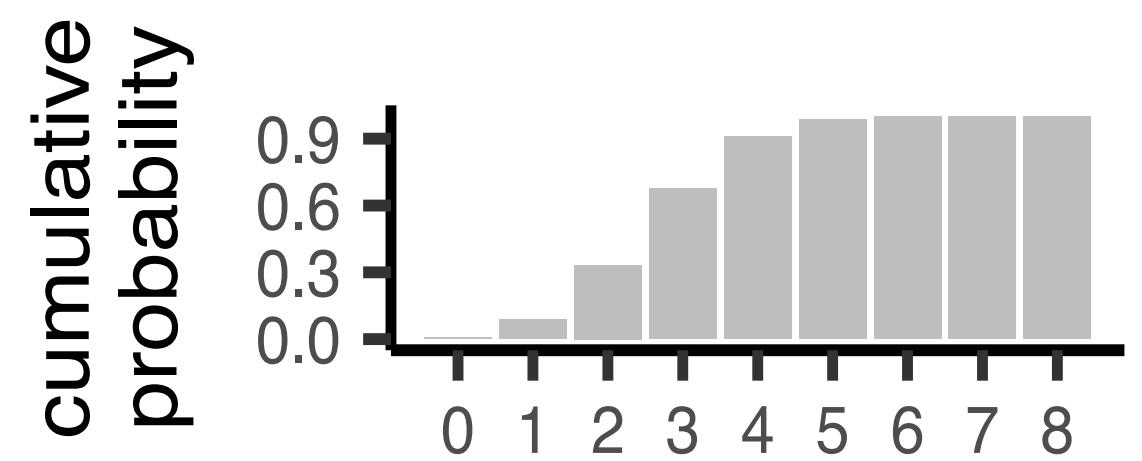
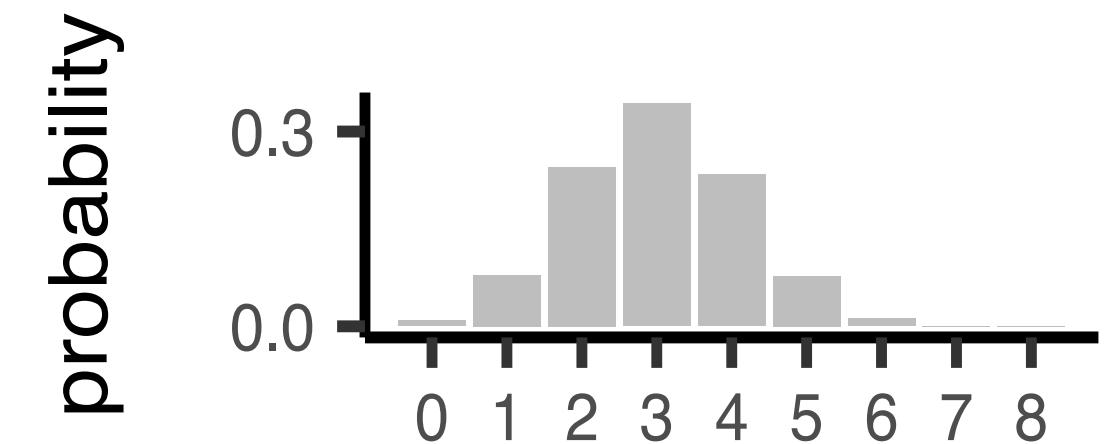
$$P(x > q | n, m, k) = 1 - \sum_{x=0}^q \frac{\binom{m}{x} \binom{n}{k-x}}{\binom{n+m}{k}}$$

```
phyper(q=6, m=9, n=15, k=8, lower.tail=FALSE)
```

```
## [1] 0.0007464604
```

How significant is q overlap in our discovery?

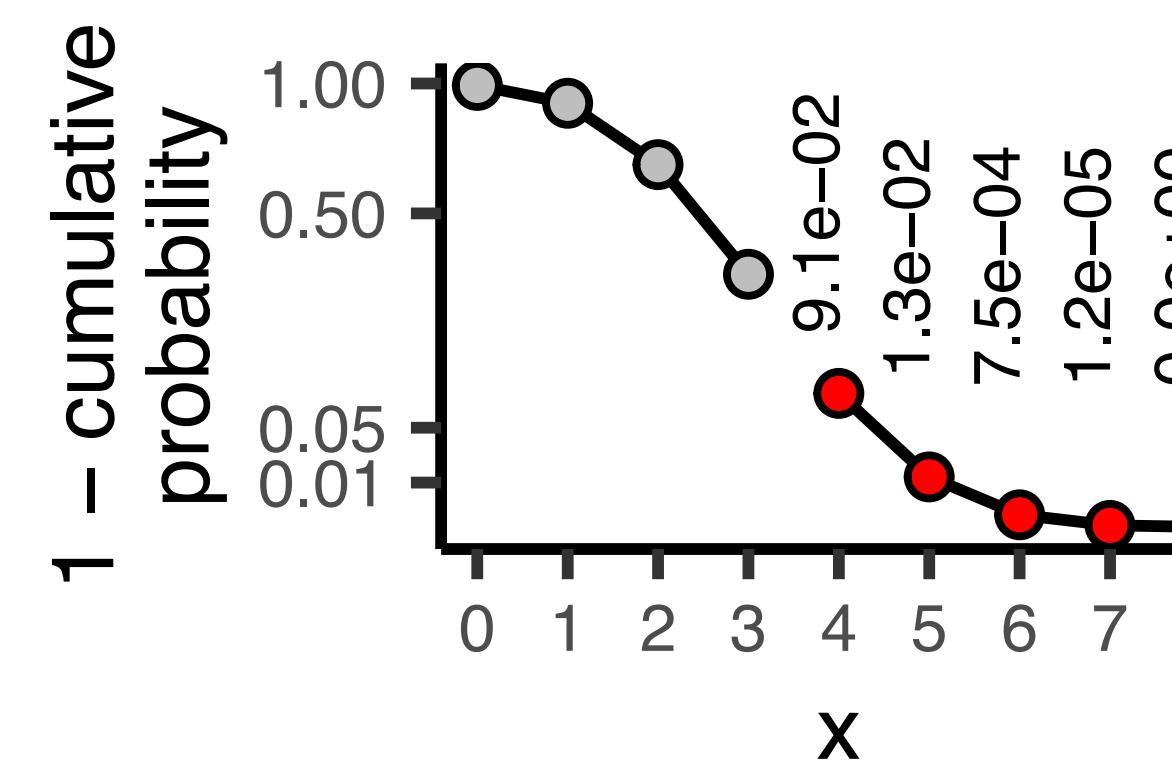
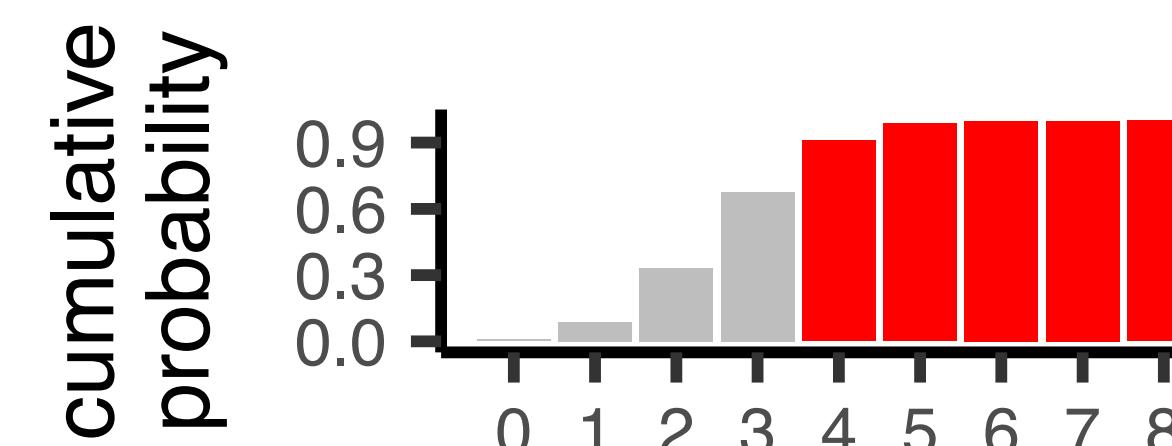
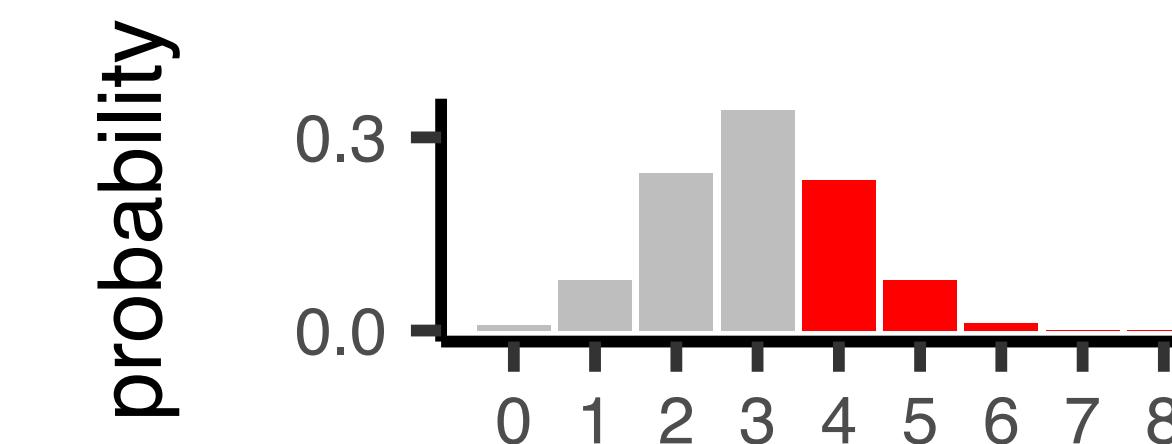
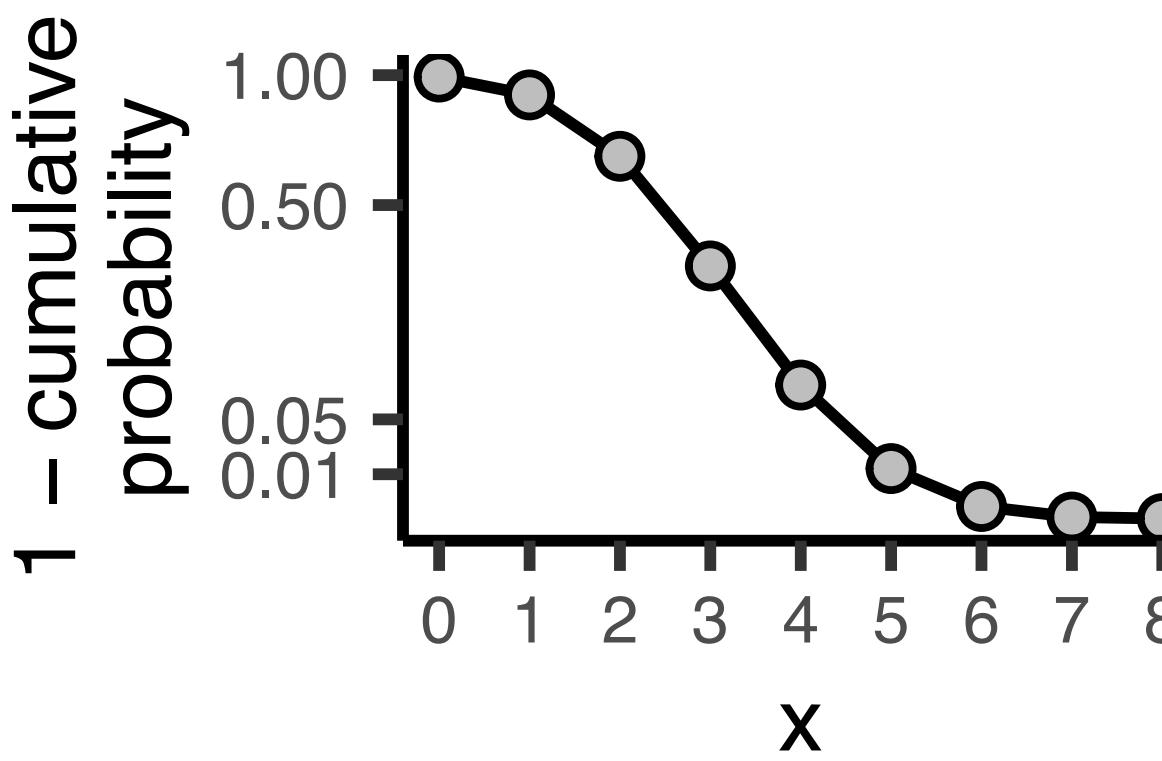
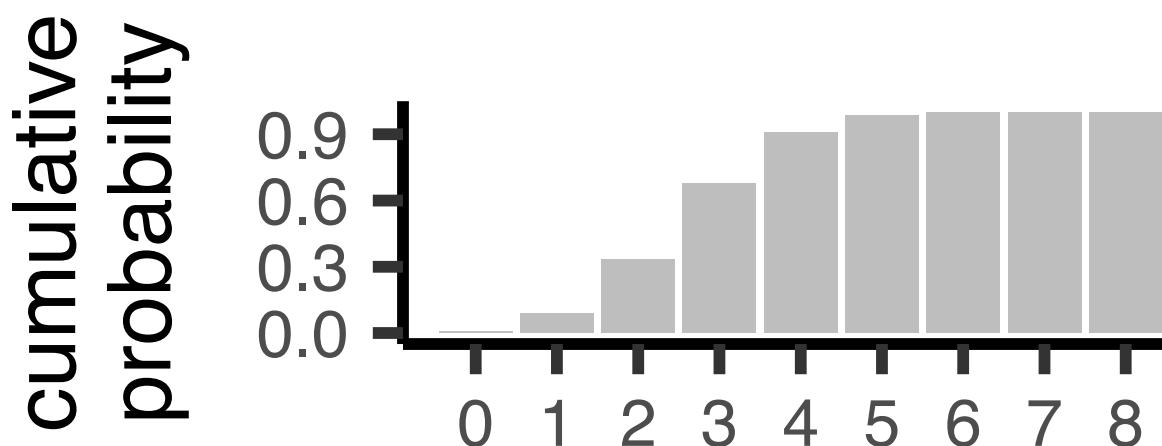
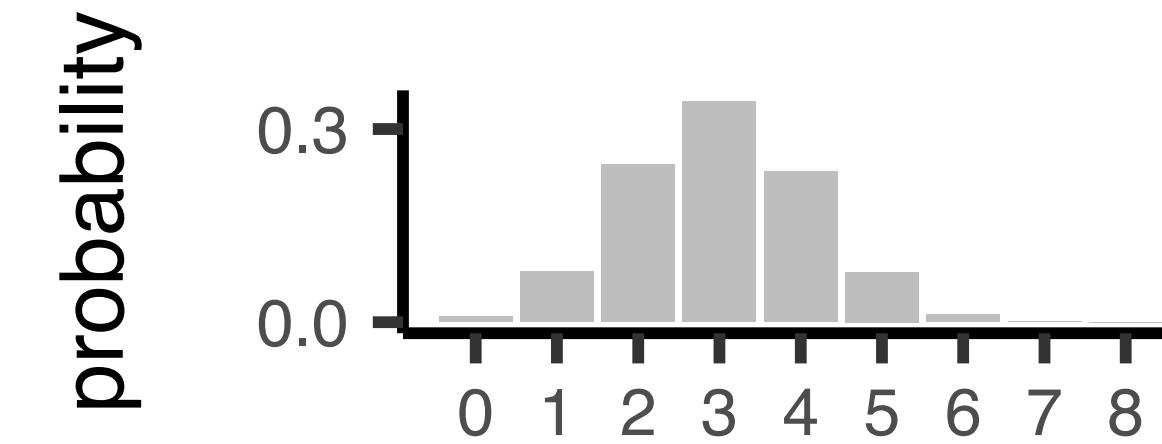
$m = 9$, $n = 15$, and $k = 8$,



How significant is q overlap in our discovery?

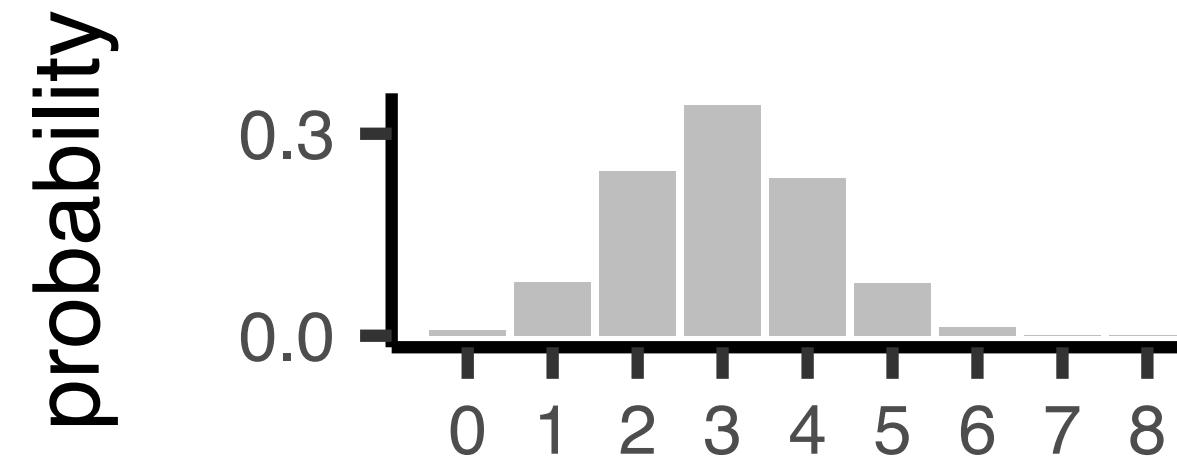
$m = 9$, $n = 15$, and $k = 8$,

If $q = 3$:

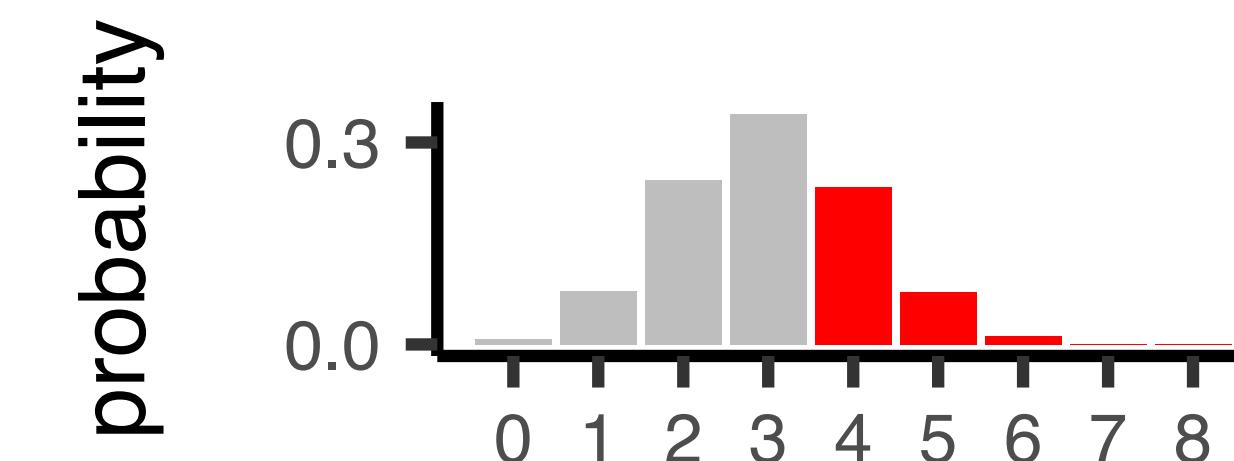


How significant is q overlap in our discovery?

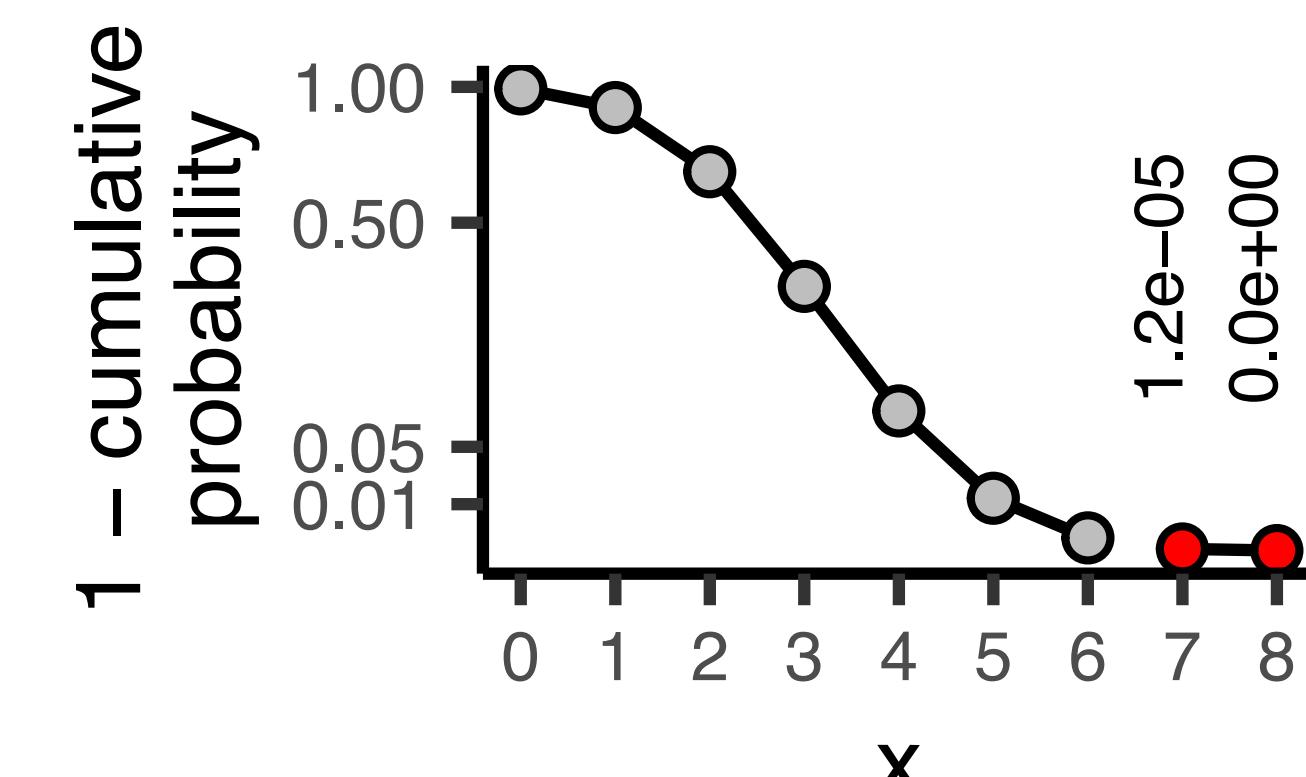
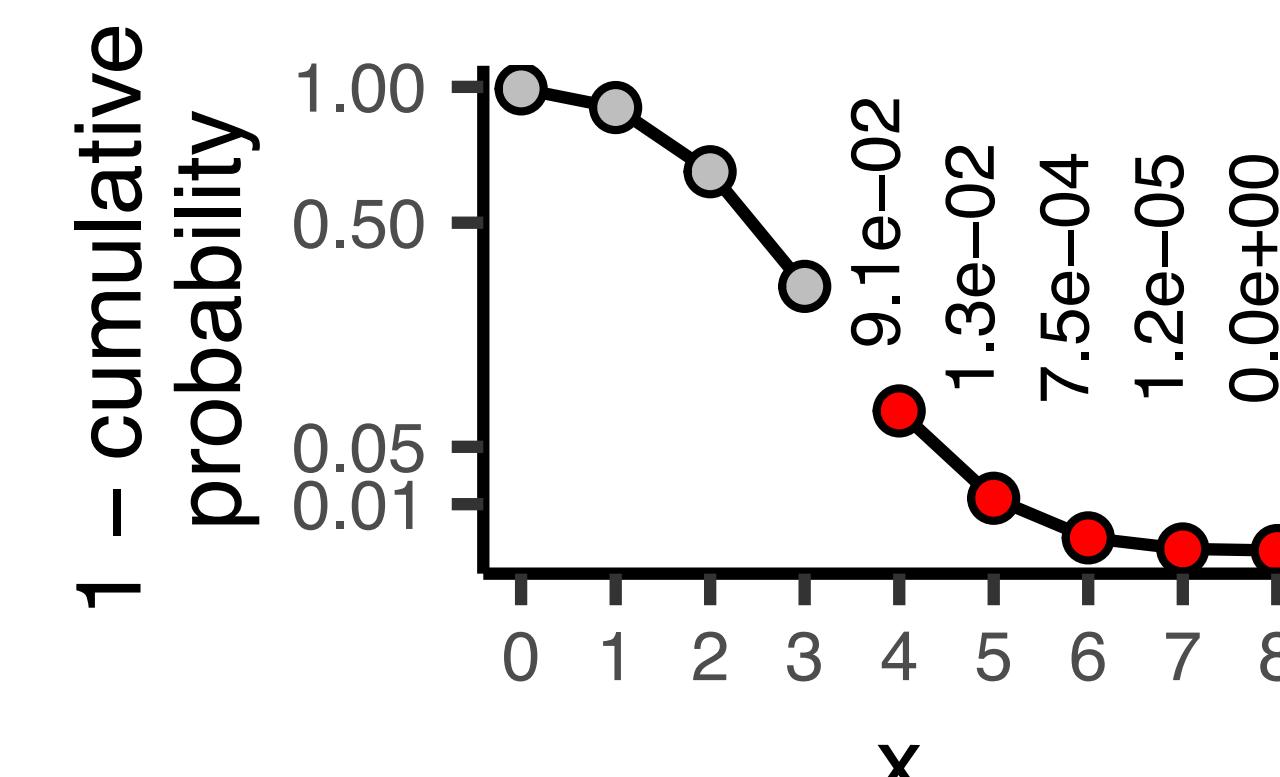
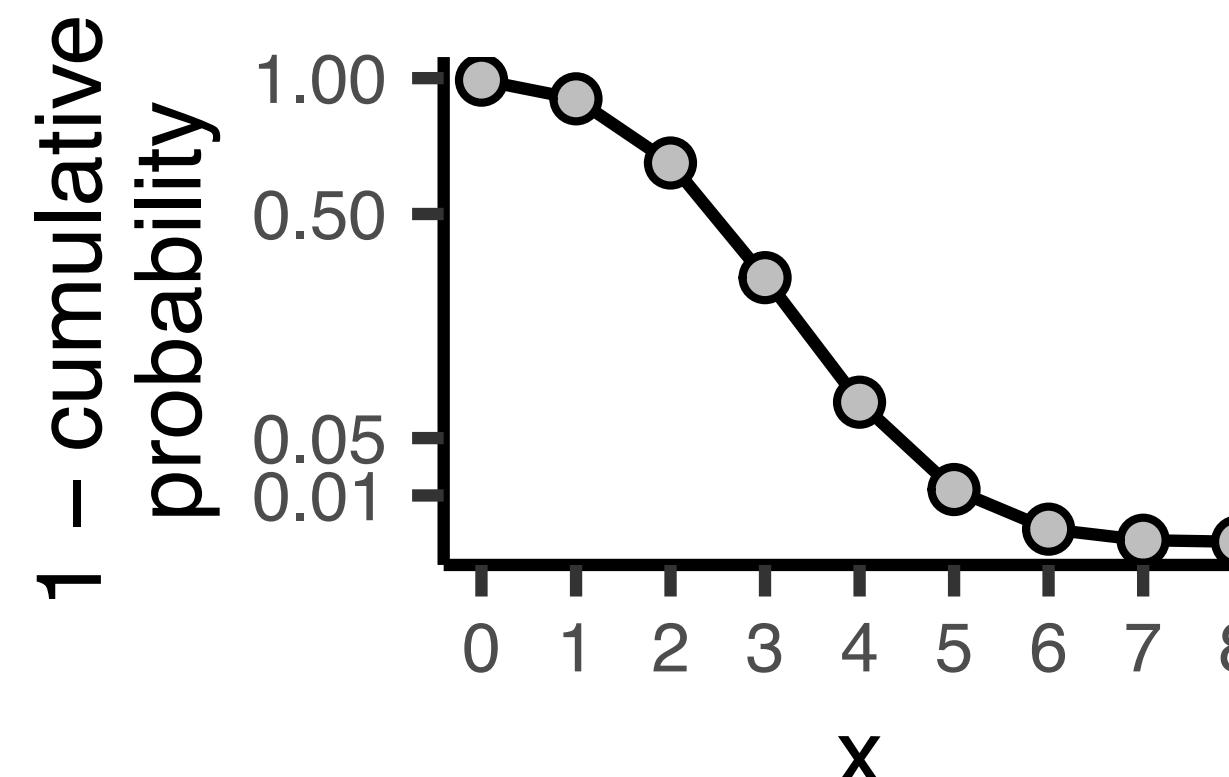
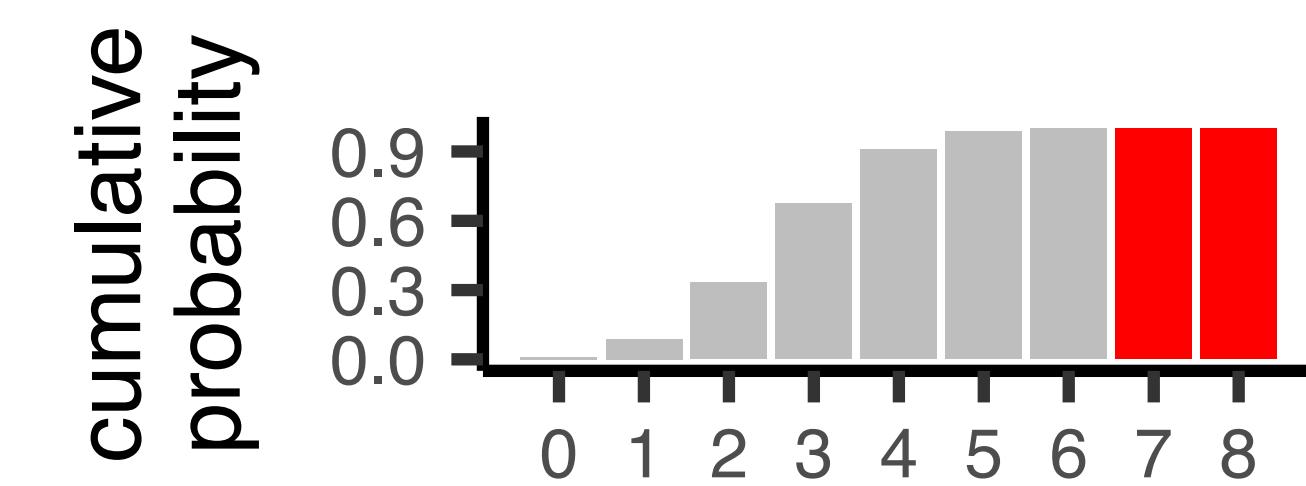
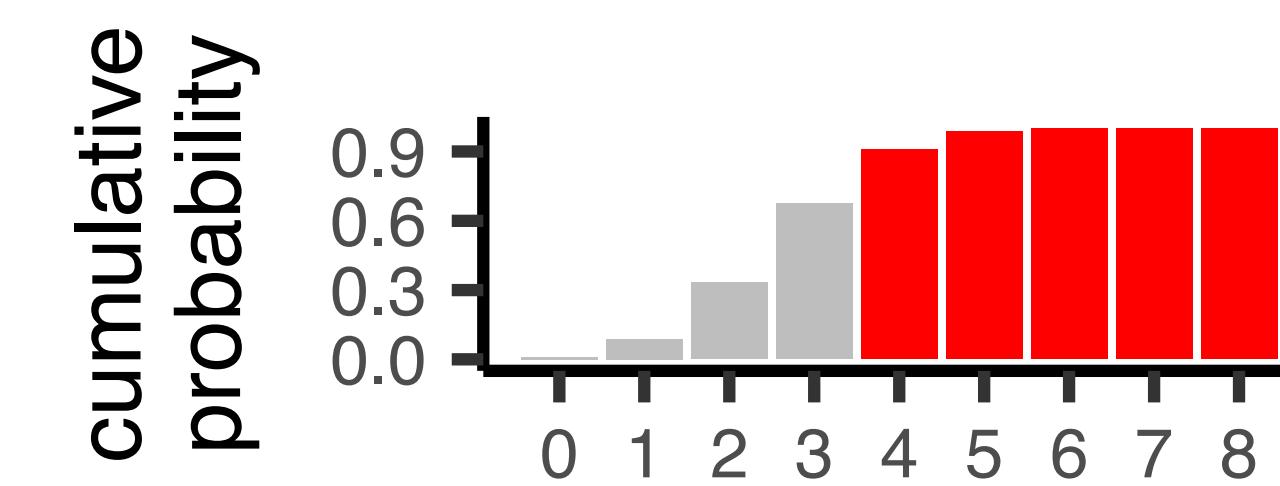
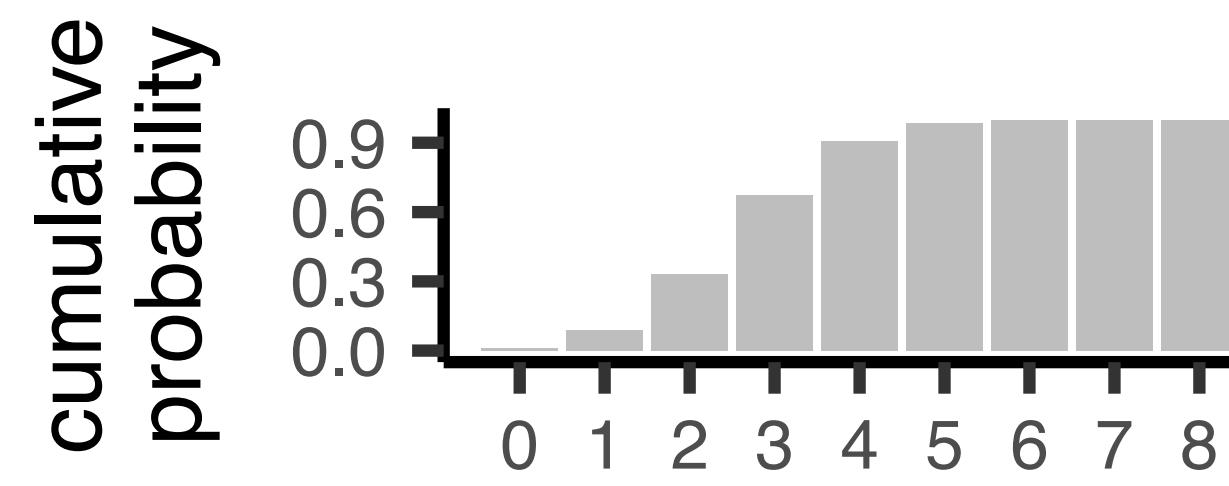
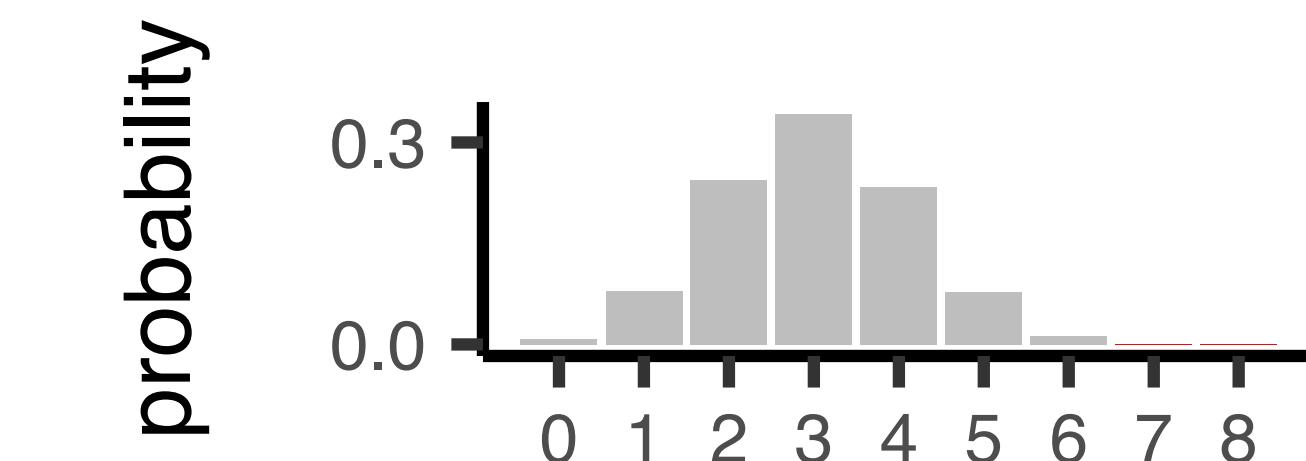
$m = 9$, $n = 15$, and $k = 8$,



If $q = 3$:



If $q = 6$:



What is Gene Set Analysis?

(Discrete) Gene Set Analysis

Input:

1. A dictionary of gene sets that map genes to sets (gene-to-set mapping)
2. A list of **top** genes identified in our own study (after FDR control)

Output:

A table of scores for all the gene sets in the dictionary.

(Rank-based) Gene Set Enrichment

Input:

1. A dictionary of gene sets that map genes to sets
2. A **full** list of gene-level **scores** (e.g., p-values)

Summary of the gene set analysis by hypergeometric test

When does it work?

- ▶ A routine to construct a set of differentially expressed genes by handling multiple hypothesis testing
- ▶ Gene sets are of similar sizes and *nearly* disjoint/independent from one another
- ▶ Genes are *nearly* independent (there is no overwhelmingly favourite genes)

When does it not work?

- ▶ **Don't** have a good way to make a set:
 - ▶ Our discovery data may lack statistical power, i.e., no (or a few) significant genes left after multiple hypothesis correction
- ▶ There is a hidden factor that can affect two steps: (1) gene set selection (annotations/knowledge) and (2) differential expression calling

Young *et al.* *Genome Biology* 2010, **11**:R14
<http://genomebiology.com/2010/11/2/R14>



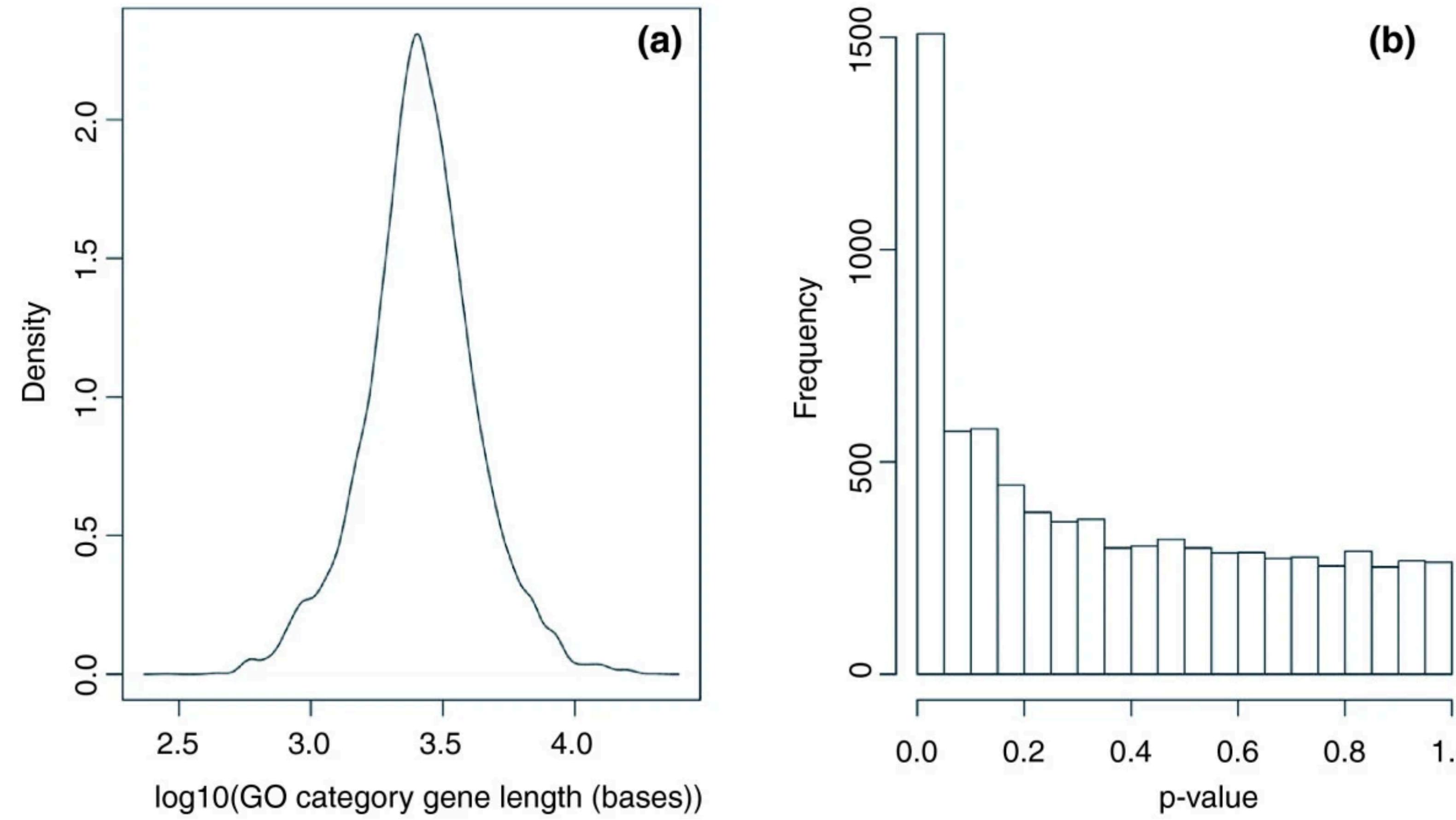
METHOD

Open Access

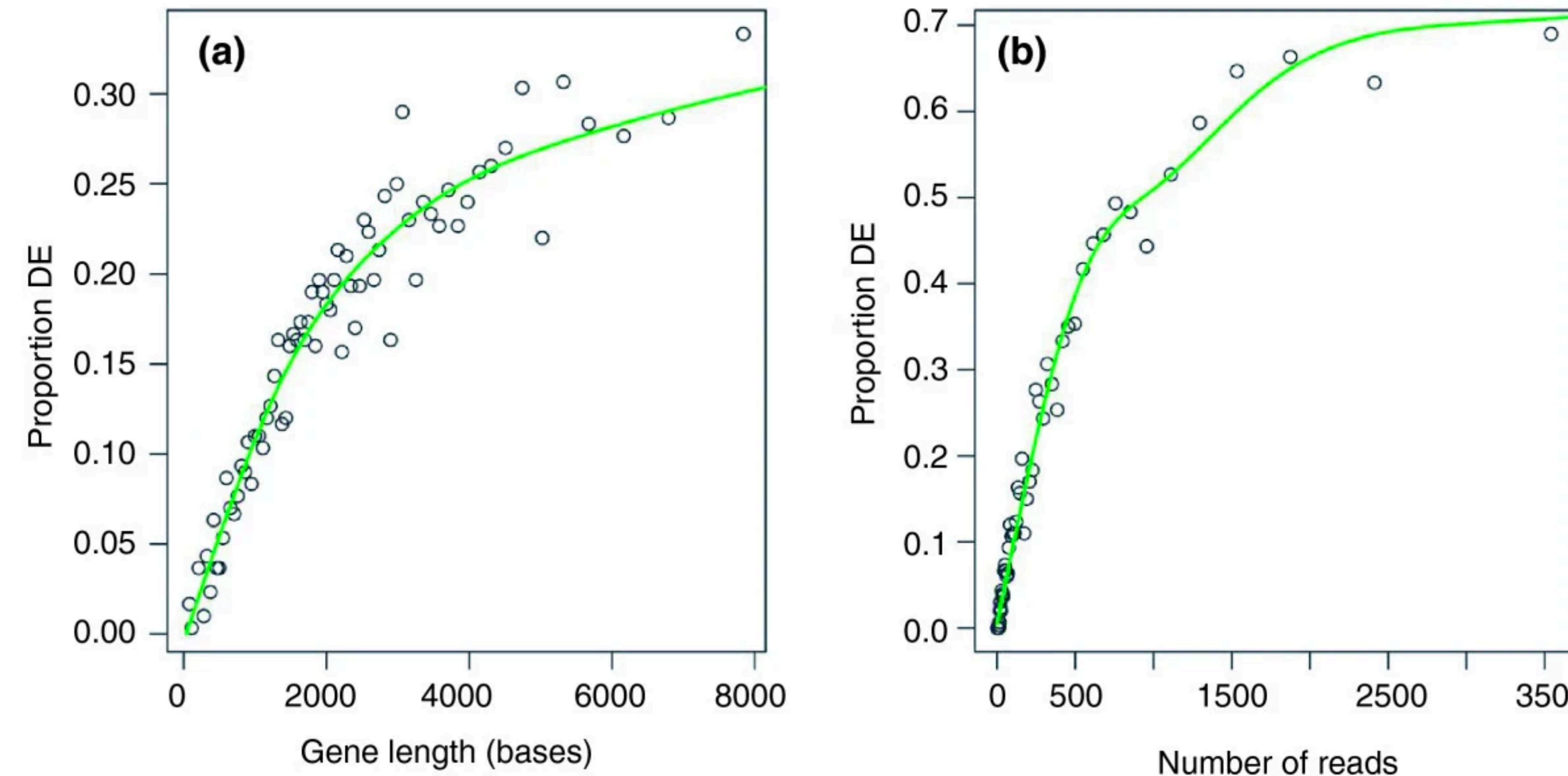
Gene ontology analysis for RNA-seq: accounting for selection bias

Matthew D Young, Matthew J Wakefield, Gordon K Smyth and Alicia Oshlack*

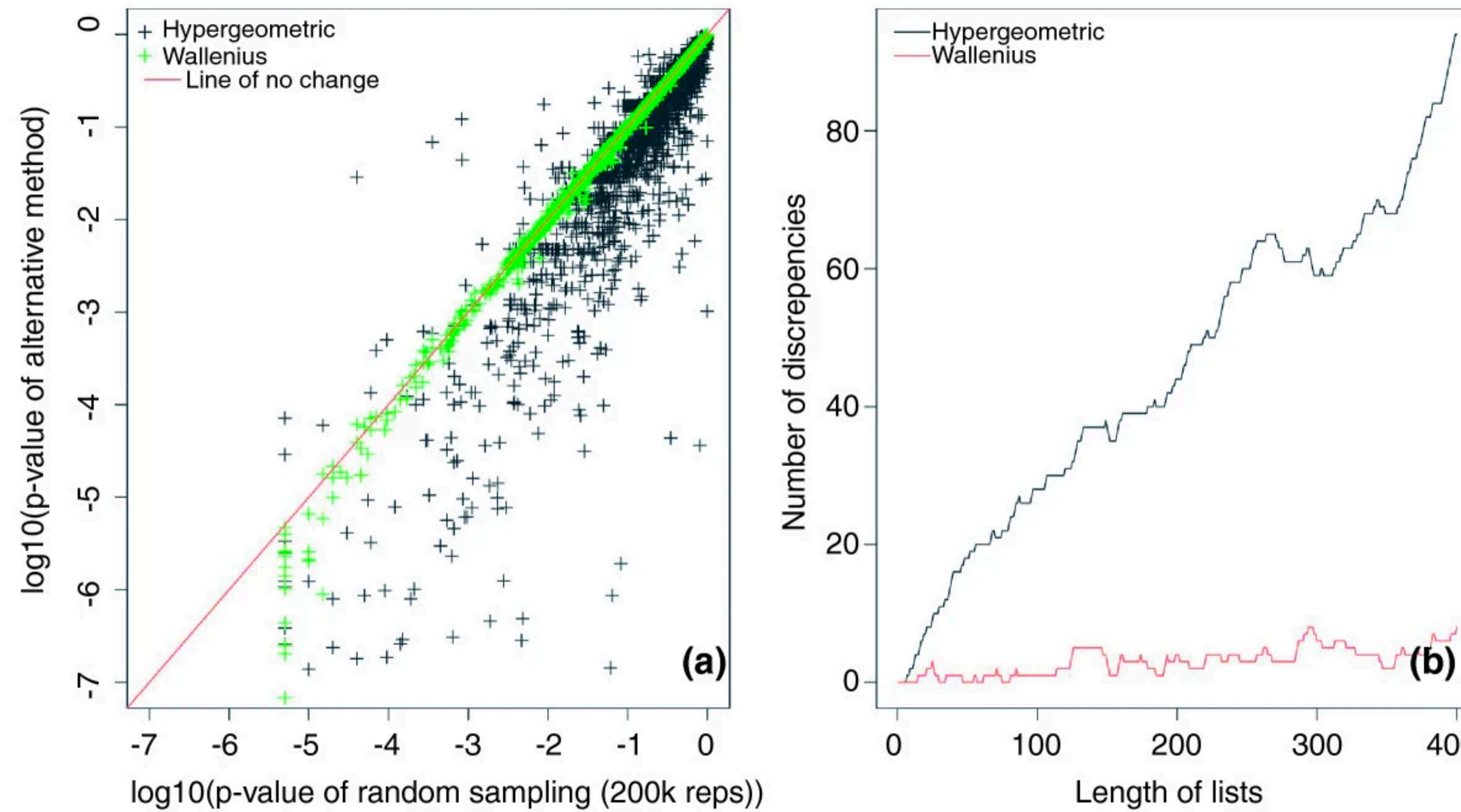
Not all gene sets are the same (gene length bias)



Longer genes tend to be more differentially expressed



A theoretical distribution of set-level scores may inflate p-values, thus wrong FDR calibration



goseq: GSA with gene length bias correction



Menu

Home > Bioconductor 3.20 > Software Packages > [goseq](#)

[goseq](#)

This is the **released** version of goseq; for the devel version, see [goseq](#).

Gene Ontology analyser for RNA-seq and other length biased data

platforms all rank 188 / 2289 support 0 / 0 in Bioc 15 years build ok updated before release

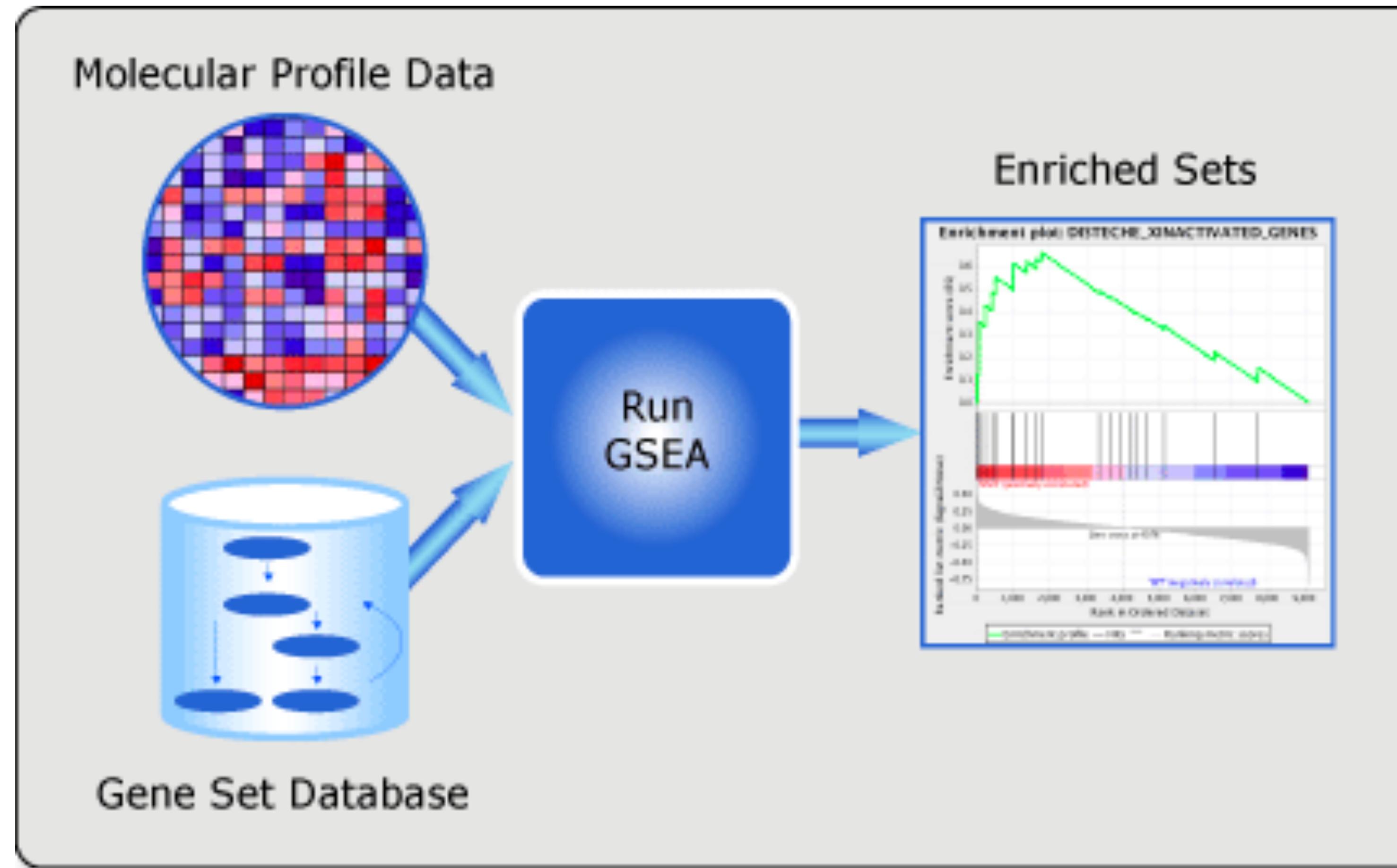
dependencies 108

DOI: [10.18129/B9.bioc.goseq](https://doi.org/10.18129/B9.bioc.goseq)

Today's lecture: Enrichment Analysis

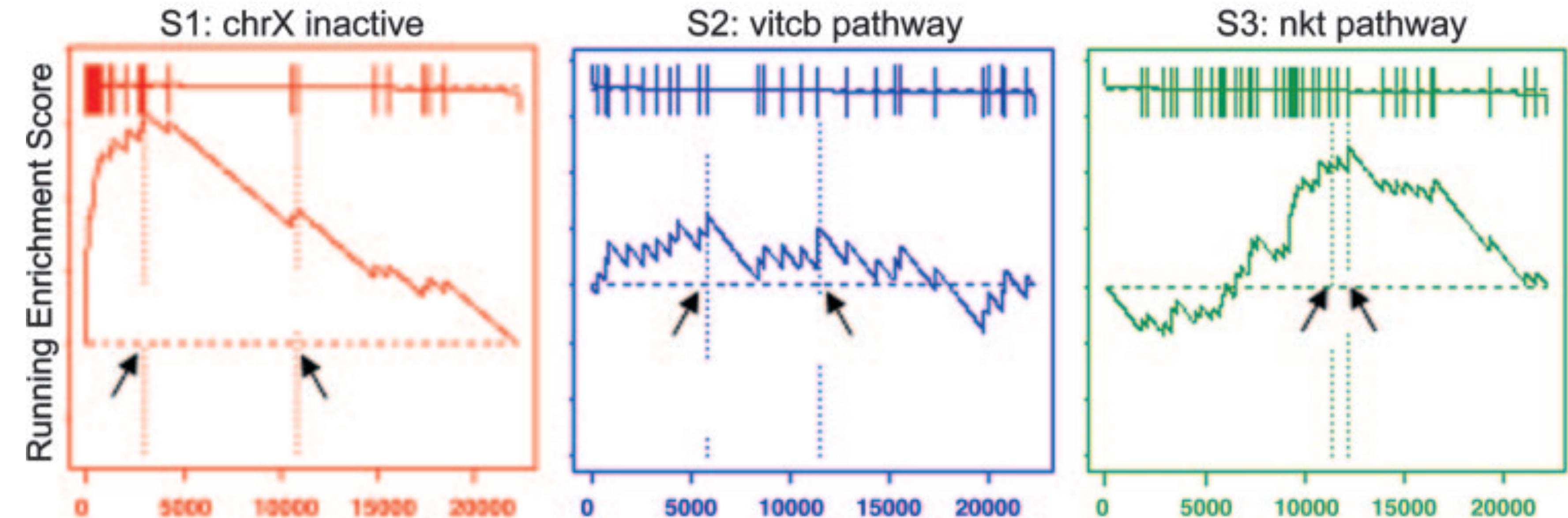
- **Motivations: What's next after genomics analysis?**
 - What have we learned?
 - How do we know that our discovery is meaningful?
- **Gene set enrichment analysis**
 - Set-based approach: Hypergeometric test
 - Rank-based approach: GSEA by KS statistic

Rank-based GSEA



Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian^{a,b}, Pablo Tamayo^{a,b}, Vamsi K. Mootha^{a,c}, Sayan Mukherjee^d, Benjamin L. Ebert^{a,e}, Michael A. Gillette^{a,f}, Amanda Paulovich^g, Scott L. Pomeroy^h, Todd R. Golub^{a,e}, Eric S. Lander^{a,c,i,j,k}, and Jill P. Mesirov^{a,k}



GSEA by ranking of genes

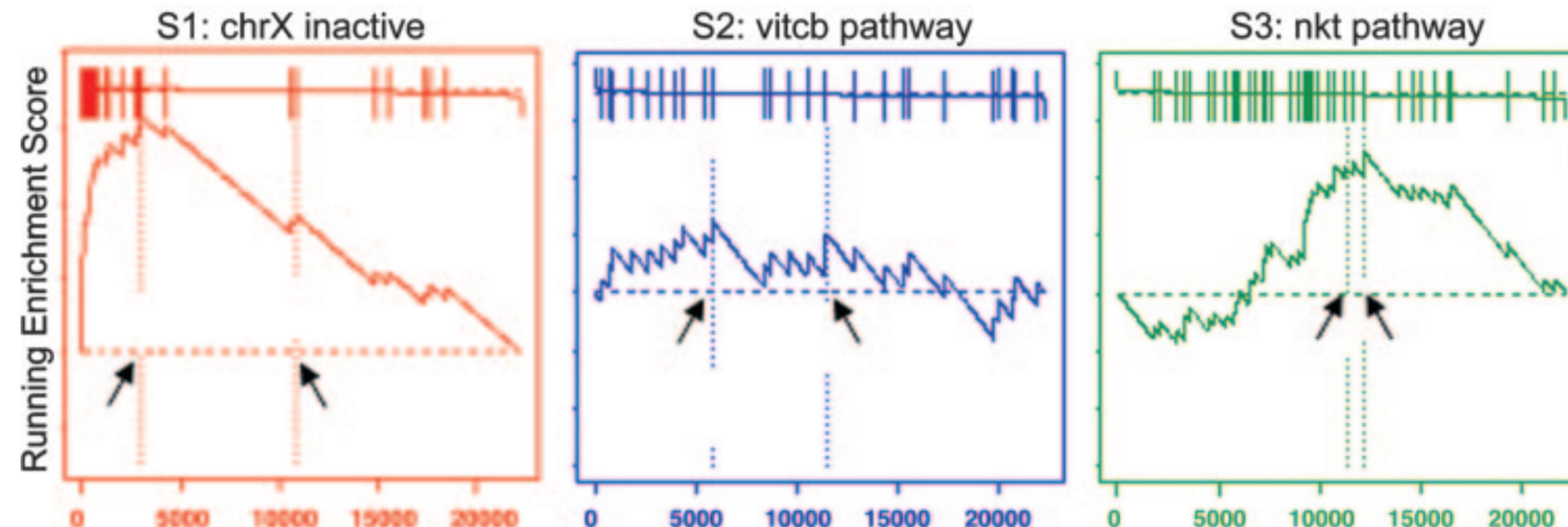
- A collection of gene sets: pathways
- A vector of gene scores (DEG p-values, or z-scores, etc)
- Question: Among the top list of our gene scores, which pathway is over-represented?

GSEA by ranking of genes

- A collection of gene sets: pathways
- A vector of gene scores (DEG p-values, or z-scores, etc)
- Question: Among the top list of our gene scores, which pathway is over-represented?

What will be a useful score for GSEA?

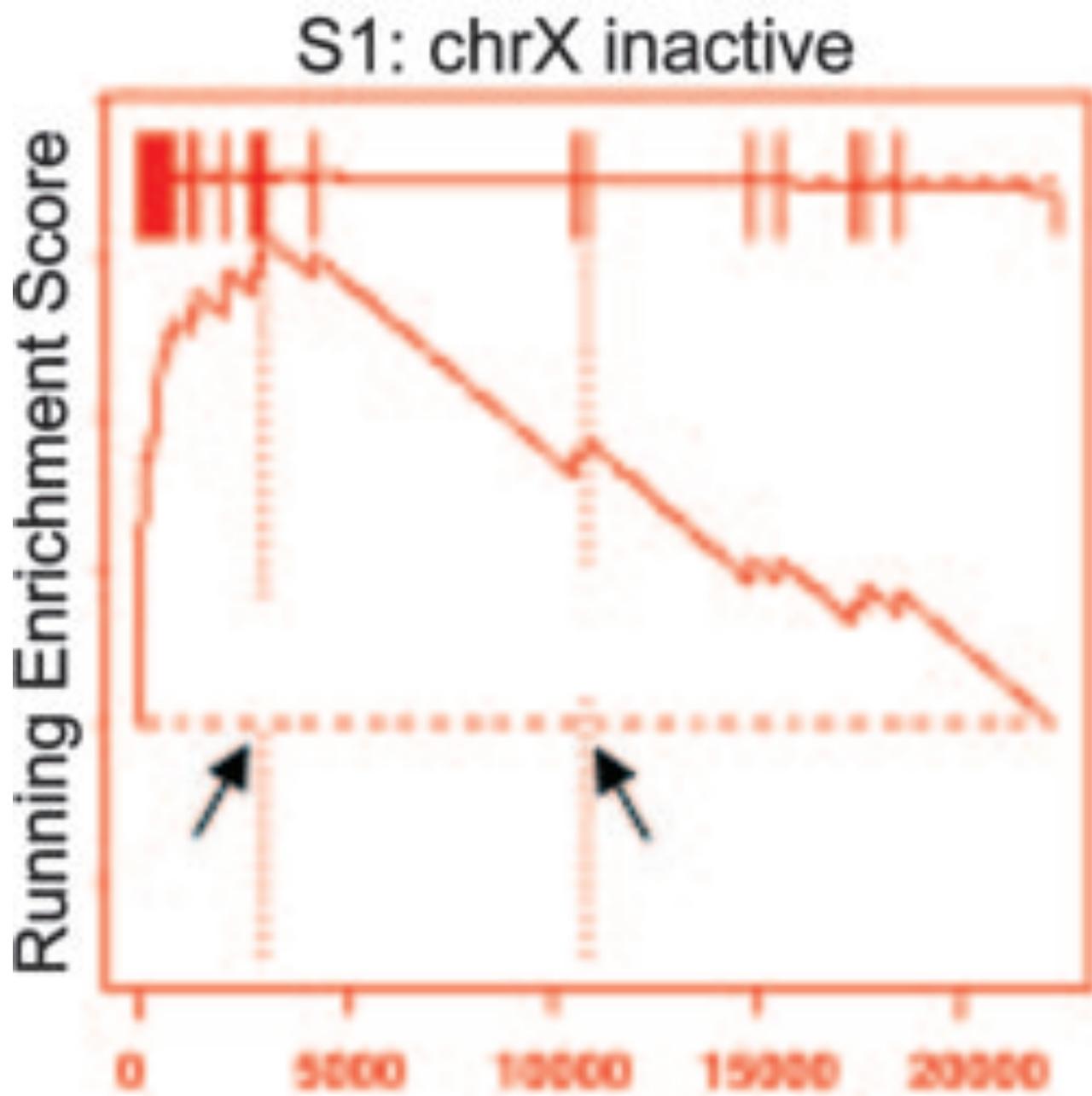
- For each gene set k :
 - Enrichment score that compare within vs. outside of this gene set



What will be a useful score for GSEA?

Enrichment Score $ES(S)$.

1. Rank order the N genes in D to form $L = \{g_1, \dots, g_N\}$ according to the correlation, $r(g_j) = r_j$, of their expression profiles with C .
2. Evaluate the fraction of genes in S (“hits”) weighted by their correlation and the fraction of genes not in S (“misses”) present up to a given position i in L .



$$P_{\text{hit}}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R}, \quad \text{where } N_R = \sum_{g_j \in S} |r_j|^p \quad [1]$$

$$P_{\text{miss}}(S, i) = \sum_{\substack{g_j \notin S \\ j \leq i}} \frac{1}{(N - N_H)}.$$

Gene Set Enrichment Analysis method (Subramanian *et al.* 2005)

- ▶ Construct null distribution of S_1, \dots, S_K by sample label (case-control) or gene-to-set membership permutation
- ▶ Using null distribution by permutation, estimate p-values and false discovery rates
- ▶ If we knew null distribution, we would not need expensive permutations.

Good:

- ▶ No cutoff/assumptions needed to estimate null distribution
- ▶ Aggregate scores across many genes! (boost the power)

Bad:

- ▶ What is an appropriate statistic?
- ▶ What should be permuted? For how long?

fgsea: fast GSEA (approximate permutation)



Menu

Home > Bioconductor 3.20 > Software Packages > **fgsea**

fgsea

This is the **released** version of fgsea; for the devel version, see [fgsea](#).

Fast Gene Set Enrichment Analysis

platforms all rank 41 / 2289 support 1 / 1 in Bioc 8.5 years build ok updated < 3 months
dependencies 49

DOI: [10.18129/B9.bioc.fgsea](https://doi.org/10.18129/B9.bioc.fgsea)

Today's lecture: Enrichment Analysis

- **Motivations: What's next after genomics analysis?**
 - What have we learned?
 - How do we know that our discovery is meaningful?
- **Gene set enrichment analysis**
 - Set-based approach: Hypergeometric test
 - Rank-based approach: GSEA by KS statistic
 - How could we come up with our own GSA?

Simulation (Efron and Tibshirani 2007)

1. Generate basal gene expression

$$X_{i,g} \sim \mathcal{N}(0, 1)$$

2. Sample case vs. control membership (the rows of X) uniformly at random
3. Sample membership gene to gene set uniformly at random
4. For the first gene set, select a certain fraction of genes to perturb
5. For the selected genes g^* , add some Δ value to X_{i,g^*} if the sample i belongs to the control group

Let's simulate some gene set data (Efron and Tibshirani 2007)

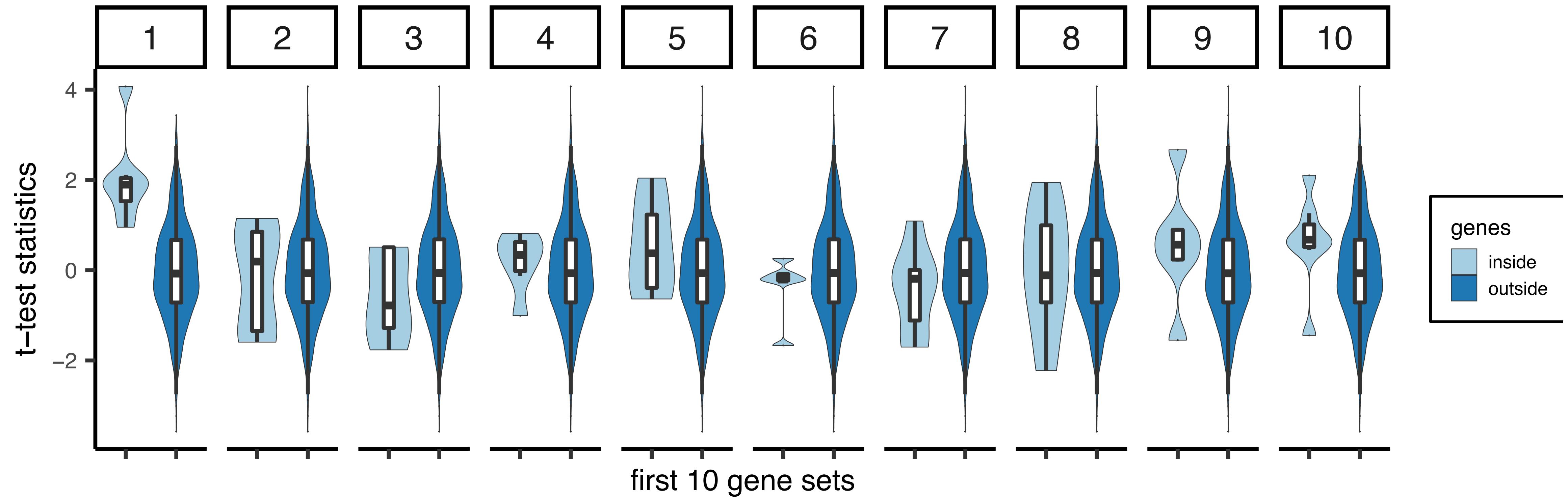
```
simulate.data <-
  function(G = 1000,           # genes
          K = 150,            # gene sets
          n.samp = 100,         # sample size
          delta = .4,           # perturbation
          p.perturb = 1) { # Pr of petrurbation
    case.control <- sample(0:1, n.samp, TRUE)
    S <- sample(K, G, TRUE)           # gene sets
    X <- rnorm(n.samp, G)             # All the other genes
    ## Perturbation of the first gene set
    .genes.1 <- which(S == 1)
    n1 <- length(.genes.1)
    n.perturb <- max(floor(n1 * p.perturb), 1)
    .genes.1 <- sample(.genes.1, n.perturb)
    .case <- case.control == 1
    X[.case, .genes.1] <- X[.case, .genes.1] + delta
    require(Matrix)
    .membership <- sparseMatrix(j=1:G, i=S, x=rep(1,G))
    list(X=X, S=.membership, Y=case.control)
  }
```

```
set.seed(1)
dat <- simulate.data()
```

Let's run t-test for each gene:

```
run.t.test <- function(X, Y){
  .case <- Y == 1
  .ctrl <- Y == 0
  .fun <- function(x){
    t.test(x[.case],
           x[.ctrl])$statistic
  }
  apply(X, 2, .fun)
}
```

The goal is to come up with a representative score for all the genes within each set



What will be a proper gene set score?

Can we simply aggregate gene-level z-scores (or t-statistics) within each set?

Irizarry *et al.* (2009), using Stouffer Z-score

$$S_k = \sum_{g \in \mathcal{C}_k} z_g / \sqrt{|\mathcal{C}_k|} \sim \mathcal{N}(0, 1)$$

if $Z_g \sim \mathcal{N}(0, 1)$, $\forall g$

Aggregating z-scores within a set to have one number for the set

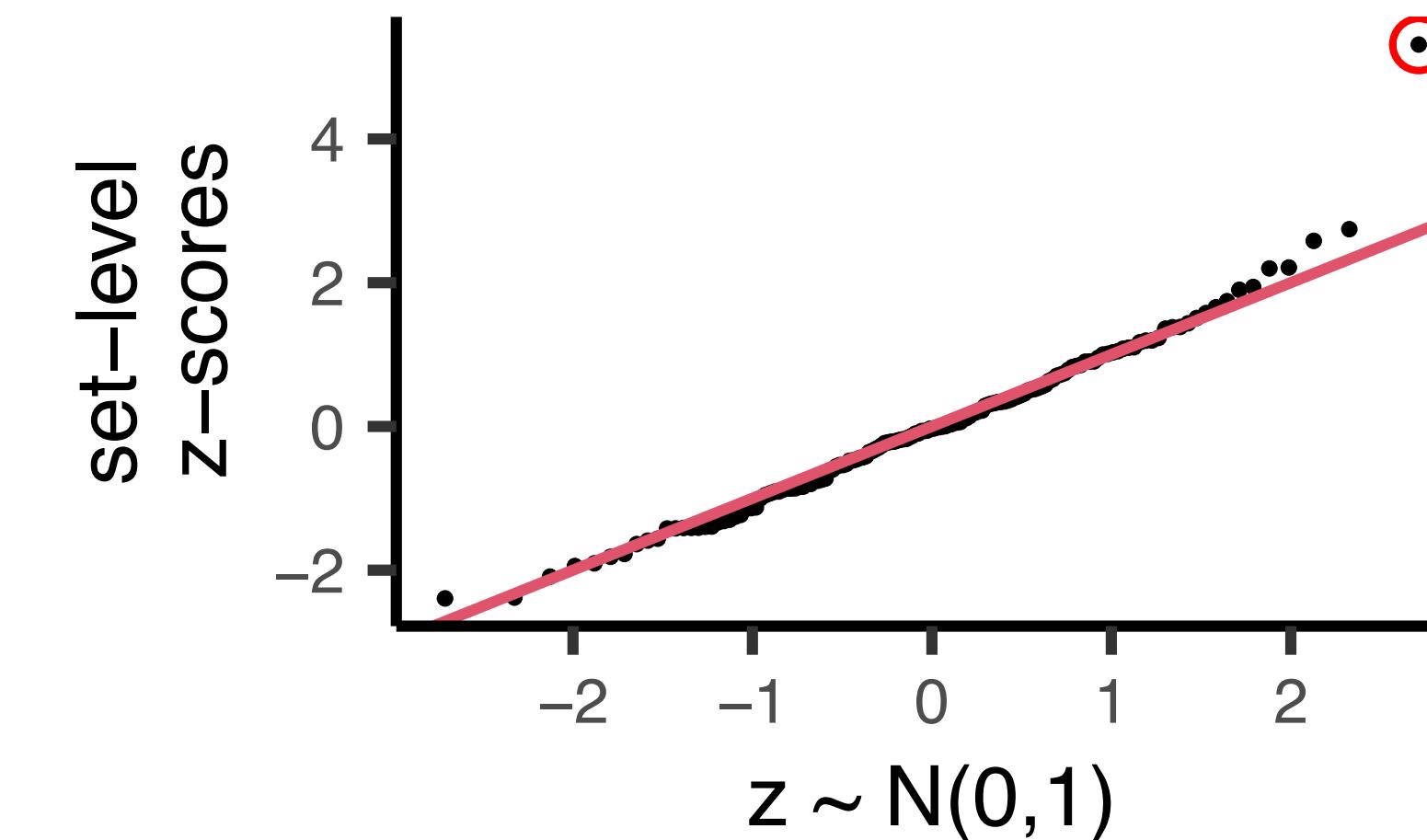
```
geneset.score <- function(X, S, Y) {  
  z.genes <- run.t.test(X, Y)  
  n.sets <- apply(S, 1, sum)  
  z.sets <- (S %*% z.genes /  
             sqrt(n.sets))  
}
```

```
z.sets <- geneset.score(dat$X, dat$S, dat$Y)
```

Aggregating z-scores within a set to have one number for the set

```
geneset.score <- function(X, S, Y) {  
  z.genes <- run.t.test(X, Y)  
  n.sets <- apply(S, 1, sum)  
  z.sets <- (S %*% z.genes /  
             sqrt(n.sets))  
}
```

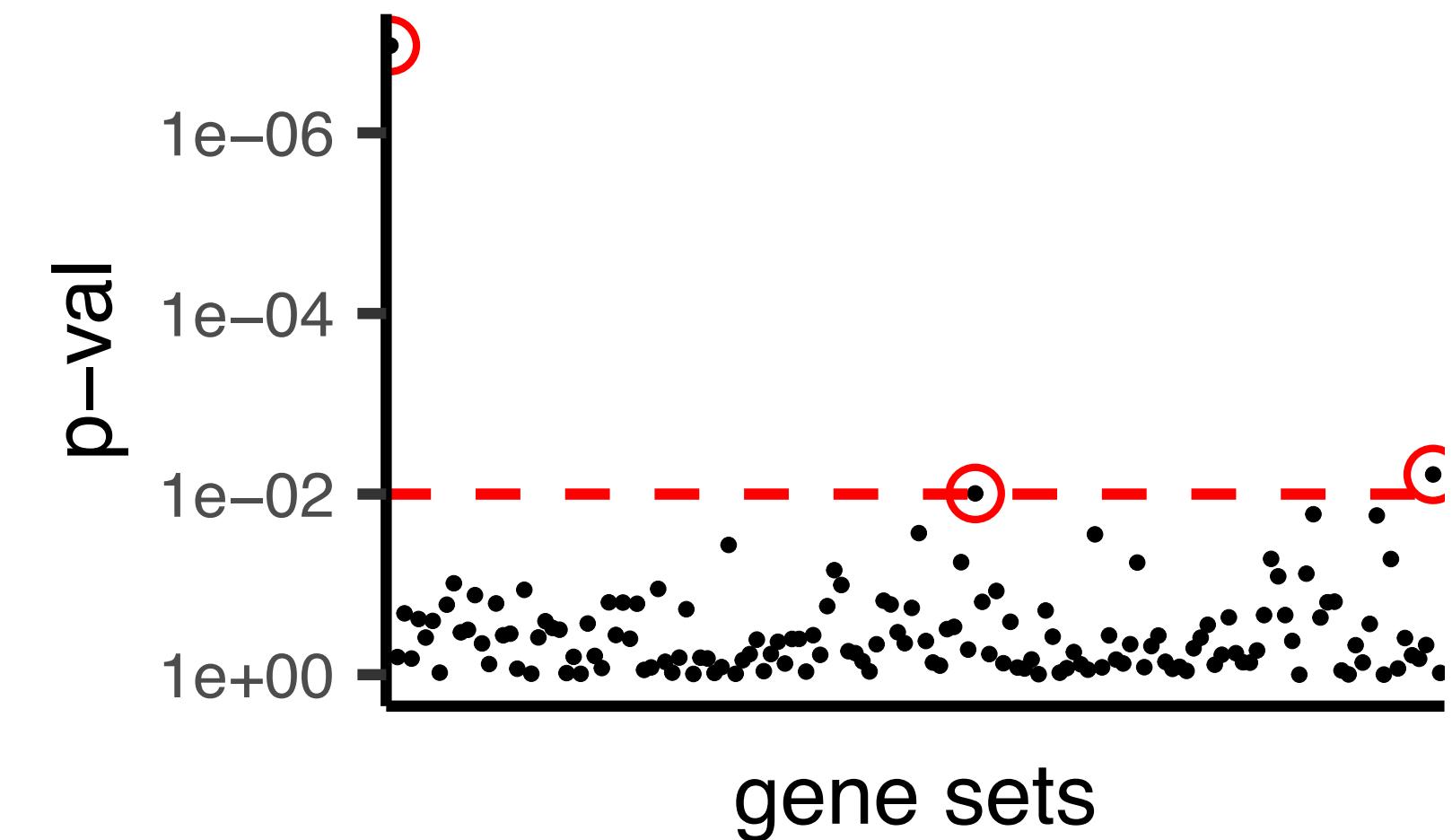
```
z.sets <- geneset.score(dat$X, dat$S, dat$Y)
```



Aggregating z-scores within a set to have one number for the set

```
geneset.score <- function(X, S, Y) {  
  z.genes <- run.t.test(X, Y)  
  n.sets <- apply(S, 1, sum)  
  z.sets <- (S %*% z.genes /  
             sqrt(n.sets))  
}
```

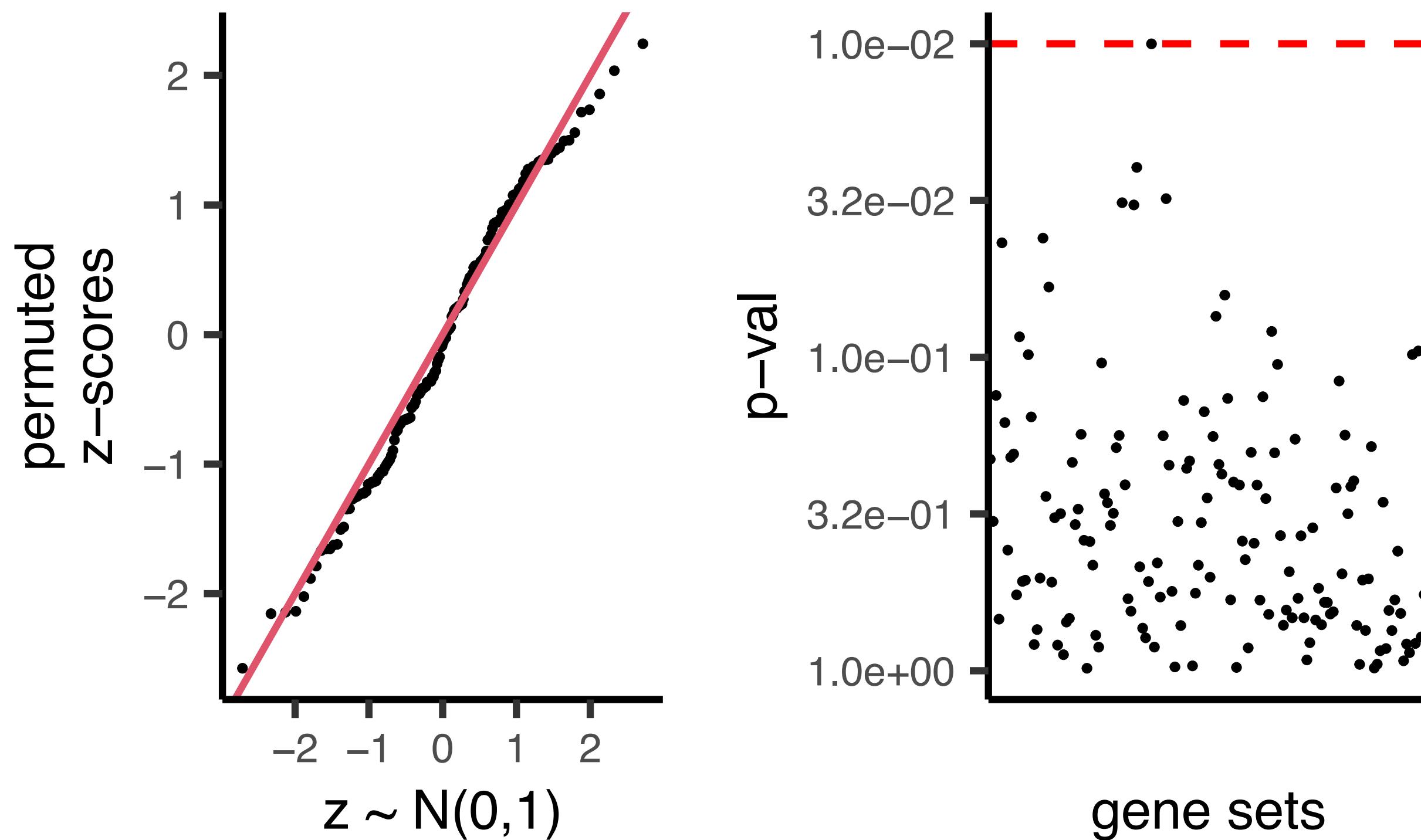
```
z.sets <- geneset.score(dat$X, dat$S, dat$Y)
```



Constructing null distribution by gene permutation

What if we don't know the distribution of set-wise scores?

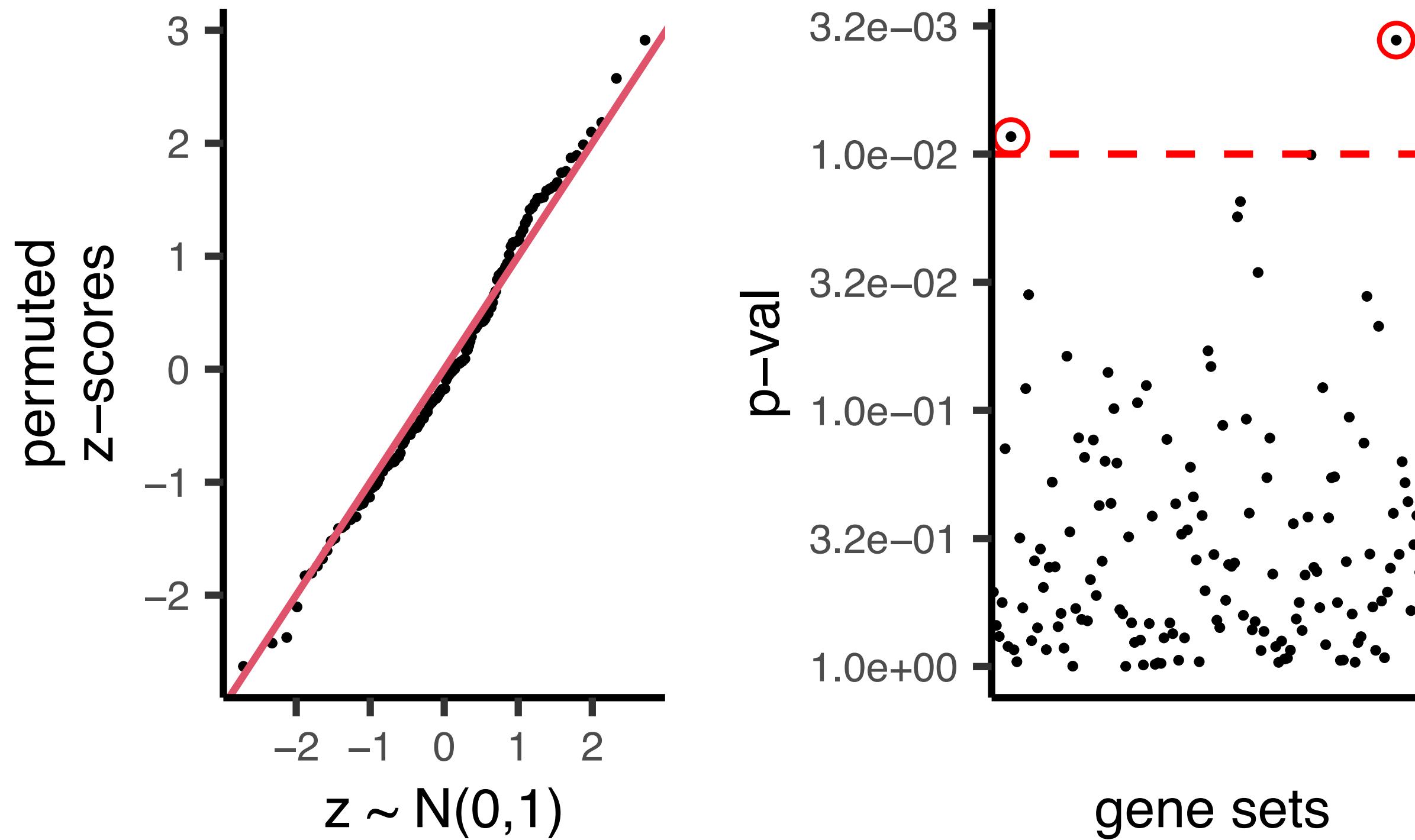
```
S.perm <- t(apply(dat$S, 1, sample))  
z.perm <- geneset.score(dat$X, S.perm, dat$Y)
```



- ▶ Repeat the permutation of gene set membership matrix while preserving the number of genes within each set
- ▶ Compute set-level z-scores (or a similar kind) and construct null distribution
- ▶ Calculate p-values by counting the frequency of observed $S_k^* > S_k^{\text{perm}}$

Constructing null distribution by sample permutation

```
Y.perm <- sample(dat$Y)  
z.perm <- geneset.score(dat$x, S.perm, Y.perm)
```



- ▶ Repeat the permutation of case-control labels while preserving the same number of cases and controls
- ▶ Compute set-level z-scores (or a similar kind) and construct null distribution
- ▶ Calculate p-values by counting the frequency of observed $S_k^* > S_k^{\text{perm}}$

ON TESTING THE SIGNIFICANCE OF SETS OF GENES

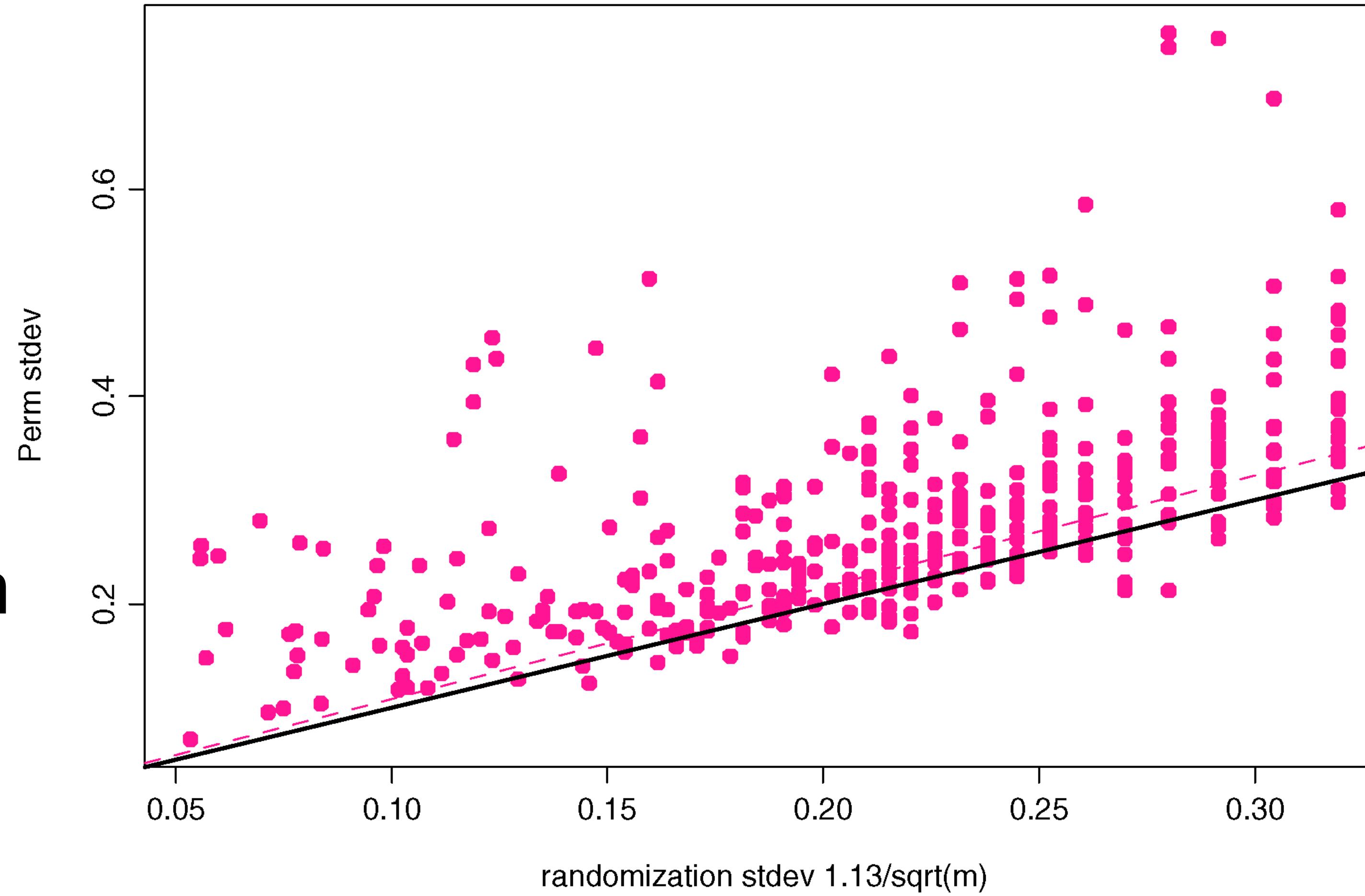
BY BRADLEY EFRON¹ AND ROBERT TIBSHIRANI²

Stanford University

This paper discusses the problem of identifying differentially expressed groups of genes from a microarray experiment. The groups of genes are externally defined, for example, sets of gene pathways derived from biological databases. Our starting point is the interesting Gene Set Enrichment Analysis (GSEA) procedure of Subramanian et al. [*Proc. Natl. Acad. Sci. USA* **102** (2005) 15545–15550]. We study the problem in some generality and propose two potential improvements to GSEA: the *maxmean* statistic for summarizing gene-sets, and *restandardization* for more accurate inferences. We discuss a variety of examples and extensions, including the use of gene-set scores for class predictions. We also describe a new R language package *GSA* that implements our ideas.

Row vs. column (gene vs. sample) permutation

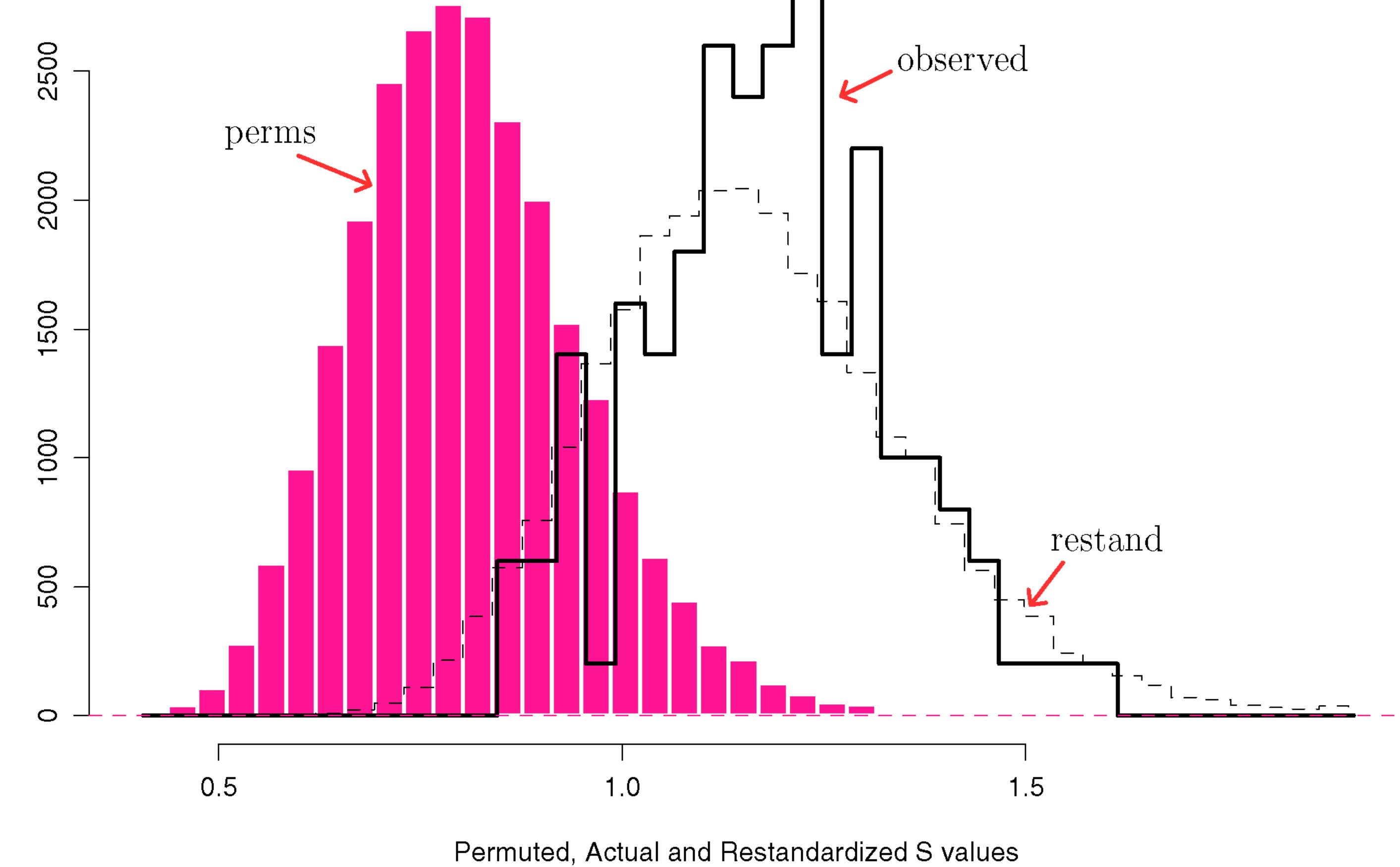
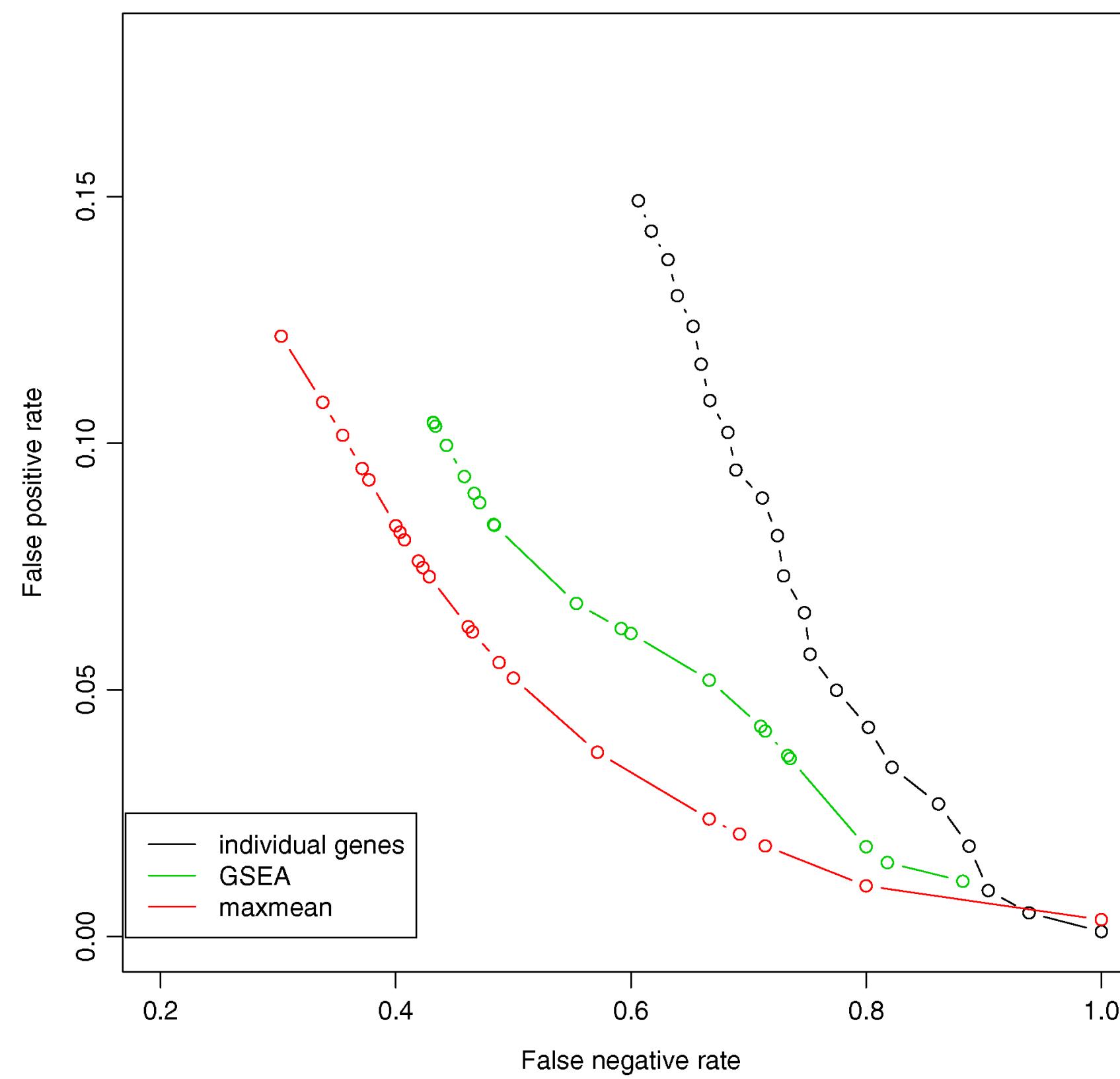
sample to
disease/
condition
label
permutation



**Under-
estimating
variance
under the
null**

gene-to-set membership permutation

Mis-calibrating the null distribution will make almost everything very significant



Today's lecture: Enrichment Analysis

- **Motivations: What's next after genomics analysis?**
 - What have we learned?
 - How do we know that our discovery is meaningful?
- **Gene set enrichment analysis**
 - Set-based approach: Hypergeometric test
 - Rank-based approach: GSEA by KS statistic
 - How could we come up with our own GSA?