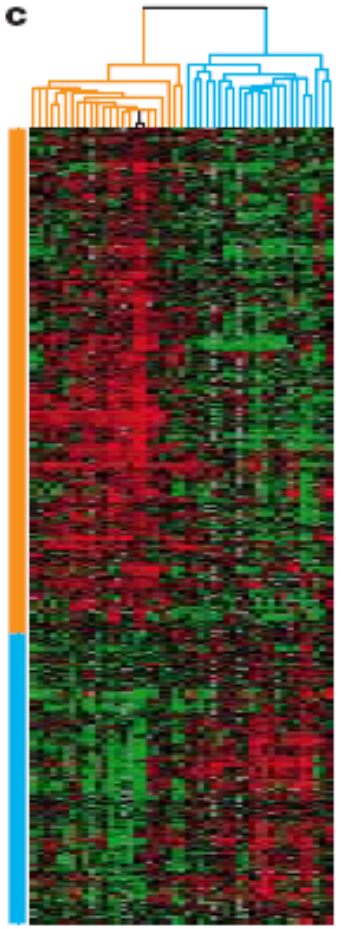
Two types of questions we might ask about expression data:



Consequences

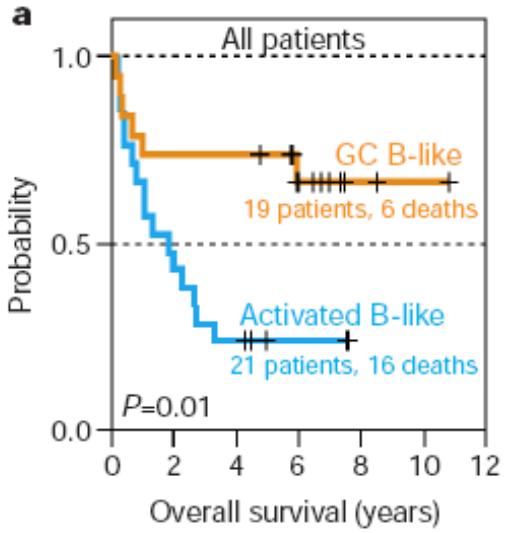
What are the biological consequences of the expression changes?

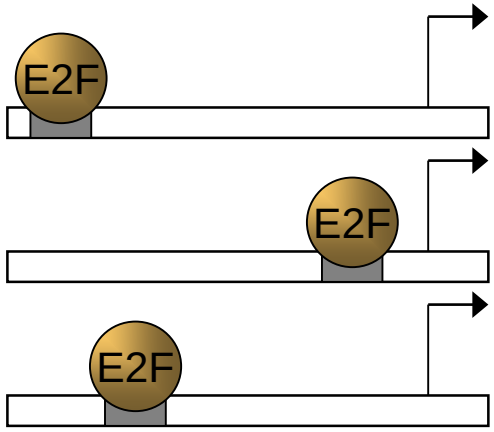
What categories of genes change in expression?

Causes

What causes these genes to change in expression?

Does a common transcription factor regulate them?

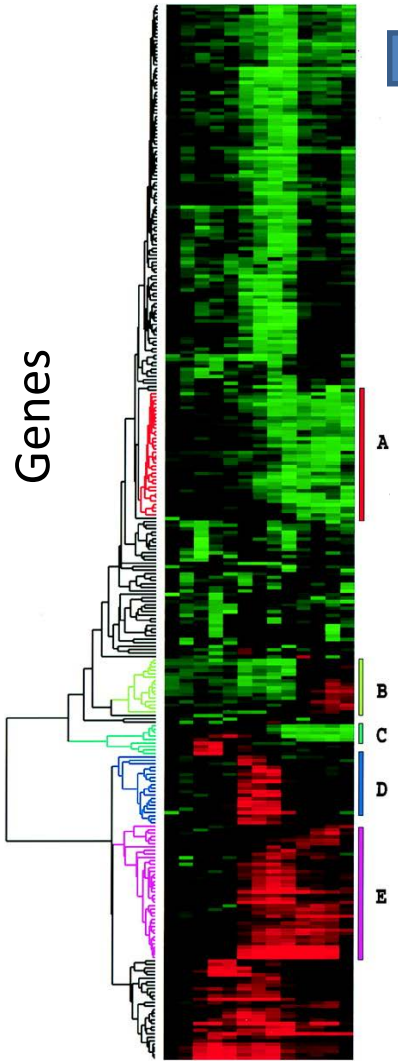




Outline

- Evaluating the statistical significance of an annotation
  - Hypergeometric distribution:
    - The null hypothesis:
  - Aggregate score statistics
  - Multiple hypotheses
  - Healthy dose of skepticism
- Applications to analysis of gene expression:
  - Consequences: Function of differentially expressed genes
  - Causes: Identity of transcriptional regulators
    - Known binding sites
    - Predicted binding sites

Recall our setting last time: Interpreting transcriptional results



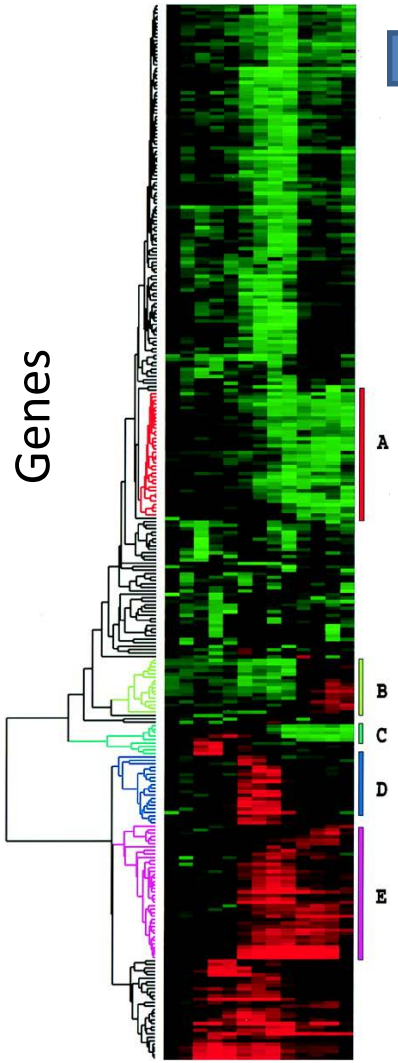
GO Terms

What do the differentially expressed genes do?

Let's say 10% of the differentially expressed genes have annotation A.
Should we investigate this annotation?

- What if this annotation contains 10% of all genes in the genome?
- What if this annotation contains 25% of all genes in the genome?

Recall our setting last time: Interpreting transcriptional results



GO Terms

What do the differentially expressed genes do?

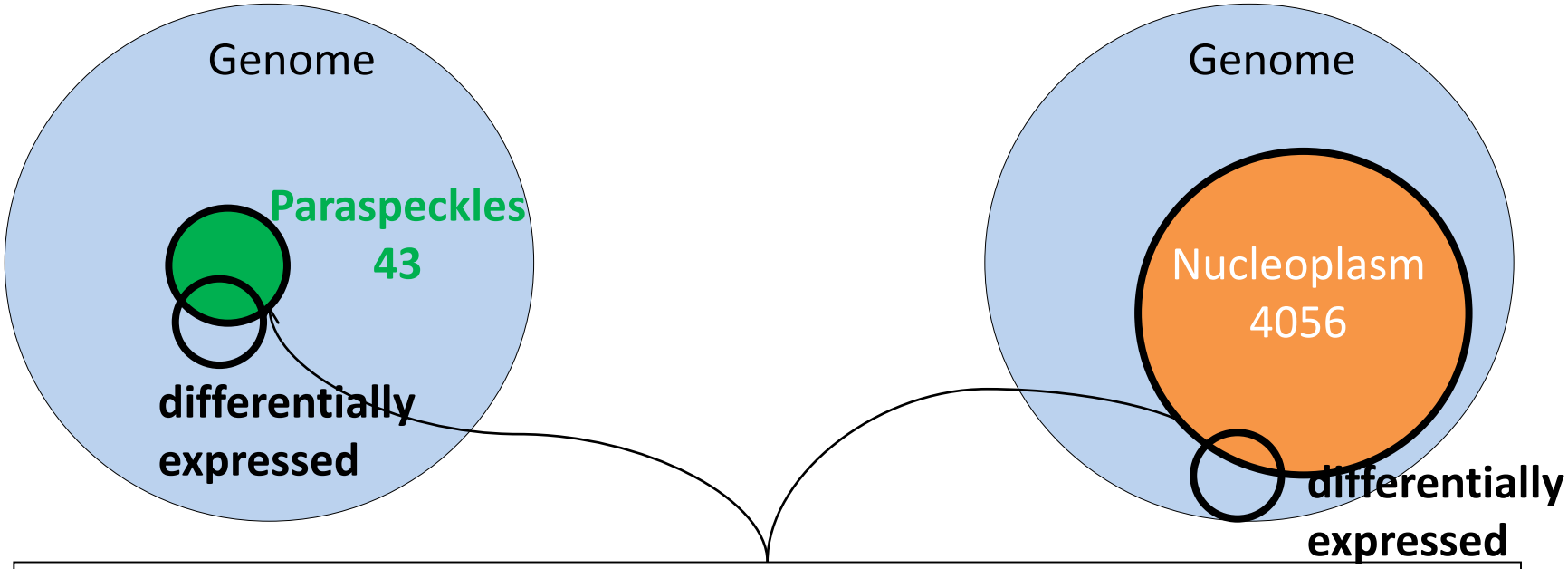
Do any annotations occur more often than expected by chance?

To answer this question, we need a ***null hypothesis***.

The simplest ***null hypothesis*** is that the occurrence of an annotation is independent of the experiment … it could have occurred by chance.

Consider two annotations: Nucleoplasm and paraspeckles

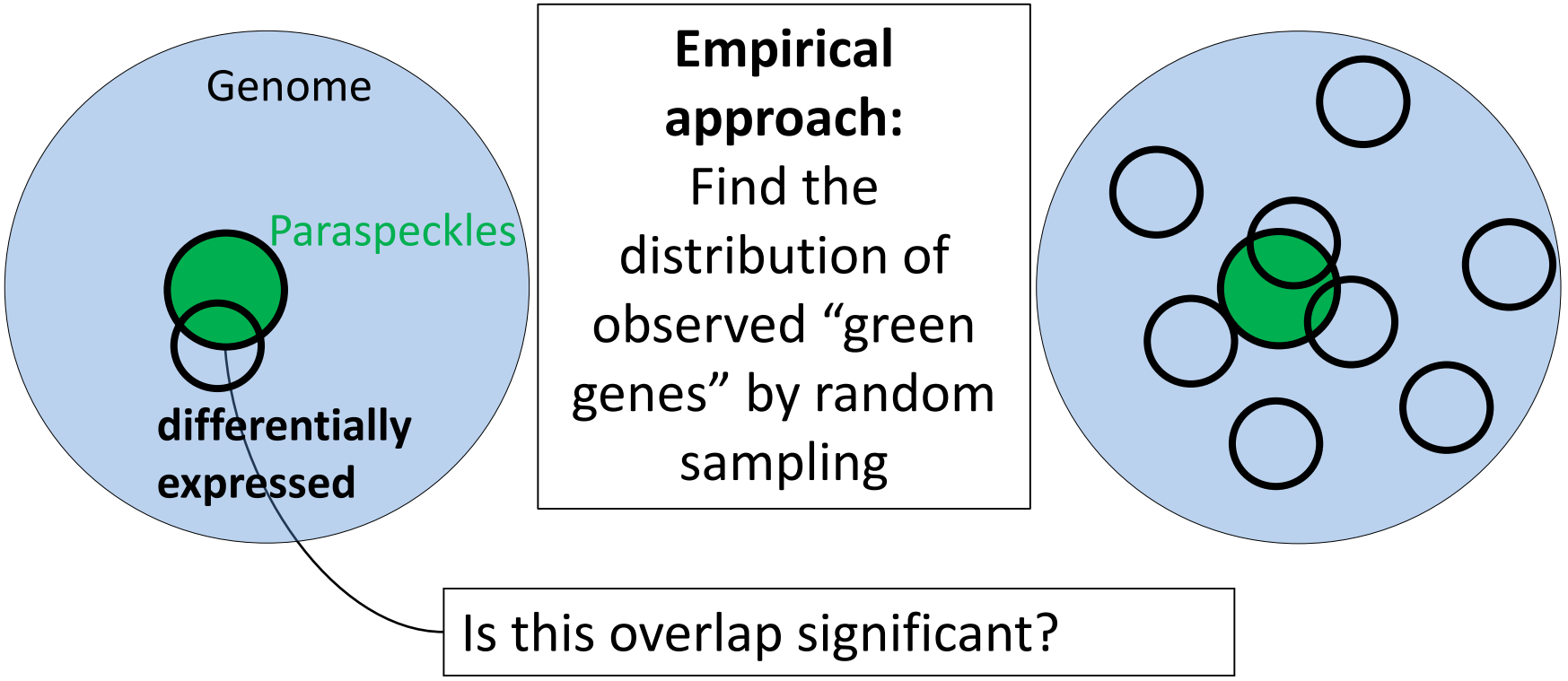
The significance depends on the size of the lists.



Very few genes are found in paraspeckles.

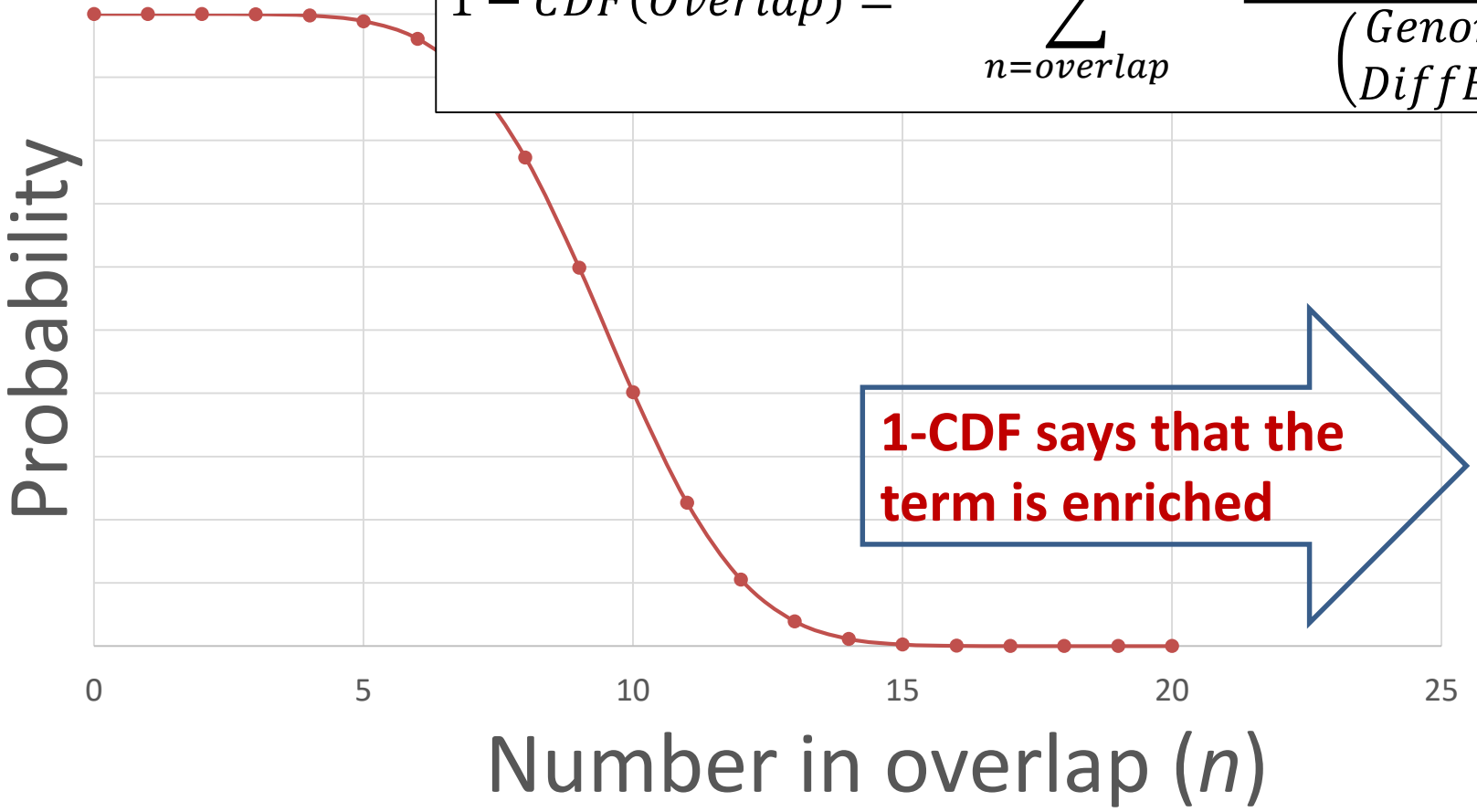
- If a lot of our differentially expressed genes have this rare annotation, it is worth exploring.
- Finding lots of nuclear genes is less interesting.

To determine statistical significance, we need to specify a null-model



(1-CDF) of the hypergeometric distribution gives the probability of observing *n or more*

$$1 - CDF(Overlap) = \sum_{n=overlap}^{N=Number\ of\ genes\ in\ category} \frac{\binom{N}{n}\binom{Genome - N}{DiffExp - n}}{\binom{Genome}{DiffExp}}$$

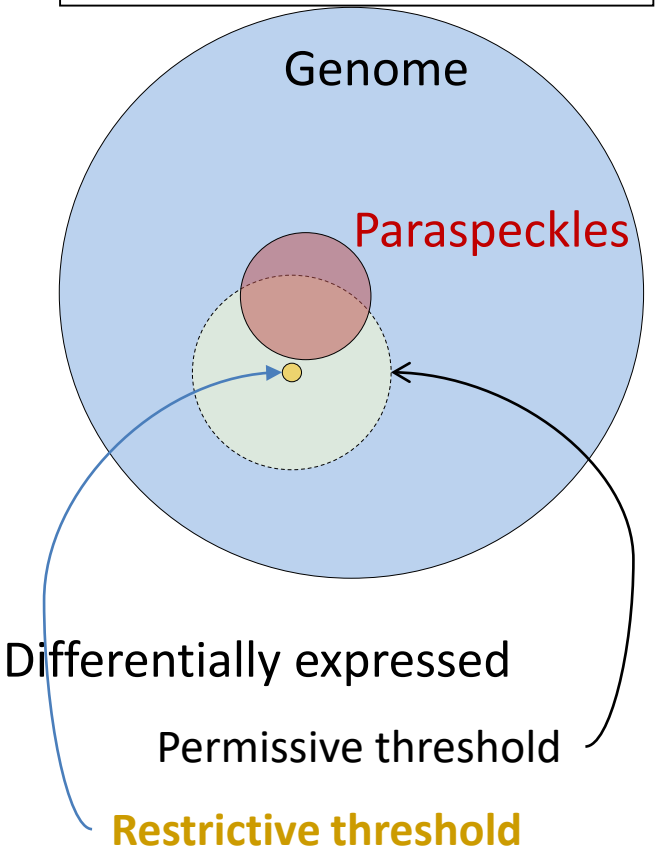


Outline

- Evaluating the statistical significance of an annotation
  - Hypergeometric distribution:
    - The null hypothesis:
  - Aggregate score statistics
  - Multiple hypotheses
  - Healthy dose of skepticism
- Applications to analysis of gene expression:
  - Consequences: Function of differentially expressed genes
  - Causes: Identity of transcriptional regulators
    - Known binding sites
    - Predicted binding sites

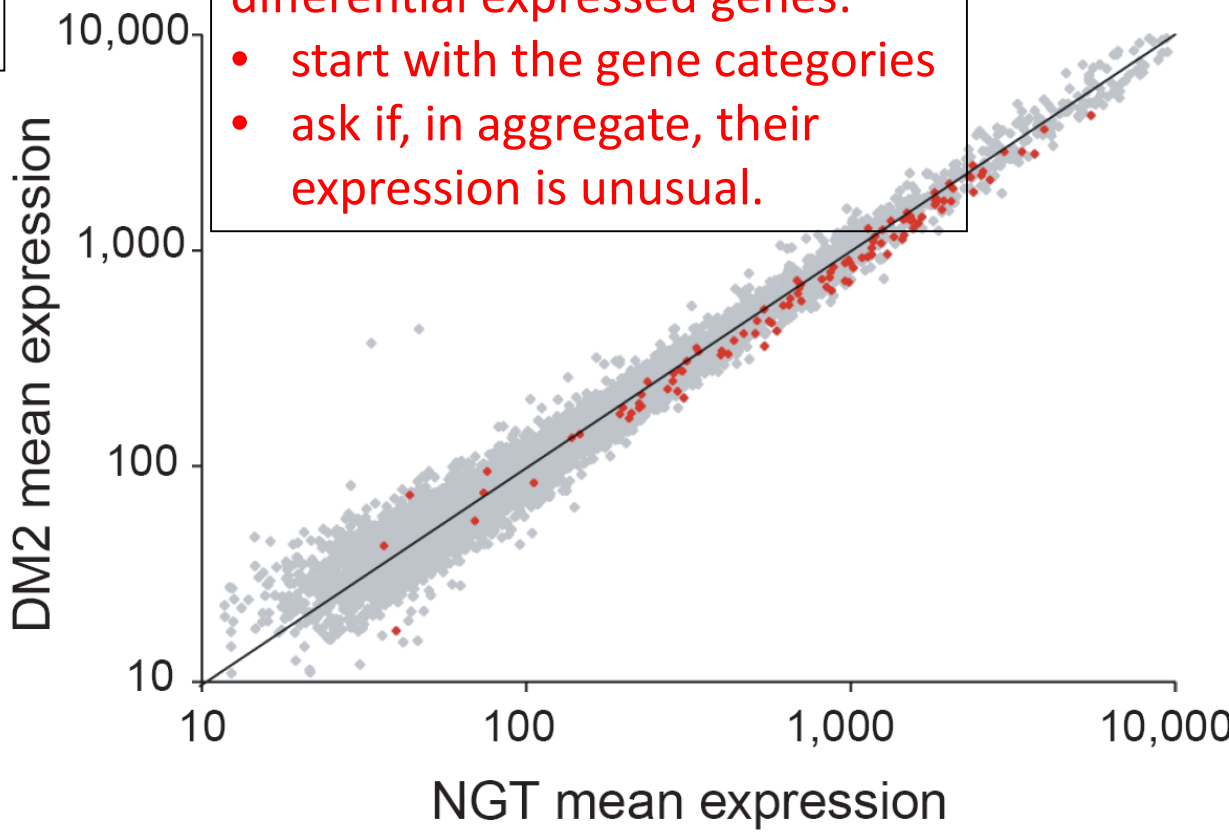
Aggregate score statistics

Hypergeometric results depend on how we define "differentially expressed"



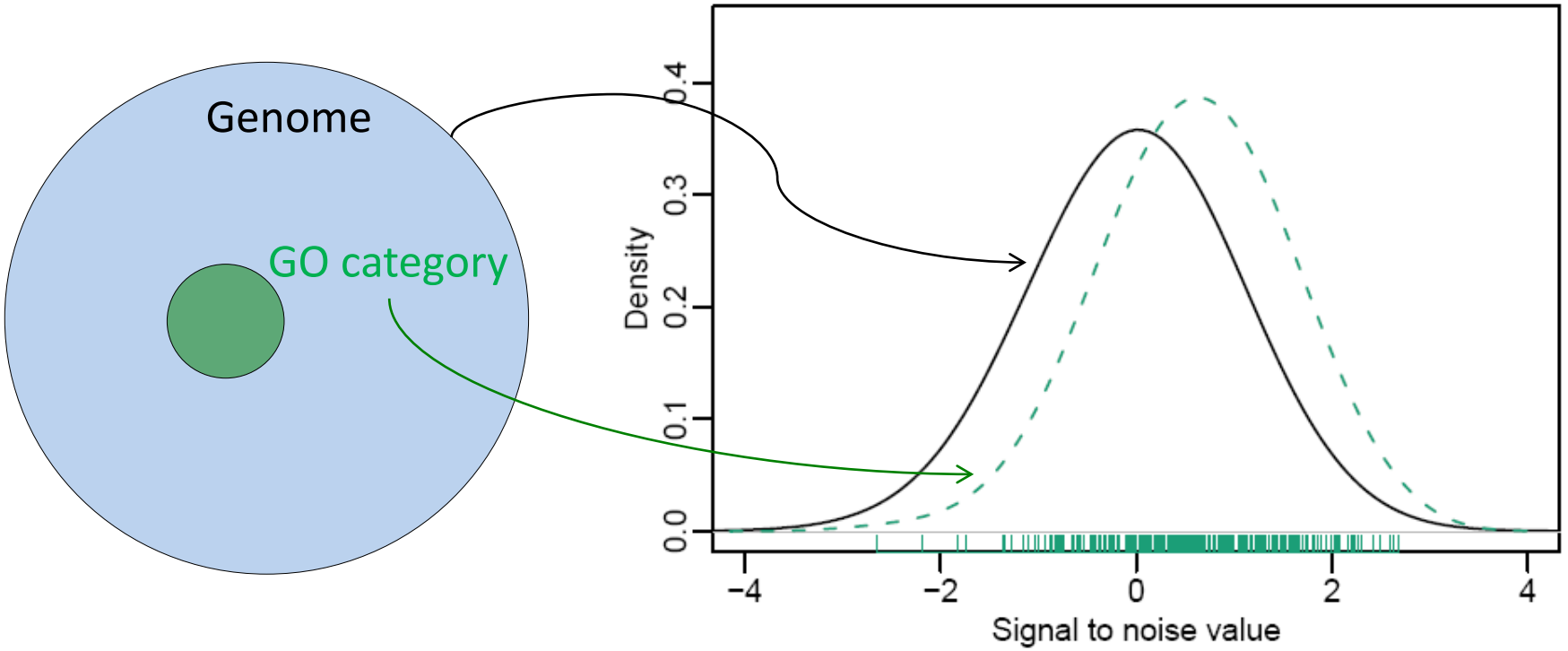
Instead of starting with differential expressed genes:

- start with the gene categories
- ask if, in aggregate, their expression is unusual.



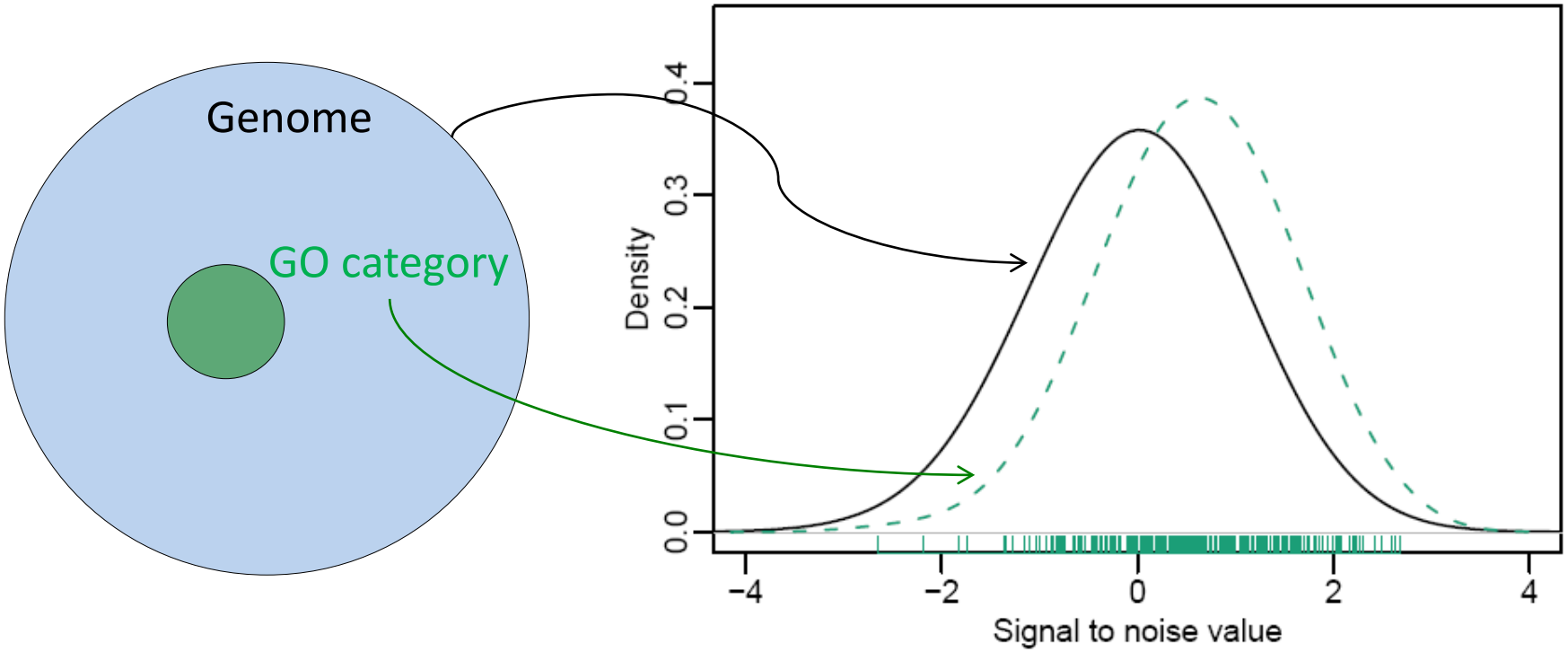
Mootha *et al.* (2003). *Nature Genetics* **34**, 267 – 273. doi:10.1038/ng1180

Aggregate score statistics



GSEA uses a Kolmogorov-Smirnov statistic to compare the distributions of t-statistics

Aggregate score statistics

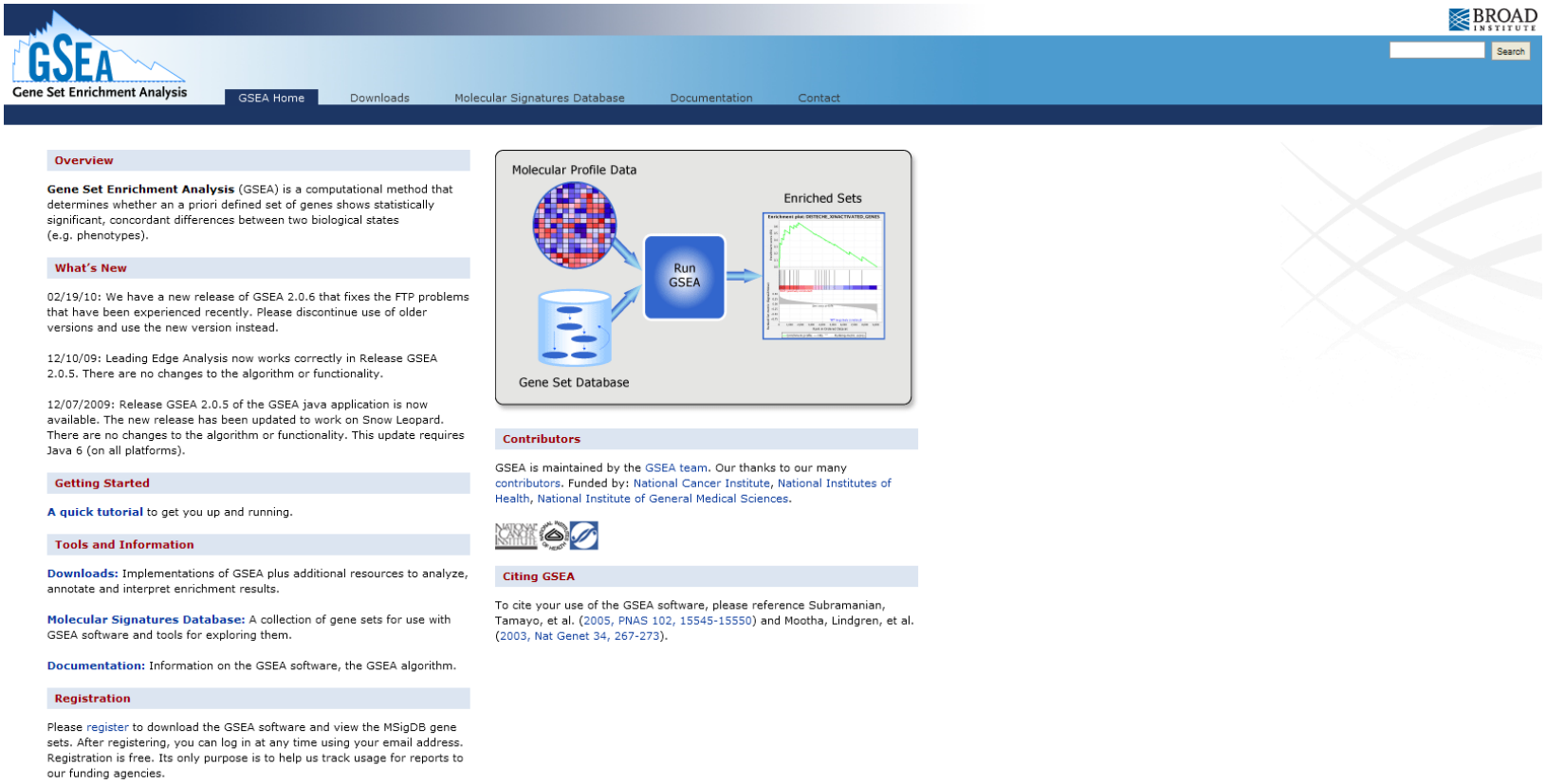


Irizarry, et al. argue for X^2 and z-test

Gene set enrichment analysis made simple. (2009) Stat Methods Med Res
http://www.bepress.com/jhubiostat/paper185/

Aggregate score statistics

http://www.broadinstitute.org/gsea/



Outline

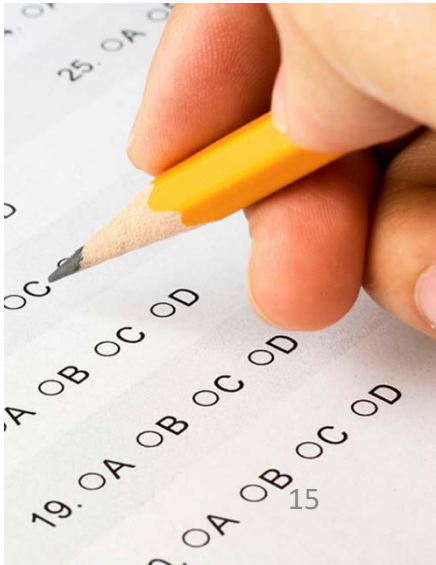
- Evaluating the statistical significance of an annotation
  - Hypergeometric distribution:
    - The null hypothesis:
  - Aggregate score statistics
  - Multiple hypotheses
  - Healthy dose of skepticism
- Applications to analysis of gene expression:
  - Consequences: Function of differentially expressed genes
  - Causes: Identity of transcriptional regulators
    - Known binding sites
    - Predicted binding sites

Testing Multiple Hypotheses

- Example:
- Filter GO terms using a $p<0.01$
- Assume there are 30,000 GO terms
- How many GO terms will look significant by chance?

Testing Multiple Hypotheses

- Example: Filter GO terms using a $p<0.01$
- By definition, the null-hypothesis has a 1% probability of being correct ***for each test.***
- There are roughly 30,000 terms in GO.
- At this level, we expect roughly 300 false positives!



Multiple Hypotheses

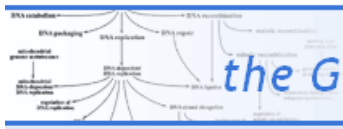
- A simple solution: require that the p-value be small enough to reduce the false positives to the desired level.
- This is called the Bonferroni correction.
- In our case, we would only accept terms with a

$$p \leq \frac{0.01}{30{,}000} = \frac{desired\ threshold}{number\ of\ tests}$$

- Since our tests are not all independent, this is very conservative, and will miss many true positives
- More sophisticated approaches exist, such as controlling the “false discovery rate”.

Outline

- Evaluating the statistical significance of an annotation
  - Hypergeometric distribution:
    - The null hypothesis:
  - Aggregate score statistics
  - Multiple hypotheses
  - Healthy dose of skepticism
- Applications to analysis of gene expression:
  - Consequences: Function of differentially expressed genes
  - Causes: Identity of transcriptional regulators
    - Known binding sites
    - Predicted binding sites



Estrogen receptor

Search
gene or protein name
go!

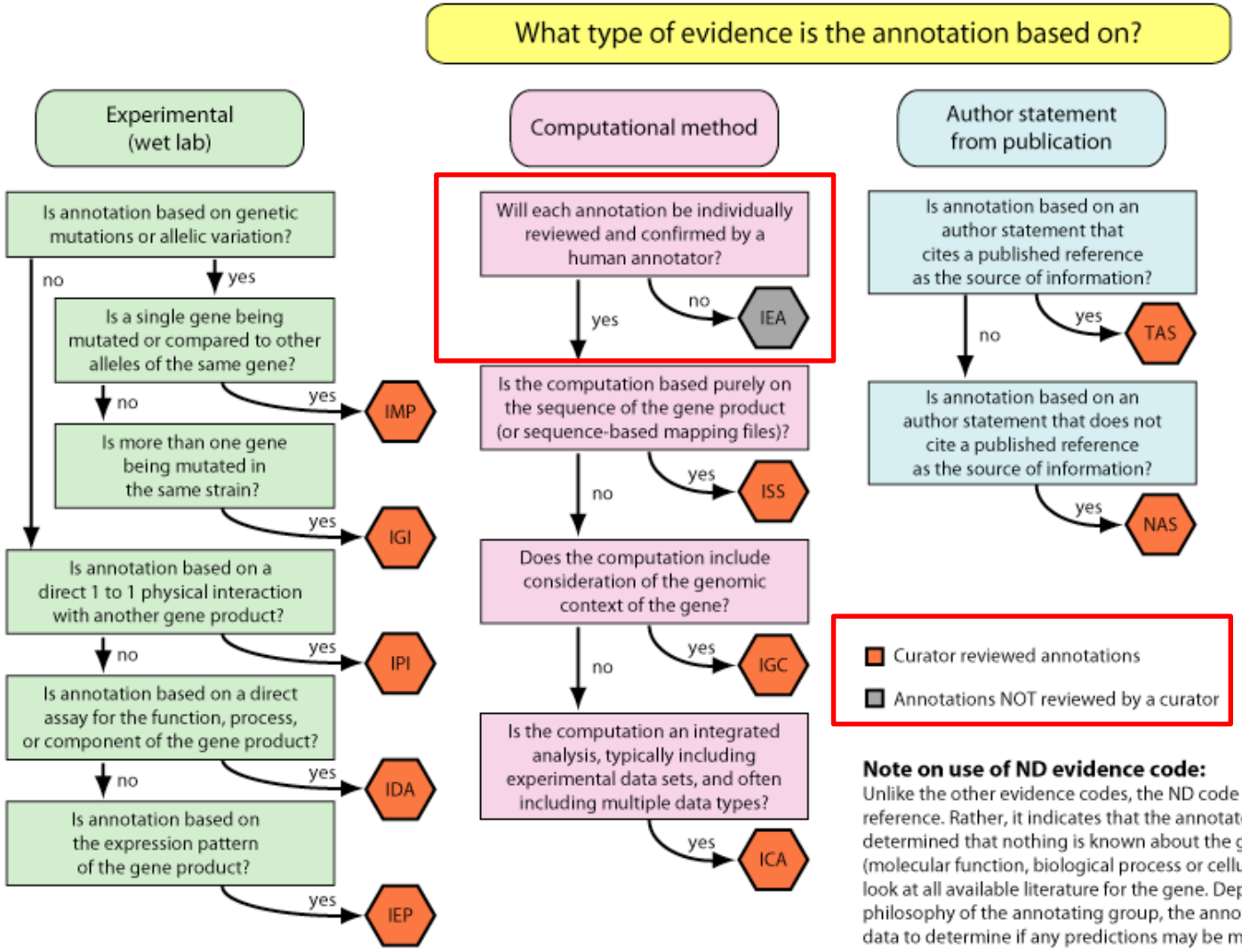
Downloads Tools Documentation Projects About Contact

Select all | Clear all | Perform an action with this page's selected terms... | Go!

Accession, Term		Ontology	Qualifier	Evidence
GO:0030520 : estrogen receptor signaling pathway	41 gene products view in tree	biological process		NAS
GO:0043526 neuroprotection	67 gene products view in tree	biological process		IEA With Ensembl:ENSRNOP00000026350
GO:0048386 : positive regulation of retinoic acid receptor signaling pathway	9 gene products view in tree	biological process		IDA
GO:0045885 : positive regulation of survival gene product expression	56 gene products view in tree	biological process		IEA With Ensembl:ENSRNOP00000026350
GO:0006355 : regulation of transcription, DNA-dependent	16904 gene products view in tree	biological process		NAS
GO:0043627 : response to estrogen stimulus	354 gene products view in tree	biological process		IEA With Ensembl:ENSRNOP00000026350
GO:0007165 : signal transduction	18490 gene products view in tree	biological process		TAS
				TAS

Not just the obvious categories

GO Evidence Code Decision Tree

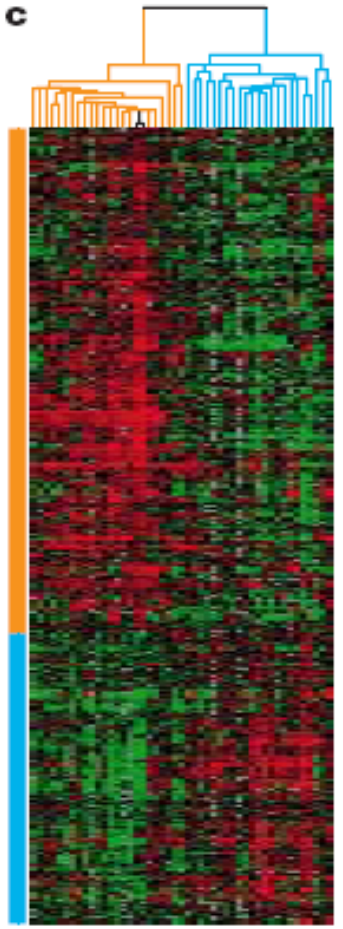


Note on use of ND evidence code:
Unlike the other evidence codes, the ND code does not indicate a method from a specific reference. Rather, it indicates that the annotator looked at the available information and determined that nothing is known about the gene product for a given aspect of GO (molecular function, biological process or cellular component). The annotator will always look at all available literature for the gene. Depending on the resources and the annotation philosophy of the annotating group, the annotator may also look at sequence comparison data to determine if any predictions may be made based on the sequence.

Outline

- Evaluating the statistical significance of an annotation
  - Hypergeometric distribution:
    - The null hypothesis:
  - Aggregate score statistics
  - Multiple hypotheses
  - Healthy dose of skepticism
- Applications to analysis of gene expression:
  - Consequences: Function of differentially expressed genes
  - Causes: Identity of transcriptional regulators
    - Known binding sites
    - Predicted binding sites

Two types of questions we might ask about expression data:



Consequences

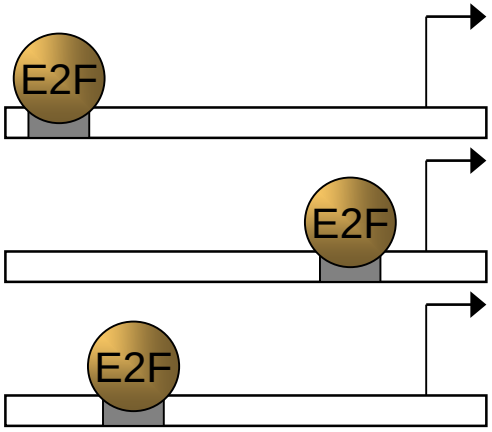
What are the biological consequences of the expression changes?

What categories of genes change in expression?

Causes

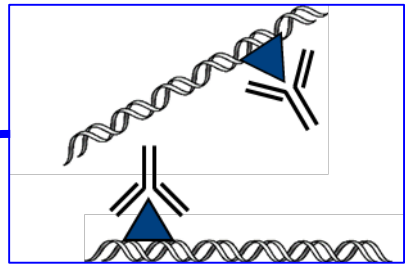
What causes these genes to change in expression?

Does a common transcription factor regulate them?

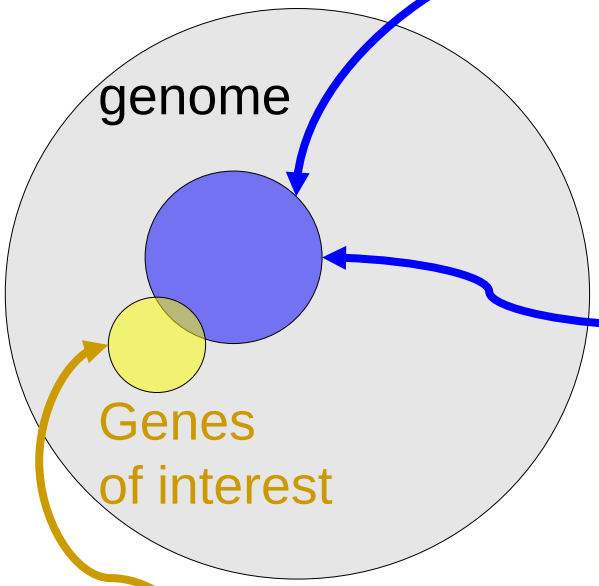


Sources of evidence for regulators

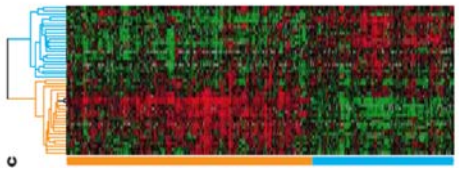
We can apply the same statistical tests to both sources of binding sites:



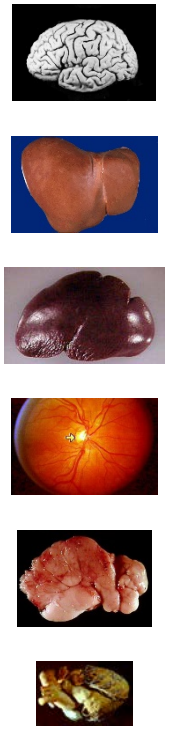
Experiments like ChIP-Seq tell us about the binding of individual proteins in specific experimental conditions

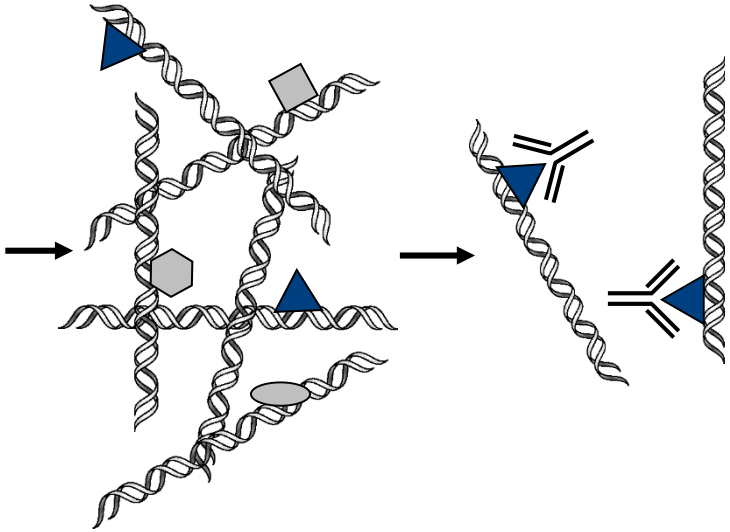


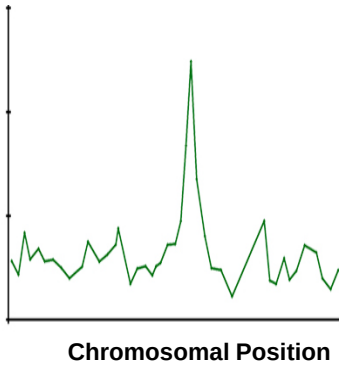
Predictions based on sequence motifs tell us about potential binding in any experimental conditions

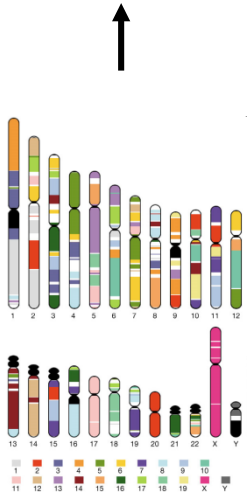


ChIP-Seq measures DNA binding in vivo for one protein of interest









Crosslink protein to binding sites in living cells

Harvest cells and fragment DNA

Enrich for protein-bound DNA fragments with antibodies

Sequence

Align to reference genome

Large databases of ChIP-Seq exist

Table 1.
Comparison of databases that are based on ChIP-seq data

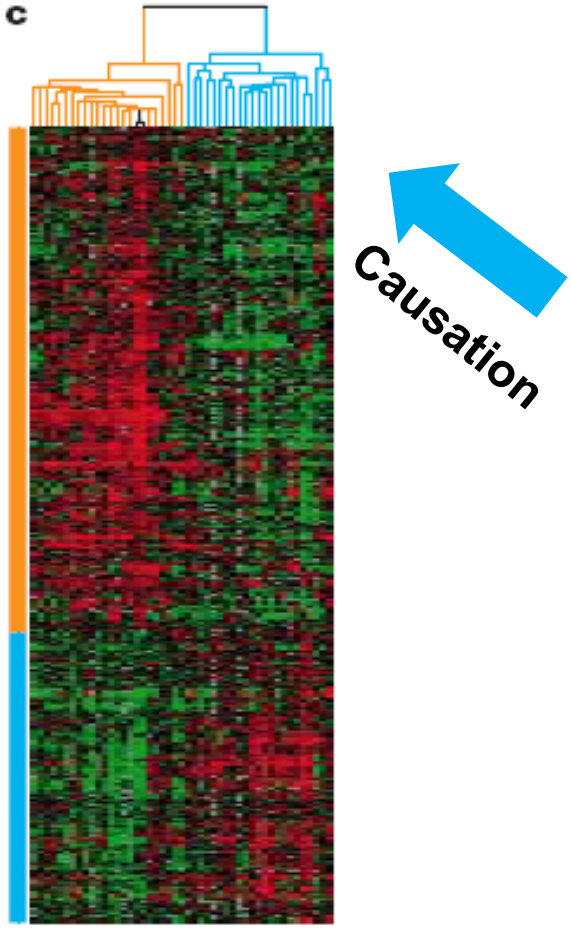
Database, URL	Source of human and mouse data	Number of samples (TF-related)*	Number of TFs
ChIPBase (http://rna.sysu.edu.cn/chipbase)	GEO, ENCODE	total 3549 human 2498 mouse 1036 rat 15	252 TFs and non-TFs for 10 species
Cistrome DB (http://dc2.cistrome.org/#/)	GEO, SRA, ENA, ENCODE	total 10 276 (TF+non-TF) human 5774 mouse 4502 rat 0	260 TFs and non-TFs
ENCODE (https://www.encodeproject.org)	ENCODE	total 1448 human 1254 mouse 194 rat 0	295 TFs and non-TFs for human, 52 TFs and non-TFs for mouse
Factorbook (http://www.factorbook.org)	ENCODE	total 1007 human 837 mouse 170 rat 0	167 TFs, co-factors and chromatin remodeling factors for human, 51—for mouse
GTRD (http://gtrd.biouml.org)	GEO, SRA, ENCODE	total 5078 human 2955 mouse 2107 rat 16	476 human and 257 mouse sequence specific TFs, corresponding to 542 TFClass classes.
ChIP-Atlas (http://chip-atlas.org)	SRA	total 10 774 human 5914 mouse 4860 rat 0	699 human and 502 mouse TFs and others.
GeneProf (http://www.geneprof.org)	SRA, ENCODE, literature	total 1692 human 693 mouse 999 rat 0	133 human and 131 mouse TFs
NGS-QC (http://www.ngs-qc.org)	GEO	total 6672 human 4234 mouse 2438 rat 0	unknown

Table taken from: “GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments”
Ivan Yevshin Ruslan Sharipov Tagir Valeev Alexander Kel Fedor Kolpakov
Nucleic Acids Research, Volume 45, Issue D1, January 2017, Pages D61–D67, https://doi.org/10.1093/nar/gkw951

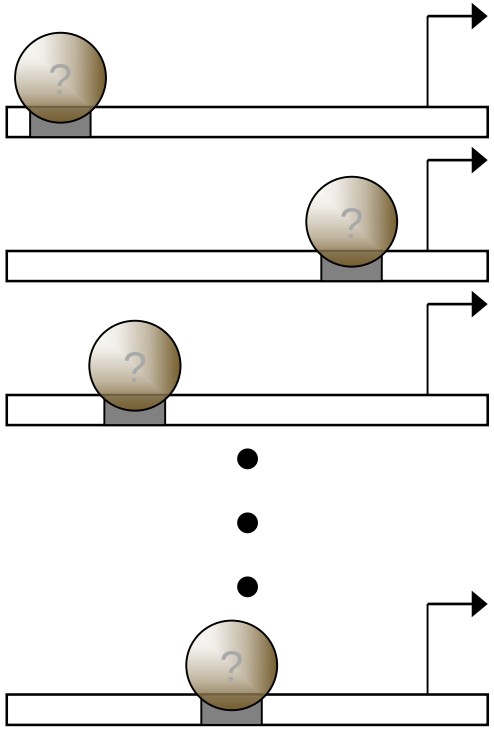
Outline

- Evaluating the statistical significance of an annotation
  - Hypergeometric distribution:
    - The null hypothesis:
  - Aggregate score statistics
  - Multiple hypotheses
  - Healthy dose of skepticism
- Applications to analysis of gene expression:
  - Consequences: Function of differentially expressed genes
  - Causes: Identity of transcriptional regulators
    - Known binding sites
    - Predicted binding sites

Sequence Motifs are Used to Predict Binding



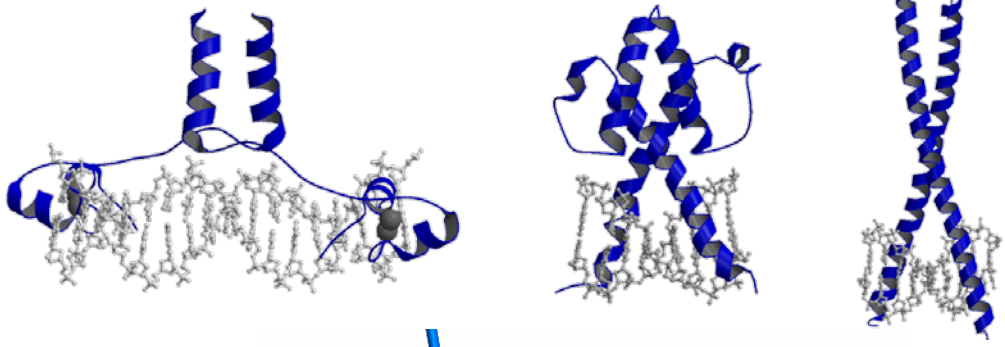
Causation

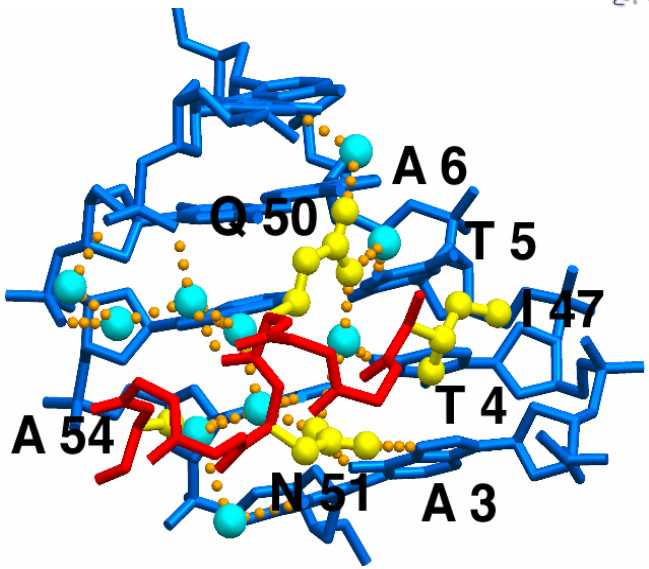


Motifs are quantitative models for the DNA-binding specificity of proteins.

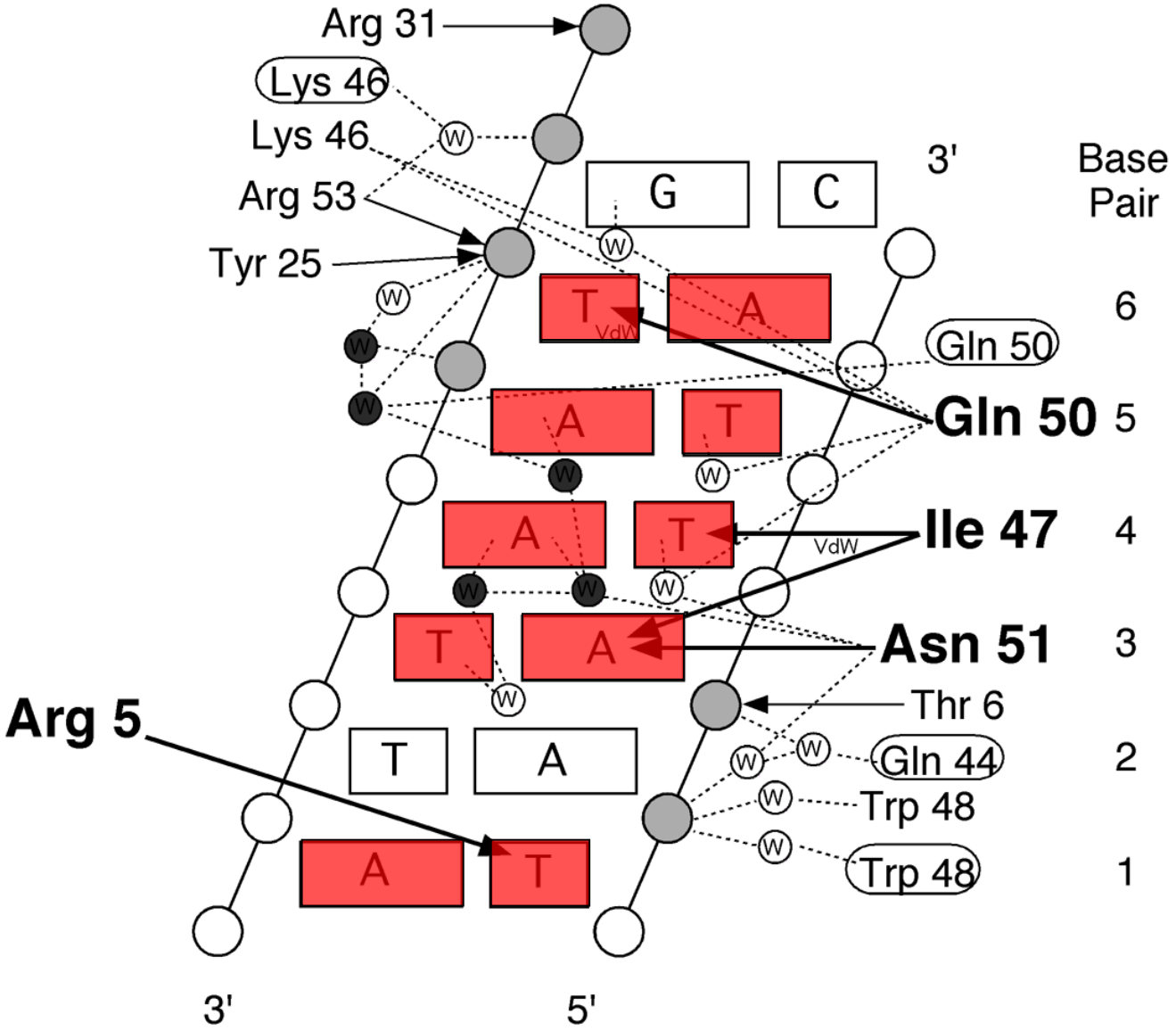
If many of the sequences match a motif, we can hypothesize that the corresponding protein binds under some condition.

Sequence Motifs Represent the Specificity of a Protein





Biophysics determines probability of binding



Some base pairs are more critical than others

Motifs can be derived from known binding sites:

If I had found these sites using ChIP-Seq, how would I describe the specificity?

TGACTCC
TGACTCA
TGACAAA
TGACTCA
TTACACA
TGACTAA
TGACTAA
TGACTCA
TGACTCA
TGACTCA

If I had found these sites using ChIP-Seq, how would I describe the specificity?

TGACTCC
TGACTCA
TGACAAA
TGACTCA
TTACACA
TGACTAA
TGACTAA
TGACTCA
TGACTCA
TGACTCA

Position Frequency Matrix (PFM)

A:	0	0	10	0	2	3	9
C:	0	0	0	10	0	7	1
G:	0	9	0	0	0	0	0
T:	10	1	0	0	8	0	0

If I had found these sites using ChIP-Seq, how would I describe the specificity?

TGACTCC
TGACTCA
TGACAAA
TGACTCA
TTACACA
TGACTAA
TGACTAA
TGACTCA
TGACTCA
TGACTCA

Position Frequency Matrix (PFM)

A:	0	0	10	0	2	3	9
C:	0	0	0	10	0	7	1
G:	0	9	0	0	0	0	0
T:	10	1	0	0	8	0	0

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

How could I use the PPM to find binding sites?

ACGTAGATCGATCCCCTGATCAAAATCGTGTTGAGCGCGCGTAATATCGCTAGCTAGCAAATTCCGATA

Match?

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

The odds ratio is used to find the most likely binding sites

- The raw probabilities can be very small.
- Say the most preferred base at each of 10 positions has p=0.8
- What is the probability of the best motif?
    - P(best match) = (0.8)^10= 0.1

Is P=0.1 good or bad?

The odds ratio is used to find the most likely binding sites

- The raw probability is very hard to interpret.
- A better question: is it more likely that this sequence is a motif match or not?
- What is the prob of any sequence in a random genome?
    - P(random)=(0.25)^10= 9.5367e-7
- The ratio of these two probabilities is called an

$$\text{odds ratio} = \frac{Model\_prob}{Background\_prob} \sim 10^5$$

The odds ratio is used to find the most likely binding sites

Odds ratio

$$\frac{Model\_prob}{Background\_prob} = \prod_{i=1}^{w} \frac{p_{model}(b,i)}{p_{background}(b)} = \prod_{i=1}^{w} odds(b,i)$$

The odds ratio quantitatively compares two hypotheses.

If the odds ratio is above an arbitrary threshold, we consider it a match

Usually each base is modeled as being independent of the others

Is a region a valid binding site?

- Steps:
  1. Define a mathematical model for matching sequences

$$Model\_prob = \prod_{i=1}^{w} p_{\text{model}}(b,i)$$

Position Probability Matrix (PPM)

A:	0.000	0.000	1.000	0.000	0.200	0.300	0.900
C:	0.000	0.000	0.000	1.000	0.000	0.700	0.100
G:	0.000	0.900	0.000	0.000	0.000	0.000	0.000
T:	1.000	0.100	0.000	0.000	0.800	0.000	0.000

Is a region a valid binding site?

- Steps:
  1. Define a mathematical model for matching sequences

$$Model\_prob = \prod_{i=1}^{w} p_{\text{model}}(b,i)$$

Position Probability Matrix (PPM)

```
A:  0.000 |  0.000 |  1.000 |  0.000 |  0.200 |  0.300 |  0.900 |
C:  0.000 |  0.000 |  0.000 |  1.000 |  0.000 |  0.700 |  0.100 |
G:  0.000 |  0.900 |  0.000 |  0.000 |  0.000 |  0.000 |  0.000 |
T:  1.000 |  0.100 |  0.000 |  0.000 |  0.800 |  0.000 |  0.000 |
```

  2. Define a model for sequences that don't match: $P_{background} = 0.25$

Is the sequence more probably a motif or a random genomic region?

- Steps:
  3. Quantitatively compare the two hypotheses

$$Model\_prob = \prod_{i=1}^{w} p_{\text{model}}(b,i)$$

$$Background\_prob = \prod_{i=1}^{w} p_{background}(b)$$

Odds ratio

$$\frac{Model\_prob}{Background\_prob} = \prod_{i=1}^{w} \frac{p_{model}(b,i)}{p_{background}(b)} = \prod_{i=1}^{w} odds(b,i)$$

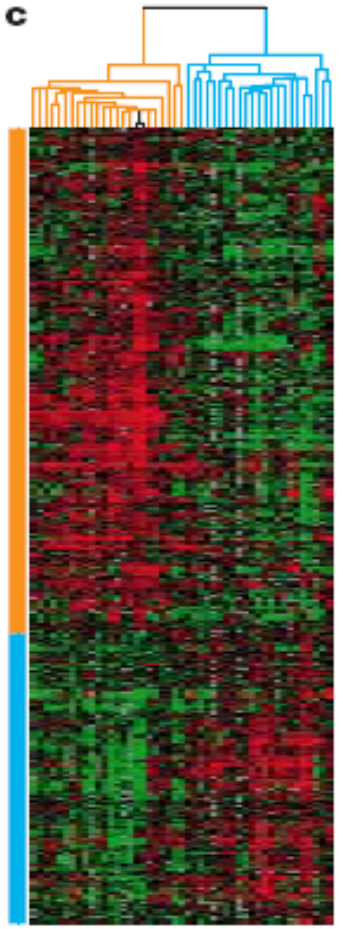
1.Define motif model | Define background model | Compare the models

Motifs are usually represented as the log-odds

$$log\left[\frac{P_{model}}{P_{background}}\right] = log[P_{model}] - log[P_{background}]$$

- The log-odds matrix is often called a:
  PWM position weight matrix or
  PSSM position-specific scoring matrix
- Taking the log helps avoid problems that computers have with very small numbers
- **Rule-of-thumb**: *60% of the maximum-possible* LLR score is a reasonable threshold for determining a match to a *PWM motif*

You now have tools to address both types of questions:



Consequences

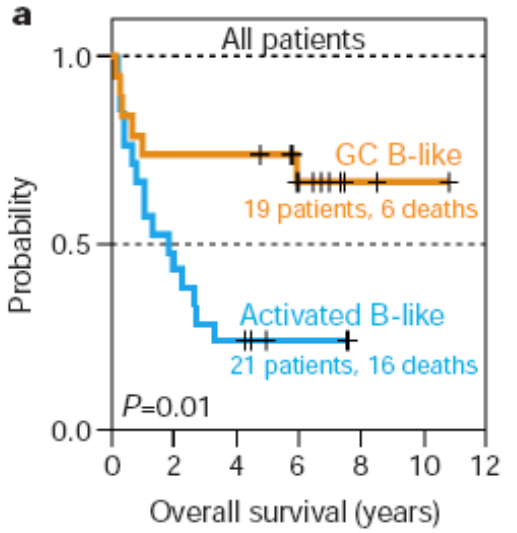
What are the biological consequences of the expression changes?

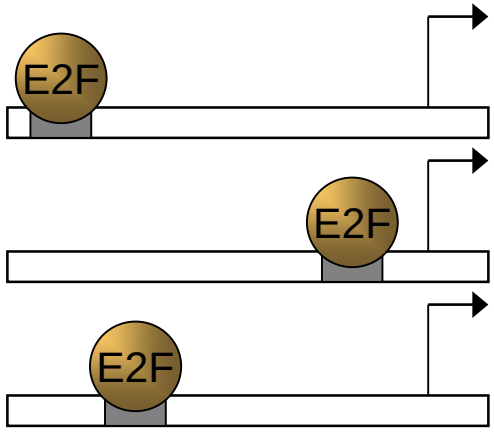
What categories of genes change in expression?

Causes

What causes these genes to change in expression?

Does a common transcription factor regulate them?





Motif Discovery

- **Given**: a set of sequences
- **Find**: the PWM for an over-represented motif

ACGTGTCTGCTACAAAATGCAAATACGATGATAAATGCAGCAATTGT

ACGTAAATGCAATTACGATGATAAATGCAGCAACCGTTATCGACTTG

ATCTTACTAGCATGGCCATCATCAACATGCAAAGCAGGTTGTGCCCT

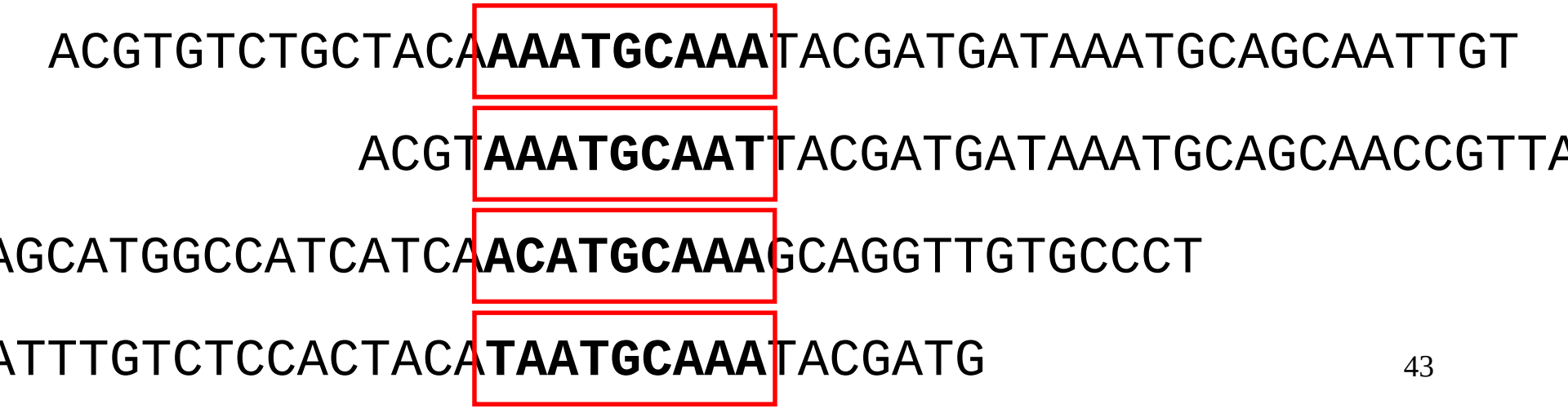
ATAAATGCCCAATTGATTTGTCTCCACTACATAATGCAAATACGATG

Motif Discovery

- **Given**: a set of sequences
- **Find**: the PWM for an over-represented motif

Motif Discovery

- **Note 1:**
  - If you know the PWM, you can easily align the sequences
- **Note 2:**
  - If the sequences are aligned, you can easily find the PWM



The Expectation Maximization (EM) Algorithm

- When we begin
  - we don't know the PWM
  - we don't know the location of the binding sites
- We iteratively:
  - assume we know the motif and look for the most likely binding site
  - assume we know the binding site and compute the best motif

Expectation Maximization

- E step – calculate expected motif locations given the current motif

Given our current best guess about the motif,
Where do we think the protein is binding?

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTGT

ACGT**TCATGTATT**TACGATGATAAATGCAGCAACCGTTATCGACTTG

ATCTTACTAGCATGGCCATCATCA**ACATGATAA**GCAGGTTGTGCCCT

ATAAATGCCCAATTGATTTGTCTCCACTACA**AAATGCAAT**TACGATG

Expectation Maximization

- M step – re-estimate the motif to maximize likelihood

Given our expectation about where binding occurs,
What is the most likely motif model?

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTG

GACATTTTGTACGT**TCATGTATT**TACGATGATAAATGCAGCAACCG

CATGGCCATCATCA**ACATGATAA**GCAGGTTGTGCCCGGTTTACTGA

TTGTCTCCACTACA**AAATGCAAT**TACGATGAGAGGGTGATGGCACT

Expectation Maximization

- M step – re-estimate the motif to maximize likelihood

Old motif

New motif

ACGTGTCTGCTACA**AAATGCAAA**TACGATGATAAATGCAGCAATTG

GACATTTTGTACGT**TCATGTATT**TACGATGATAAATGCAGCAACCG

CATGGCCATCATCA**ACATGATAA**GCAGGTTGTGCCCGGTTTACTGA

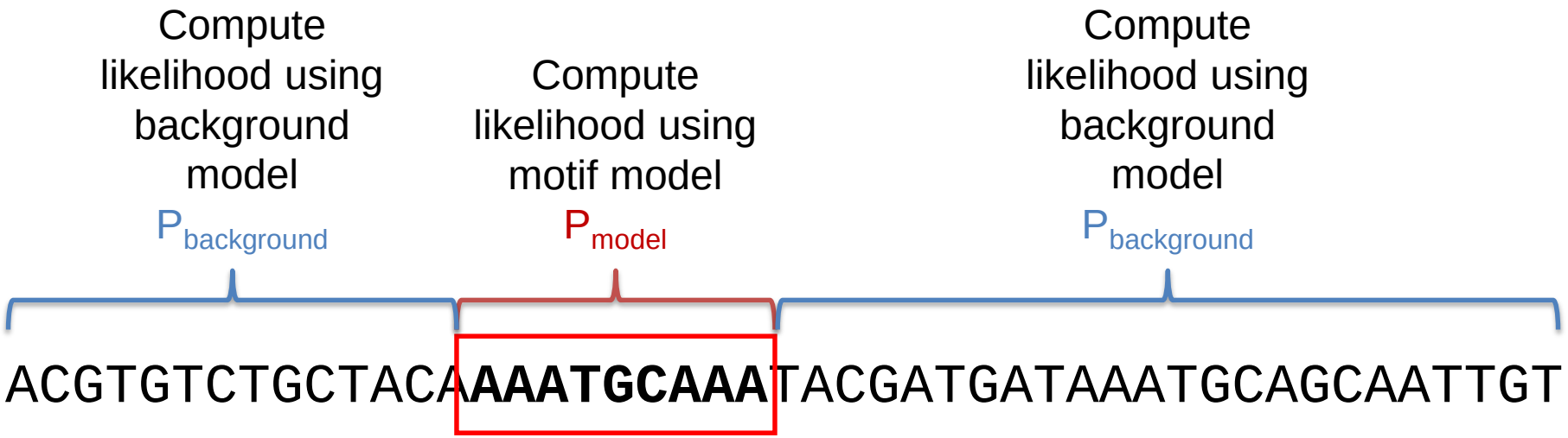
TTGTCTCCACTACA**AAATGCAAT**TACGATGAGAGGGTGATGGCACT

Properties of the EM algorithm

- EM is guaranteed to converge
  - at each step our overall score improves
- EM is not guaranteed to give the right answer
  - had we started with a different initial guess, we might have found a better answer

What do we maximize?

- We maximize the likelihood of the full sequences given our current motif model.



- Remember that each element of the motif is

$$log\left[\frac{P_{model}}{P_{background}}\right] = log[P_{model}] - log[P_{background}]$$