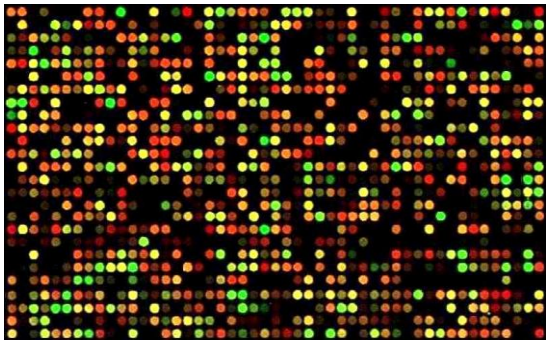
10-810 / 02-710

Computational Genomics

Functional Genomics: Microarrays

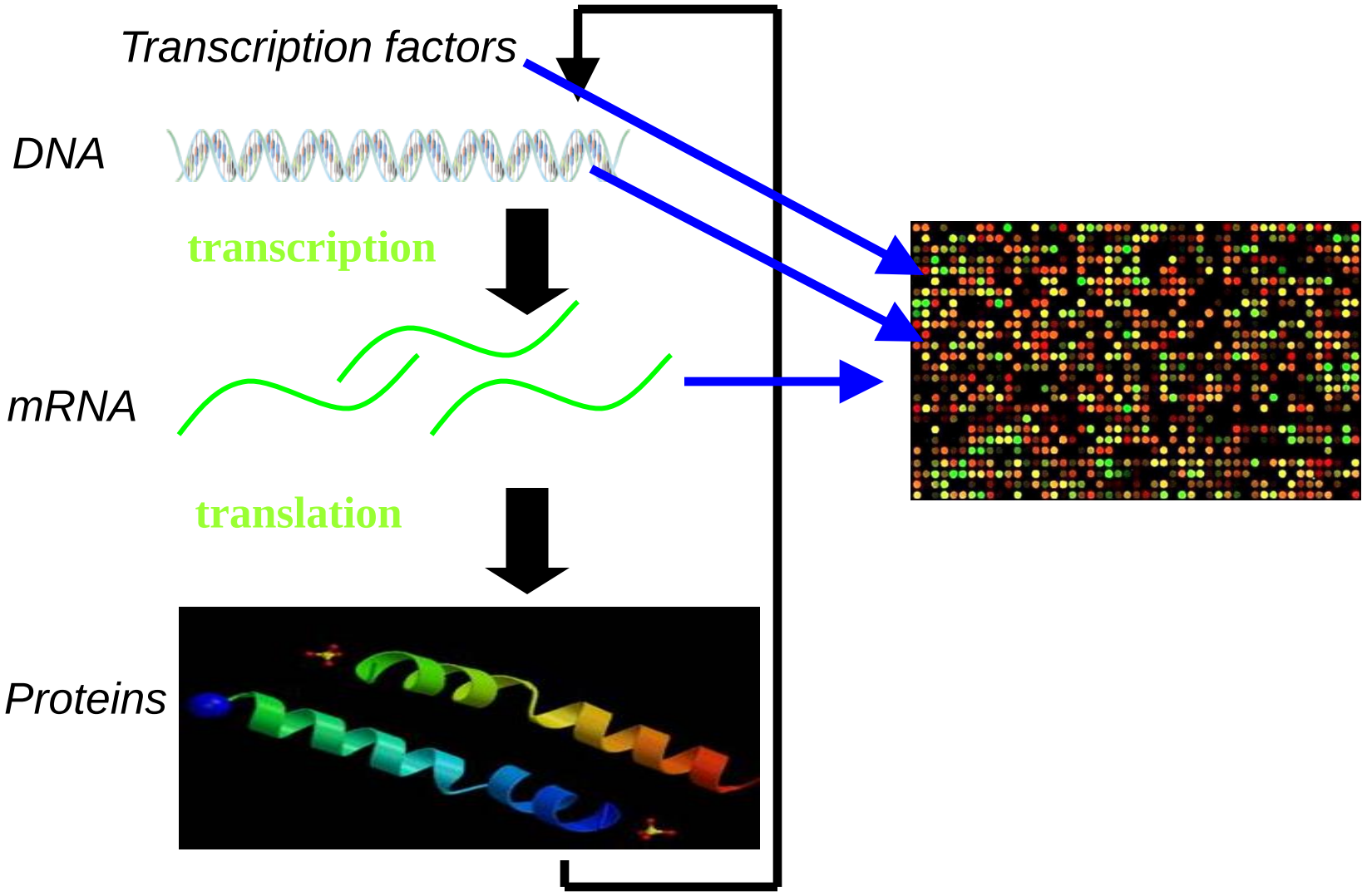


Why sequence is not enough

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

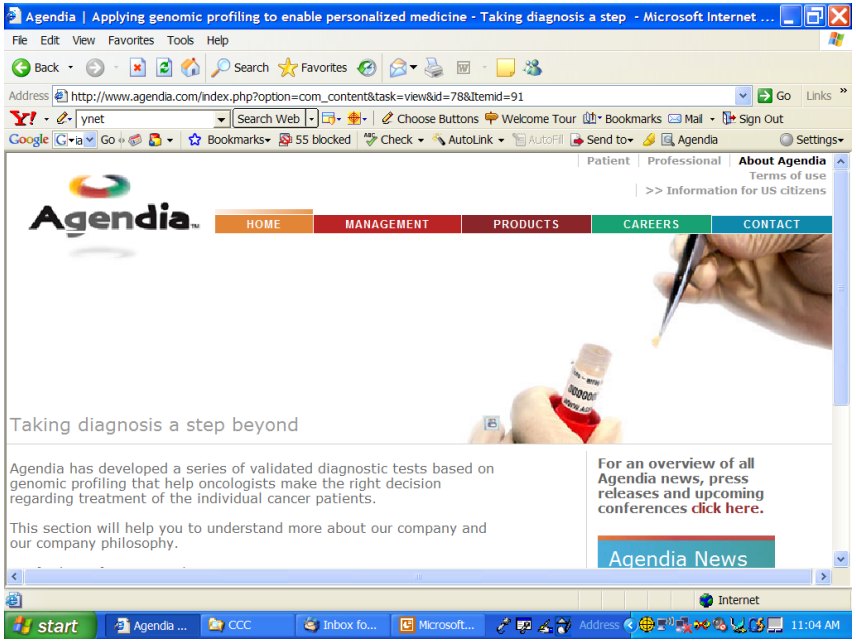
- We need to determine the function of genes.
- We would like to determine which genes are activated in which cells and under which conditions.
- We would like to know the relationships between genes (protein-DNA, protein-protein interactions etc.).
- We would like to model the various dynamic systems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

Microarrays for molecular biology



FDA Approves Gene-Based Breast Cancer Test*

“ MammaPrint is a DNA microarray-based test that measures the activity of 70 genes... The test measures each of these genes in a sample of a woman's breast-cancer tumor and then uses a specific formula to determine whether the patient is deemed low risk or high risk for the spread of the cancer to another site.”



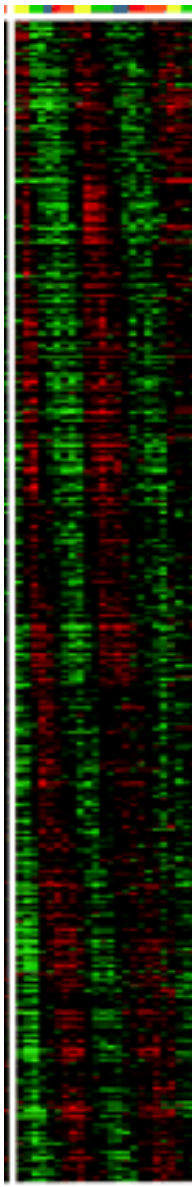
*Washington Post, 2/06/2007

What is gene expression?

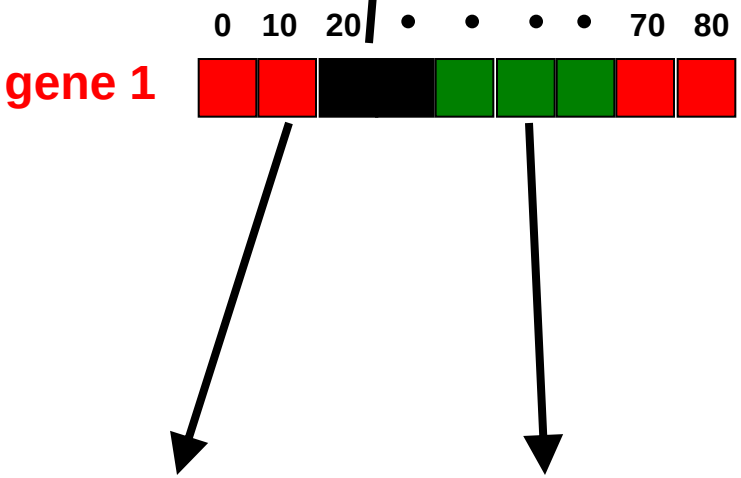
Expression = level of gene in this experiment

Experiments (over time)

genes



baseline expression



Higher expression compared to baseline

Lower expression compared to baseline

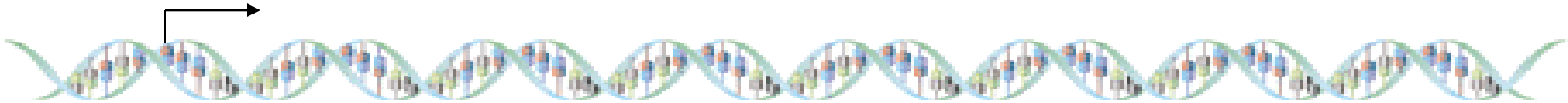
Spellman *et al Mol. Biol. Cell* 1998

Genes and Gene Expression

Technology

Display of Expression Information

Promoter Protein coding sequence Terminator



Genomic DNA

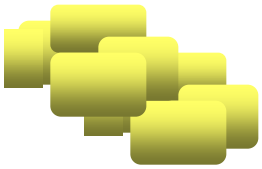
How are Genes Regulated?
DNA-binding Activators Are Key To Specific Gene Expression

Activators

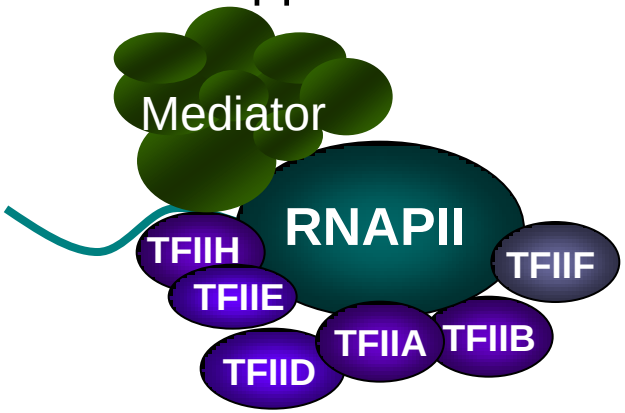
Gene

How are Genes Regulated?
DNA-binding activators are key, but there are additional factors

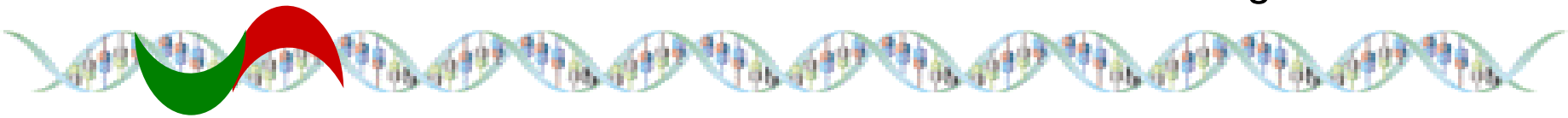
Chromatin modification complexes



Transcription initiation apparatus



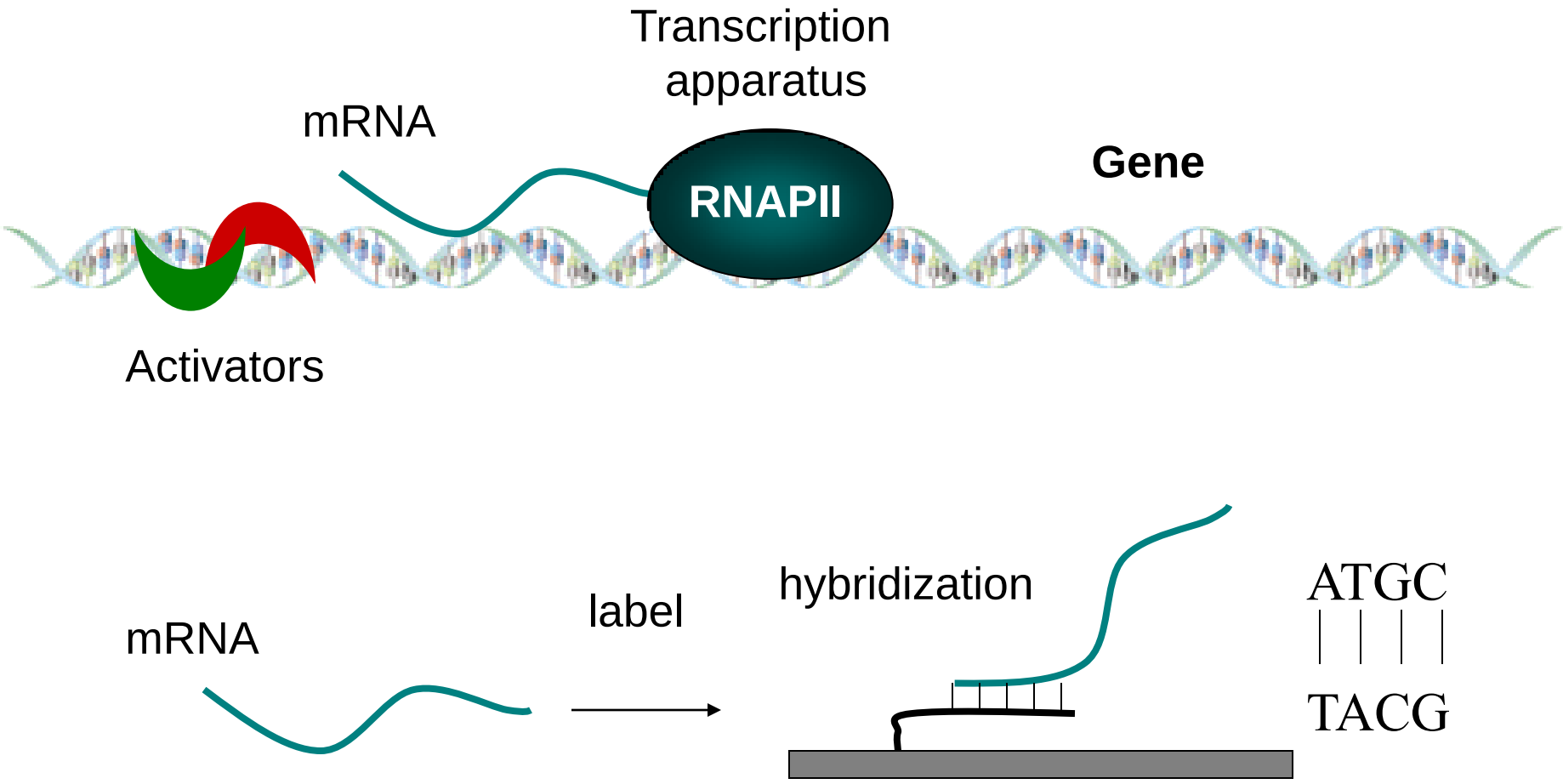
activators
repressors
coactivators
corepressors
transcription apparatus
chromatin factors
RNA processing
RNA transport
RNA degradation



Activators

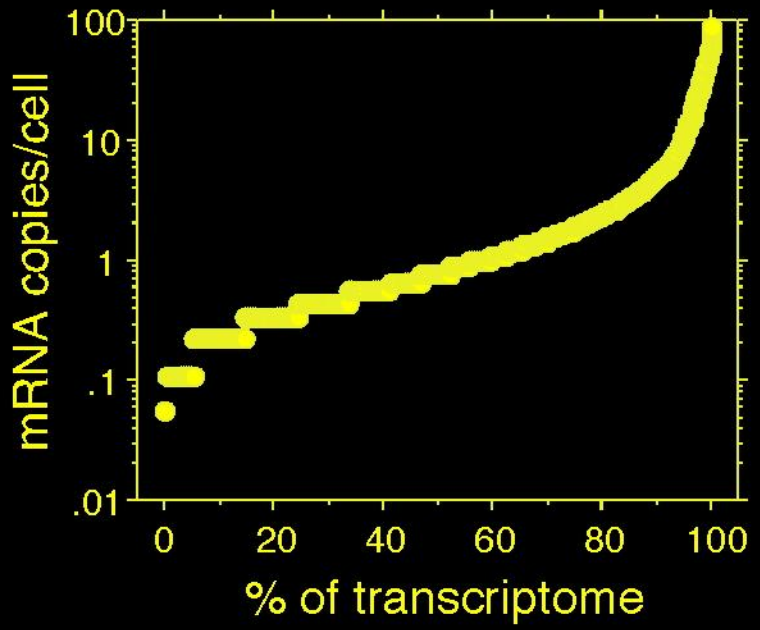
Gene

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

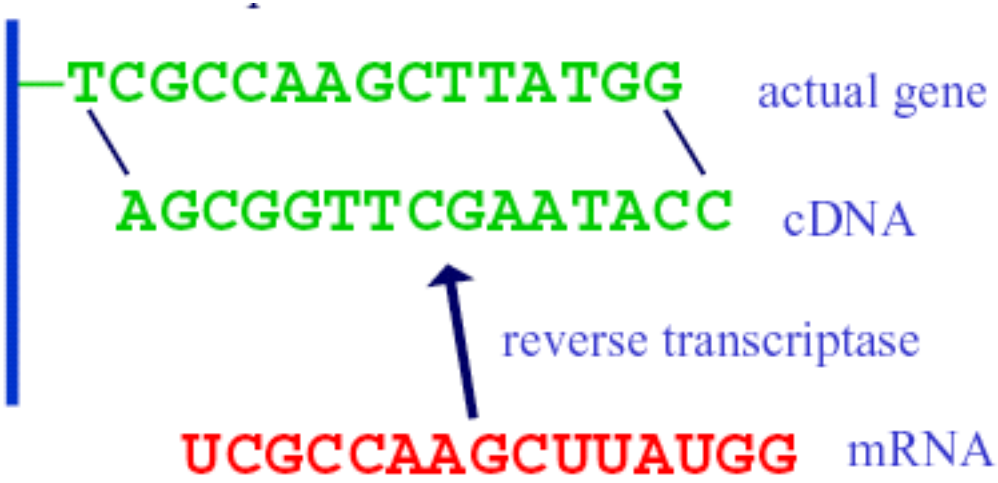
Genes and Gene Expression

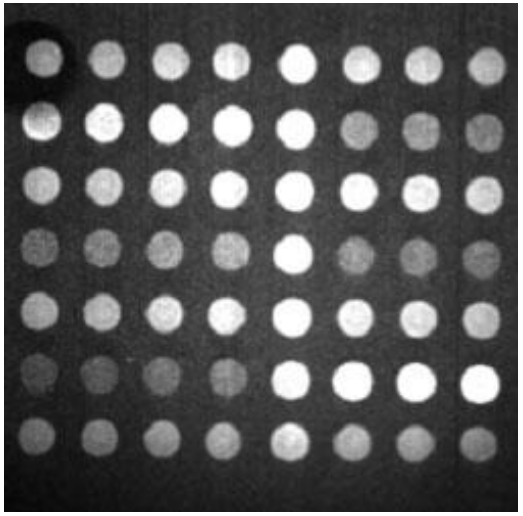
Technology

Display of Expression Information

Microarray Hybridization

- Watson-Crick base pairing of complementary DNA sequences.





- Microarrays have tens of 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, intensity is correlated with mRNA transcript abundance.

Technologies for measuring gene expression

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

Cheap but less popular these days

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

Still widely used

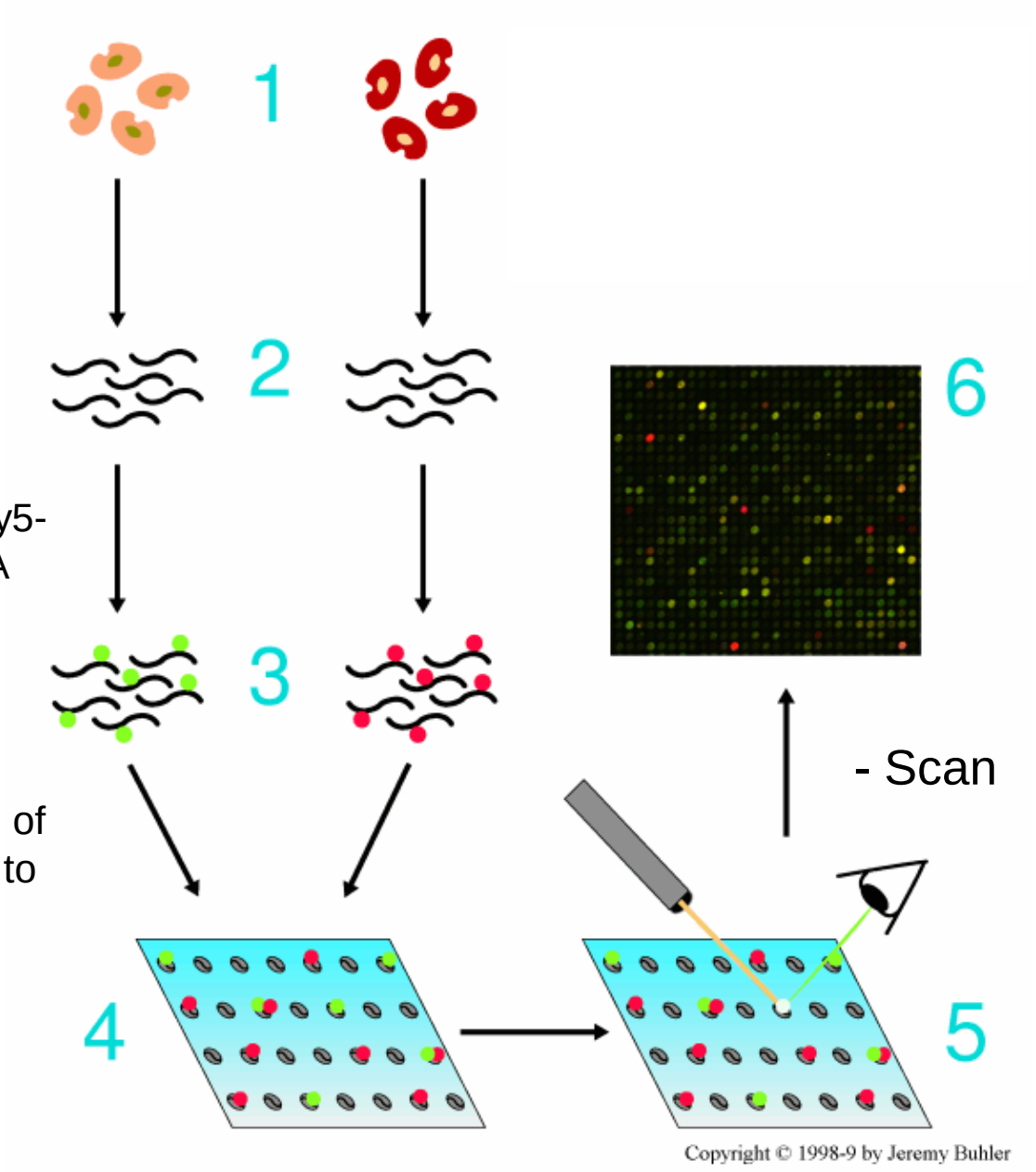
- RNA-Seq
  - Sequencing based technology
  - Rather than sequencing DNA sequence RNA

Starting to penetrate the market but expensive

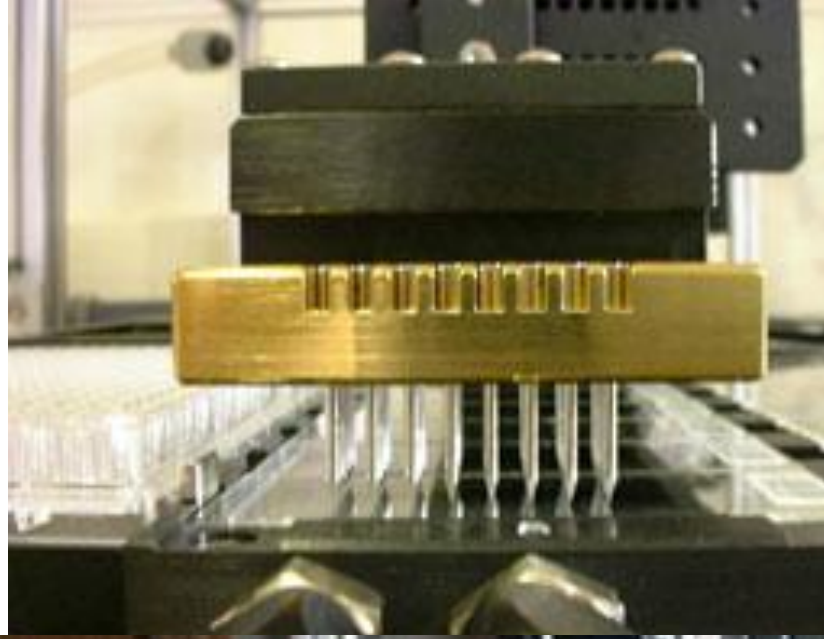
Hybridization and Scanning— cDNA arrays

- Prepare Cy3, Cy5-labeled ss cDNA

- Hybridize 600 ng of labeled ss cDNA to glass slide array

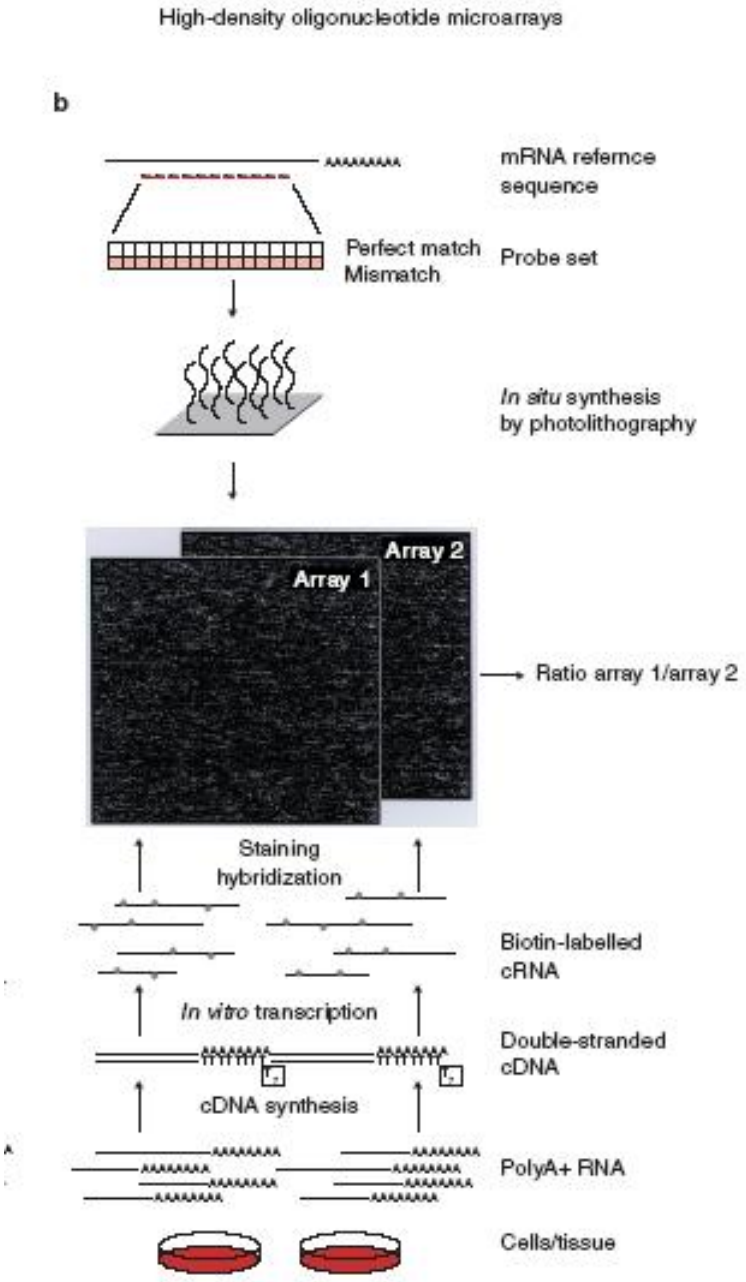


Cartesian PixSys 5500 with quill printing technology



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

Hybridization and Scanning—oligo arrays



cDNA vs. Oligo: Pros and Cons

Even though you are not likely to perform cDNA array experiments, you may be using prior data that was generated by this platform

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

cDNA vs. Oligo: Pros and Cons

Some arrays (including from Agilent) have relatively long probes for each gene (60 bp)

- Does not require sequence
- Cheap
- Direct comparisons

Affymetrix arrays contain an extra 'mismatch' probe designed for internal control on a probe by probe basis

Oligo

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

cDNA vs. Oligo: Pros and Cons

cDNA

- Does
- Chea
- Direct comparisons
- Inaccurate
- Cannot measure individual samples

While this may be personal, all experiments I have been involved with over the last five years used Oligo arrays. Still, there are a lot of cDNA results published and often one needs to reanalyze these for their research.

mize

ol

- 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

Note that many of these (but not all) are eliminated when using RNA-Seq

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:

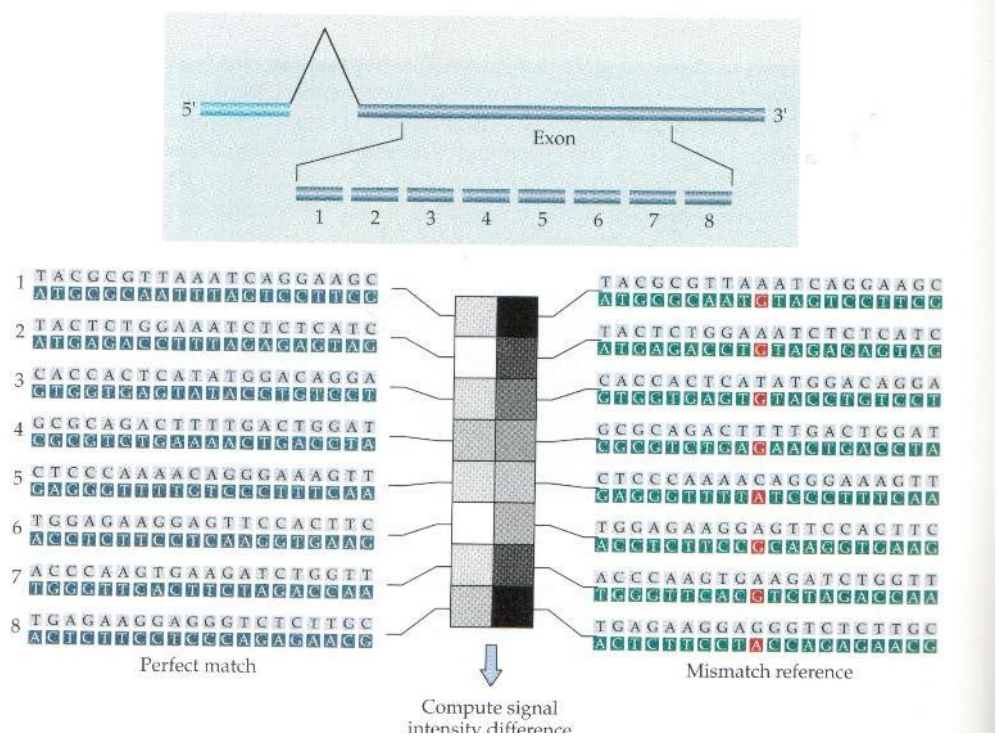
$Y_i = R_i / G_i$

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.

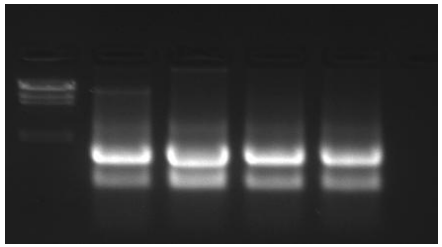
Handling the Different Errors

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

Normalization (later)

mRNA-Seq vs microarrays (1)

Total RNA Extraction



Total RNA

Same source of RNA was hybridized on microarrays

Extraction of Poly A mRNA
ds cDNA synthesis, amplification

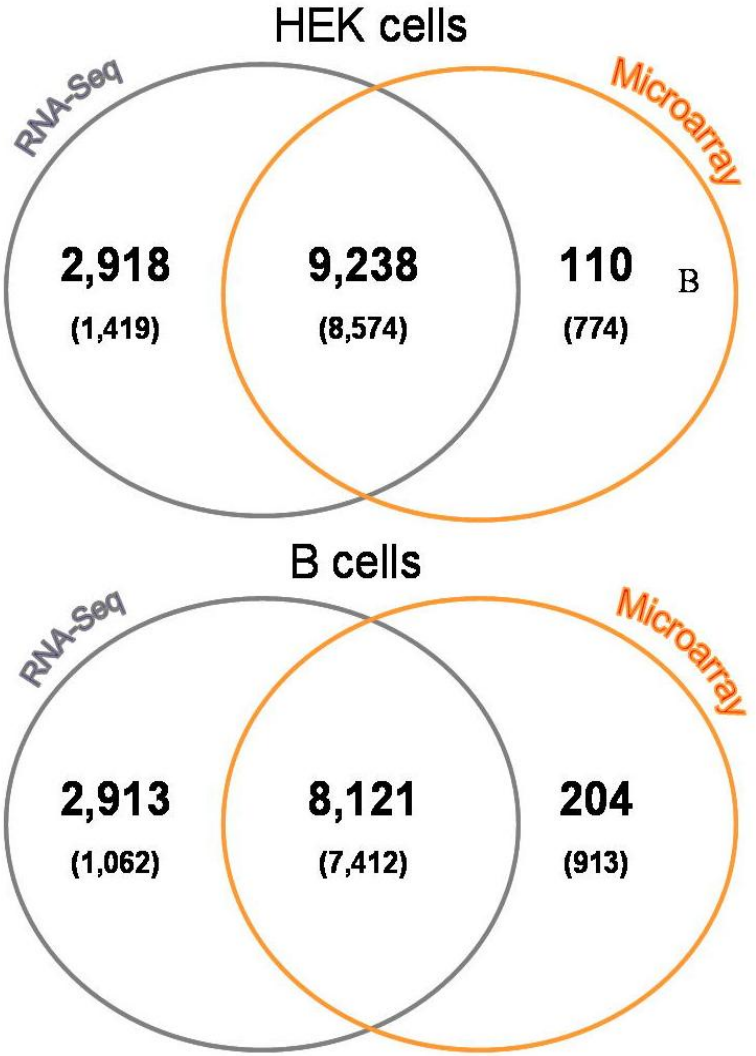
dscDNA synthesis, cRNA

8 channels device

Hybridization on human
ref 6 bead-chip (Illumina)

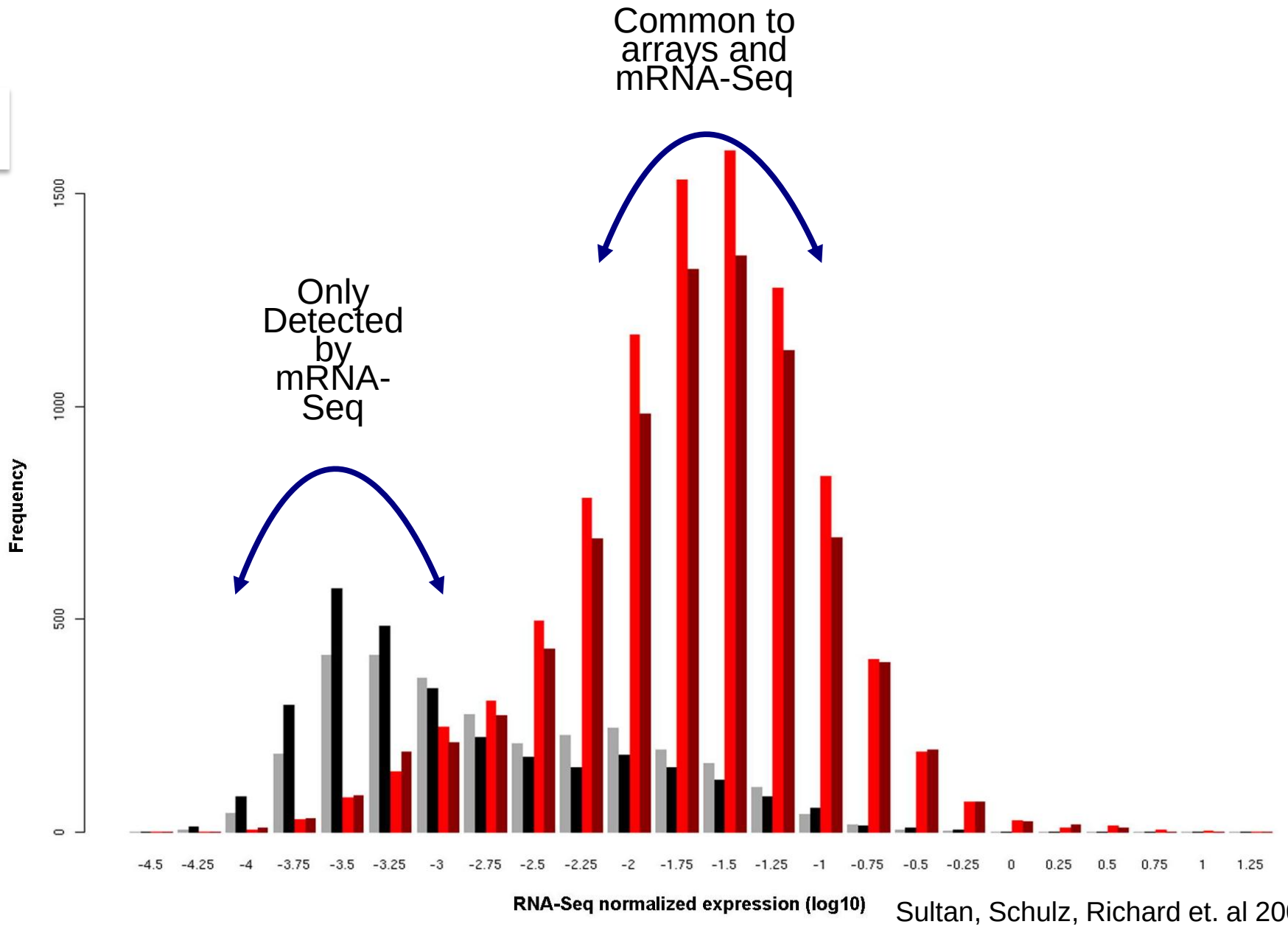
Sultan, Schulz, Richard et. al 2008

mRNA-Seq vs microarrays (2)



Sultan, Schulz, Richard et. al 2008

mRNA-Seq vs microarrays (3)



Sultan, Schulz, Richard et. al 2008

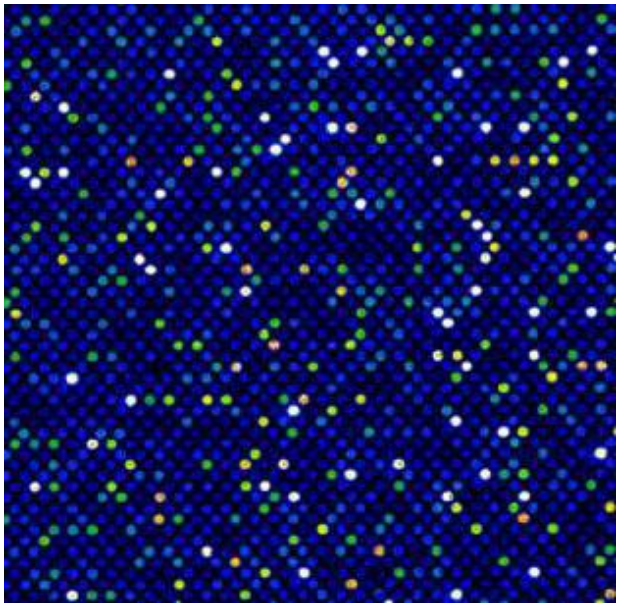
Protein-DNA Binding arrays

- 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.

As with mRNA, sequencing can be used for protein binding studies as well. This type of experiment is termed ChIP-Seq

Protein Binding Arrays (PBMs)

- The arrays we discussed so far where measuring *in-vivo* activity.
- There are also arrays that are used for *in-vitro* studies.
- In these arrays researchers use purified proteins and comprehensive set of *k-mers* (currently covering all 8 mers).
- This allows them to zoom in not only on the actual motif the protein is binding to.
- The arrays are universal, they can be used for any organism.



Genes and Gene Expression

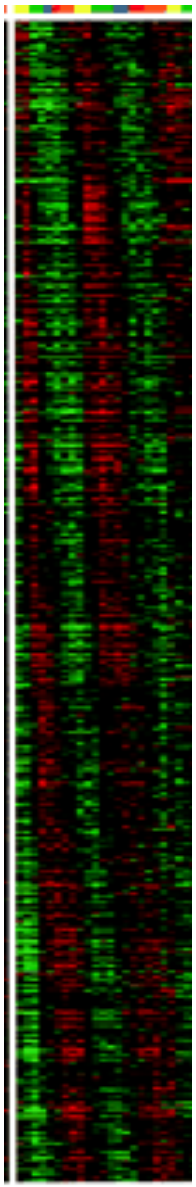
Technology

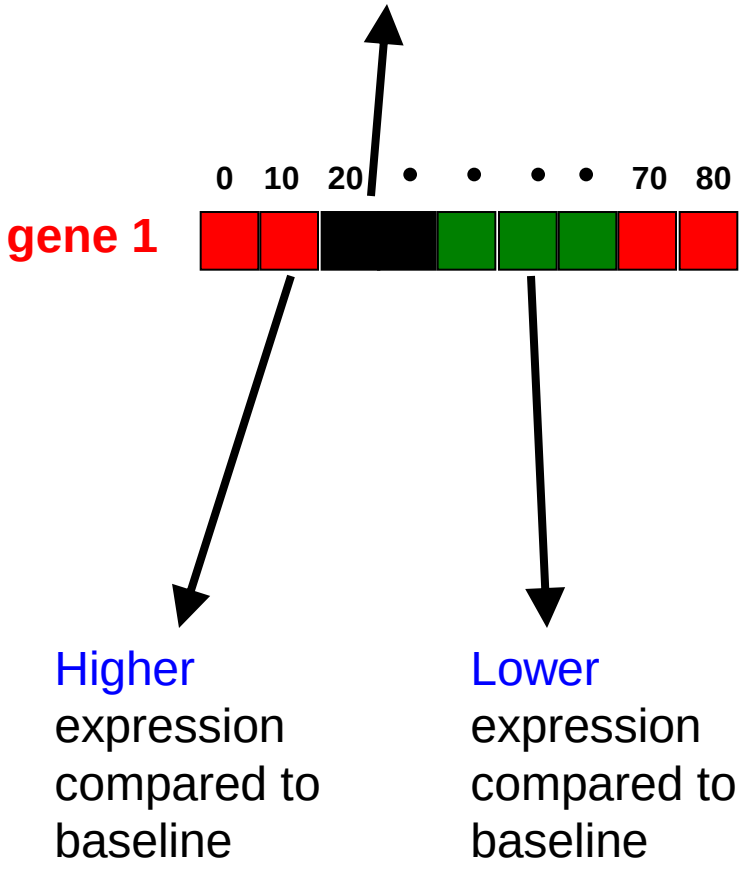
Display of Expression Information

Yeast cell cycle expression program

Experiments (over time)

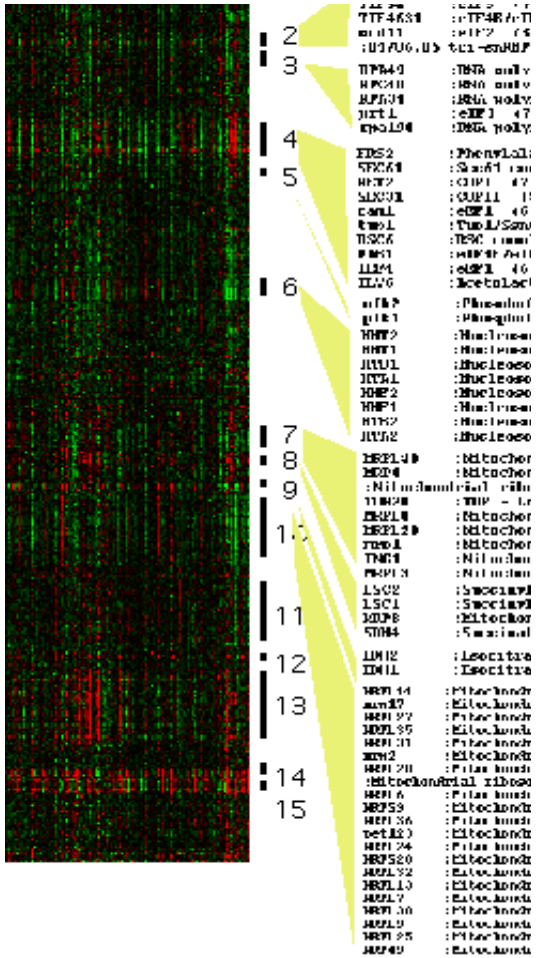
genes





Spellman *et al Mol. Biol. Cell* 1998

600 Conditions/Mutations



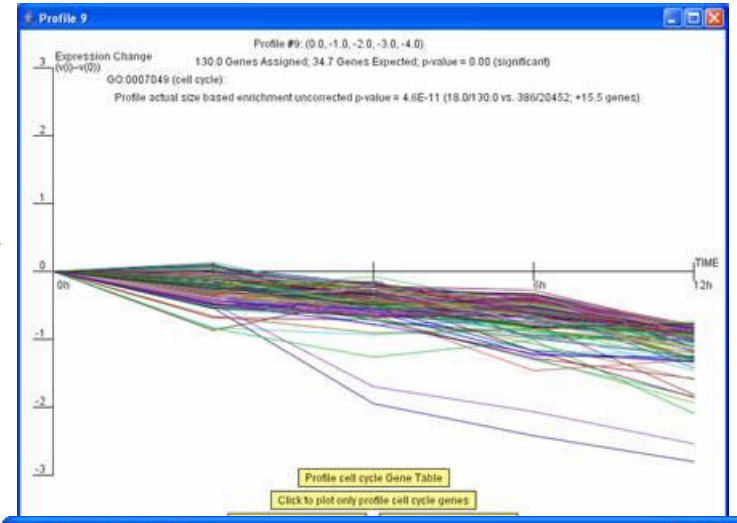
6200 Genes

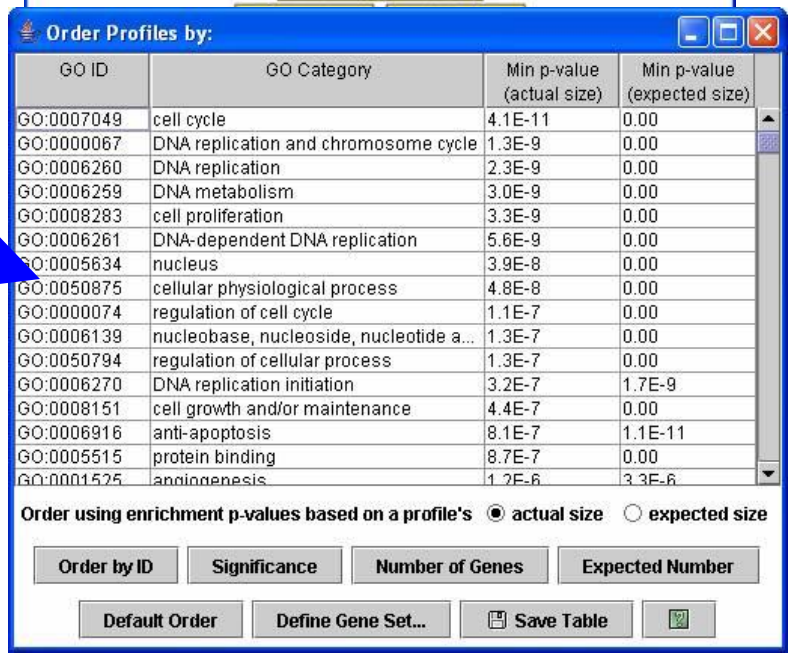
Environment

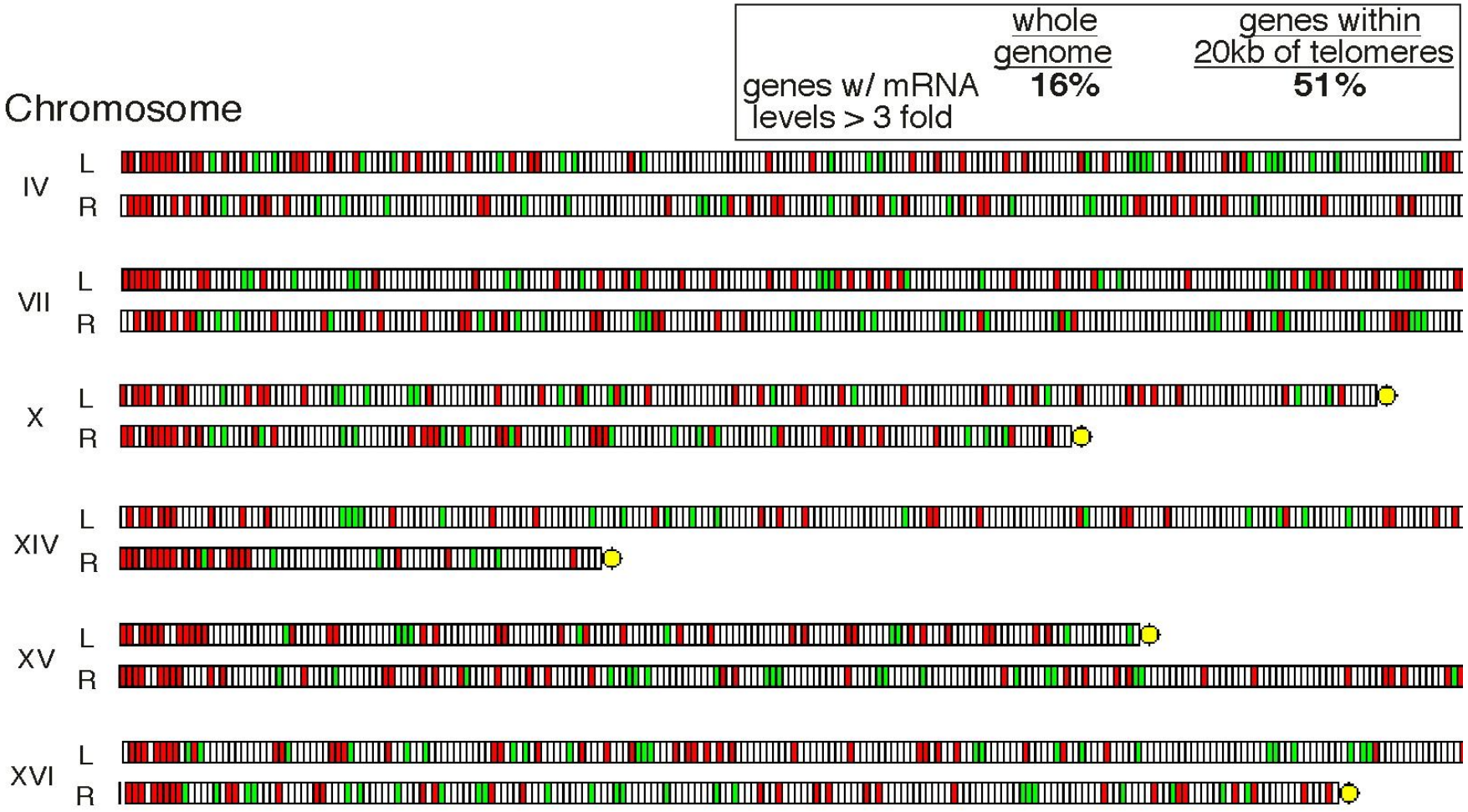
Single-gene Mutations

Using annotation databases

- Statistical tests to identify the overlap with various functional categories



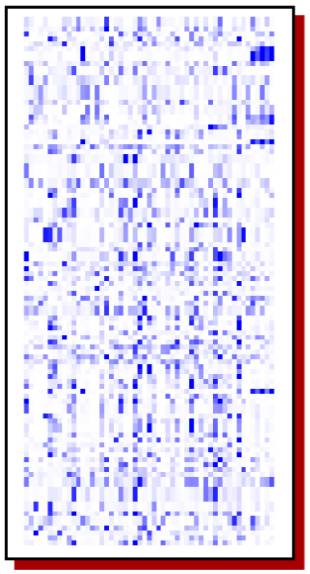



genes w/ mRNA levels > 3 fold
whole genome
16%
genes within 20kb of telomeres
51%
Chromosome
IV
L
R
VII
L
R
X
L
R
XIV
L
R
XV
L
R
XVI
L
R

Genome wide binding

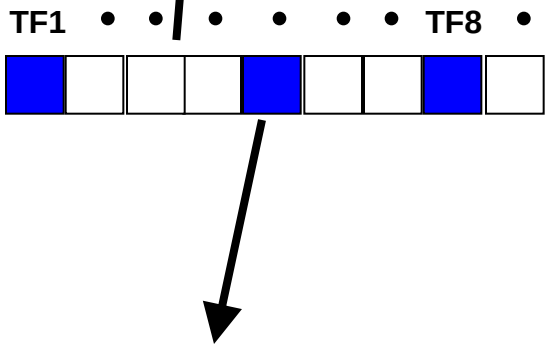
experiments (transcription factors)

genes



no binding by this TF

gene 1



Probably bound by this TF

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)