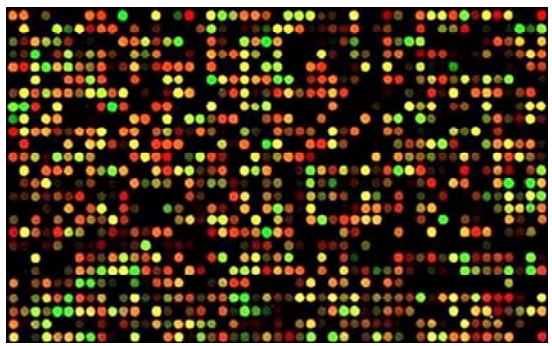
10-810: Advanced Algorithms and Models for Computational Biology

Microarrays

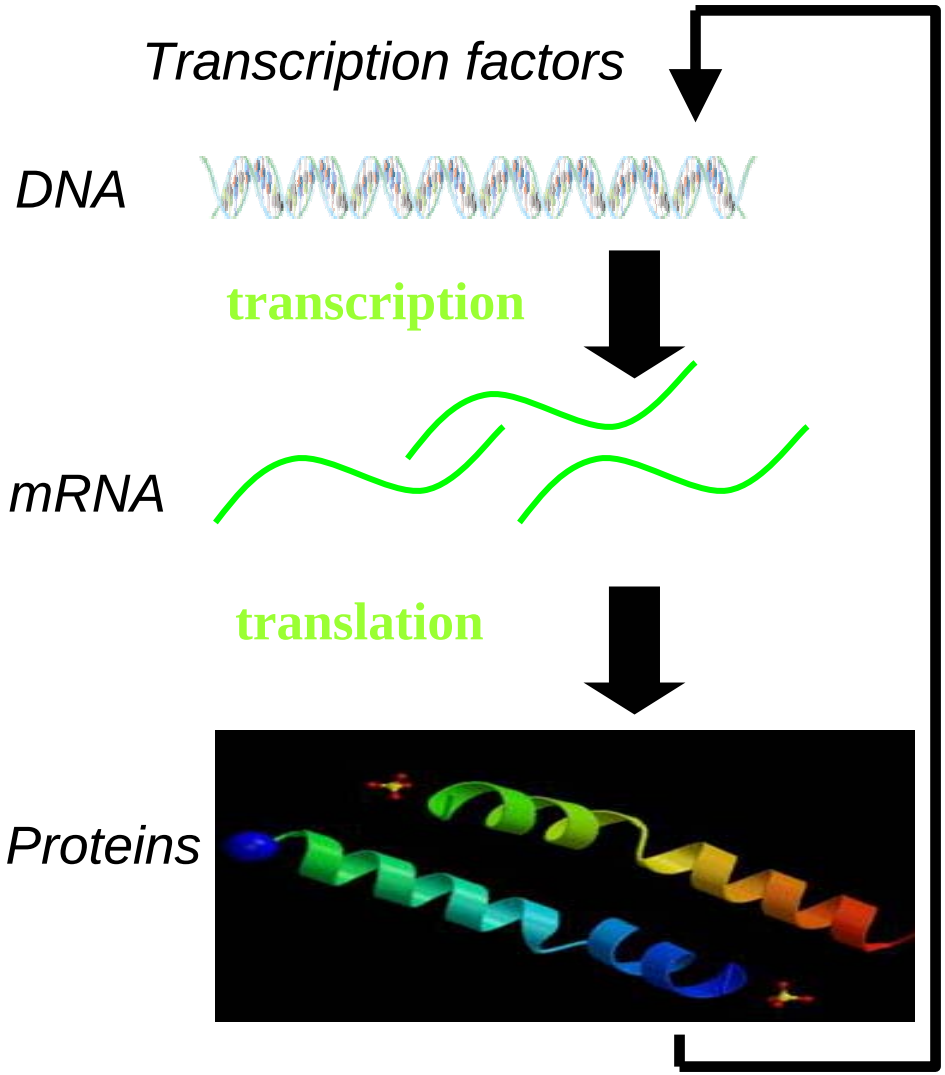


Why sequence is not enough

Identifying genes and control regions is not enough to decipher the inner workings of the cell:

- We need to determine the function of genes.
- We would like to determine which genes are activated in which cells and under which conditions.
- We would like to know the relationships between genes (protein-DNA, protein-protein interactions etc.).
- We would like to model the various dynamic systems in the cell

Central Dogma



Genes and Gene Expression

Technology

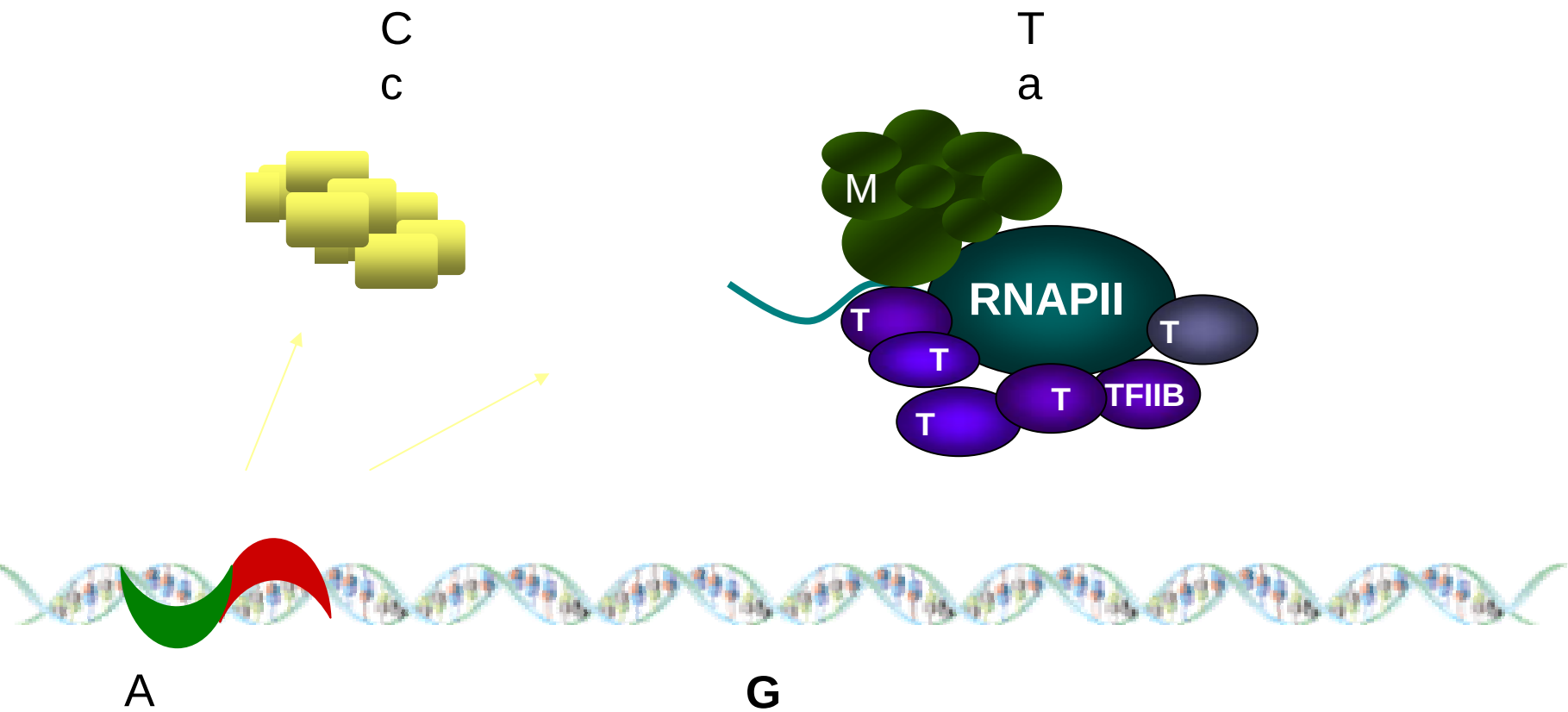
Display of Expression Information

What is a gene?

Promoter Protein coding sequence Terminator

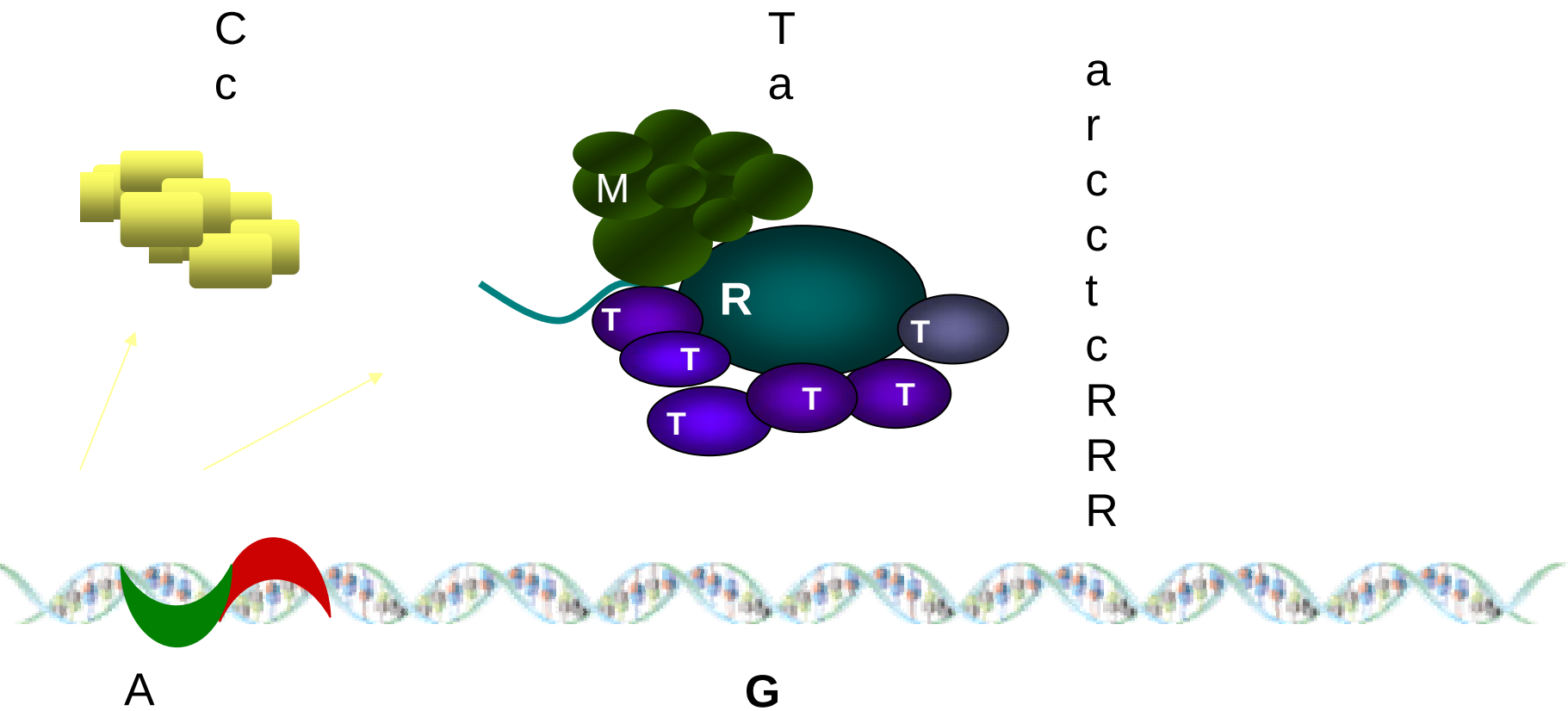
Genomic DNA

How are Genes Regulated?
DNA-binding Activators Are Key To Specific Gene Expression

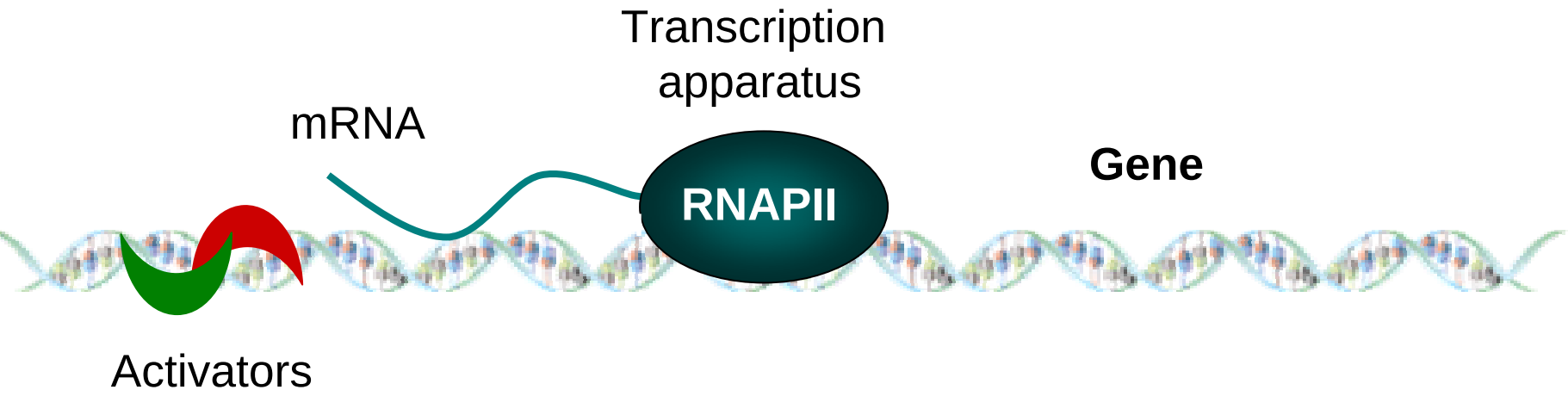


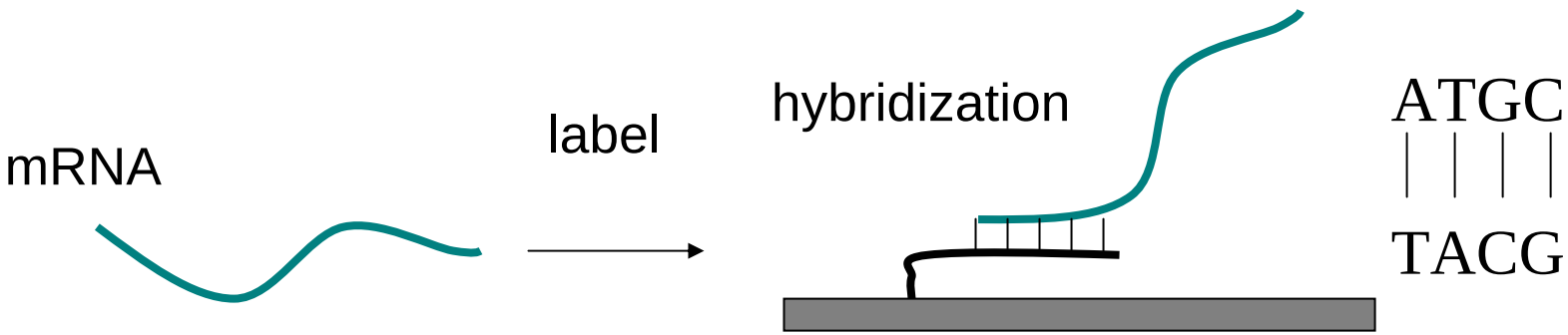
How are Genes Regulated?

DNA-binding activators are key, but there are additional factors



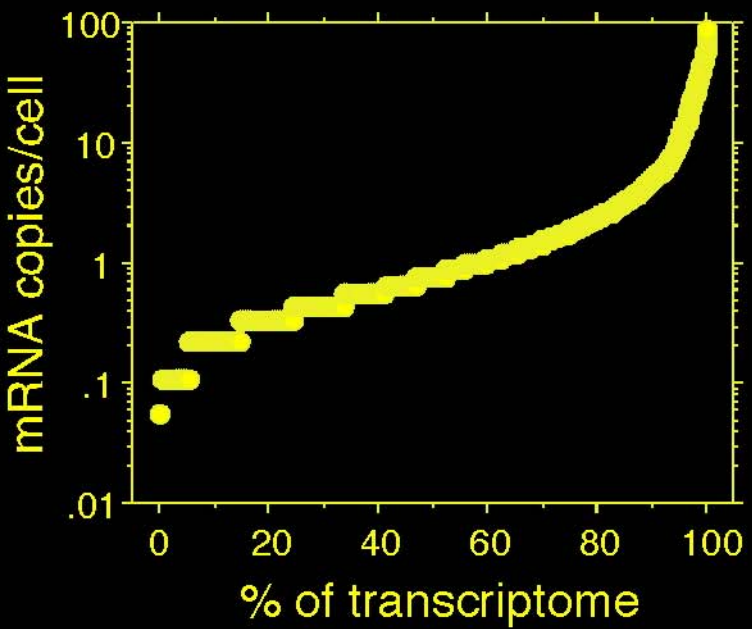
Genome-wide Gene Expression (mRNA) can be Measured with DNA Microarrays





Yeast Transcriptome (Glucose)

5460 mRNA species
average level: 2.8 copies/cell
median level: 0.8 copies/cell

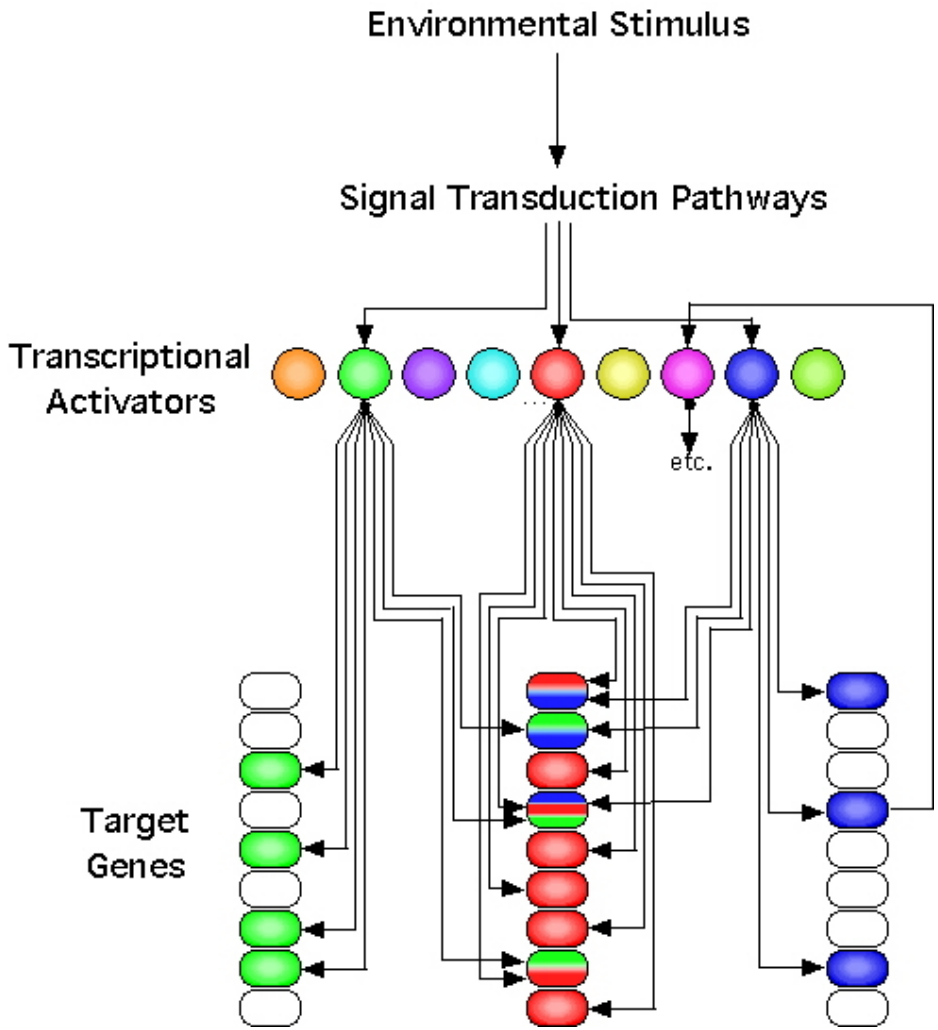


80% of the transcriptome is
expressed at 0.1 - 2 mRNA copies/cell

Some Additional Numbers

Yeast:
6200 genes
~200 transcriptional regulators

Human:
30,000 - 50,000 genes
~1700 transcriptional regulators
>100 cell types



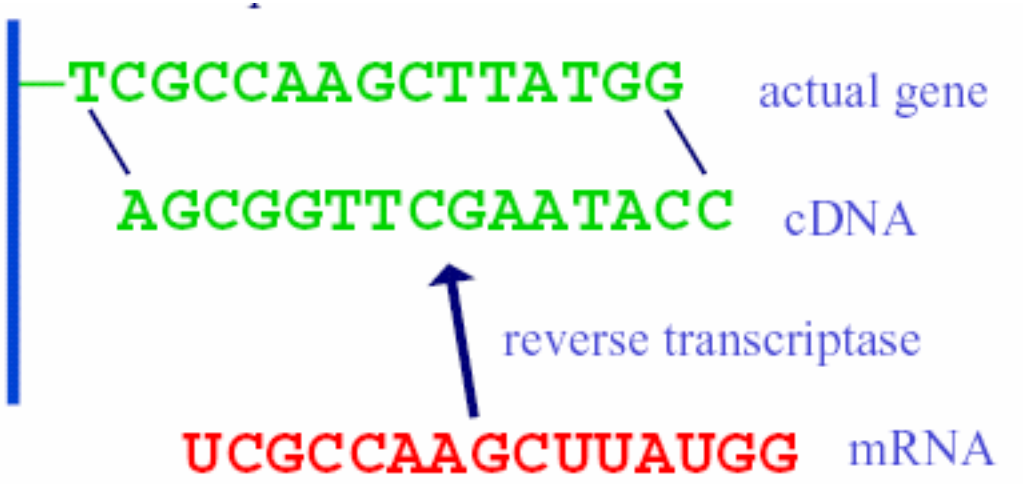
Genes and Gene Expression

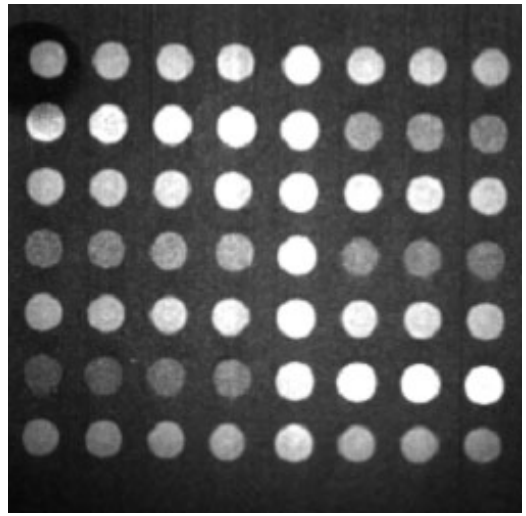
Technology

Display of Expression Information

Microarray Hybridization

- Watson-Crick base pairing of complementary DNA sequences.





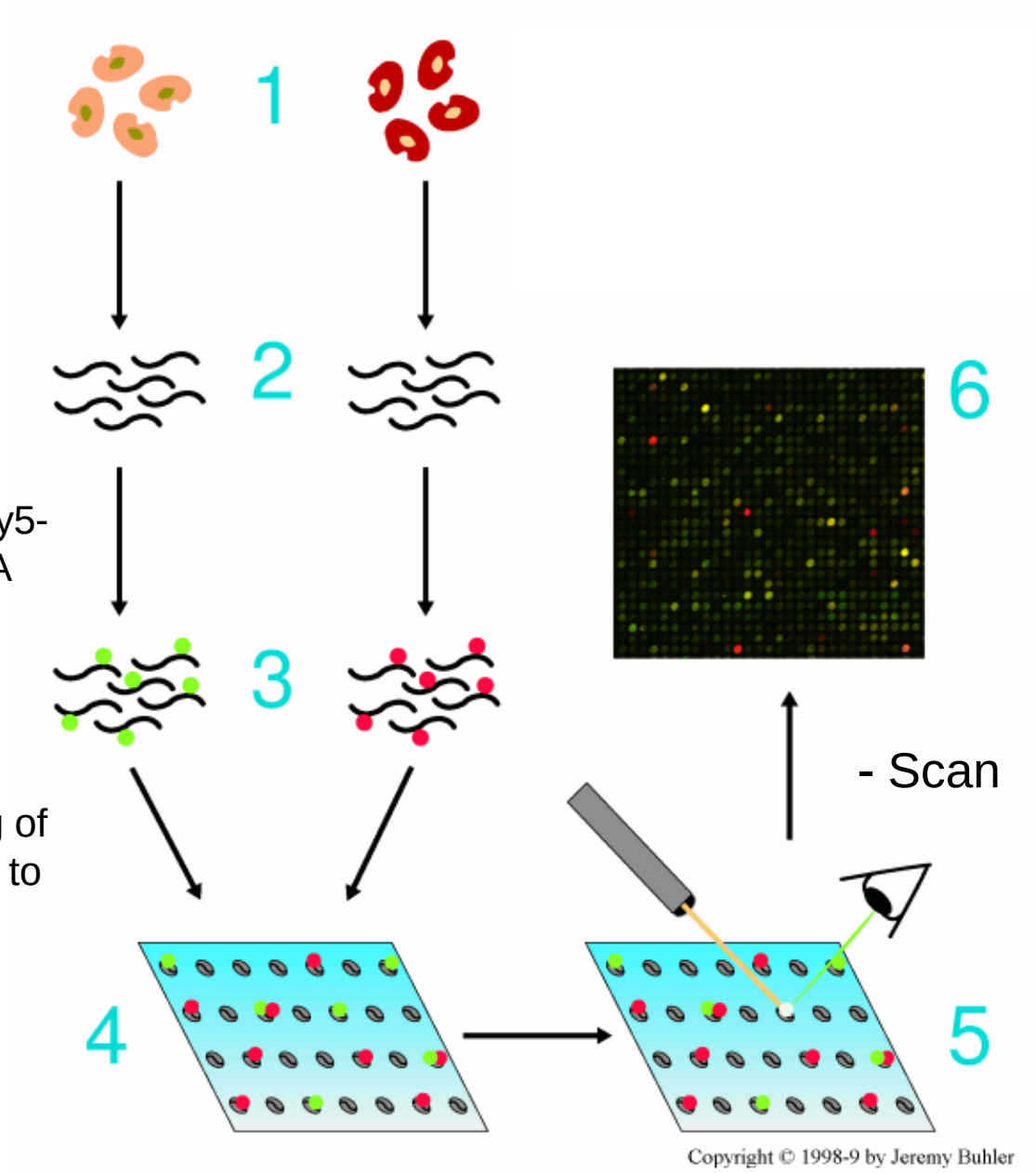
- Microarrays have thousands of spots, each representing a piece of one gene, immobilized on a glass slide.
- The intensity (or intensity ratio) of each spot indicates the amount of labeled cDNA hybridized, thus, representing the starting mRNA transcript abundance.

Two major technologies

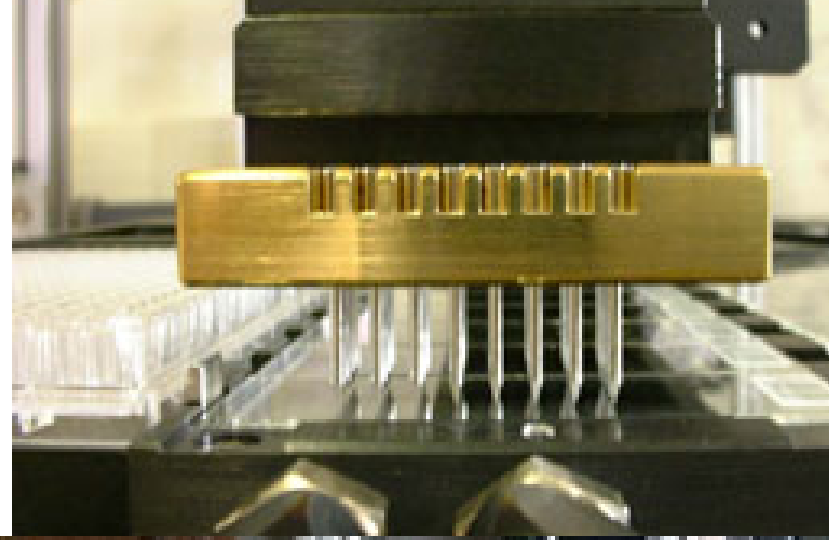
- cDNA arrays
  - probes are placed on the slides
  - allows comparison of different cell types

- Oligonucleotide arrays
  - partial sequences are printed on the array
  - measure values in one tissue type

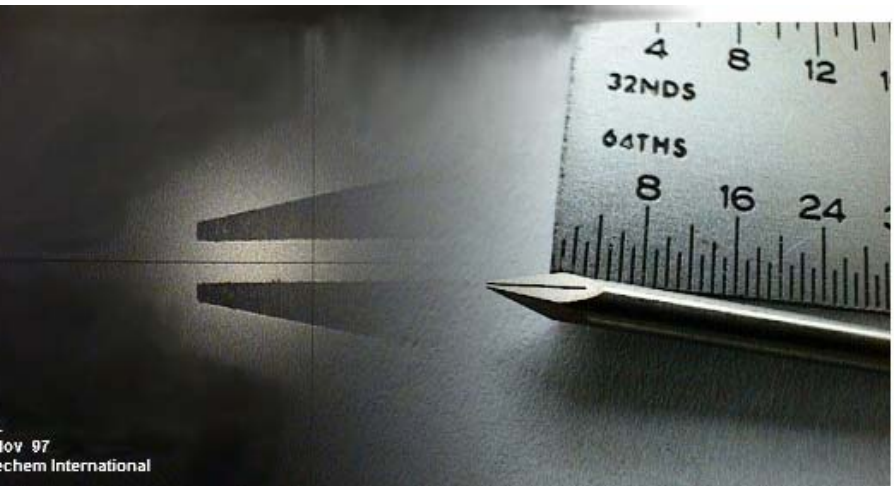
Hybridization and Scanning – cDNA arrays



Cartesian PixSys 5500 with quill printing technology



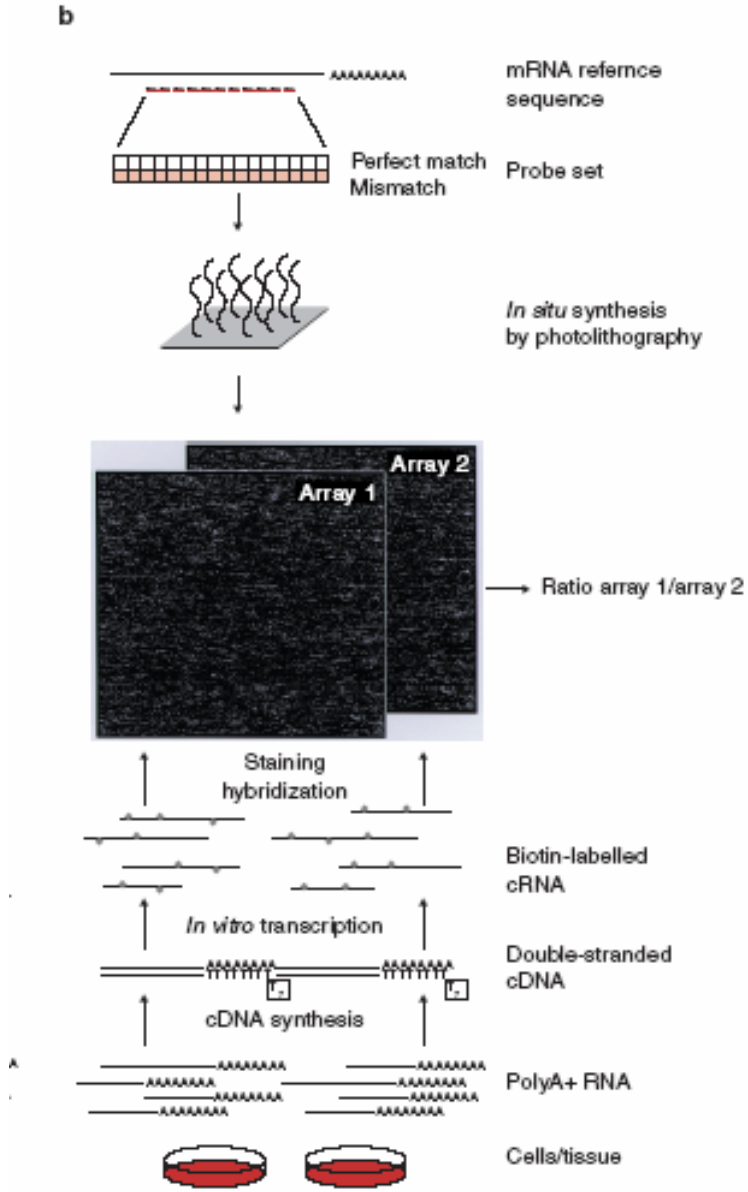
- Complete subsequences are printed on the array
- 10,000 spots/slide
- Spots are 100-200 µm in diameter
- Hybridization volumes: 20-100ul



Array Scanning

Laser based - fluorescent emission

Hybridization
and
Scanning—
oligo arrays



cDNA vs. Oligo: Pros and Cons

cDNA

- Does not require sequence
- Cheap
- Direct comparisons
- Inaccurate
- Cannot measure individual samples

Oligo

- Can be designed to minimize cross hybridization
- Allows for internal control
- Both lead to better accuracy
- expensive
- limited to certain species

Errors

Microarrays introduce many errors which should be taken into account when working with measured expression values:

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

Error types

Microarrays introduce many types of errors which should be taken into account when working with measured expression values:

- Scanning errors additive + multiplicative
- Spotting errors multiplicative
- Cross hybridization multiplicative
- Errors related to day / reading device / experimentalist

additive + multiplicative

- Background differences between slides additive

Handling the Different Errors

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

Analysis of image data (we assume it was performed)

Handling the Different Errors

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

Use ratio instead of individual values:

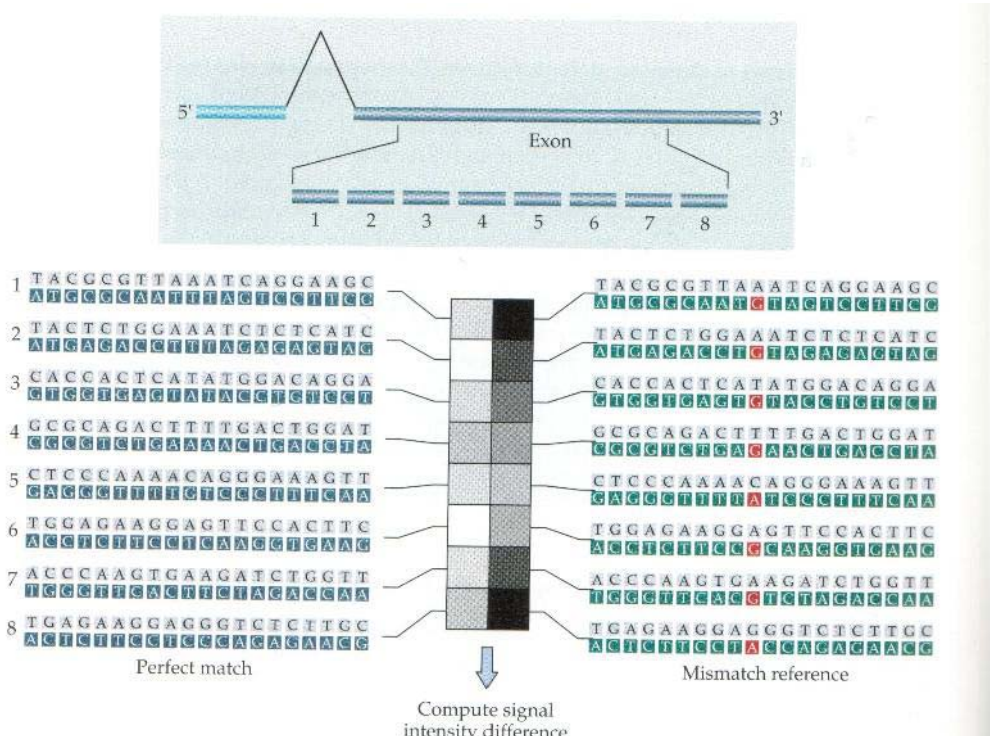
$Y_i = R_i / G_i$

Handling the Different Errors

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

For Oligo arrays, use the match / mismatch spots

Match / Mismatch



- Presence and absent calls can be made using the Match / Mismatch information.

- However, it has been reported that in some cases the mismatch was higher than the match.

Handling the Different Errors

- Scanning errors
- Spotting errors
- Cross hybridization
- Errors related to day / reading device / experimentalist
- Background differences between slides

Normalization (next lecture)

Binding arrays

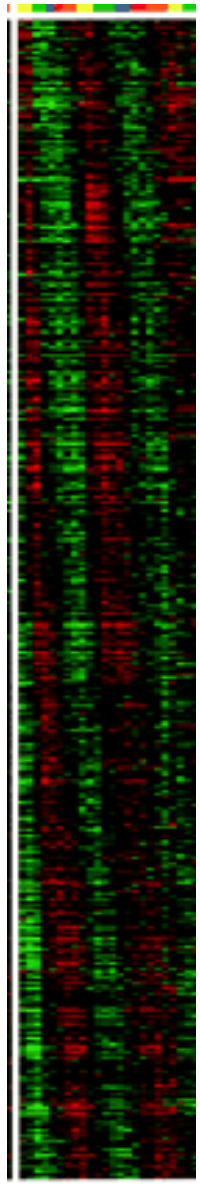
- Instead of printing the genes on the microarray, we can print the intergenic region (an area upstream of the gene).
- We tag a protein of interest (a transcription factor) and fuse all proteins to DNA.
- Next, we hybridize the extracted portions of DNA onto the array, resulting in areas that are bound by the TF being spotted on the microarray.

Genes and Gene Expression Technology

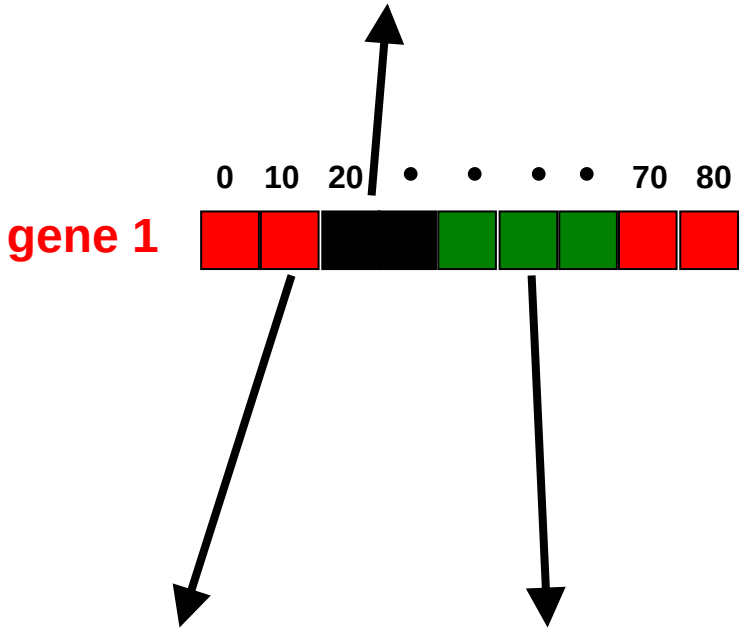
Display of Expression Information

Yeast cell cycle expression program

genes



baseline expression

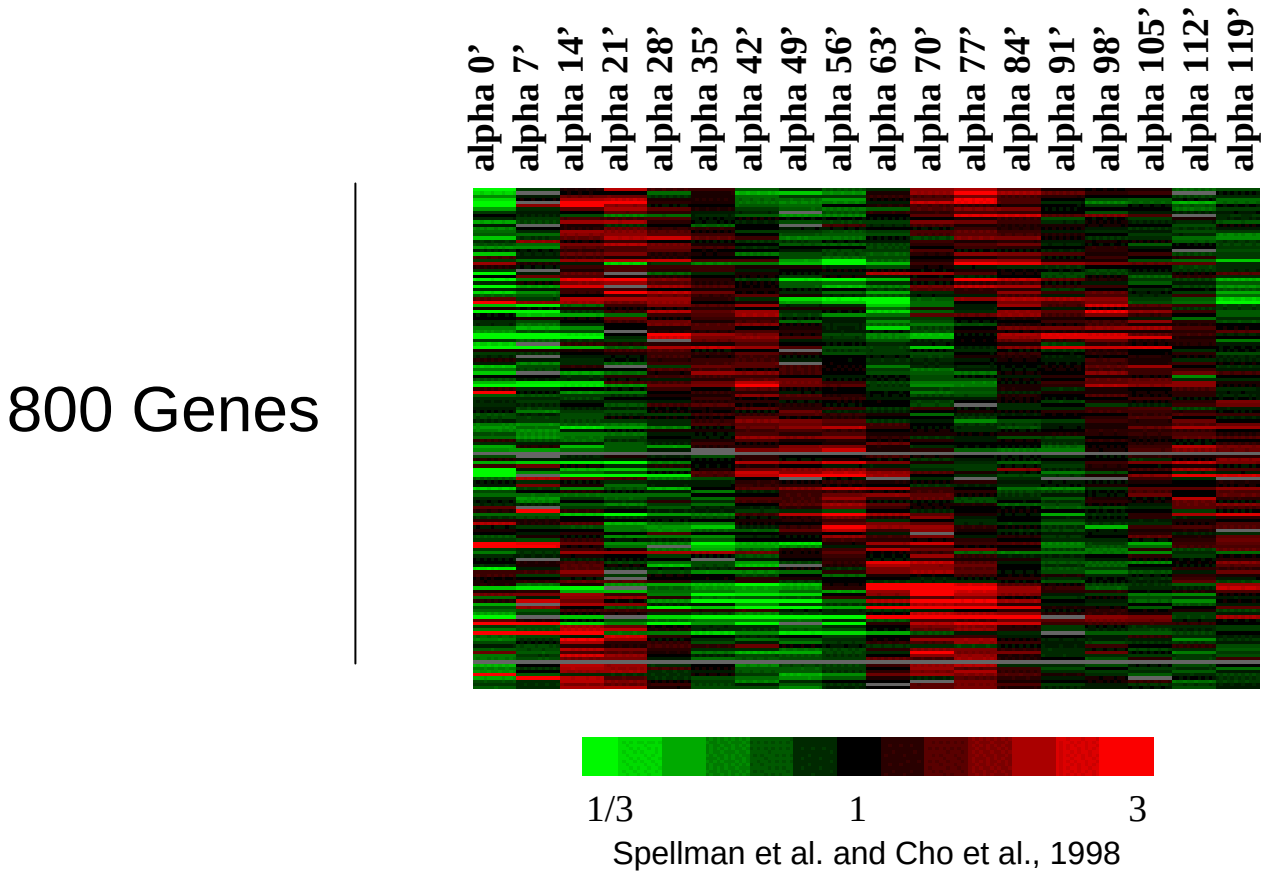


Higher expression compared to baseline

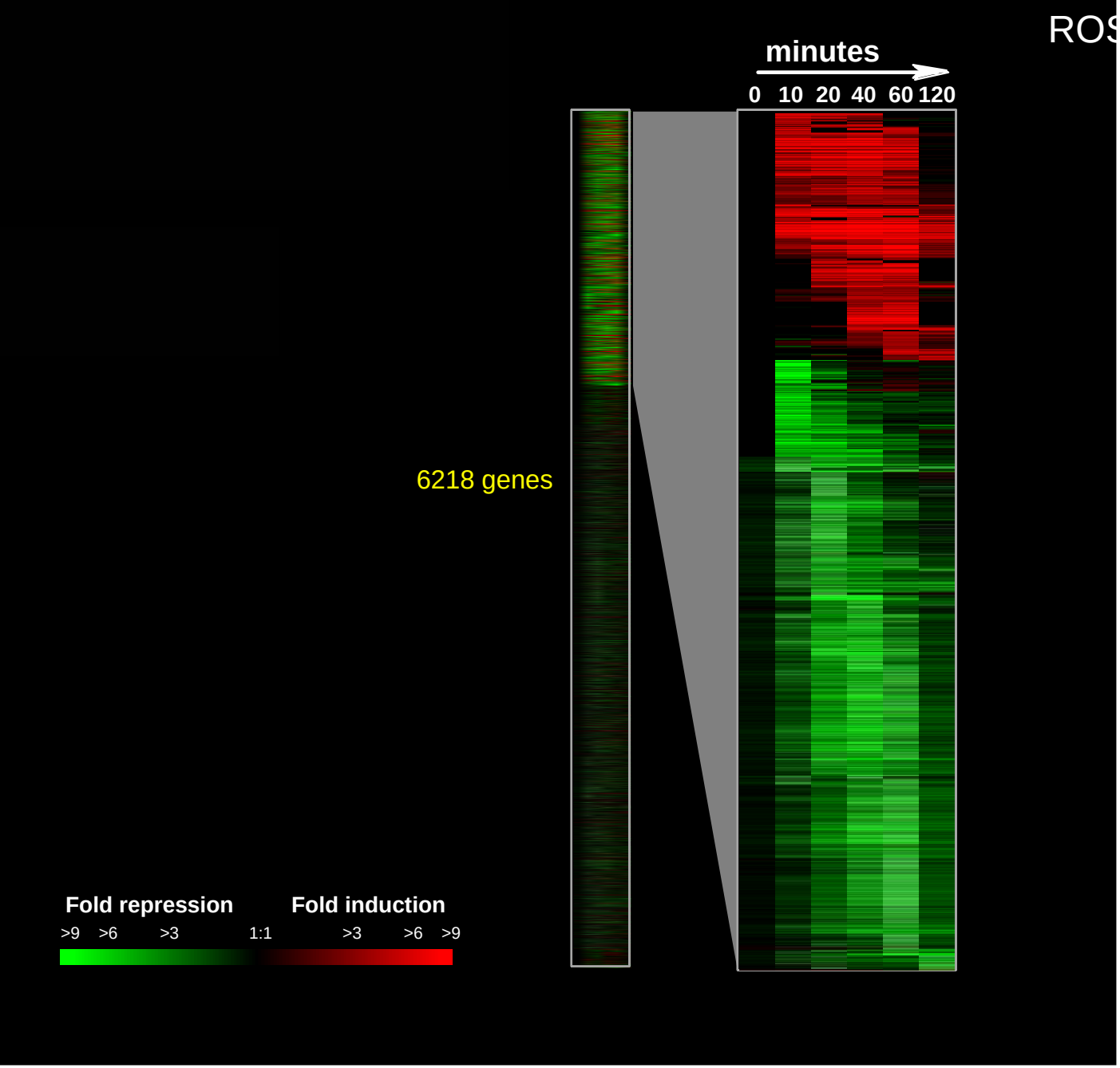
Lower expression compared to baseline

Spellman *et al Mol. Biol. Cell* 1998

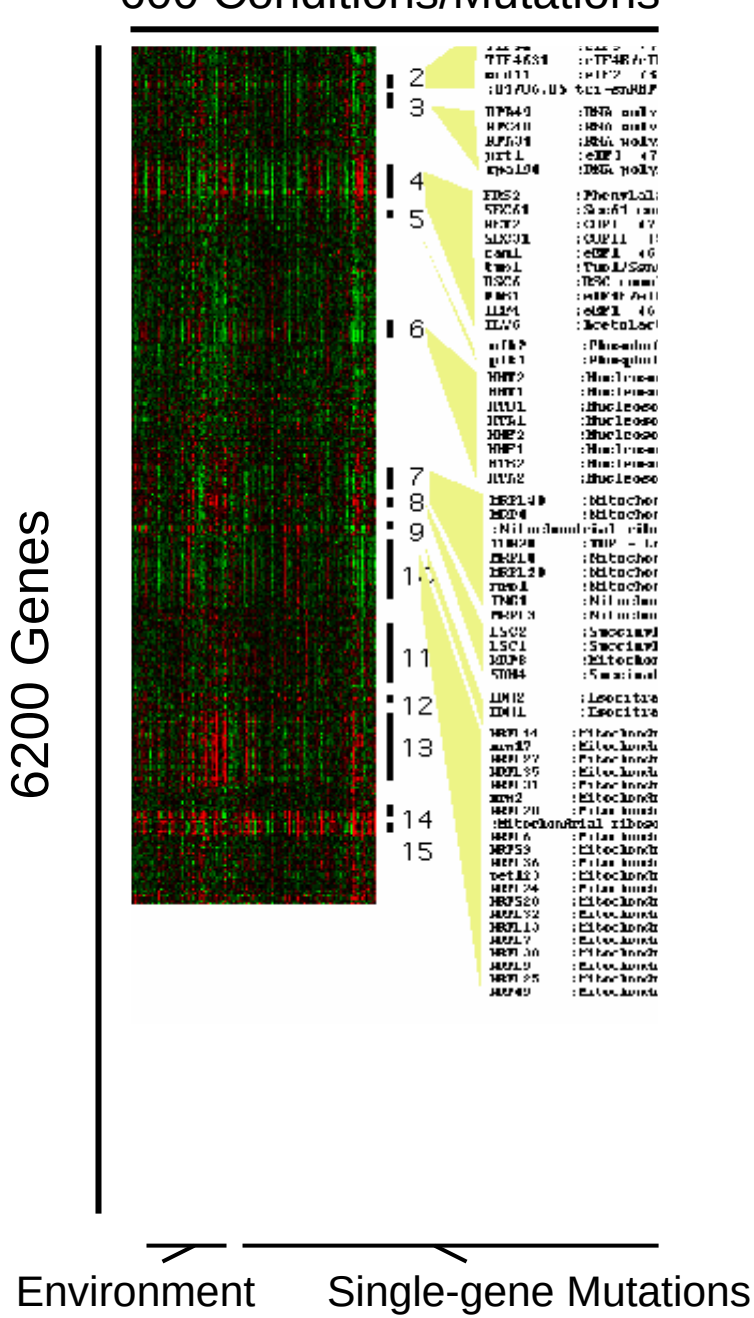
Yeast Cell Cycle Gene Expression Program



Spellman et al. and Cho et al., 1998

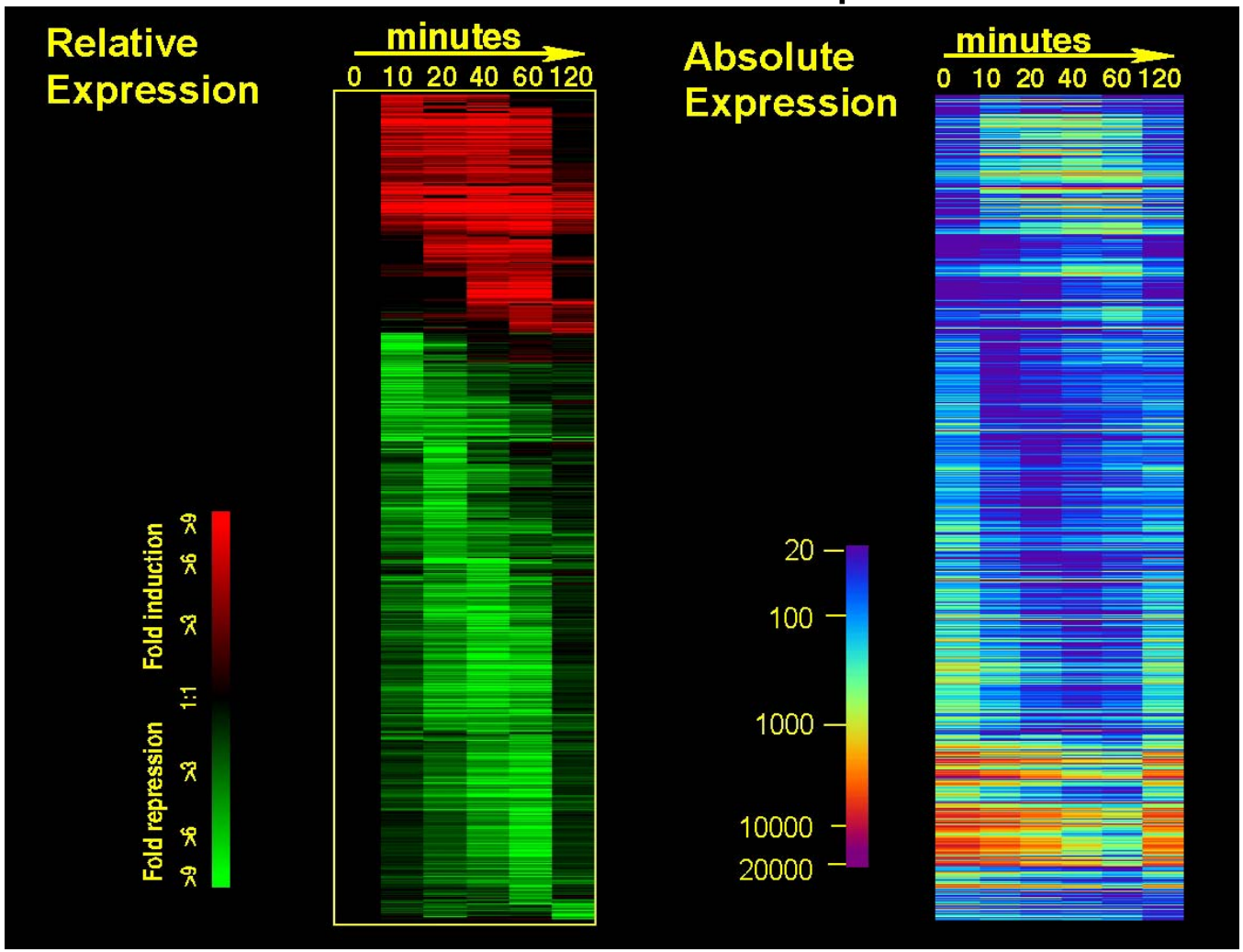

ROS
minutes
0 10 20 40 60 120
6218 genes
Fold repression
Fold induction
>9 >6 >3 1:1 >3 >6 >9

Exercising the Genome



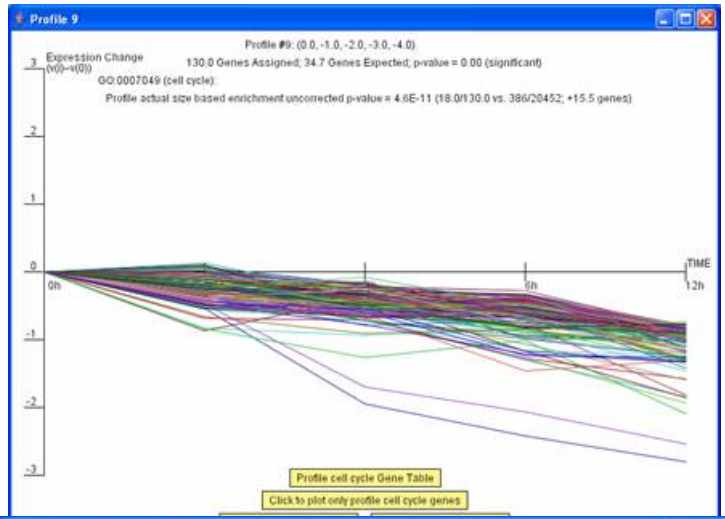
Visualization:

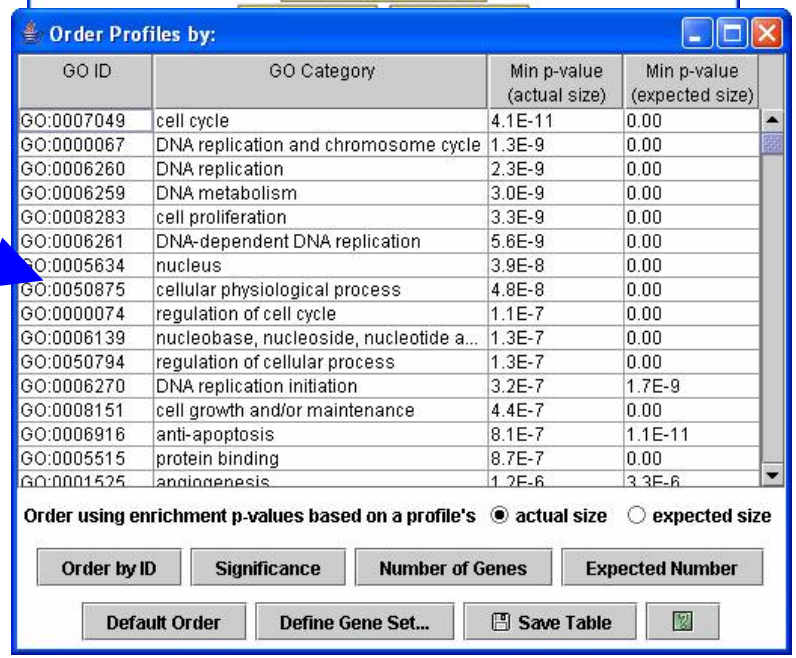
Relative vs. absolute expression



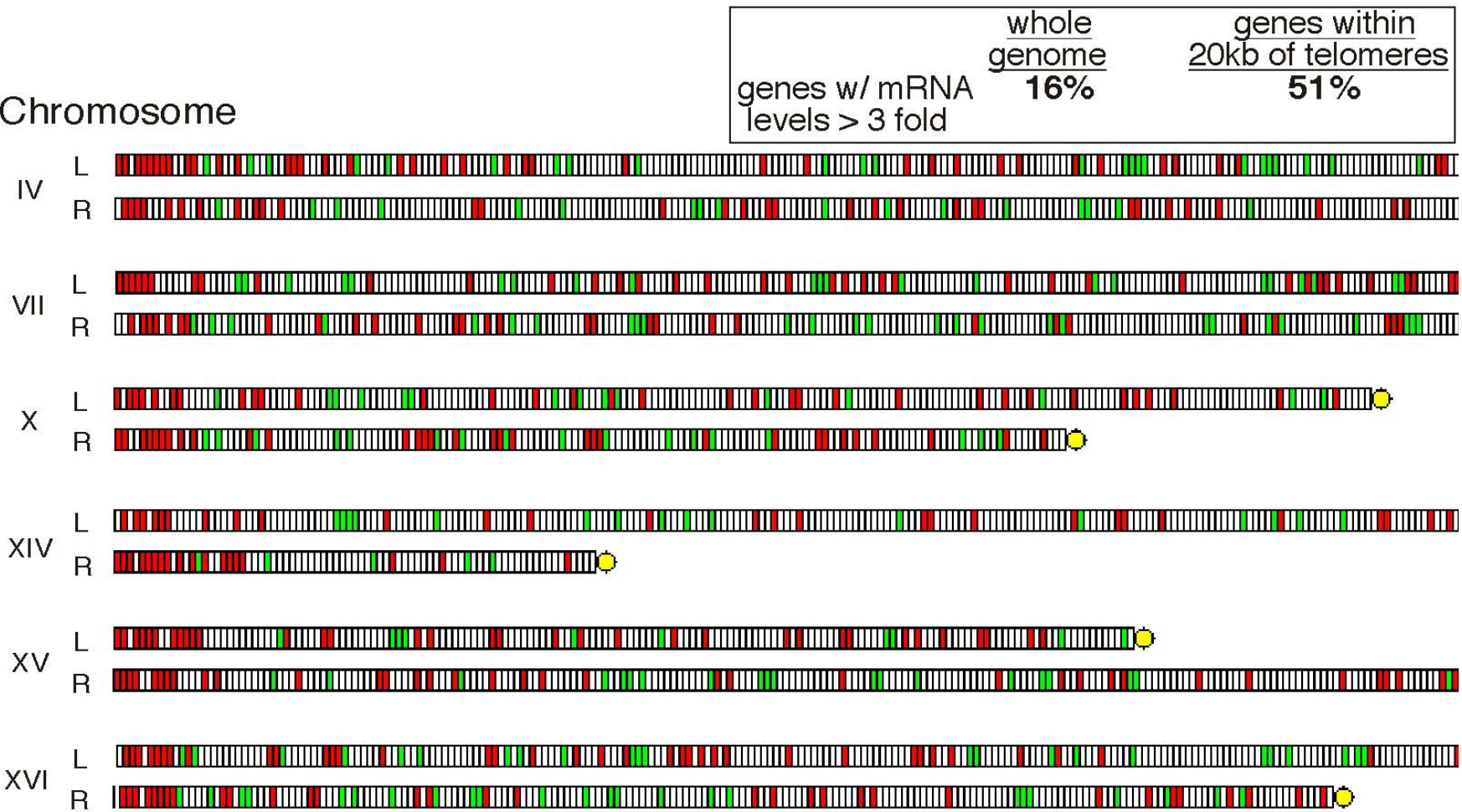
Using annotation databases

- Statistical tests to identify the overlap with various functional categories





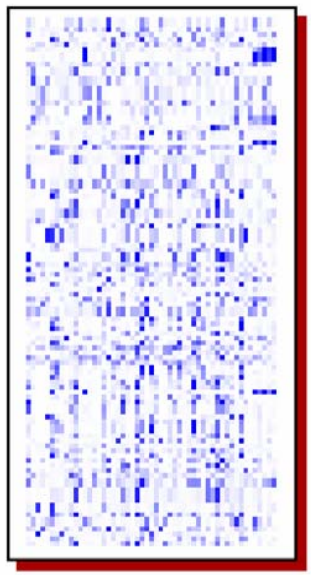
GO ID	GO Category	Min p-value (actual size)	Min p-value (expected size)
GO:0007049	cell cycle	4.1E-11	0.00
GO:0000067	DNA replication and chromosome cycle	1.3E-9	0.00
GO:0006260	DNA replication	2.3E-9	0.00
GO:0006259	DNA metabolism	3.0E-9	0.00
GO:0008283	cell proliferation	3.3E-9	0.00
GO:0006261	DNA-dependent DNA replication	5.6E-9	0.00
GO:0005634	nucleus	3.9E-8	0.00
GO:0050875	cellular physiological process	4.8E-8	0.00
GO:0000074	regulation of cell cycle	1.1E-7	0.00
GO:0006139	nucleobase, nucleoside, nucleotide a...	1.3E-7	0.00
GO:0050794	regulation of cellular process	1.3E-7	0.00
GO:0006270	DNA replication initiation	3.2E-7	1.7E-9
GO:0008151	cell growth and/or maintenance	4.4E-7	0.00
GO:0006916	anti-apoptosis	8.1E-7	1.1E-11
GO:0005515	protein binding	8.7E-7	0.00
GO:0001525	angiogenesis	1.2E-6	3.3E-6

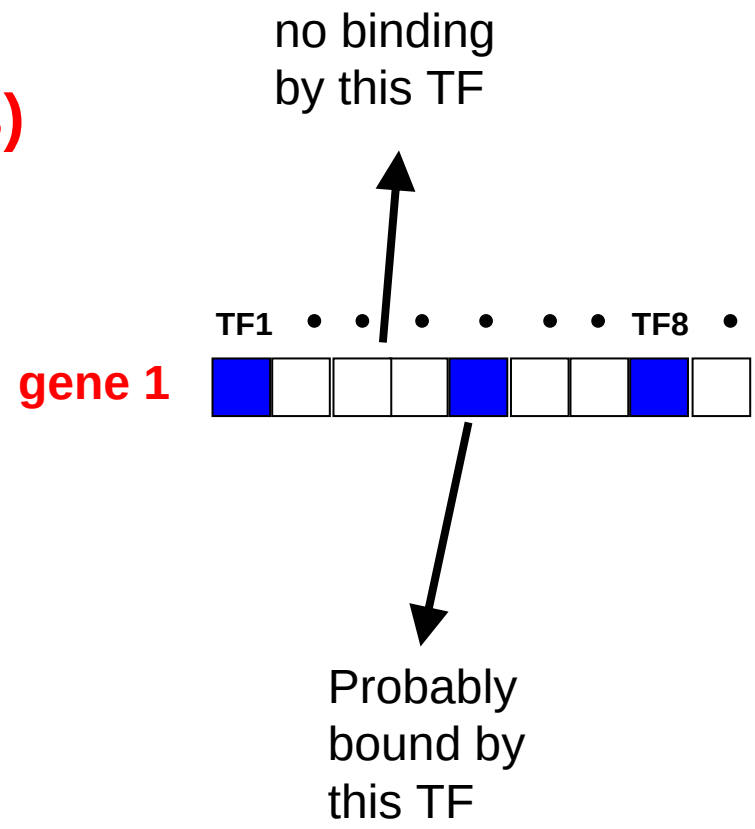

Chromosome
genes w/ mRNA levels > 3 fold
whole genome
16%
genes within 20kb of telomeres
51%
IV
L
R
VII
L
R
X
L
R
XIV
L
R
XV
L
R
XVI
L
R

Genome wide binding

experiments (transcription factors)

enes





Lee *et al Science* 2002

What you should know

- The basic idea behind microarray profiling
- The two different microarray technologies
- Pros and cons for each
- Noise factors in microarray experiments (more next time)