Computational Systems Biology

What is “Systems Biology”?

The study of the mechanisms underlying complex biological processes as integrated systems of many interacting components. Systems biology involves (1) collection of large sets of experimental data (2) proposal of mathematical models that might account for at least some significant aspects of this data set, (3) accurate computer solution of the mathematical equations to obtain numerical predictions, and (4) assessment of the quality of the model by comparing numerical simulations with the experimental data.

-(Leroy Hood, 1999)

What are Biological Systems?

Popular Notion:

It is a complex system consisting of very many simple and identical elements interacting to produce what appears to be complex behavior

Example: Cells, Proteins

What are Biological Systems?

- Complex systems of simple elements have functions that emerge from the properties of the networks they form

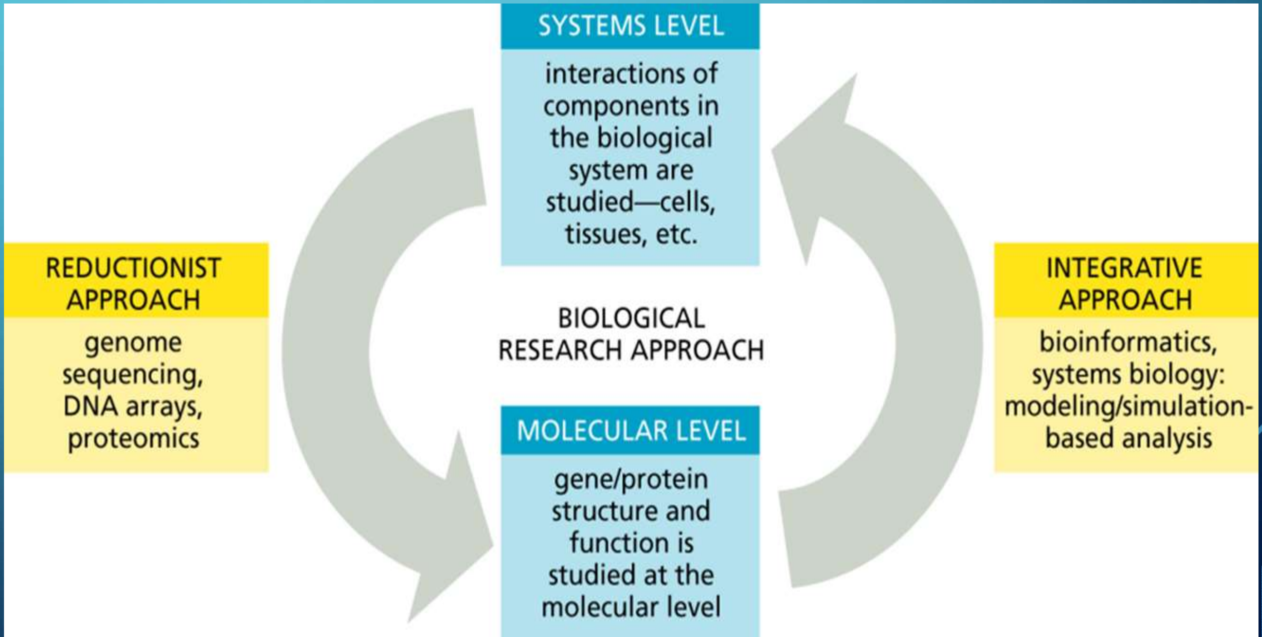
- Biological systems have functions that rely on a combination of the network and the specific elements involved

Two ways of looking a problem

- Top down or bottom up
- Either look at the whole organism and abstract large portions of it
- Or try to understand each small piece and then after understanding every small piece assemble into the whole
- Both are used, valid and complement each other

Molecular Biology vs. Systems Biology

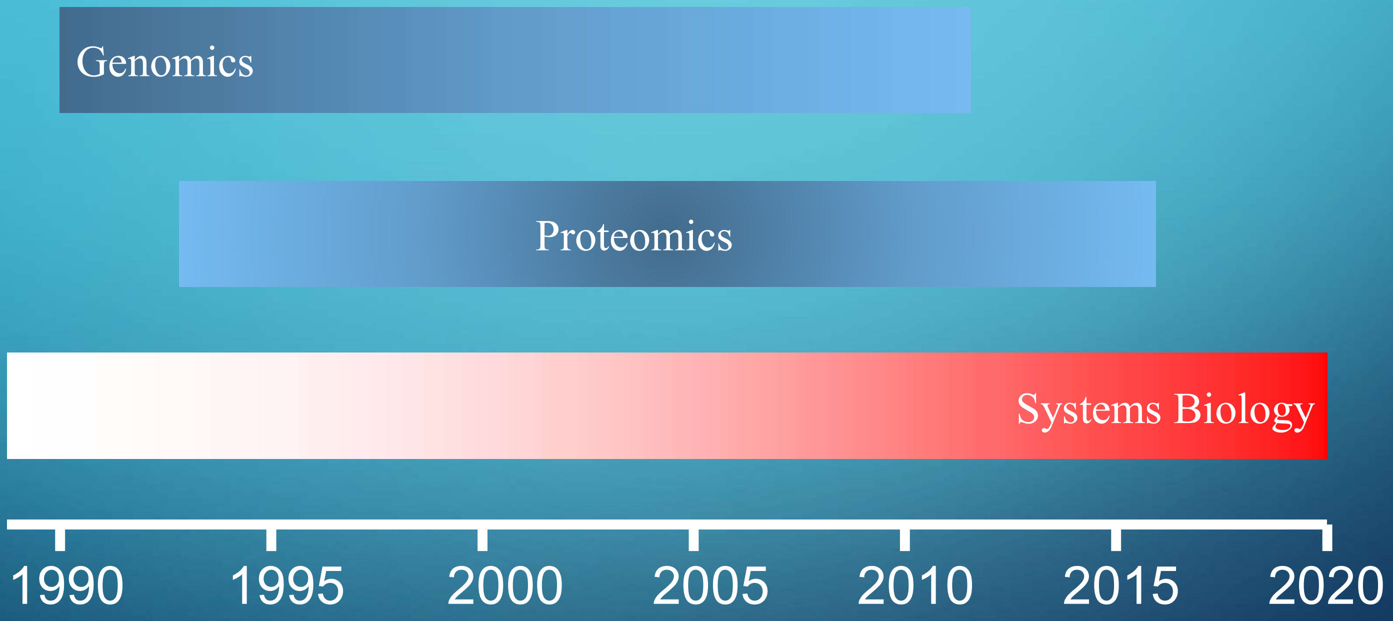
- In molecular biology, gene structure and function is studied at the molecular level
- In systems biology, specific interactions of components in the biological system are studied – cells, tissues, organs, and ecological webs.


SYSTEMS LEVEL
interactions of components in the biological system are studied—cells, tissues, etc.
REDUCTIONIST APPROACH
genome sequencing, DNA arrays, proteomics
BIOLOGICAL RESEARCH APPROACH
INTEGRATIVE APPROACH
bioinformatics, systems biology: modeling/simulation-based analysis
MOLECULAR LEVEL
gene/protein structure and function is studied at the molecular level

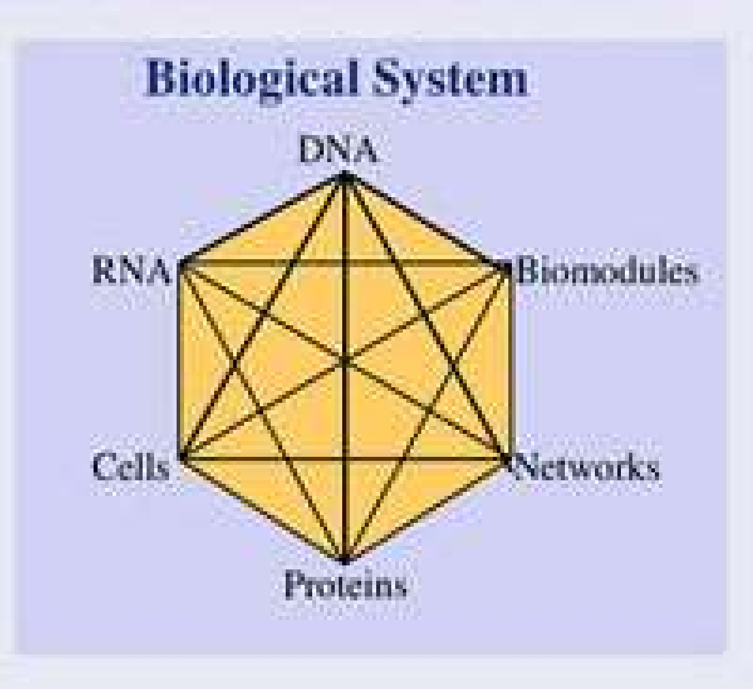
Systems Biology vs. traditional cell and molecular biology

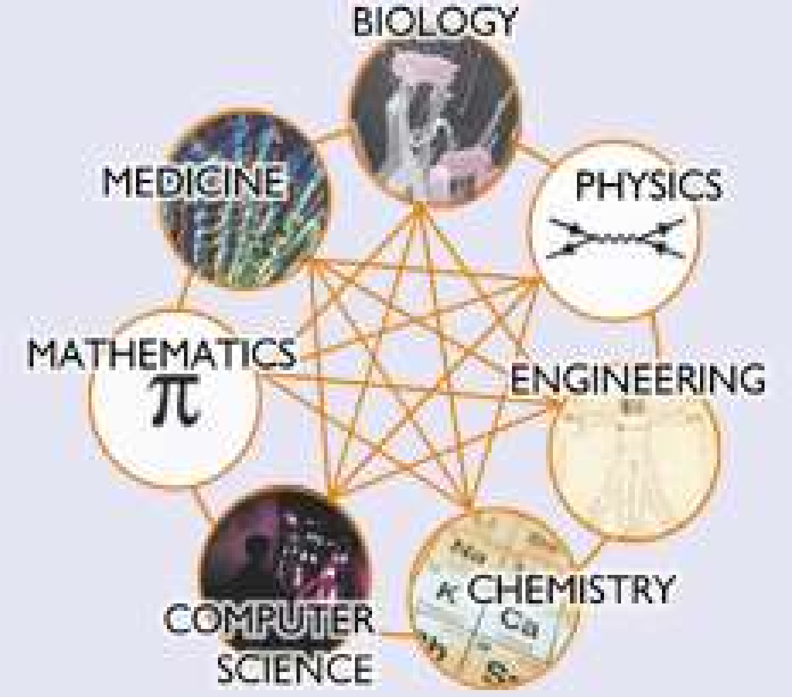
- Experimental techniques in systems biology are high throughput.
- Intensive computation is involved from the start in systems biology, in order to organize the data into usable computable databases.
- Exploration in traditional biology proceeds by successive cycles of hypothesis formation and testing; data accumulates during these cycles.
- Systems biology initially gathers data without prior hypothesis formation; hypothesis formation and testing comes during post-experiment data analysis and modeling.

Genomics, Proteomics & Systems Biology



Systems Biology is an integration of data & approaches





Technologies to study systems at different levels

- Genomics (HT-DNA sequencing)
- Mutation detection (SNP (single nucleotide polymorphisms) methods)
- Transcriptomics (Gene/Transcript measurement, gene chips, microarrays)
- Proteomics (MS, 2D-PAGE, protein chips, Yeast-2-hybrid, X-ray, NMR)
- Metabolomics

From Systems Biology to Computational Biology

- Biological Systems are complex, thus, a combination of experimental and computational approaches are needed

- Linkages need to be made between molecular characteristics and systems biology results

High-throughput techniques

- Large-scale identification of components (genes, RNAs, and proteins)
- Their expression patterns
- Their biochemical and genetic interactions
- Provide valuable information about the functions of individual components and unexpected relationships between components and cellular processes
- A variety of large-scale data sets have been identified and used to assemble different networks.

biological networks

- Interaction data from individual studies and large-scale screens can be assembled into a network format
- five types of biological networks:

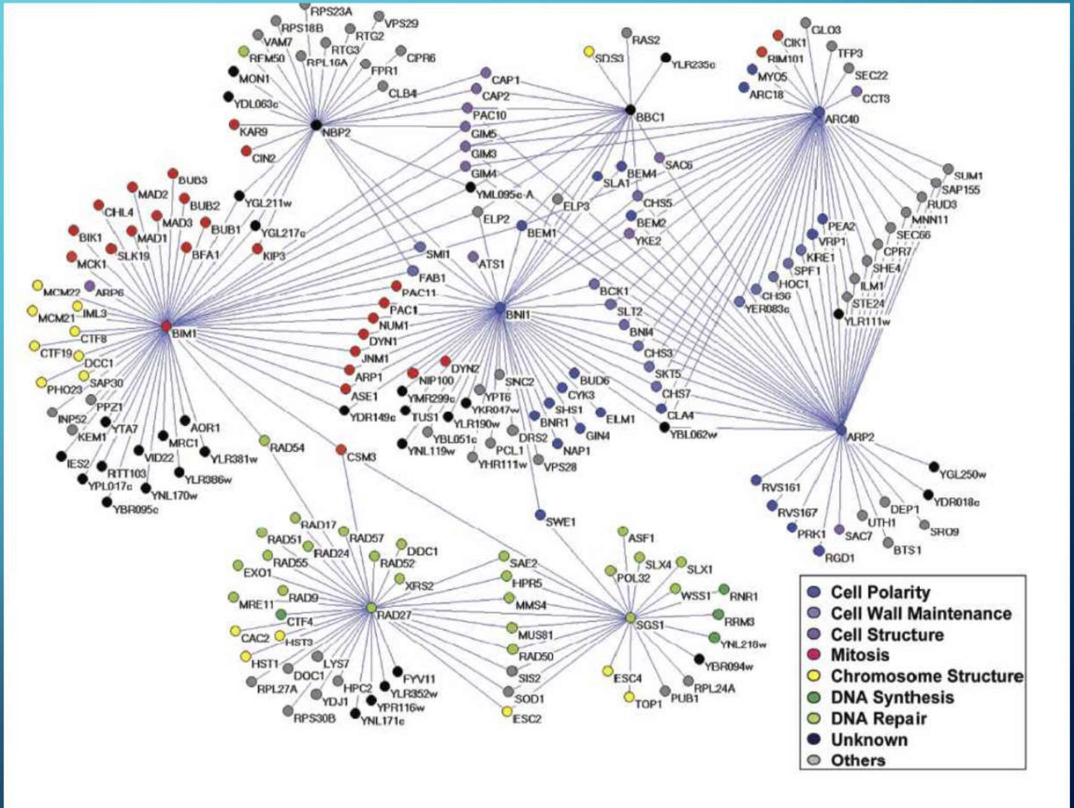
➢ transcription factor binding

➢ protein–protein interaction

➢ protein phosphorylation

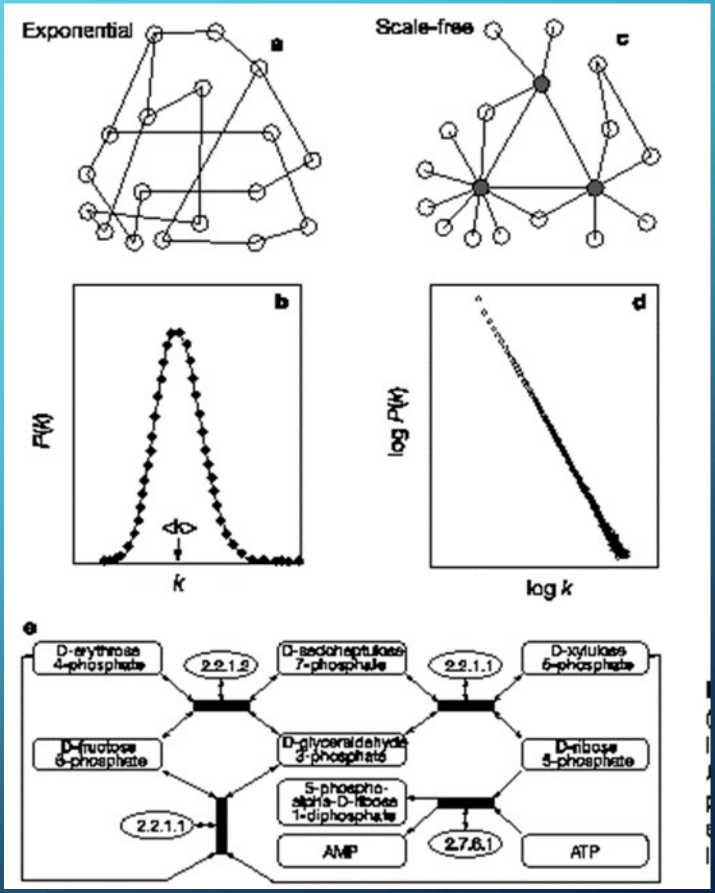
➢ metabolic interaction

➢ genetic interaction network



Graph theory, networks

- Two types of networks
- Exponential and scale free
- Most cellular networks are scale free
- It makes the most sense to study the interactions of the central nodes not the outer nodes

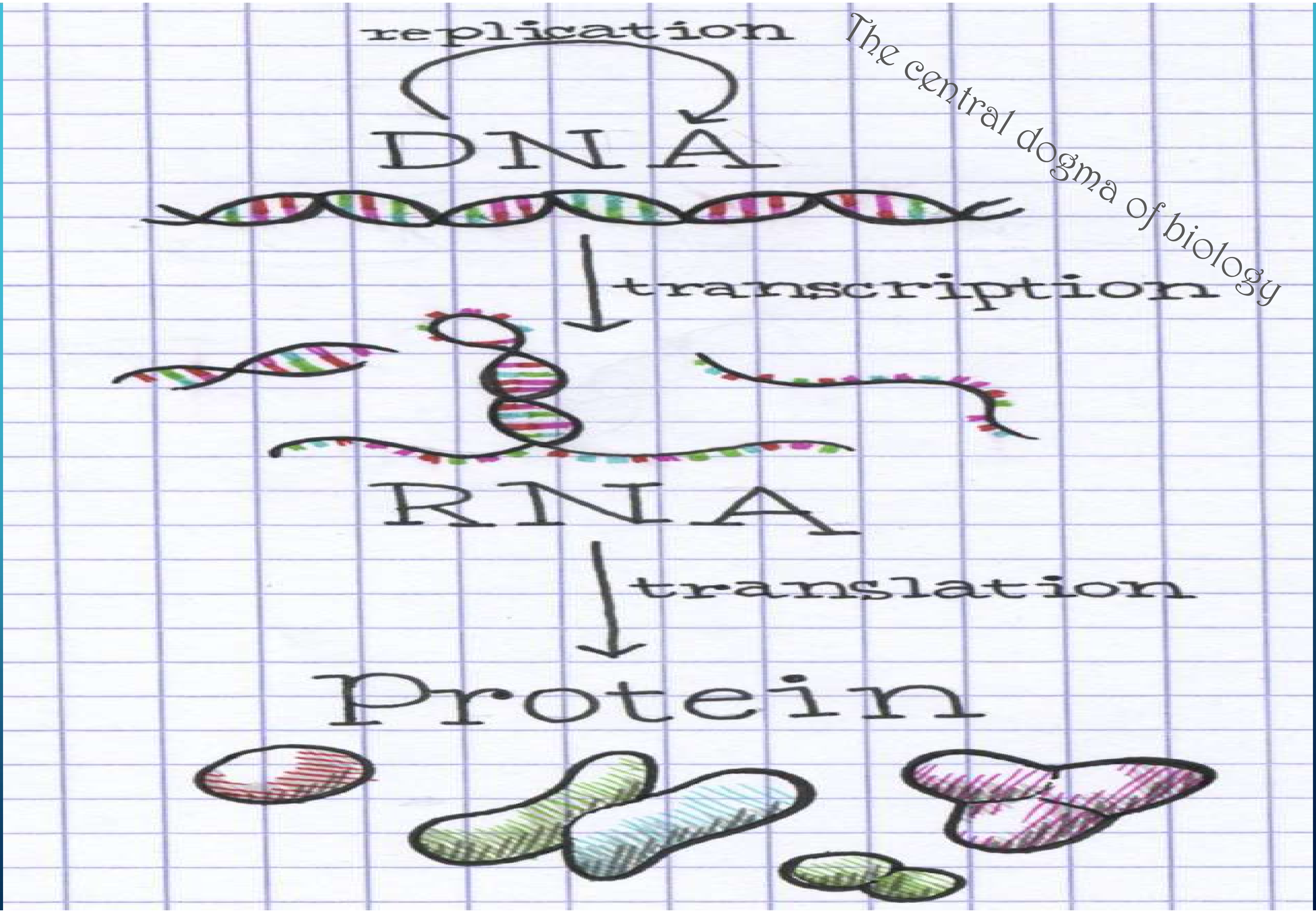


Genetic Interaction Networks

- functionally related genes tend to exhibit genetic interactions
- finding genetic interactions has been crucial to geneticists
- systematic high-throughput mapping of genetic interactions by microarray
- In genetic interaction network nodes represent genes and edges represent interaction between genes

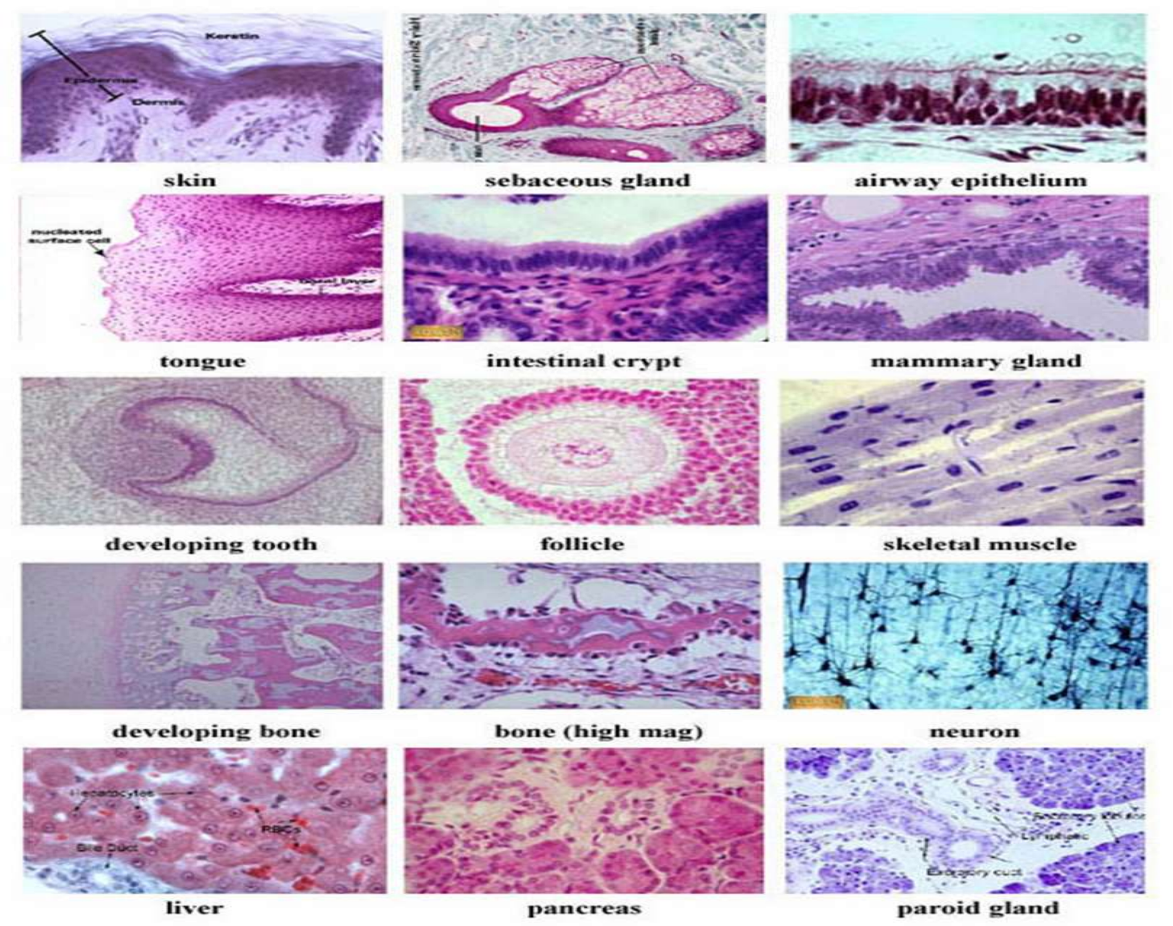
Microarray technology

- Gene expression biology
- Measuring gene expression levels
- two technologies: Two-color cDNA arrays and single-color Affymetrix genechips
- Finding and understanding differentially expressed genes
- Advanced analysis (clustering and classification)



The central dogma of biology

Genome information is complete for hundreds of organisms...
...but the complexity and diversity of the resulting phenotype is challenging


skin
sebaceous gland
airway epithelium
tongue
intestinal crypt
mammary gland
developing tooth
follicle
skeletal muscle
developing bone
bone (high mag)
neuron
liver
pancreas
paroid gland

The dramatic consequences of gene regulation in biology

Same genome ➡

- Different tissues
- Different physiology
- Different proteome
- Different expression pattern

Gene expression distinguishes...

- ...physiological status (nutrition, environment)
- ...sex and age
- ...various tissues and cell types
- ...response to stimuli (drugs, signals, toxins)
- ...health and disease

✓ underlying pathogenic diversity

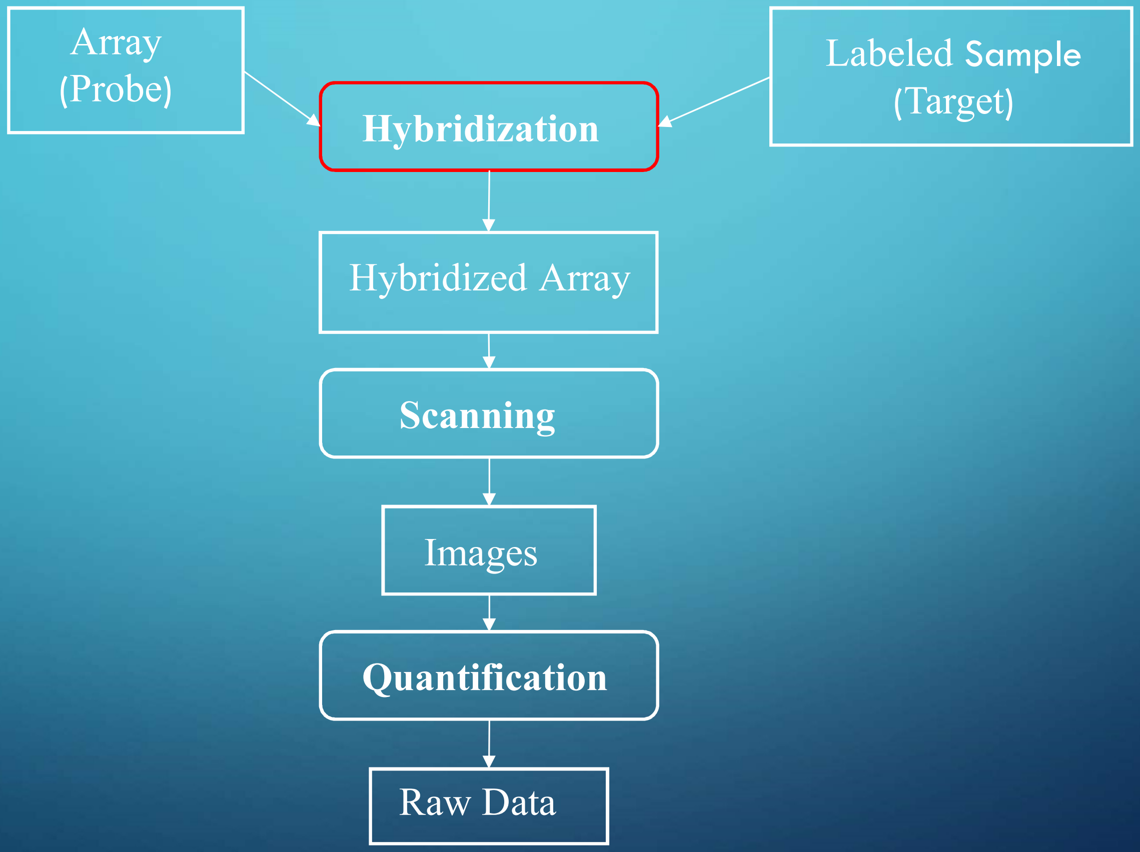
✓ progression and response to treatment

✓ patient classes of varying prospects

Measuring gene expression levels

1. total amount of mRNA = optical density at appropriate (UV) wavelength
2. mass separation and specific probing, one gene at a time = Northern blot
3. comprehensive "molecular sorting" = microarray technology
    1. two-color cDNA or oligo arrays
    2. single-color Affymetrix genechips

Core Technology



Array
(Probe)

Labeled Sample
(Target)

Hybridization

Hybridized Array

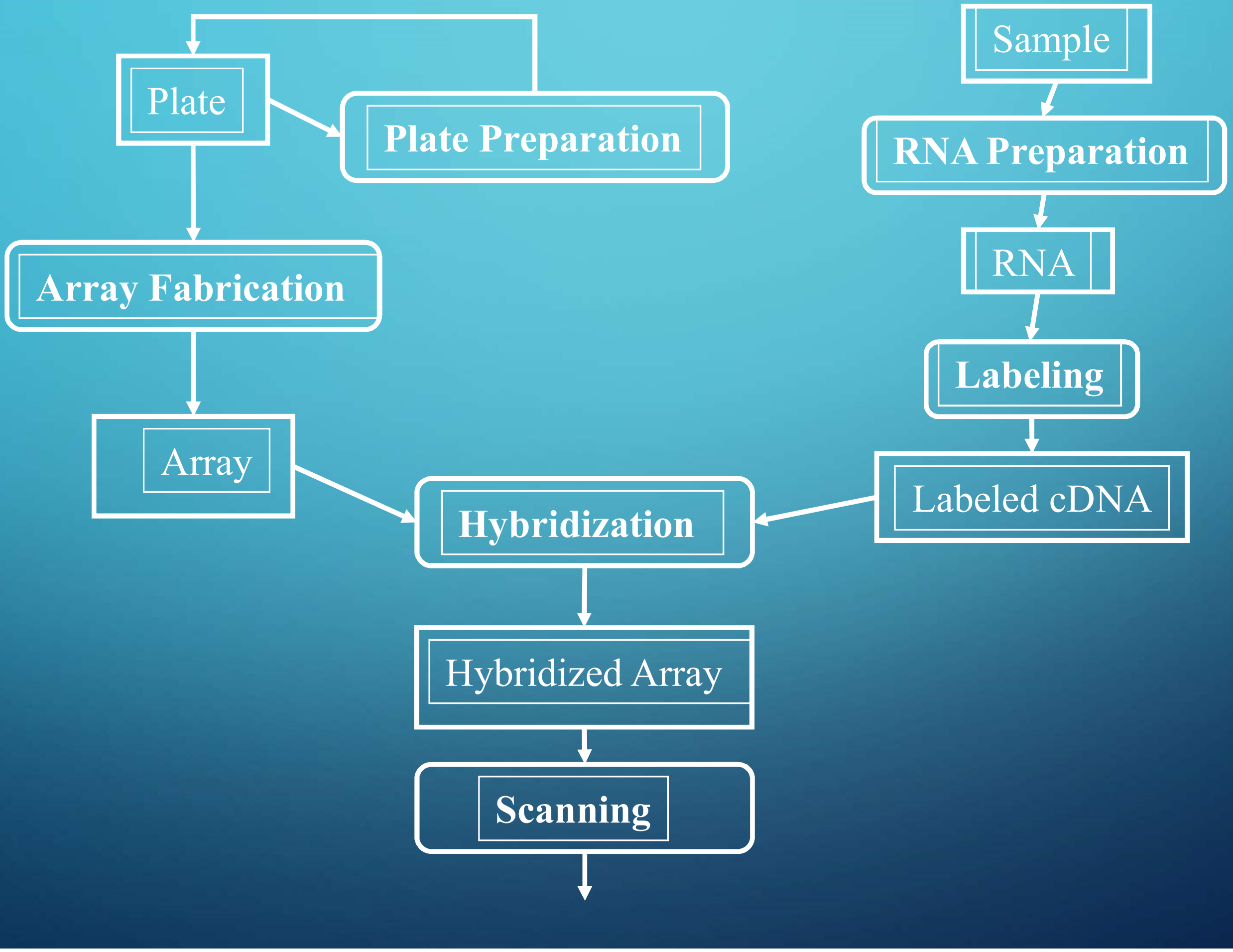
Scanning

Images

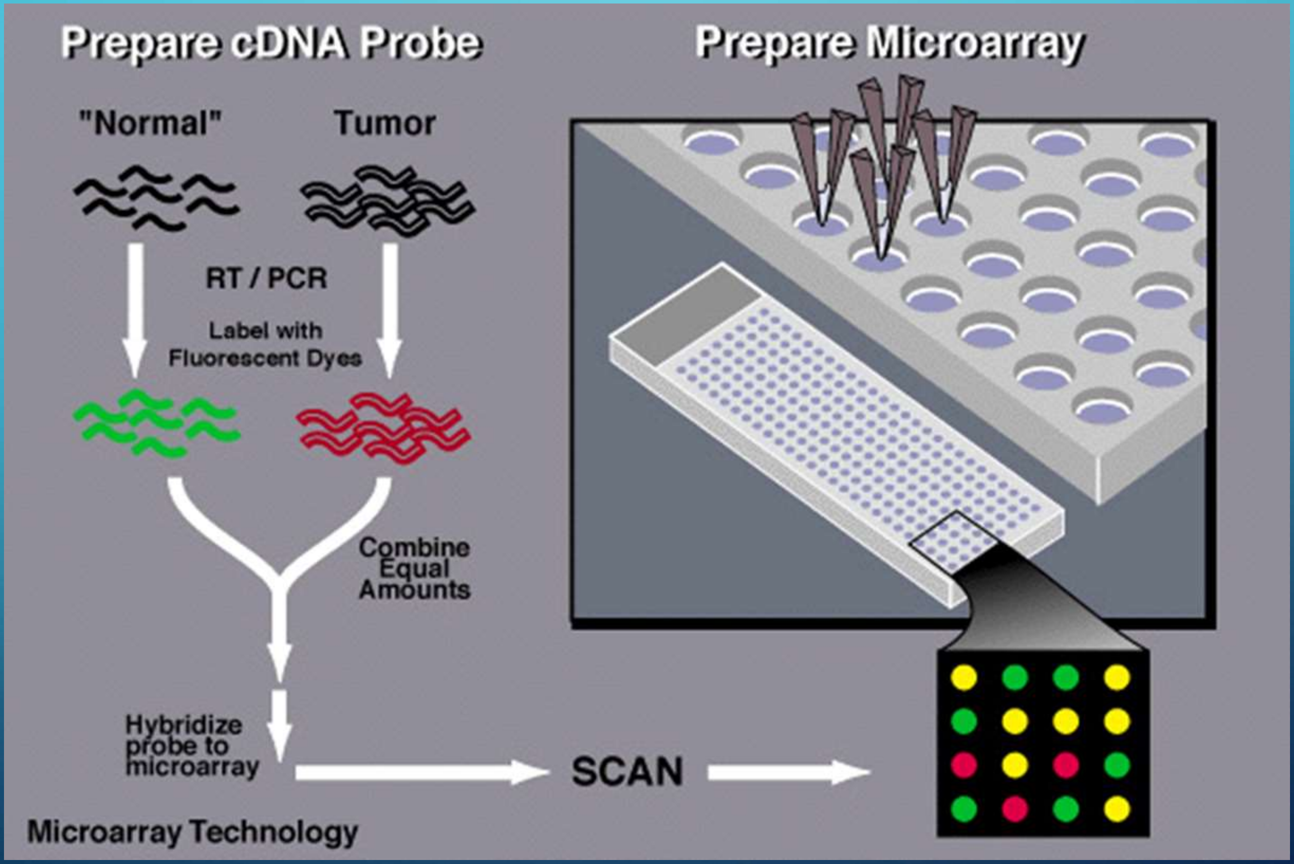
Quantification

Raw Data

cDNA Spotted Array

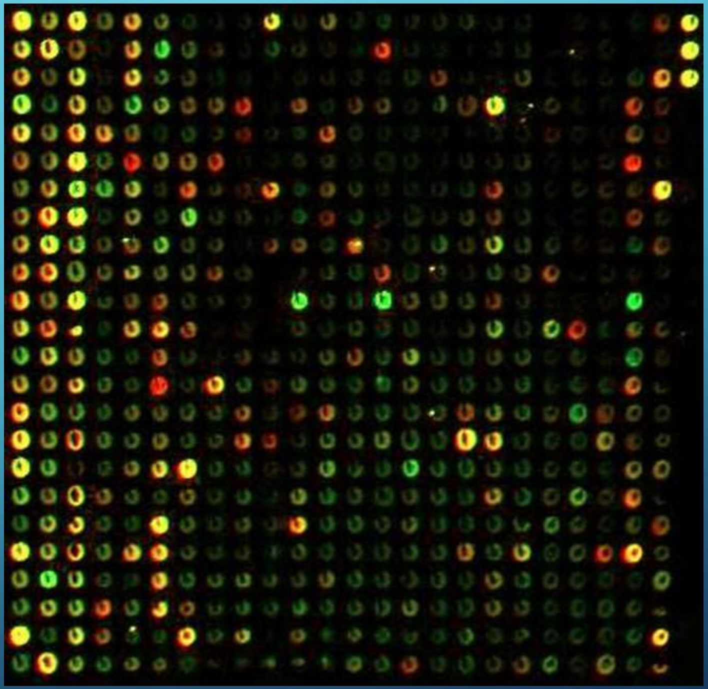


cDNA Spotted Array



Scanning

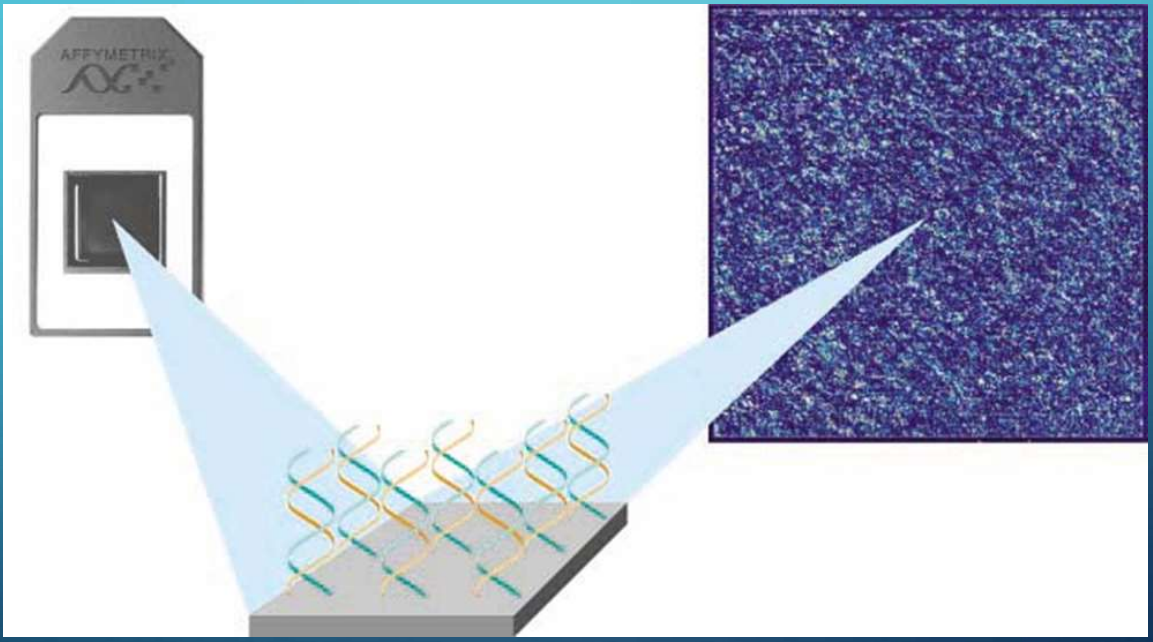
- Scanner: fluorescent light detection
- two channel (Cy3/Cy5)



Affymetrix GeneChip

Affymetrix GeneChip

an array of oligonucleotide (20~80-mer oligos) probes is synthesized in situ (on-chip)

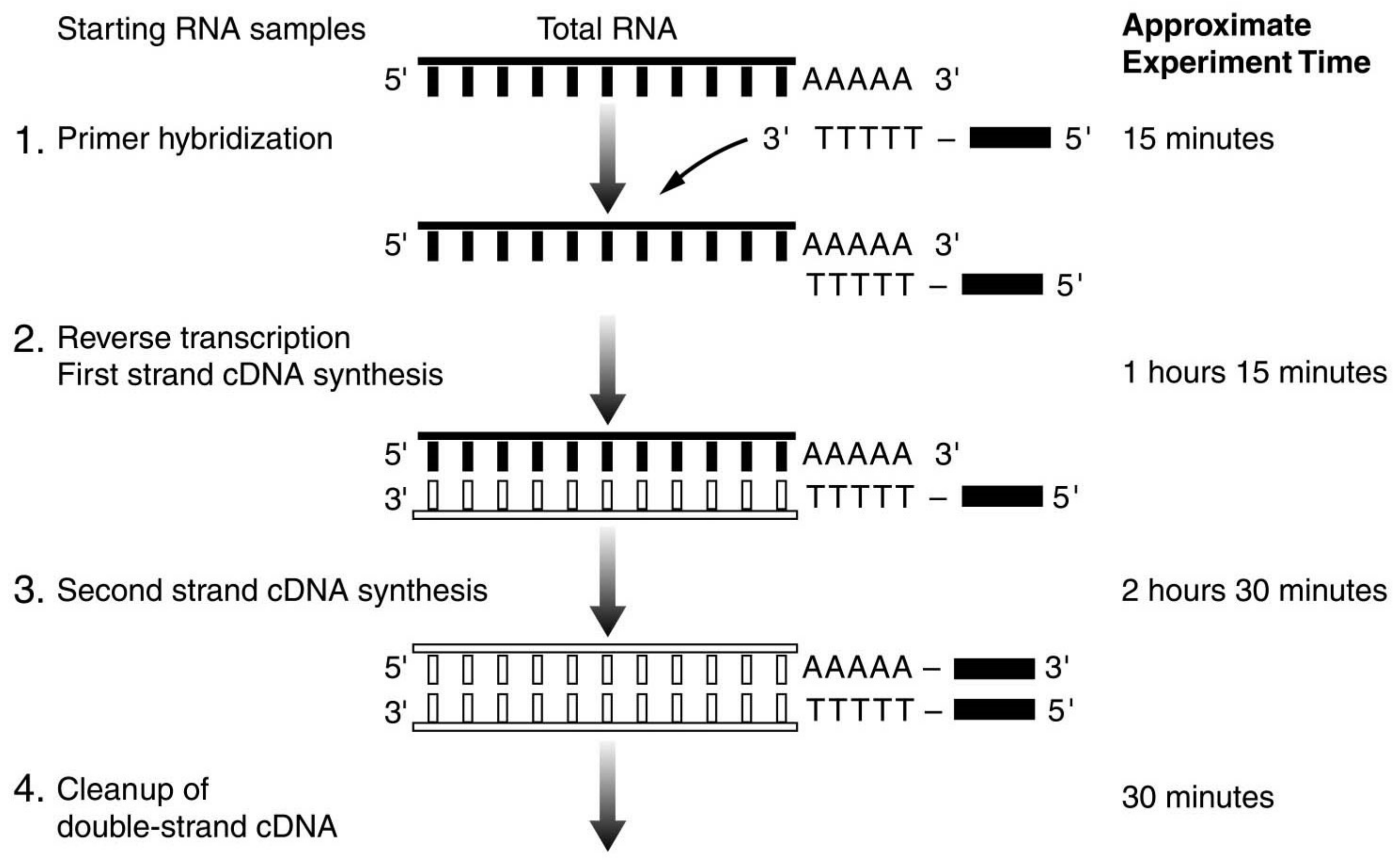


Affymetrix GeneChip

Sample Preparation

1. extract total RNA
2. convert mRNA to cDNA
   - reverse transcription with poly(T) primer
3. amplify cDNA into labeled cRNA
   - T7 RNA polymerase with biotin labeled CTP and UTP
4. break cRNA into fragments of 35-200mers
5. hybridize and wash
6. scan the chip

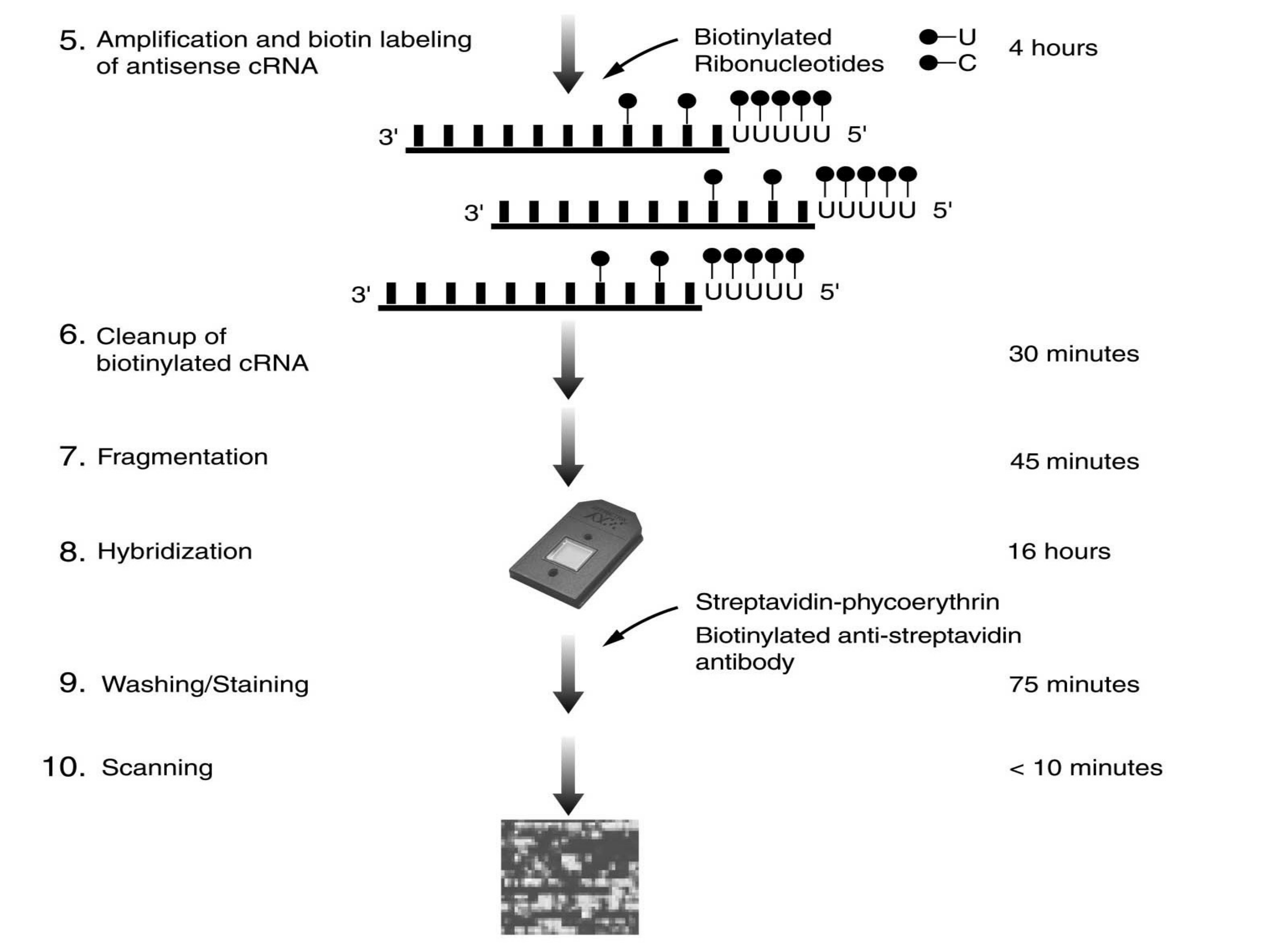
Eukaryotic Target Labeling for GeneChip® Probe Arrays



Starting RNA samples

Approximate Experiment Time

1. Primer hybridization — 15 minutes
2. Reverse transcription
   First strand cDNA synthesis — 1 hours 15 minutes
3. Second strand cDNA synthesis — 2 hours 30 minutes
4. Cleanup of
   double-strand cDNA — 30 minutes



5. Amplification and biotin labeling of antisense cRNA — 4 hours

6. Cleanup of biotinylated cRNA — 30 minutes

7. Fragmentation — 45 minutes

8. Hybridization — 16 hours

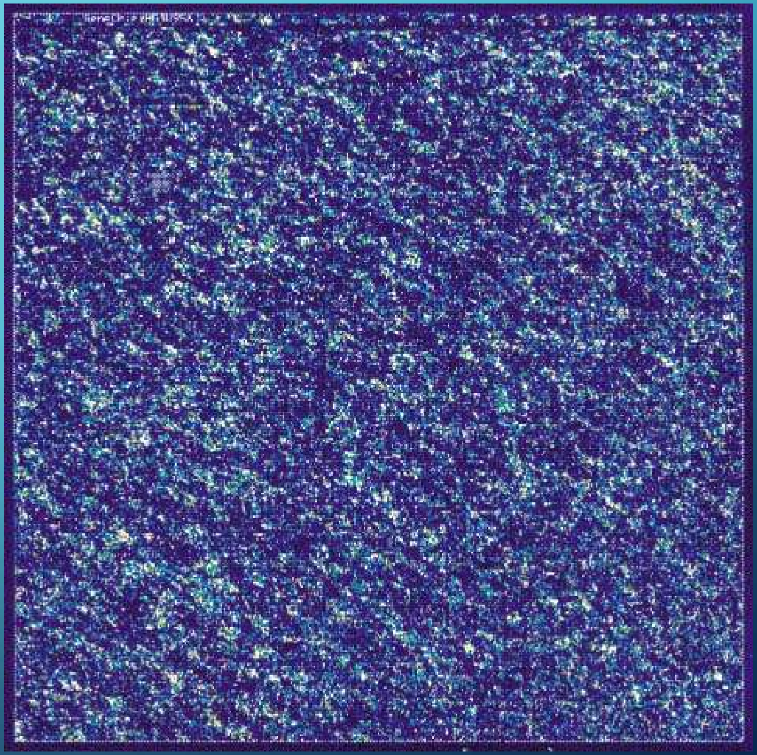
9. Washing/Staining — 75 minutes

10. Scanning — < 10 minutes

Biotinylated
Ribonucleotides
●—U
●—C
3'
UUUUU
5'
Streptavidin-phycoerythrin
Biotinylated anti-streptavidin antibody

Affymetrix GeneChip

- Scanning
- Quantification



Affymetrix *vs* tow color microarray

- Affymetrix chip
- ➢ Fluorescently tagged cRNA
- ➢ One chip per sample
- ➢ One for control
- ➢ One for each experiment
- Other methods include two dyes/one chip
- ➢ Red dye
- ➢ Green dye
- ➢ Control and experiment on same chip

Definitions

Probe – a single-stranded DNA oligonucleotide complementary to a specific sequence. Each probe cell consists of millions of probe molecules.

Probe Array – a collection of probes sets.

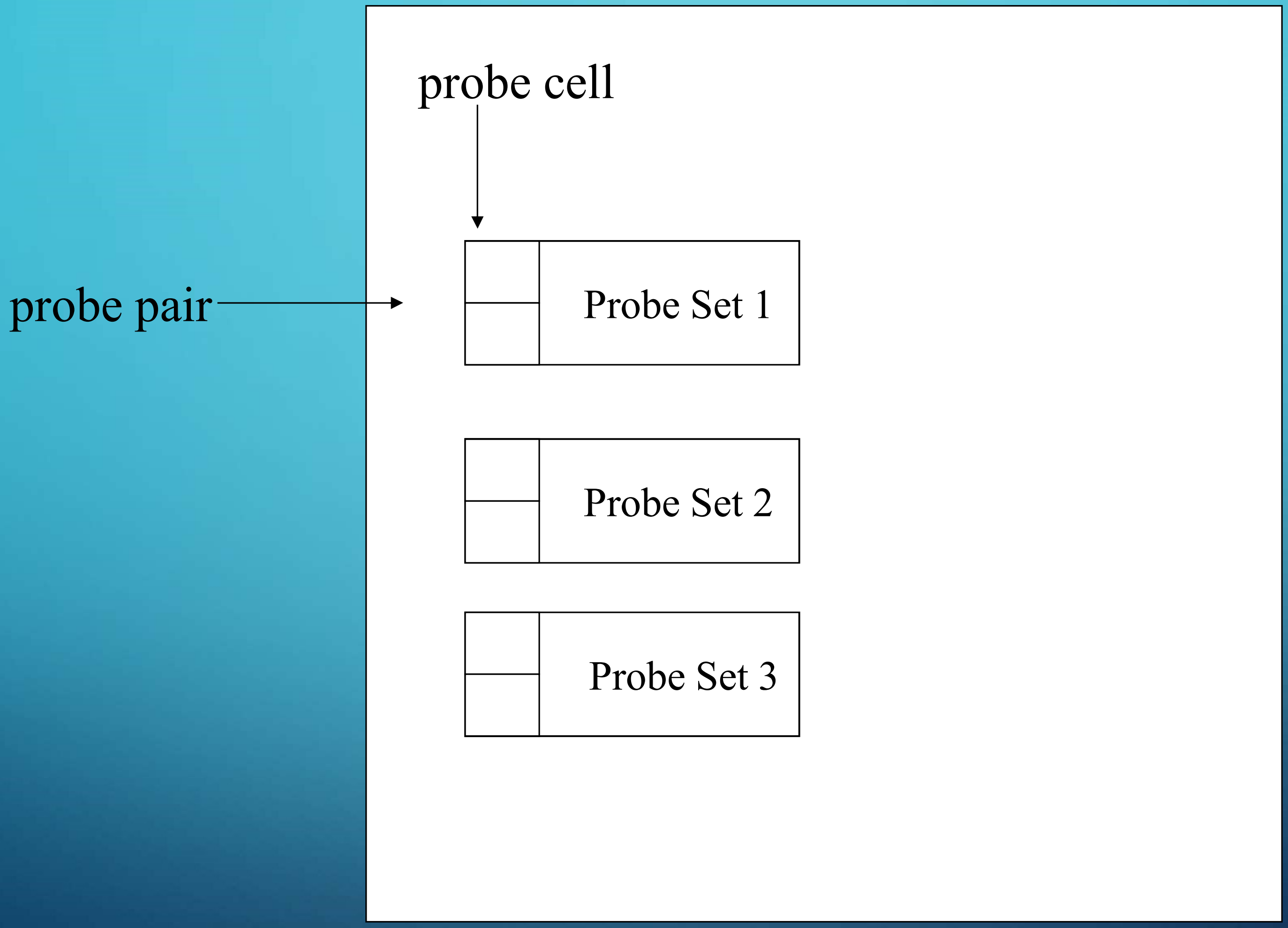
Probe Set – a set of probes designed to detect one transcript. 16-20 probe pairs. A 20 probe pair set is made up of 20 PM and 20 MM for a total of 40 probe cells.

Probe Pair – Two probe cells, a PM and its corresponding MM.

Perfect Match(PM) – probes that are designed to be complementary to the reference sequence.

Mis Match(MM) – probes that are designed to be complementary to the reference sequence except for 1 base.

Target – sequence from your sample.



probe cell

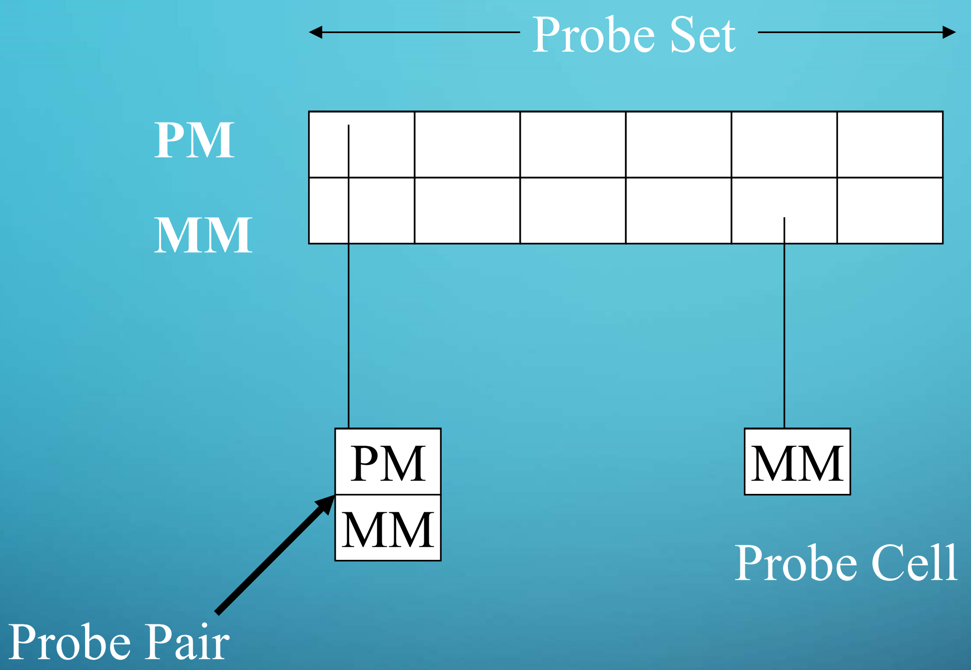
probe pair

Probe Set 1

Probe Set 2

Probe Set 3

Probe Array (chip)



Probe Set

PM

MM

PM

MM

MM

Probe Cell

Probe Pair

Each pair represents a different *subsequence* of the gene

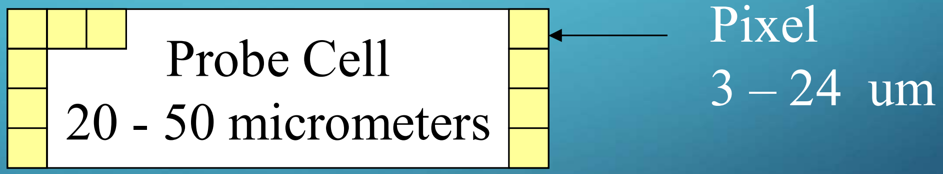
GeneChip Hierarchy

Probe Array = Chip

- Probe Set – 16-20 probe pairs(to detect particular gene)
  - Probe Pair
    - Probe Cell (Mis Match) 20
    - Probe Cell (Perfect Match) 20
      - Probes <= 25 bases
      - (millions of copies)
        - Pixels 24 sq. um

Probe – a single-stranded DNA oligonucleotide complementary to a specific sequence. Each probe cell consists of millions of the same probe molecules.

The intensity of each cell is an average of each of its scanned pixels.



Probe Cell

20 - 50 micrometers

Pixel

3 – 24 um

Affymetrix File Types

DAT file:

Raw (TIFF) optical image of the hybridized chip

CDF File (Chip Description File):

Provided by Affy, describes layout of chip

CEL File:

Processed DAT file (intensity/position values)

CHP File:

Experiment results created from CEL and CDF files

TXT File:

Probe set expression values with annotation (CHP file in text format)

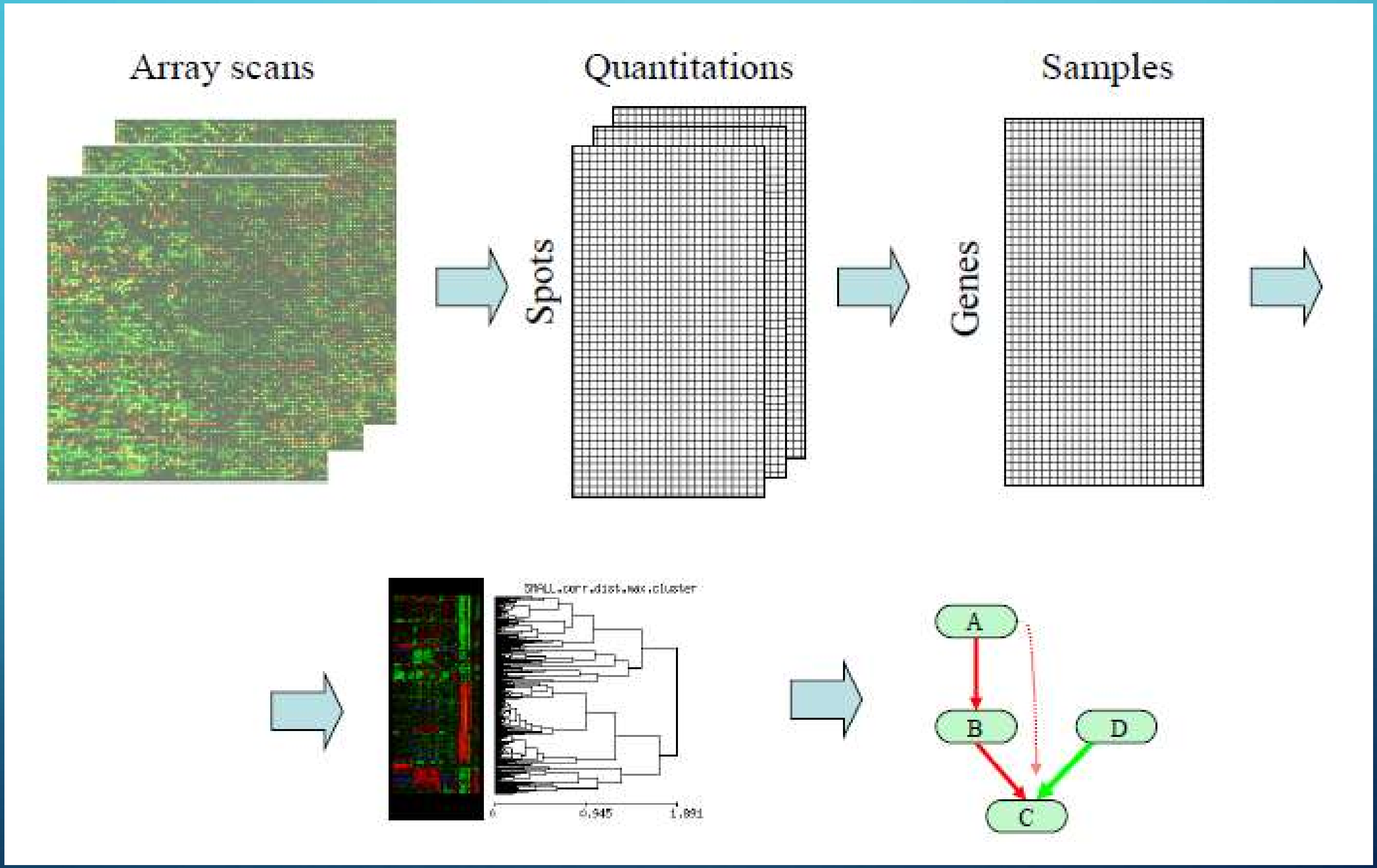
EXP File

Small text file of Experiment details (time, name, etc)

RPT File

Generated by Affy software, report of QC info

Steps in microarray data processing



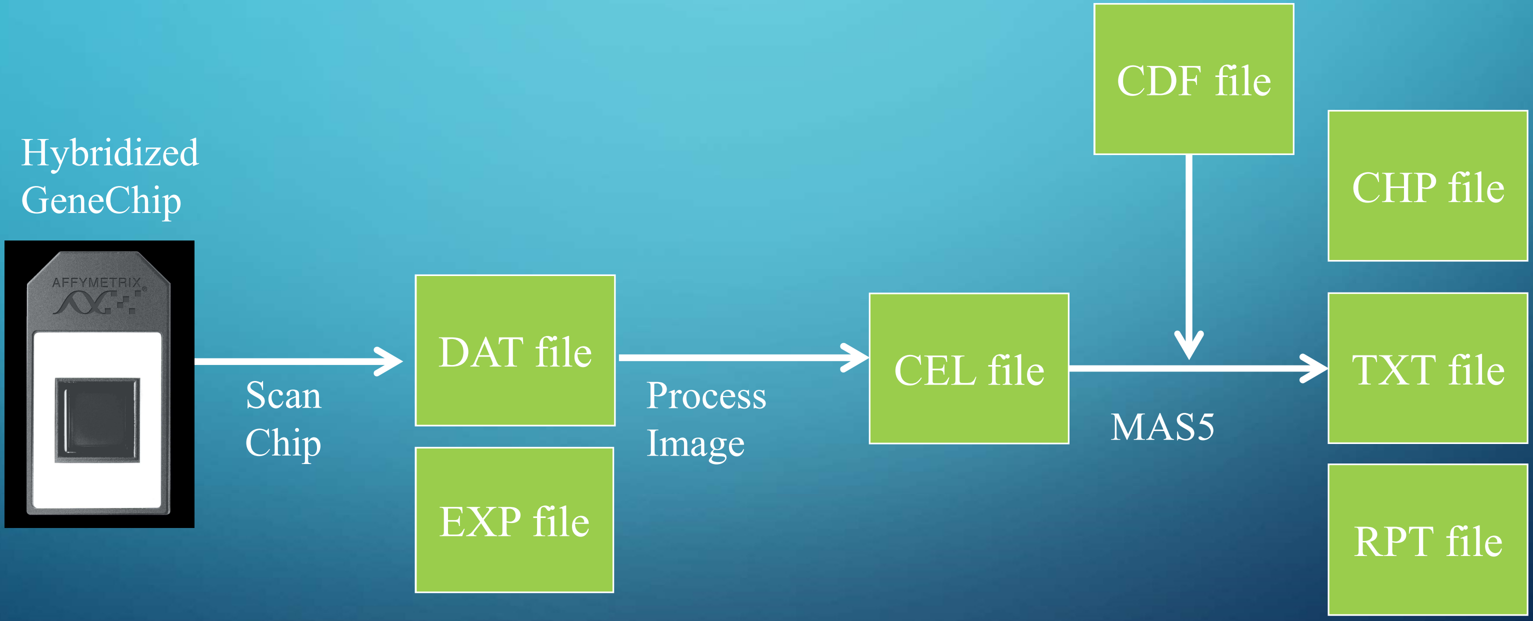
Gene expression

- A human organism has over 250 different cell types (e.g., muscle, skin, bone, neuron), most of which have identical genomes, yet they look different and do different jobs
- It is believed that less than 20% of the genes are ‘expressed’ (i.e., making RNA) in a typical cell type
- Apparently the differences in gene expression is what makes the cells different

Some questions for the golden age of genomics

- How gene expression differs in different cell types?
- How gene expression differs in a normal and diseased (e.g., cancerous) cell?
- How gene expression changes when a cell is treated by a drug?
- How gene expression changes when the organism develops and cells are differentiating?
- How gene expression is regulated – which genes regulate which and how?

AFFYMETRIX DATA FLOW



Hybridized GeneChip

Scan Chip

DAT file

EXP file

Process Image

CEL file

CDF file

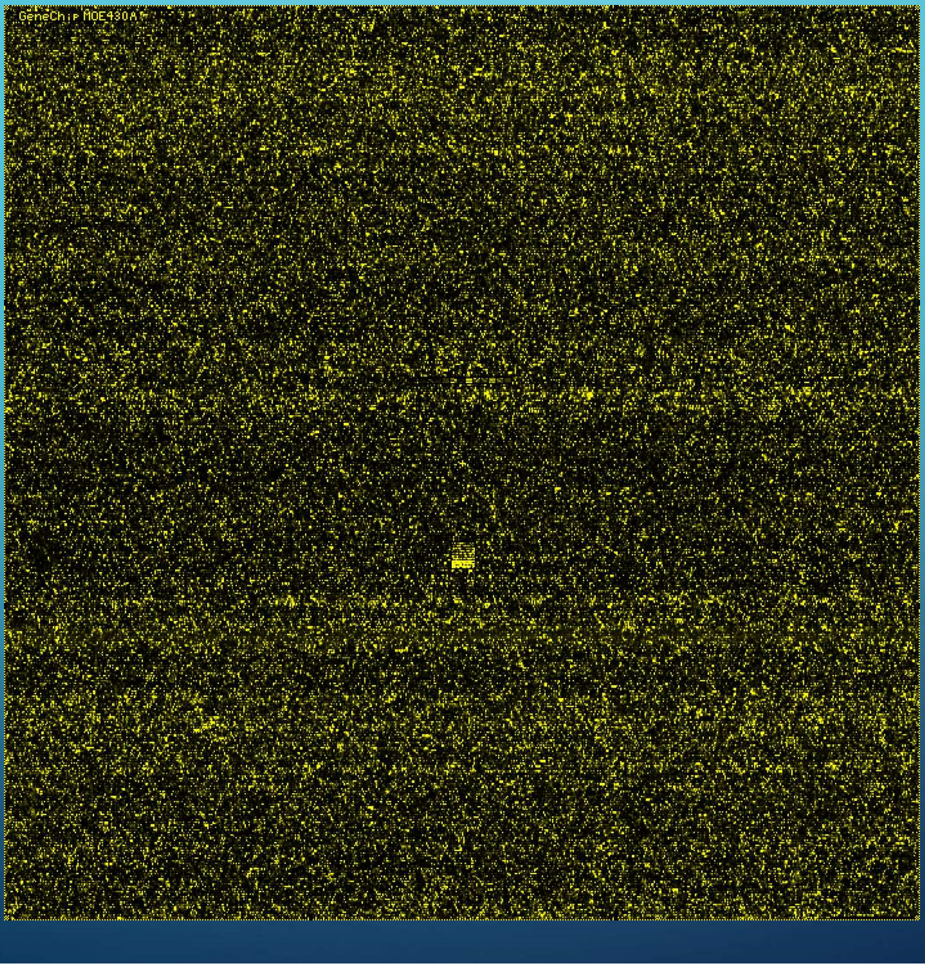
MAS5

CHP file

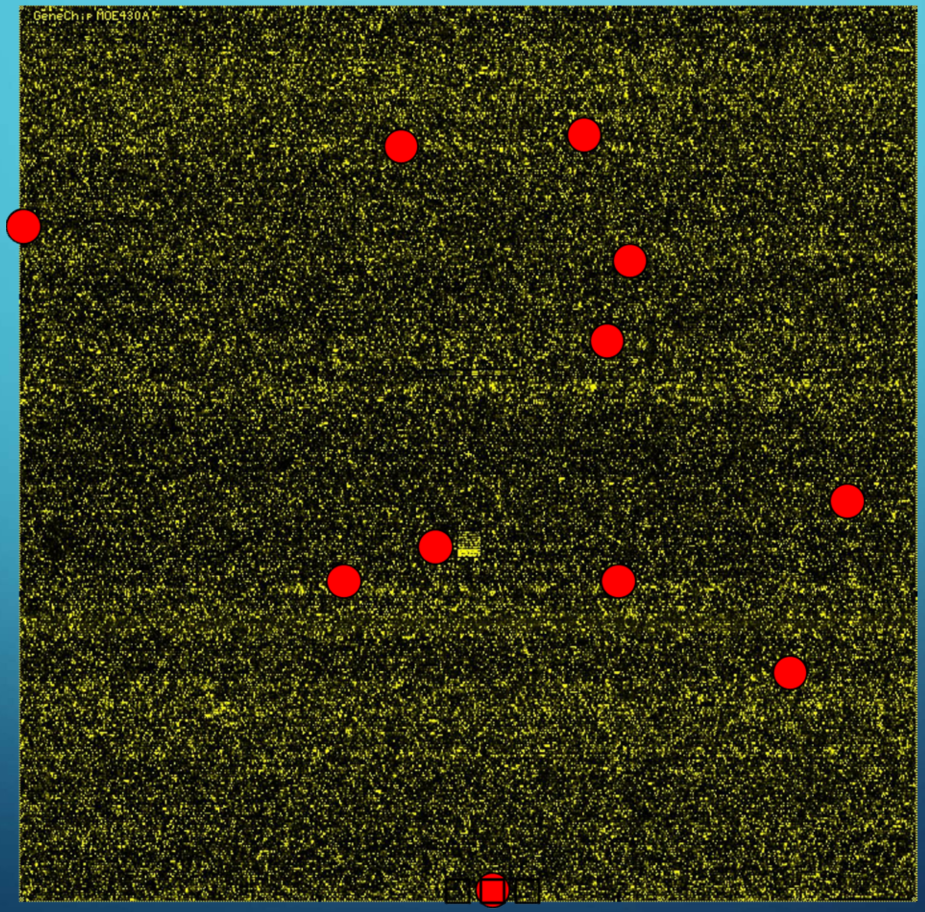
TXT file

RPT file

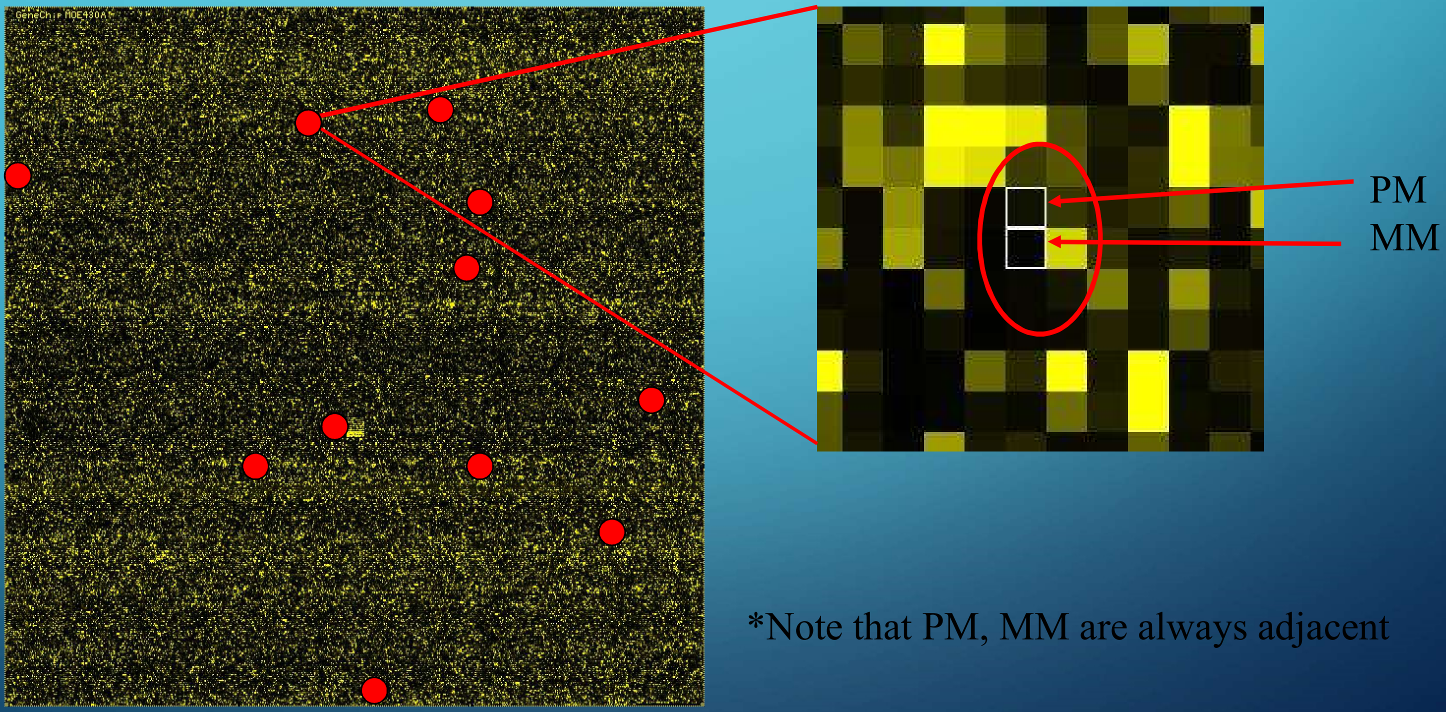
Affymetrix Chip Pseudo-image



1415771_at on MOE430A

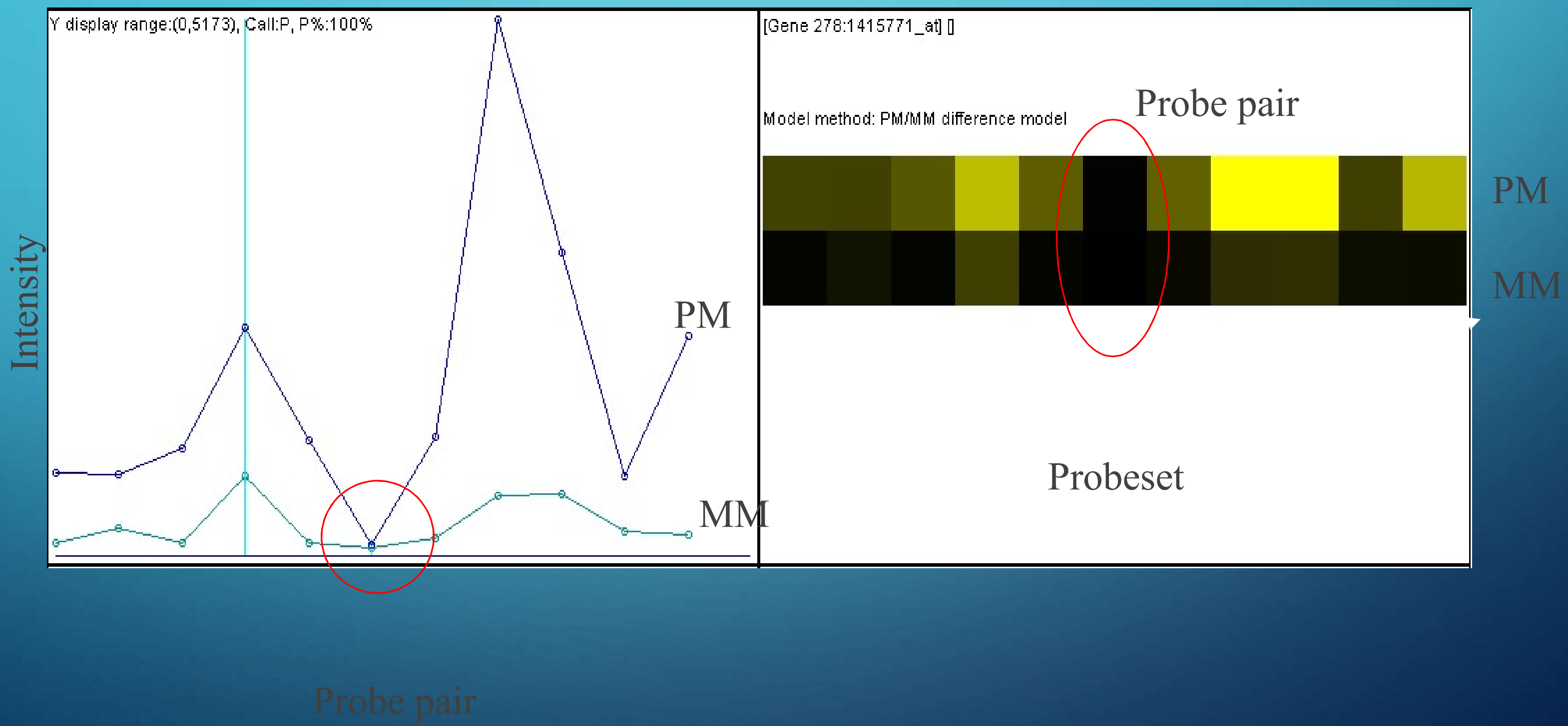


1415771_at on MOE430A



*Note that PM, MM are always adjacent

1415771_at on MOE430A


Y display range:(0,5173), Call:P, P%:100%
Intensity
PM
MM
Probe pair
[Gene 278:1415771_at] []
Model method: PM/MM difference model
Probe pair
PM
MM
Probeset

Intensity to Expression

- Now we have thousands of intensity values associated with probes, grouped into probe sets.
- How do you transform intensity to expression values?
  - ➢ Algorithms
    - ✓ MAS5

      Affymetrix proprietary method
    - ✓ RMA/GCRMA

      Irizarry, Bolstad

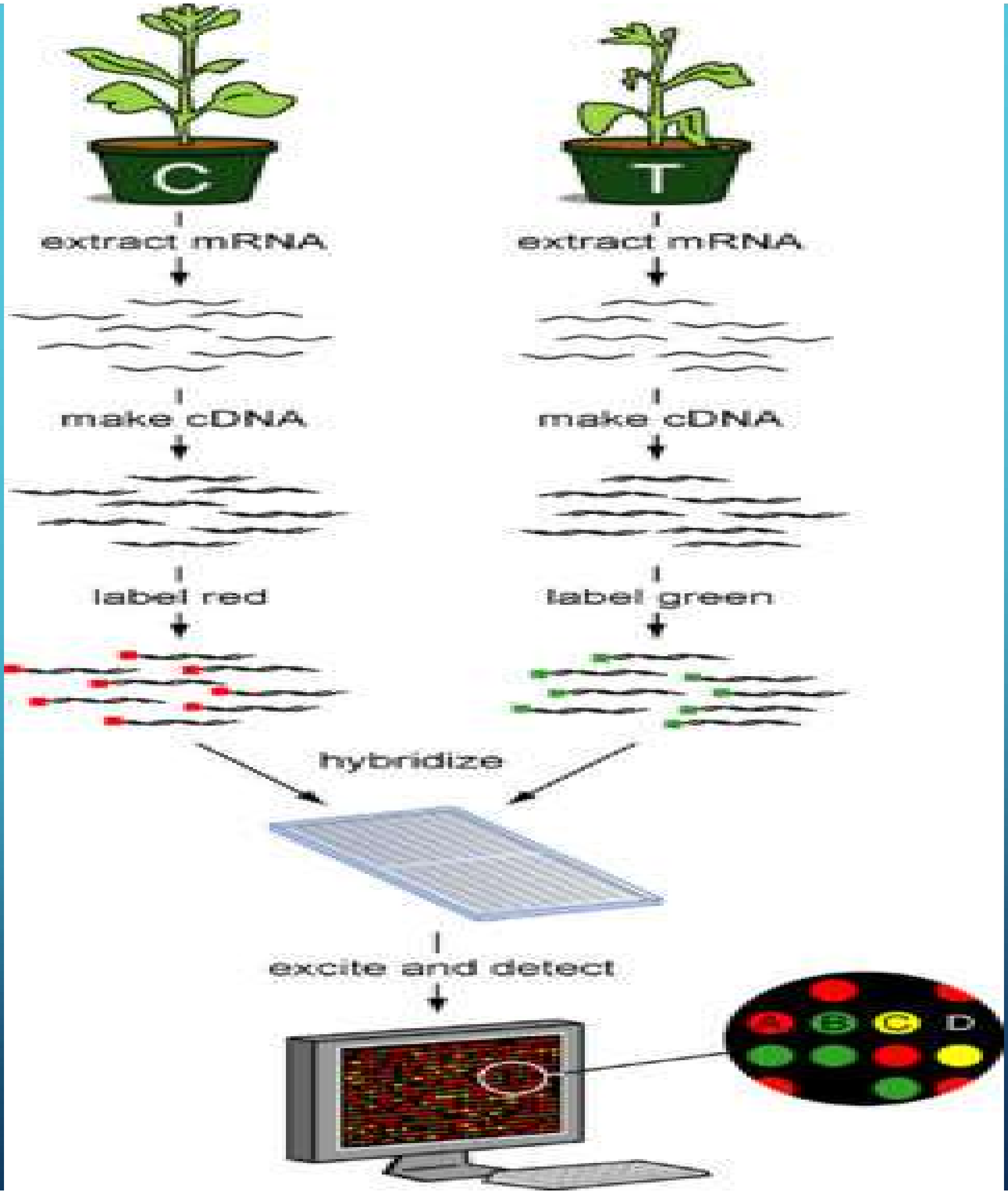
    ..many others
- Often called “normalization”

Common elements of different techniques

- All techniques do the following:
  - Background adjustment
  - Scaling
  - Aggregation
- The goal is to remove non-biological elements of the signal

Introduction to the Statistical Analysis of Two-Color Microarray Data

- Up to 100,000 genes on a single1.5 cm × 5 cm slide
- huge amounts of data that require the use of biostatistics for analysis and validation
- elementary ideas behind the statistical analysis of microarray data
- Why would a researcher want to do microarray experiments?
- For example, which genes become active if a plant is subjected to prolonged drought stress?


C
T
extract mRNA
extract mRNA
make cDNA
make cDNA
label red
label green
hybridize
excite and detect
A
B
C
D

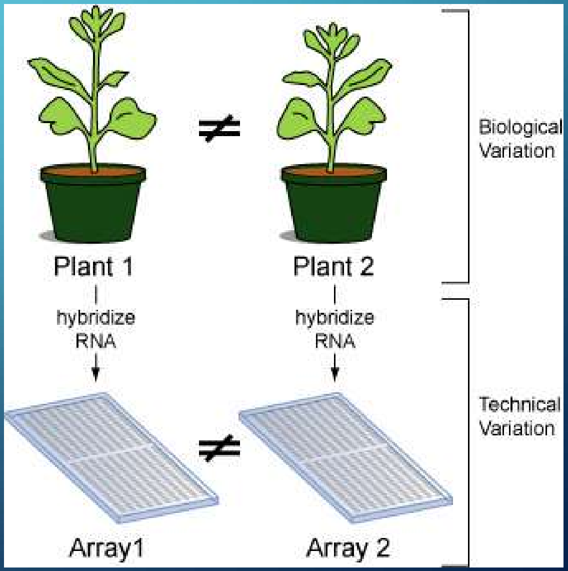
Uncertainty in statics

- Many experiments that address the same problem or question can differ in their outcomes when conducted by different people or with different materials
- two main sources of variation

➢ Biological Variation

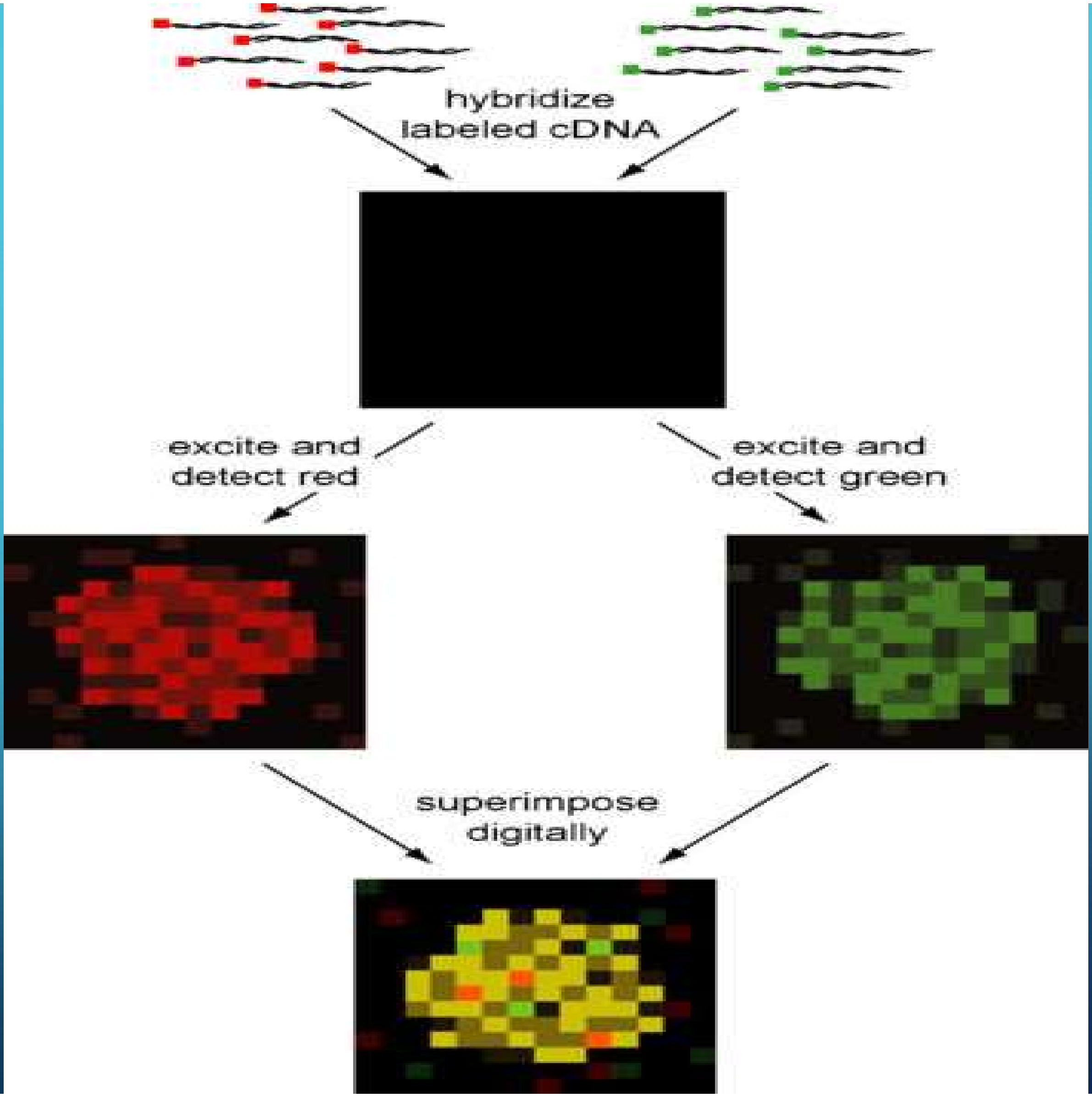
➢ Technical Variation

(Technical variation refers to the differences resulting from human and manufacturing error)



Overview of Microarray Experiments

- goal of microarray experiments: compare the gene expression levels of a treatment group with those of a control group
- mRNA is extracted from the cells
- The samples are labeled with red and green fluorescent dyes
- The scanner divides the features on the array into pixels and for each pixel a computer records the scanned red and green intensity
- Usually, the spots for each microarray cell are not uniform
- The spots with very low intensity are called the background


hybridize
labeled cDNA
excite and
detect red
excite and
detect green
superimpose
digitally

Microarray Data Analysis

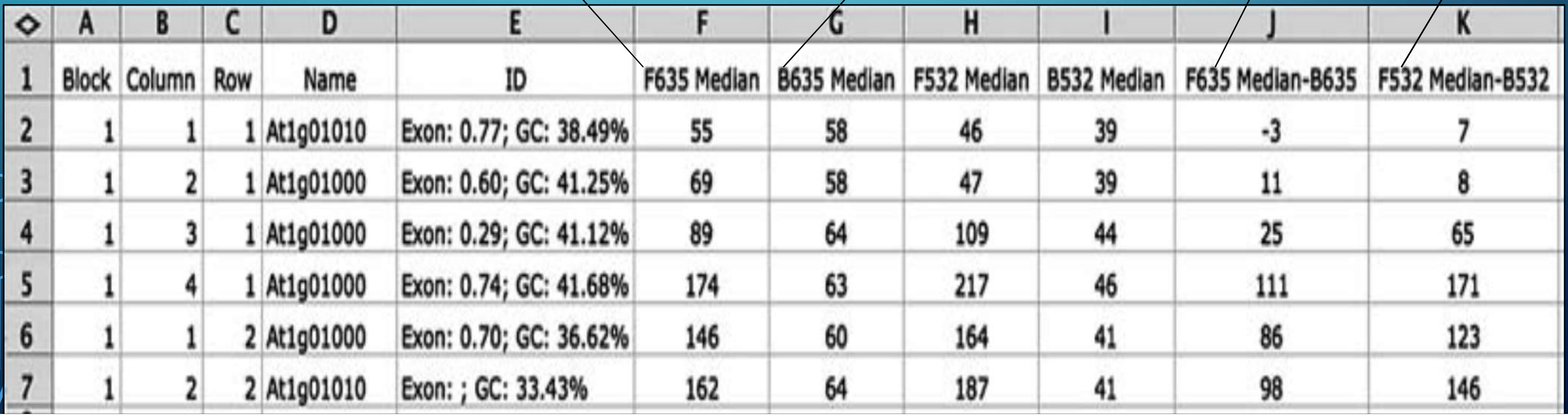
- The first three columns tell us the position of the scanned spot
- D ⟶ name of the gene
- E ⟶ contains information pertaining to the exact part of the gene that was used
- background, in essence, the probe that binds to the silicate of the glass slide falsely increasing the signal of each spot

F for foreground median

B for background median

red wavelength

green wavelength



◇	A	B	C	D	E	F	G	H	I	J	K
1	Block	Column	Row	Name	ID	F635 Median	B635 Median	F532 Median	B532 Median	F635 Median-B635	F532 Median-B532
2	1	1	1	At1g01010	Exon: 0.77; GC: 38.49%	55	58	46	39	-3	7
3	1	2	1	At1g01000	Exon: 0.60; GC: 41.25%	69	58	47	39	11	8
4	1	3	1	At1g01000	Exon: 0.29; GC: 41.12%	89	64	109	44	25	65
5	1	4	1	At1g01000	Exon: 0.74; GC: 41.68%	174	63	217	46	111	171
6	1	1	2	At1g01000	Exon: 0.70; GC: 36.62%	146	60	164	41	86	123
7	1	2	2	At1g01010	Exon: ; GC: 33.43%	162	64	187	41	98	146

Adjustment

- For some genes the background-corrected intensity is negative
- for some spots the median spot intensity is actually lower than the median background intensity
- negative gene expression values do not make biological sense (a gene can have no expression but not negative expression!)
- Usually, negative values are replaced by zeros or small positive constants

Normalization

- The results of a microarray experiment are obviously influenced by technical variation
- Normalization means to mathematically manipulate the data to make it uniform
- In microarray experiments, the results are reported as log ratios of the two intensity
- If *Gi* represents the green intensity for gene *i* and *Ri* represents the red intensity for gene *i*, then two quantities commonly used are

$$M_i = \log_2 \left(\frac{R_i}{G_i} \right)$$

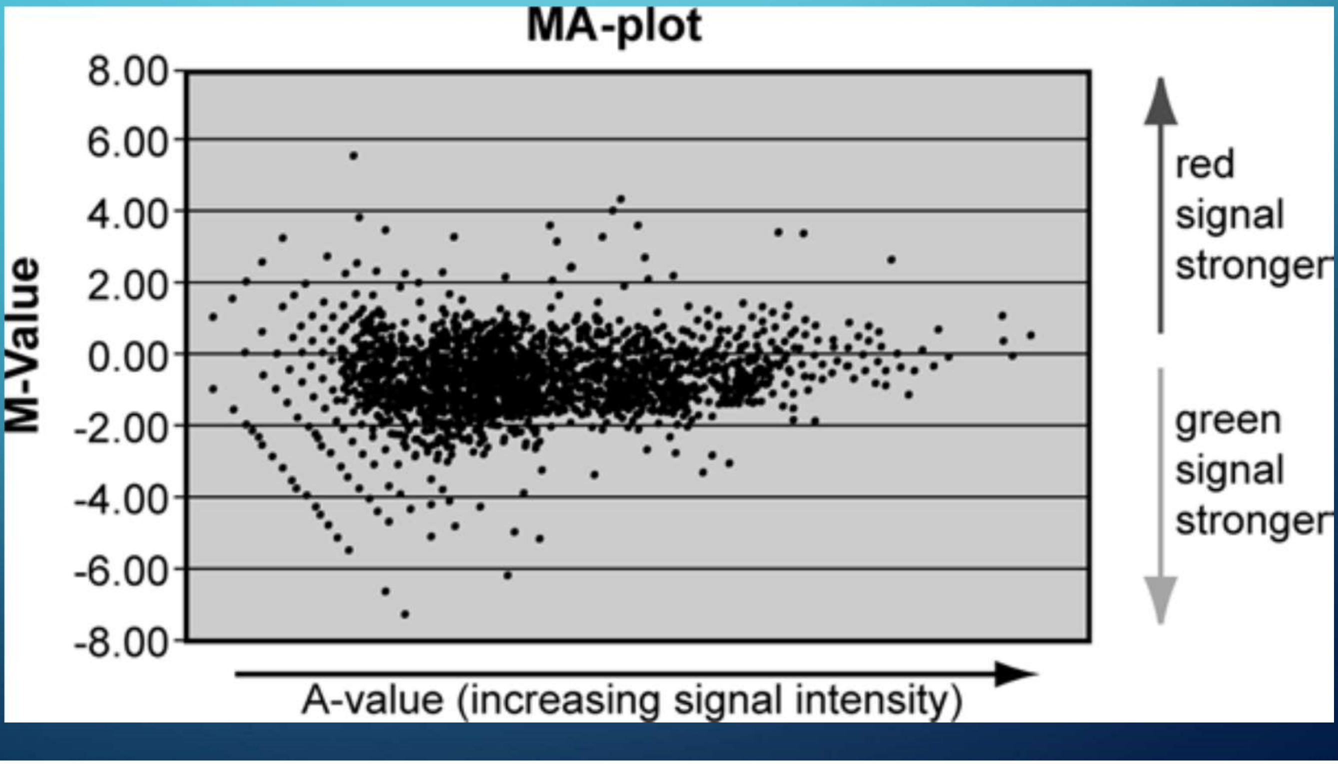
$$A_i = \frac{1}{2} (\log_2 R_i + \log_2 G_i)$$

- The quantity *Mi* (the log ratio) describes the relationship between the two groups
- (the intensity in the red- and green labeled group is the same, then *Mi* will be zero. If the red intensity is twice as big as the green then *Mi* will be equal to 1. If, on the other hand, the green intensity is twice as big as the red, then *Mi* will be equal to –1)
- The quantity *Ai* describes the overall intensity observed for gene i

MA plot

For every feature *i* on the array, the values *Mi* and *Ai* are computed and are plotted in an xy plot

Highly up- or down regulated genes are above or below the x-axis.



MA plot

- Experimenters are usually most interested in those genes with:

✓ high fold change (large positive or negative *M* value)

✓ high intensity (large *A* value)

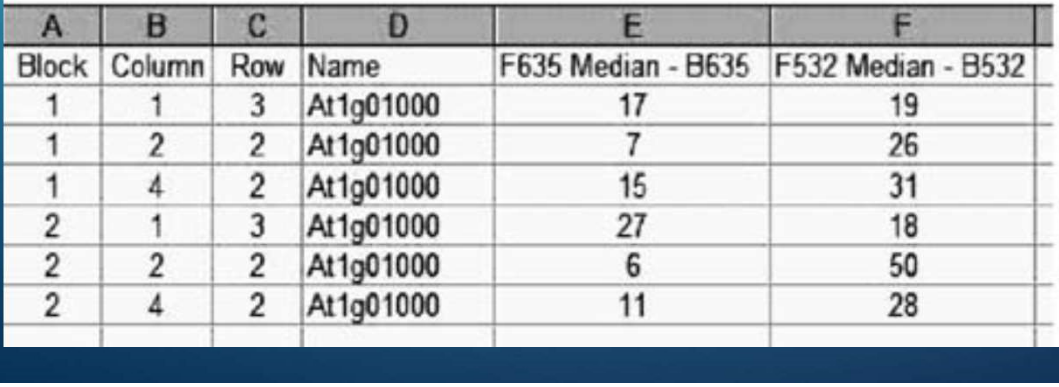
- For spots that have high fold change but very low intensity it is hard to distinguish if the fold change is due to a biological effect or due to technical variation in the measurements
- On average, we would like the intensities for both dyes to be about the same
- That means, on average, the log ratios *Mi* should be about zero

Normalization

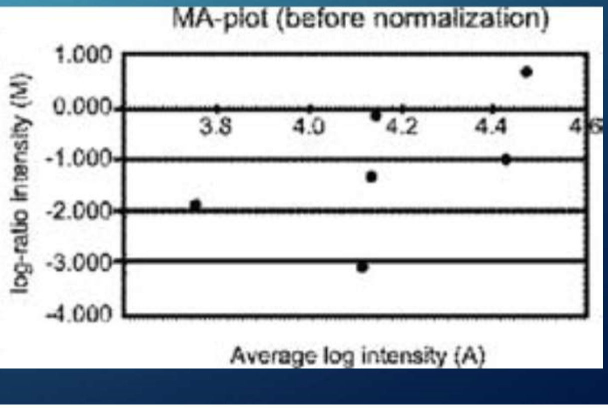
- The main purpose of normalization is to mathematically remove as much systematic variation not caused by biological effects as possible

- Normalization means that one computes the average *M* value for all genes spotted on the array and then makes sure that the average will be zero

- In the small-scale example below, the six features shown all represent the same gene

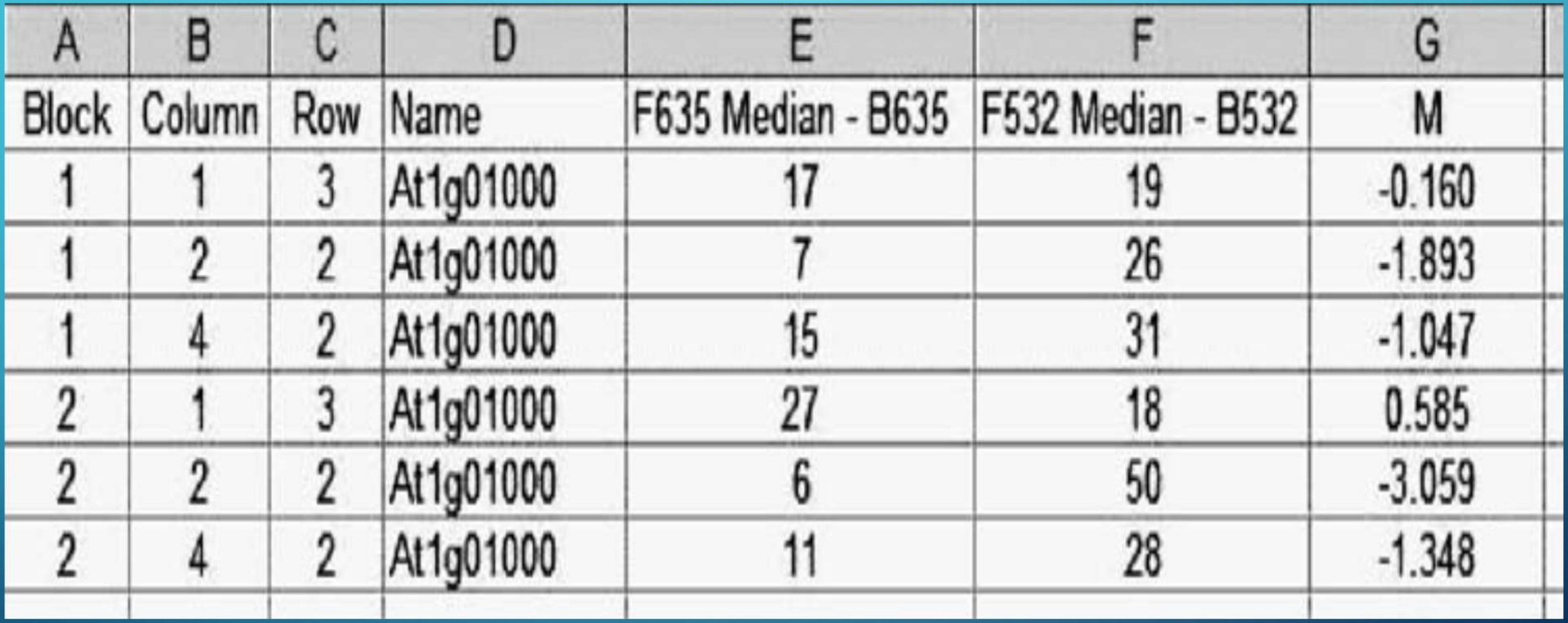


A	B	C	D	E	F
Block	Column	Row	Name	F635 Median - B635	F532 Median - B532
1	1	3	At1g01000	17	19
1	2	2	At1g01000	7	26
1	4	2	At1g01000	15	31
2	1	3	At1g01000	27	18
2	2	2	At1g01000	6	50
2	4	2	At1g01000	11	28



Normalization

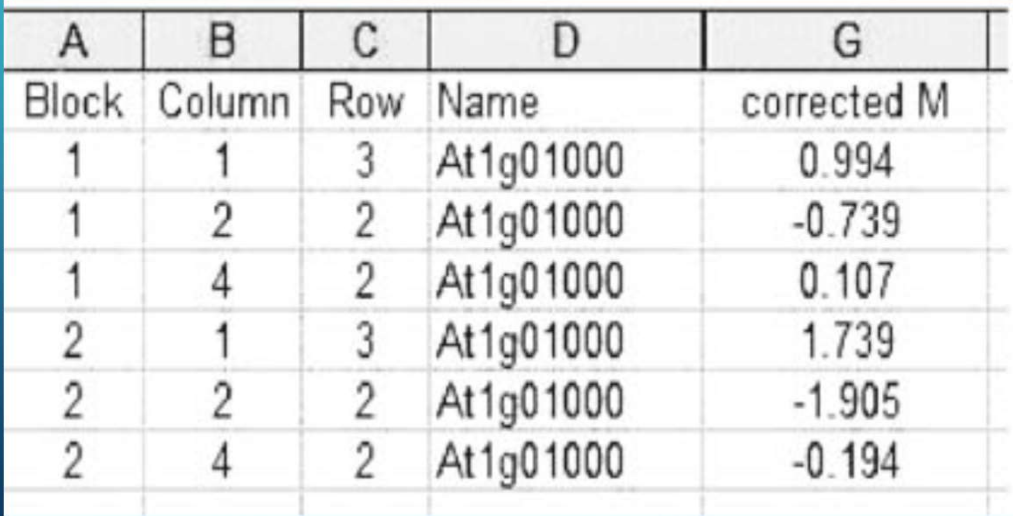
In the first step the log ratios of red (635) to green (532) signal are computed



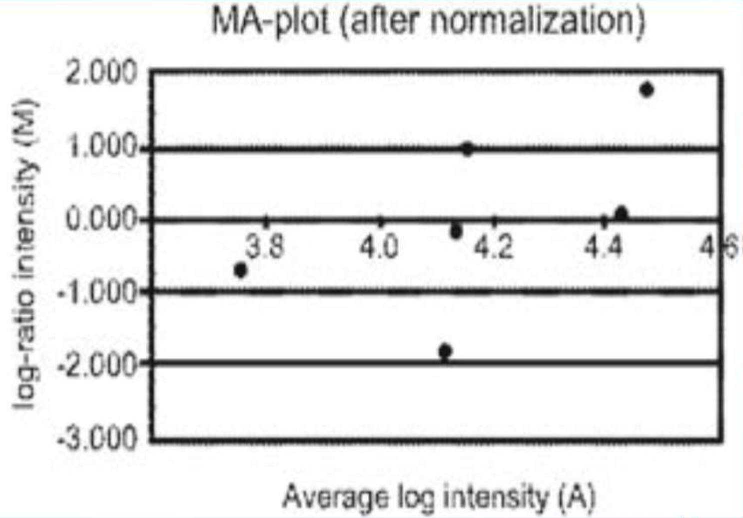
A	B	C	D	E	F	G
Block	Column	Row	Name	F635 Median - B635	F532 Median - B532	M
1	1	3	At1g01000	17	19	-0.160
1	2	2	At1g01000	7	26	-1.893
1	4	2	At1g01000	15	31	-1.047
2	1	3	At1g01000	27	18	0.585
2	2	2	At1g01000	6	50	-3.059
2	4	2	At1g01000	11	28	-1.348

Normalization

The goal of normalization is to assure that the average of the *M* values is set to zero Computing the average of the six *M* values above yields –1.154. This average value (here –1.154) is then subtracted from the *M* values for all genes in the example above to “correct” them



A	B	C	D	G
Block	Column	Row	Name	corrected M
1	1	3	At1g01000	0.994
1	2	2	At1g01000	-0.739
1	4	2	At1g01000	0.107
2	1	3	At1g01000	1.739
2	2	2	At1g01000	-1.905
2	4	2	At1g01000	-0.194



Normalization by Dye-Swap Design

A better way to deal with uneven binding of the dyes to certain genes

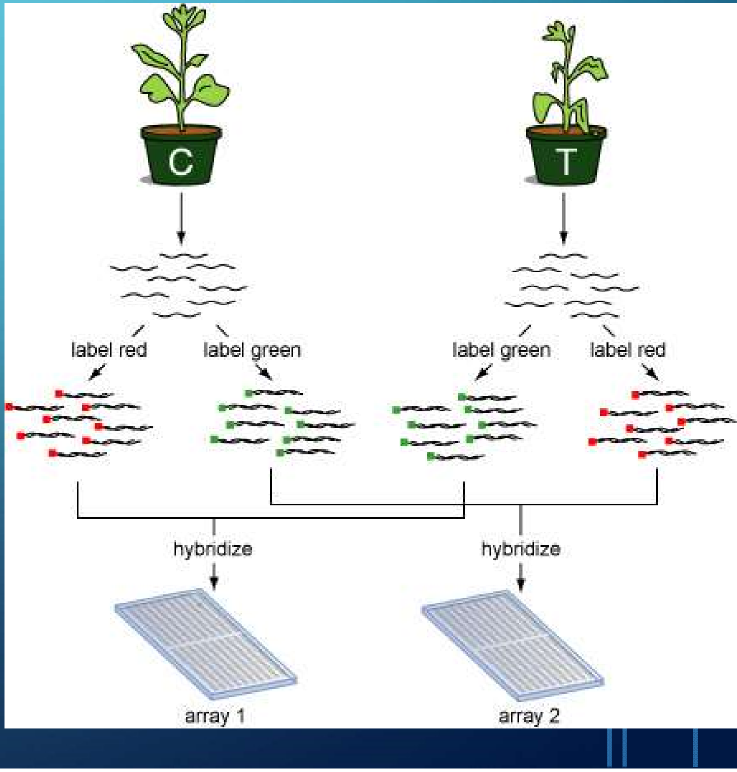
In them, the probe from each group (treatment and control) is split into two portions and labeled with different dyes

compute M values for the two arrays once as red/green log ratio and for the other slide as the green/red log ratio

Then average the M values from the two arrays for each feature:

$$M_i = \frac{1}{2}\left(M_i^{(1)} + M_i^{(2)}\right)$$

Here $M_i^{(1)}$ is the log ratio for feature i on array 1 and $M_i^{(2)}$ is the log ratio for the same feature on array 2

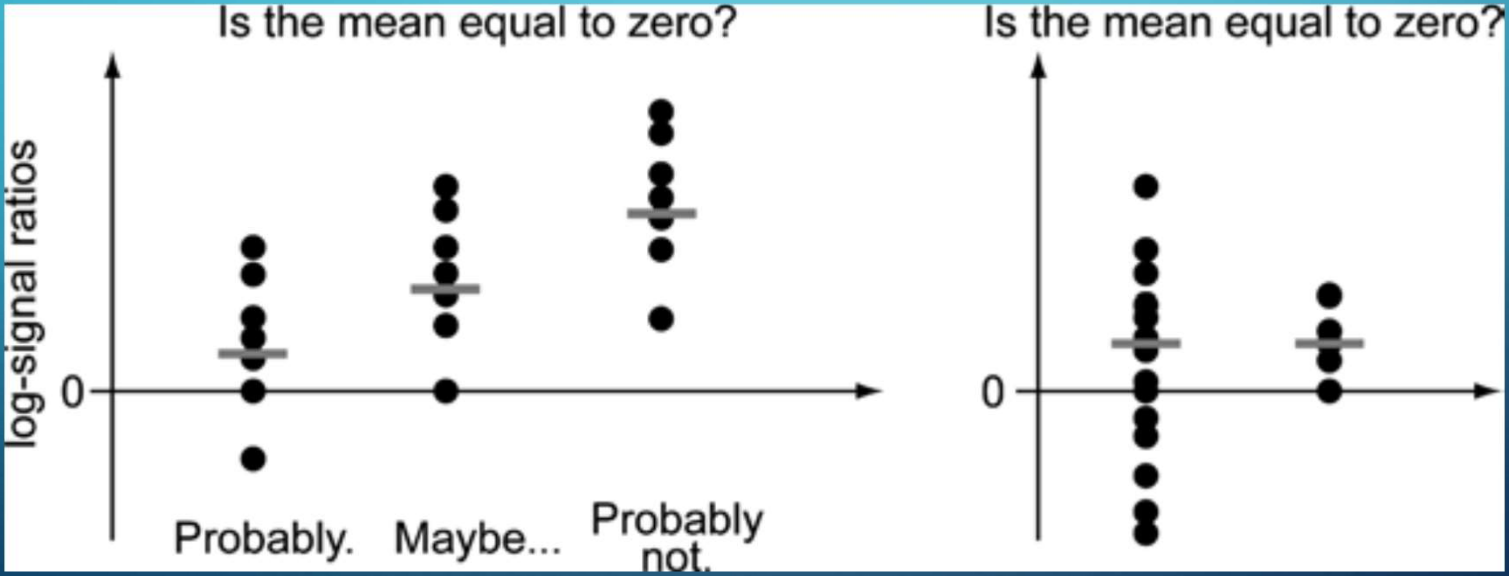


Drawing Conclusions

- the goal of microarray experiments to identify genes that become either more or less expressed
- For every gene on the microarray we want to decide whether the expression levels in the two experimental groups are (significantly) different or not
- If the green and red intensities are different, their quotient R/G is not equal to one. If the quotient is not equal to one, then the *M* value for the spot is not equal to zero
- The challenge in analysis is to determine whether the observed differences are due to biological or technical variation
- This is achieved by a statistical analysis of the *M* values
- How far away from zero do these *M* values have to be so that we are convinced that the result is due to a real difference

Drawing Conclusions

- Only if the distance of the absolute value of the mean is large compared to the variation within the measurements will we declare the mean significantly different from zero
- If the variation is small, we may be more inclined to assume a non-zero mean than if the variation is large for the same absolute value of mean.



Statistical Decision Making

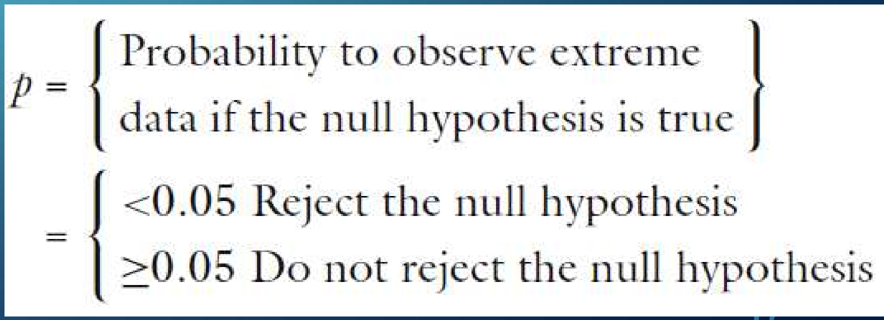
- Hypothesis tests are an important tool for statistical decision making
- They are used to answer a “Yes/No” question about a population
- Suppose that we repeatedly observe one gene under both treatment conditions
- If the average difference of expression levels is large, then the question that has to be answered is:
- whether it is plausible that this large difference is explainable only by biological and technical noise?
- If the answer is “Yes,” then we have no reason to be very excited
- but if the answer is “No,” then the difference that we observed is likely due to a treatment effect

Hypothesis Testing

- A statistical hypothesis test always follows the same scheme
- This is done in the form of two opposing statements about a population parameter
- The null hypothesis is always of the form “there is nothing unusual happening here”
- In the case of a microarray experiment this translates into “the gene is not differentially expressed.”
- The alternative hypothesis is a contradiction to the null hypothesis
- The scientist now takes on the role of skeptic
- If the null hypothesis were true, and there truly is no effect, we will compute the probability to see an outcome as extreme as the one we observed purely through error variation
- “Extreme” in this context is any observation (such as a large difference in gene expressions) that supports the alternative hypothesis more strongly than the null hypothesis

Hypothesis Testing

- If it is unlikely to see an effect as extreme or more extreme than the one observed from error variation alone, then we conclude that there likely is a treatment effect and we then reject the null hypothesis in favor of the alternative hypothesis
- This does not mean that the alternative has been proved to be true
- The probability to observe an apparent "effect" (i.e., a large difference between treatment and control measurements) if there is only nuisance variation is called the p-value of a hypothesis test
- The smaller a p-value is, the less likely it is to observe data such as the one you observed in the experiment if the null hypothesis were true



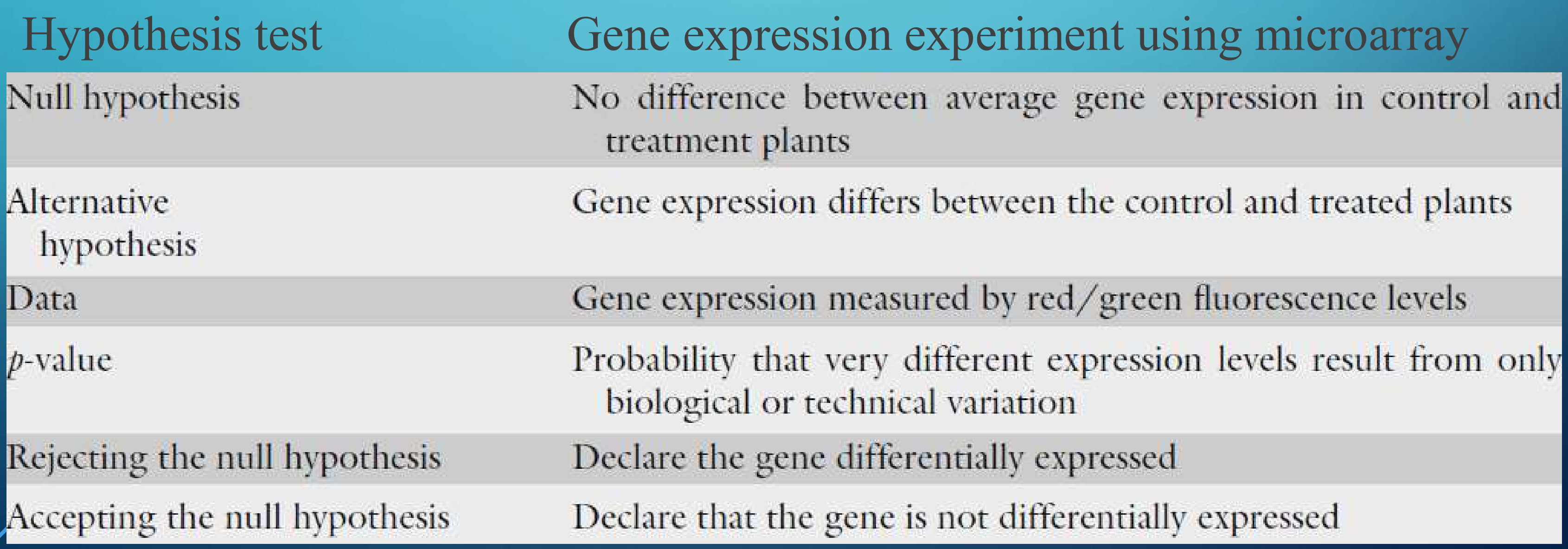
$$p = \left\{ \begin{array}{l} \text{Probability to observe extreme} \\ \text{data if the null hypothesis is true} \end{array} \right\}$$

$$= \left\{ \begin{array}{l} <0.05 \text{ Reject the null hypothesis} \\ \geq 0.05 \text{ Do not reject the null hypothesis} \end{array} \right.$$

Hypothesis Testing

- To identify differentially expressed genes in a microarray experiment

  separate hypothesis tests are performed for each gene spotted on the array



Hypothesis test	Gene expression experiment using microarray
Null hypothesis	No difference between average gene expression in control and treatment plants
Alternative hypothesis	Gene expression differs between the control and treated plants
Data	Gene expression measured by red/green fluorescence levels
p-value	Probability that very different expression levels result from only biological or technical variation
Rejecting the null hypothesis	Declare the gene differentially expressed
Accepting the null hypothesis	Declare that the gene is not differentially expressed

Hypothesis Test for Log Ratios

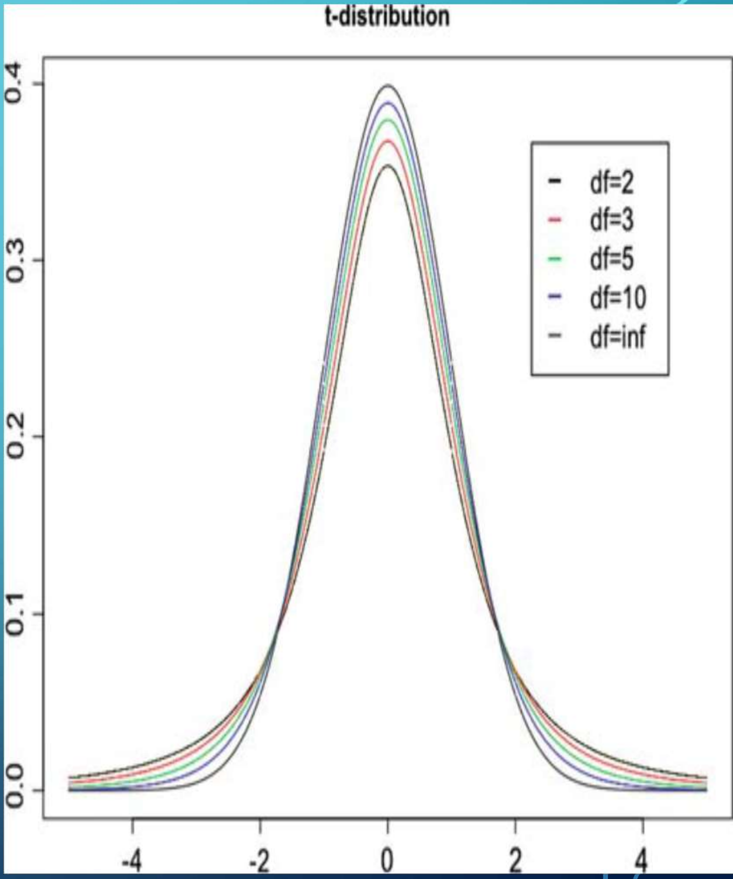
- The data file will contain one *M* value for every spot on the array
- Ideally, each gene is spotted several times on the same array, so that there are several *M* values for each gene
- To conclude whether a gene is expressed differently in the two groups, we will decide whether the *M* values for that gene are close to zero (on average) or not
- To make this decision, we will also have to take the variance of the observations into account
- A t-test will be used to decide whether a gene is differentially expressed
- Suppose that a random characteristic with mean zero is measured repeatedly. For *n* measurements *x1, . . . , xn* with average *x* and standard deviation *s*, the quantity

$$t = \frac{\overline{x}}{\sqrt{\frac{s^2}{n}}}$$

has a *t* distribution with *df = n − 1*

Hypothesis Test for Log Ratios

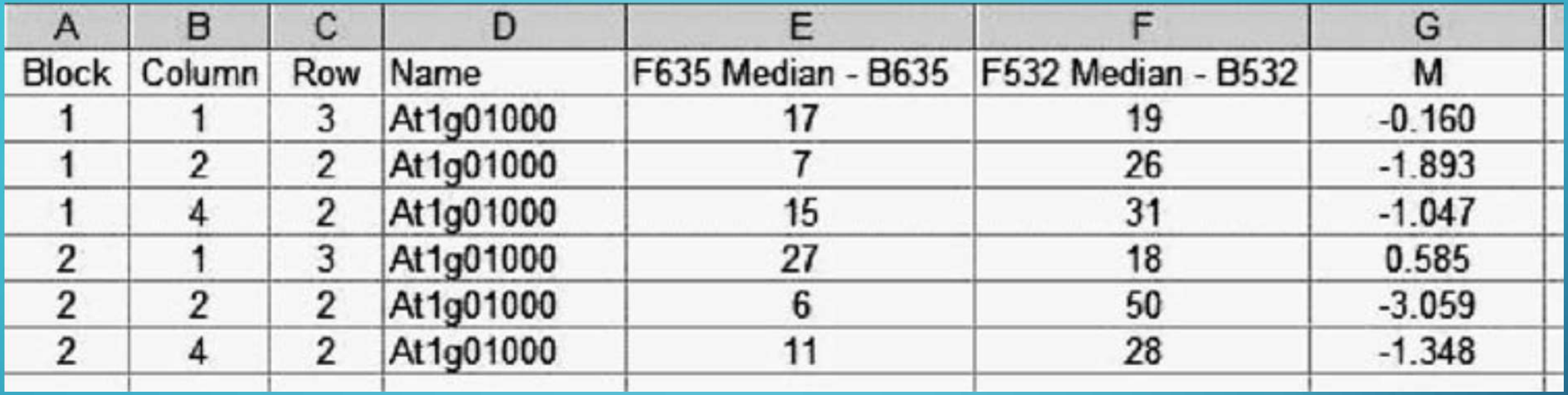
- Since we assumed the characteristic to have mean zero, the “normal” (or typical) values are those close to zero
- The unusual values are the ones in the tails of the distribution, either large positive or large negative numbers
- Large positive values of the test statistic mean that the log ratio is positive, which means that the red intensity is much higher than the green
- Large negative values of the test statistic mean that the log ratio is negative, which means that the green intensity is much higher than the red



How Large Is “Large”?

- How large (or small) will a test statistic value need to be so that we can call it unusual?
- Most researchers work with a significance level of 5%
- They call an observation unusual, if its p-value is smaller than 5%
- That means that the test statistic value falls into the outer 5% tail area of the distribution in the graph of the t distribution above
- If that occurs, one may safely argue that the two values (for red and green) differ from each other in a “statistically significant” manner between the treatment and the control group

t-test for Microarray Data



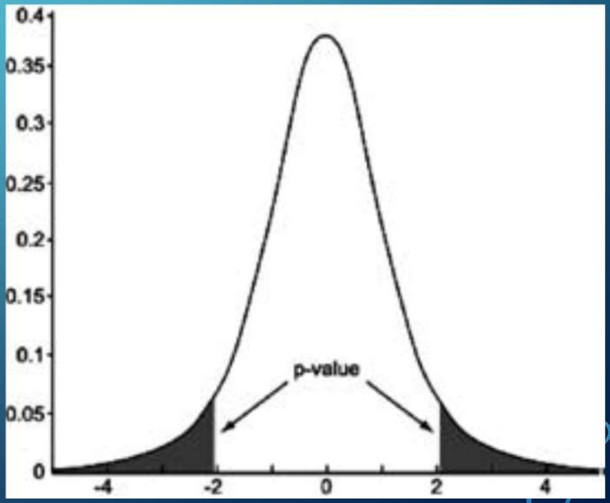
A	B	C	D	E	F	G
Block	Column	Row	Name	F635 Median - B635	F532 Median - B532	M
1	1	3	At1g01000	17	19	-0.160
1	2	2	At1g01000	7	26	-1.893
1	4	2	At1g01000	15	31	-1.047
2	1	3	At1g01000	27	18	0.585
2	2	2	At1g01000	6	50	-3.059
2	4	2	At1g01000	11	28	-1.348

- We have six *M* values for the gene At1g01000
- We can compute the average of the six observations $x = -1.15$ and the standard deviation $s = 1.28$. $n = 6$, since we have six observations
- Now we can compute the value of the test statistic as

$$t = \frac{\overline{x}}{\sqrt{\frac{s^2}{n}}} = \frac{-1.15}{\sqrt{\frac{1.28^2}{6}}} = -2.08$$

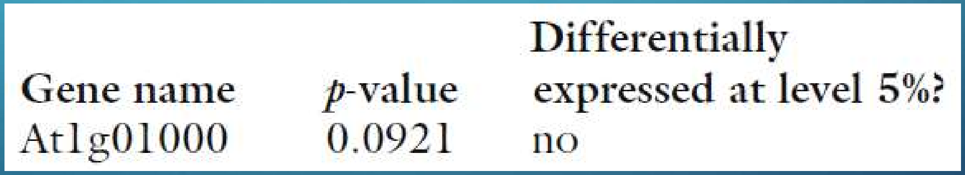
t-test for Microarray Data

- To find the p-value, we have to find the percentage of cases, in which the t-test statistic with *df = 5* would take on more extreme values than the –2.08 that we observed
- Extreme values are the ones far away from zero
- The area under the t distribution curve corresponding to the extreme values (smaller than –2.08 or larger than 2.08) is shaded in the graph below
- In the past, these values had to be looked up in tables. Today, Excel and other software programs have them stored in their statistics package
- In our example, the exact p-value (shaded tail area of the distribution) is *0.0921* or *9.21%*



t-test for Microarray Data

- What conclusion can we draw?
- The p-value is the probability to observe data as extreme/unusual as the one we saw if the gene expression in the two groups were the same
- Our p-value of 9.21% is quite large (bigger than 5%)
- That means that we would get observations such as these by random chance and not due to real difference in gene expression almost 10% of the time
- Hence, our data are nothing unusual and we cannot reject the null hypothesis (equal expression in both groups) for gene At1g01000



Gene name	*p*-value	Differentially expressed at level 5%?
At1g01000	0.0921	no

- To determine the p-values for the other genes spotted on the microarray, repeat the steps described above

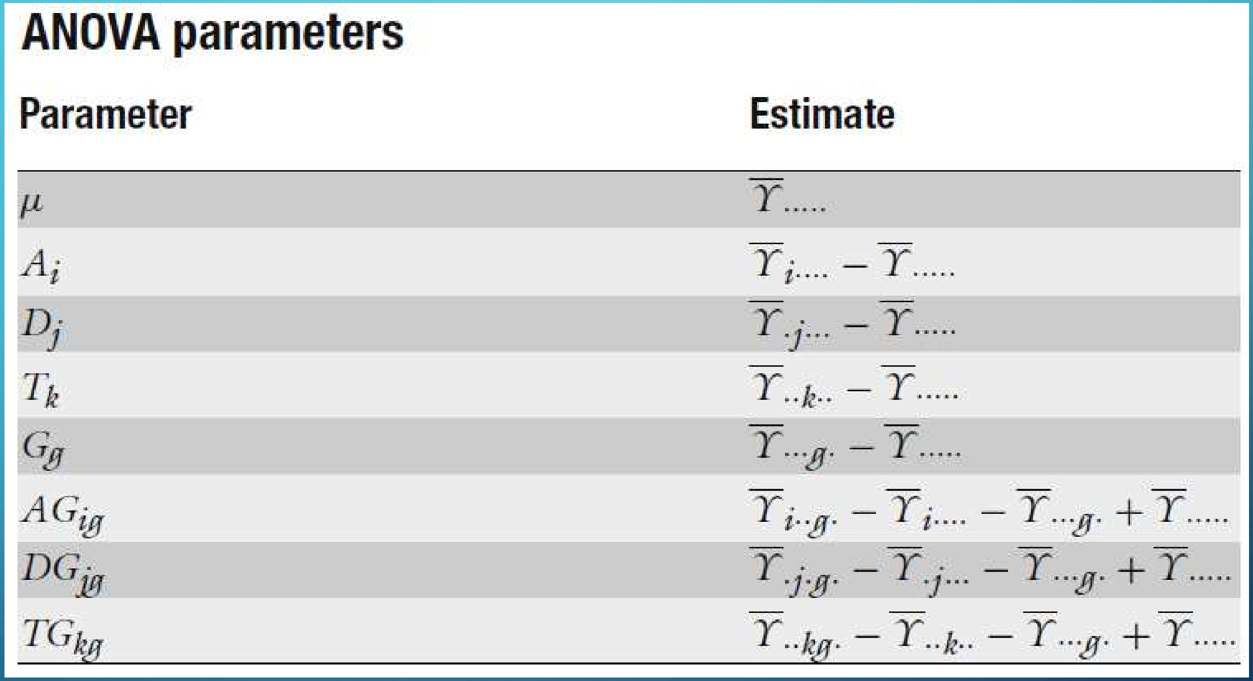
ANOVA Model for Gene Expression

- Instead of the normalization and t-test procedure these programs are based on statistical ANOVA models
- ANOVA stands for *analysis of variance*
- For each gene on the array it is to be decided whether the observed differences between treatment and control group are large enough compared to the variation in the experiment to declare the gene differentially expressed
- Other than for the t-test now all observations are combined in just one statistical model
- $\Upsilon ijkgr = \mu + \mathrm{A}i + \mathrm{D}j + \mathrm{T}k + \mathrm{G}g + \mathrm{AG}ig + \mathrm{DG}jg + \mathrm{TG}kg + \varepsilon ijkgr$
- Υ stands for the logarithm of background corrected intensity
  $\Upsilon = \log$ (Foreground median − Background)
- $\Upsilon ijkgr$ is the log-intensity for the *rth* replication of gene *g* under treatment *k* labeled with dye *j* on array *i*
- *A* stands for the array effect, *D* stands for dye effect, *T* stands for treatment effect, and *G* stands for gene effect
- *AG*, *DG*, and *TG* stand for the array–gene, the dye–gene, and the treatment–gene interaction effects

ANOVA Model for Gene Expression

- μ represents the overall log-intensity mean
- ε are called the errors
- The errors represent the variation in the experiment that cannot be explained in a systematic manner (through different dyes, treatments, arrays, or genes)
- The effects (array, dye, treatment, and gene) in the microarray ANOVA model describe the average contribution that the respective factors have on the log intensities
- For example, consider once more the experiment in which the gene expression of treatment is compared to control. In this case the factor "treatment" takes on two levels (k = 1 or k = 2) and the treatment effect *Tk* describes the average difference of *log-intensities* between the two groups
- The interaction effects allow us to consider that not all combinations of factors will influence log-intensity equally
- All parameters in an ANOVA model can be estimated by averaging over the original observations.

ANOVA Model for Gene Expression



ANOVA parameters

Parameter	Estimate
μ	$\overline{\Upsilon}_{.....}$
A_i	$\overline{\Upsilon}_{i....} - \overline{\Upsilon}_{.....}$
D_j	$\overline{\Upsilon}_{.j...} - \overline{\Upsilon}_{.....}$
T_k	$\overline{\Upsilon}_{..k..} - \overline{\Upsilon}_{.....}$
G_g	$\overline{\Upsilon}_{...g.} - \overline{\Upsilon}_{.....}$
AG_{ig}	$\overline{\Upsilon}_{i..g.} - \overline{\Upsilon}_{i....} - \overline{\Upsilon}_{...g.} + \overline{\Upsilon}_{.....}$
DG_{jg}	$\overline{\Upsilon}_{.j.g.} - \overline{\Upsilon}_{.j...} - \overline{\Upsilon}_{...g.} + \overline{\Upsilon}_{.....}$
TG_{kg}	$\overline{\Upsilon}_{..kg.} - \overline{\Upsilon}_{..k..} - \overline{\Upsilon}_{...g.} + \overline{\Upsilon}_{.....}$

ANOVA Hypotheses

- For each gene *g* on the microarray the null hypothesis corresponding to the ANOVA model for differential expression is

$$H0 : T1 + TG1\, g = T2 + TG2\, g$$

- In the test statistic, the estimate of the treatment effect

$$\overline{\Upsilon}_{..1g.} - \overline{\Upsilon}_{..2g.}$$

- is compared to its standard deviation ˆσ/√r where ˆσ is the estimate of the residual standard deviation and *r* is the number of times the gene is spotted under the same conditions on the array

$$t = \frac{\overline{\Upsilon}_{..1g.} - \overline{\Upsilon}_{..2g.}}{\hat{\sigma} / \sqrt{r}} \sim t(df = r - 1)$$

- For each gene spotted repeatedly on the array, the value of the test statistic is computed and a corresponding p-value obtained
- The resulting list of p-values (one for each gene) will be used to make the decision about differential expression of each gene

Variance Estimation in ANOVA

- There are two possible ways to estimate the standard deviation in the ANOVA model for differential expression
- One can assume that the observed variation has the same magnitude for all genes on the array this is known as the common gene variance model
- Statistically, this method is powerful, since the standard deviation estimate is based on very many observations
- biologically this method may not be very meaningful, because it is known that genes with very low expression across treatments vary less than genes with very high expression
- Alternatively, the residual standard deviation estimate can also be computed repeatedly and separately for each gene. This is known as a per-gene variance model
- the estimate is statistically much less powerful
- biologically is more appropriate, since it allows for the possibility of different genes having different magnitudes of standard deviation

Multiple Testing Issues

- In real microarray experiments, many more than two or six genes are spotted on an array
- Regardless of whether we use t-tests or an ANOVA model for the analysis very many decisions will have to be made
- If we use a significance level of 5% then for *each gene* there is up to a 5% chance that we falsely declare the gene differentially expressed
- If there is a small (5%) chance of a mistake in every decision, then overall in the very many decisions we will have to make, many mistakes
- Different procedures exist to correct this problem

Bonferroni Method

- microarray has spots that represent 1000 genes
- each of these 1000 genes differentially expressed
- If you work with a level of 5% for each individual test, then the probability that you make at least one wrong decision is

$$1-(1-0.05)^{1000}=0.9999999\ldots999(\text{twenty-three 9s})$$

- This means that it is virtually certain that your analysis will contain at least one error
- The Bonferroni correction method says, if you would like the probability of making at least one mistake to be less than α, (so called family-wise error rate (FWER)), then use a significance level of α/n
- In the above example this would mean that if we want to keep the probability of making at least one mistake under 10%, then we should declare only those genes differentially expressed, whose p-values are smaller than 0.10/1000 = 0.0001

False Discovery Rate

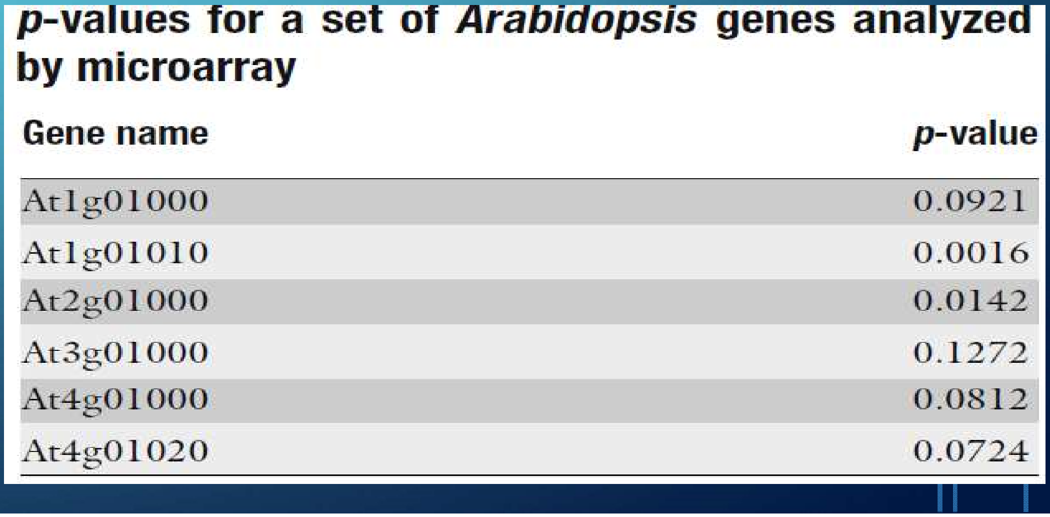
the expected proportion of the falsely rejected null hypothesis

$$FDR = \text{Average}\left(\frac{\text{Falsely rejected null hypotheses}}{\text{rejected null hypothesis}}\right)$$

"linear step-up procedure" that pushes the false discovery rate below a given level q

That means that you can make sure that the proportion of "false discoveries" stays, for example, under 10% (if you set $q = 0.10$)

To conduct a linear step-up test:



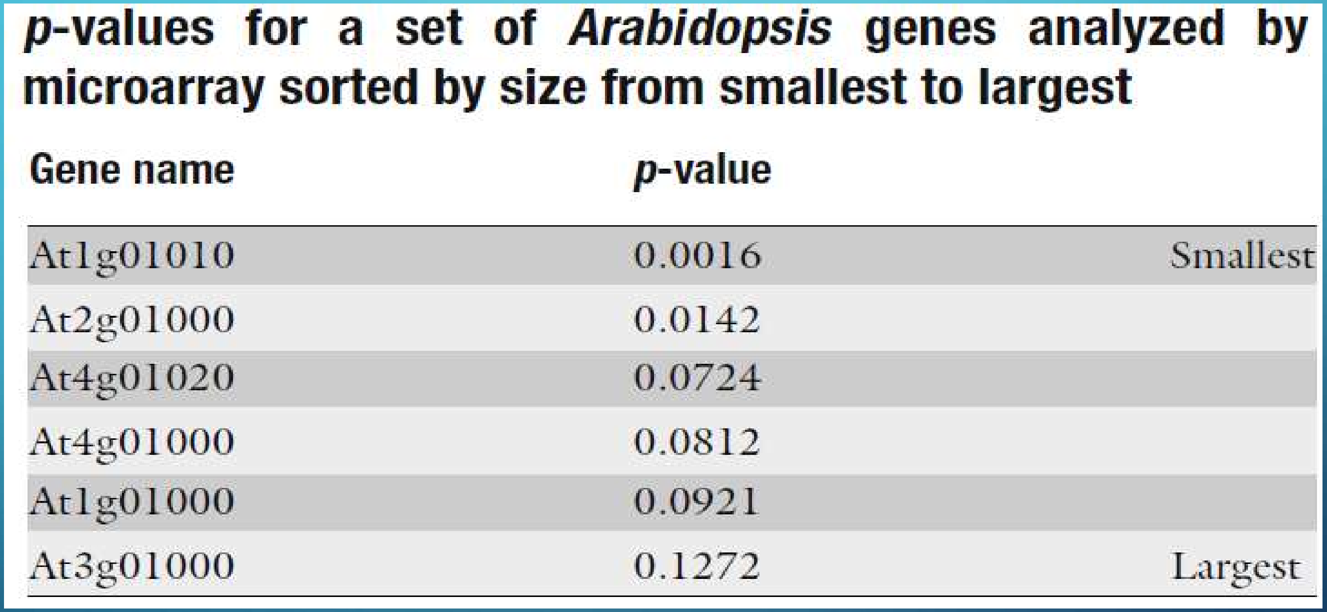
***p*-values for a set of *Arabidopsis* genes analyzed by microarray**

Gene name	*p*-value
At1g01000	0.0921
At1g01010	0.0016
At2g01000	0.0142
At3g01000	0.1272
At4g01000	0.0812
At4g01020	0.0724

False Discovery Rate

- these *p-values* are shown sorted by size (smallest to largest)
- Pick a level q under which you want to control the false discovery rate. Common values for q are 0.10 (10%) or 0.05 (5%)
- Start at the top of the list and check for each gene, whether its *p-value* is bigger than $q \cdot \frac{i}{n}$
- Here, q is the level under which you want to control the *FDR*, n is the total number of genes you have, and i is the number of the test you conduct
- When you find a gene for which the p-value is bigger than $q \cdot \frac{i}{n}$, declare this gene and all other genes with larger p-values not differentially expressed

False Discovery Rate

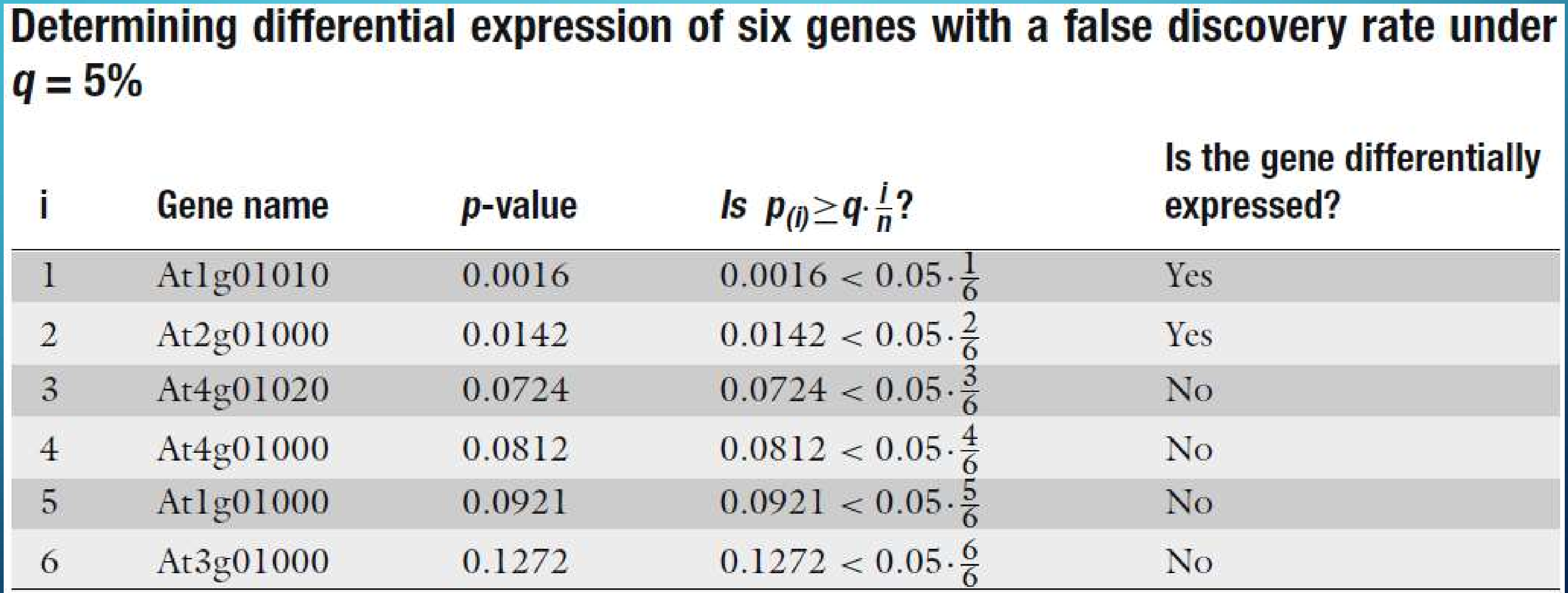


***p*-values for a set of *Arabidopsis* genes analyzed by microarray sorted by size from smallest to largest**

Gene name	*p*-value	
At1g01010	0.0016	Smallest
At2g01000	0.0142	
At4g01020	0.0724	
At4g01000	0.0812	
At1g01000	0.0921	
At3g01000	0.1272	Largest

False Discovery Rate

In this example, the smallest *i* for which the p-value is larger than 0.05 · $q \cdot \frac{i}{n}$ is $i = 3$. In this case we declare only the first two genes differentially expressed



Determining differential expression of six genes with a false discovery rate under *q* = 5%

i	Gene name	*p*-value	*Is* $p_{(i)} \geq q \cdot \frac{i}{n}$?	Is the gene differentially expressed?
1	At1g01010	0.0016	$0.0016 < 0.05 \cdot \frac{1}{6}$	Yes
2	At2g01000	0.0142	$0.0142 < 0.05 \cdot \frac{2}{6}$	Yes
3	At4g01020	0.0724	$0.0724 < 0.05 \cdot \frac{3}{6}$	No
4	At4g01000	0.0812	$0.0812 < 0.05 \cdot \frac{4}{6}$	No
5	At1g01000	0.0921	$0.0921 < 0.05 \cdot \frac{5}{6}$	No
6	At3g01000	0.1272	$0.1272 < 0.05 \cdot \frac{6}{6}$	No

با تشکر از توجه شما