Lesson 11

Functional Genomics I: Microarray Analysis

Transcription of DNA and translation of RNA vary with biological conditions

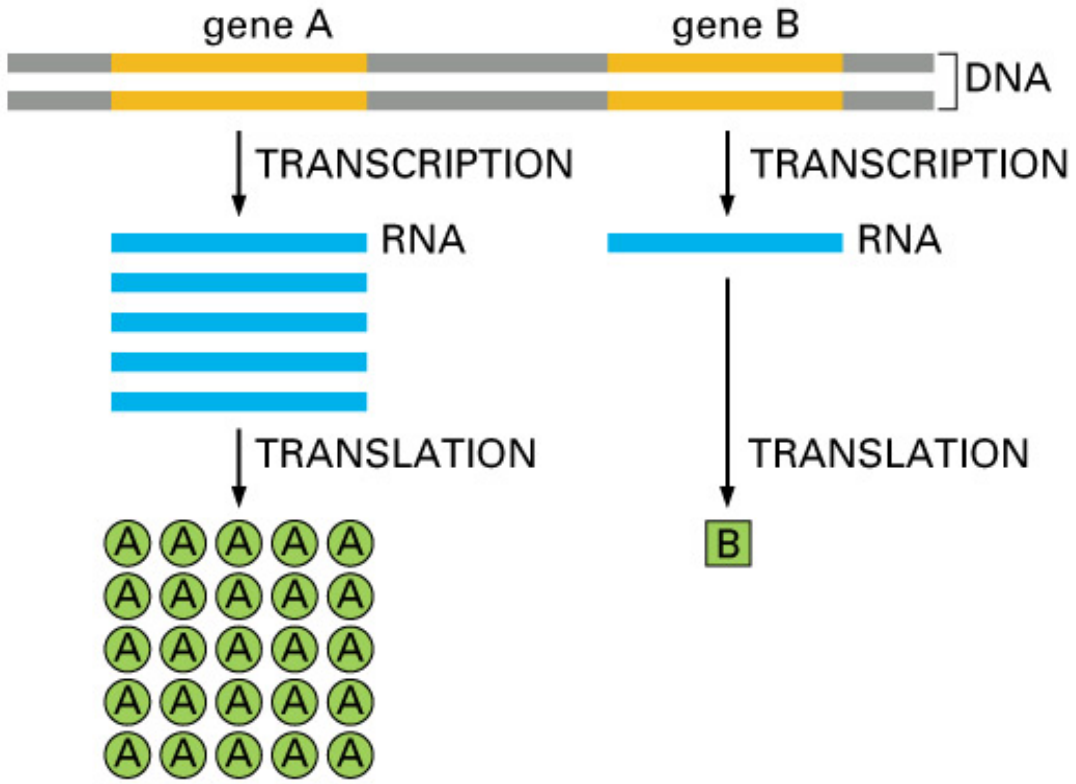


Figure 6–3. Molecular Biology of the Cell, 4th Edition.

3 kinds of microarray platforms

- Spotted Array - 2 color - Pat Brown (Stanford)
- Synthesized Oligonucleotide - 1 color – Affymetrix
- Synthesized Oligonucleotide - 2 color – Agilent

Spotted (2 Color) Arrays

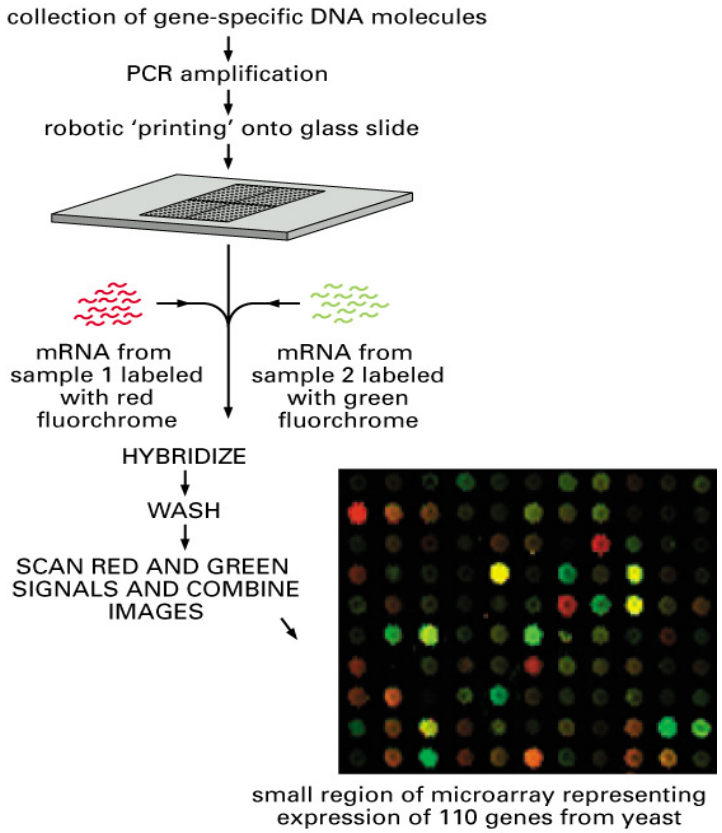


Figure 8–62. Molecular Biology of the Cell, 4th Edition.

Agilent 2 color Arrays

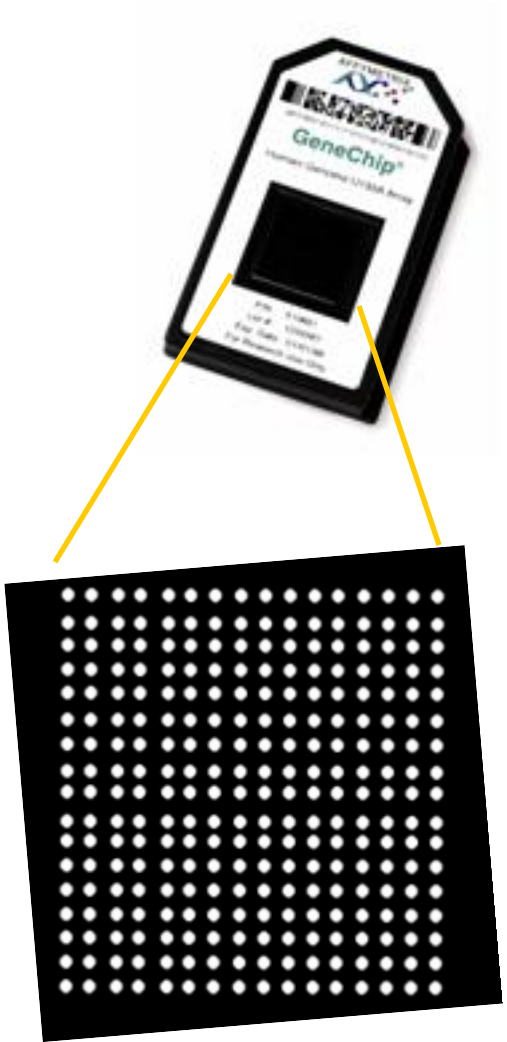
Synthesized oligonucleotides like Affymetrix arrays but 2 colors like spotted arrays.

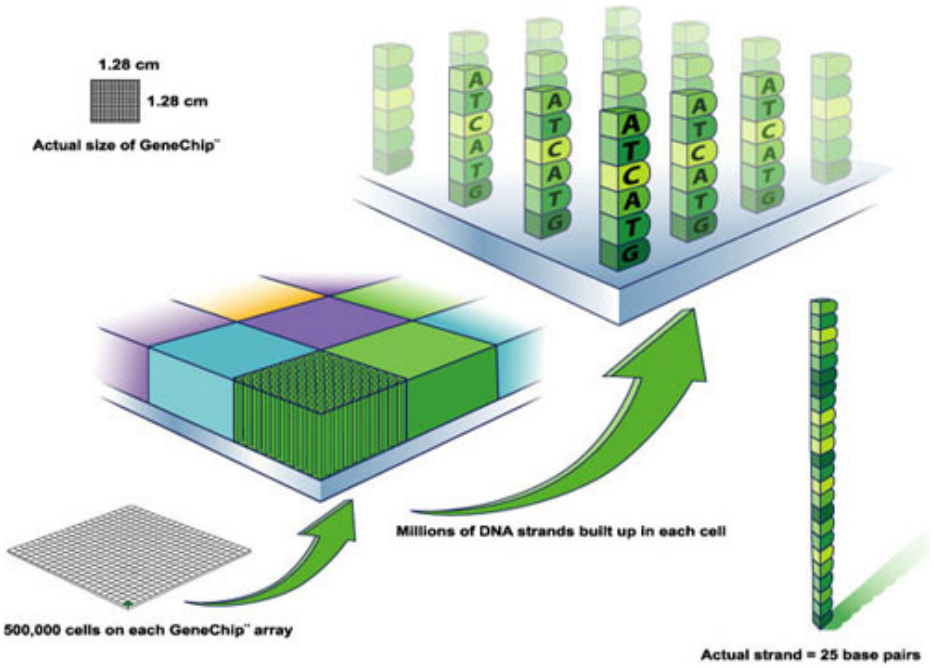
Concentrations from 2 Color Experiments

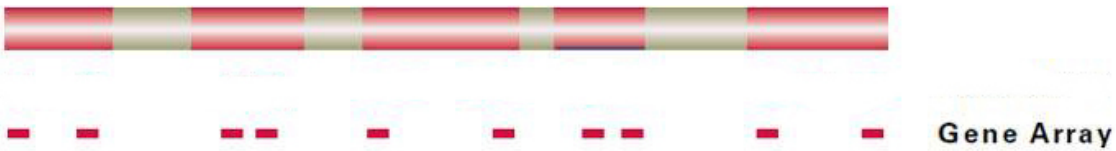
$$\frac{RNA(Experiment)}{RNA(Control)} = \frac{Intensity\,\mathrm{Re}\,d}{IntensityGreen}$$

Gene Expression Arrays

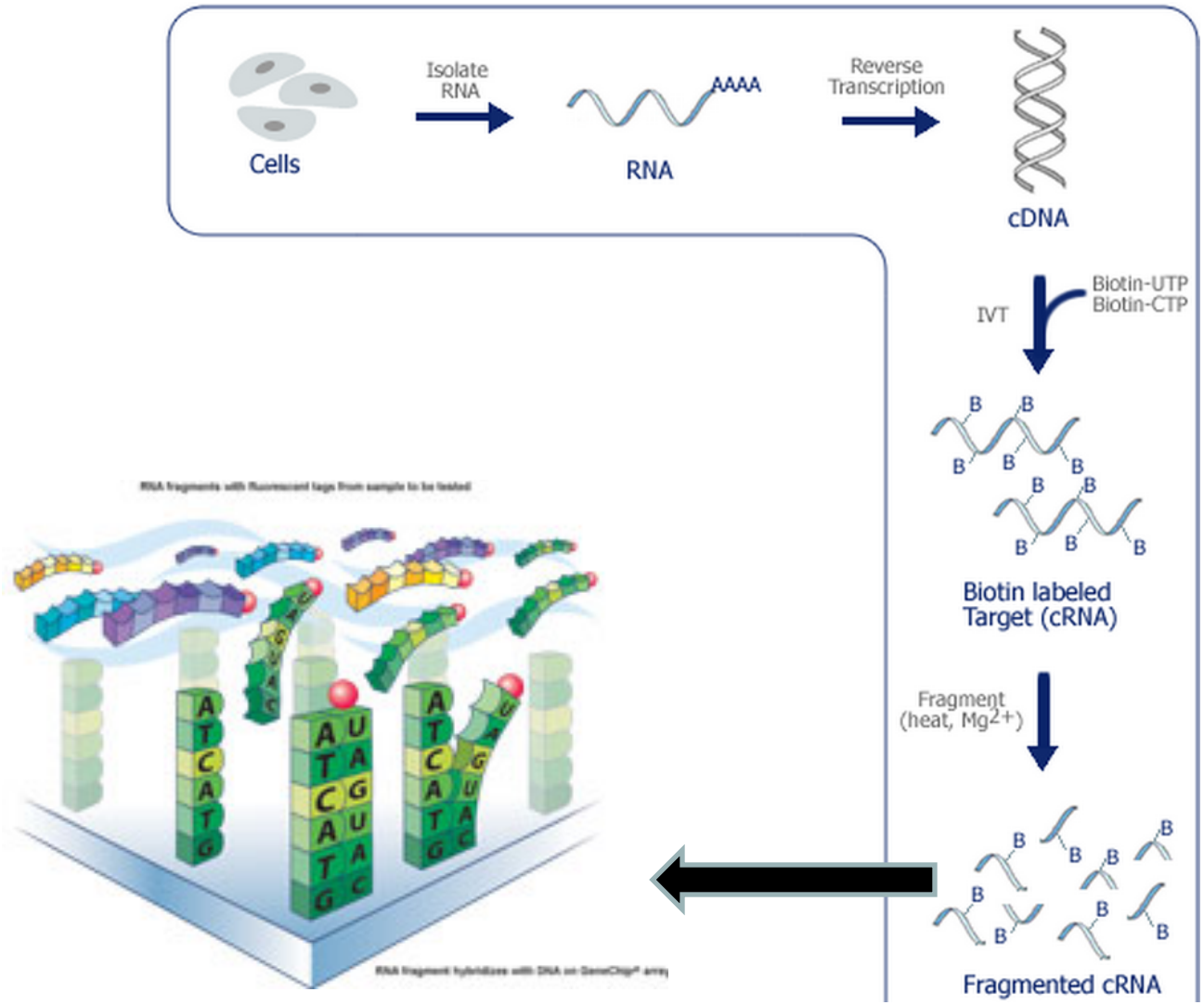
Measurement of mRNA levels for all genes.



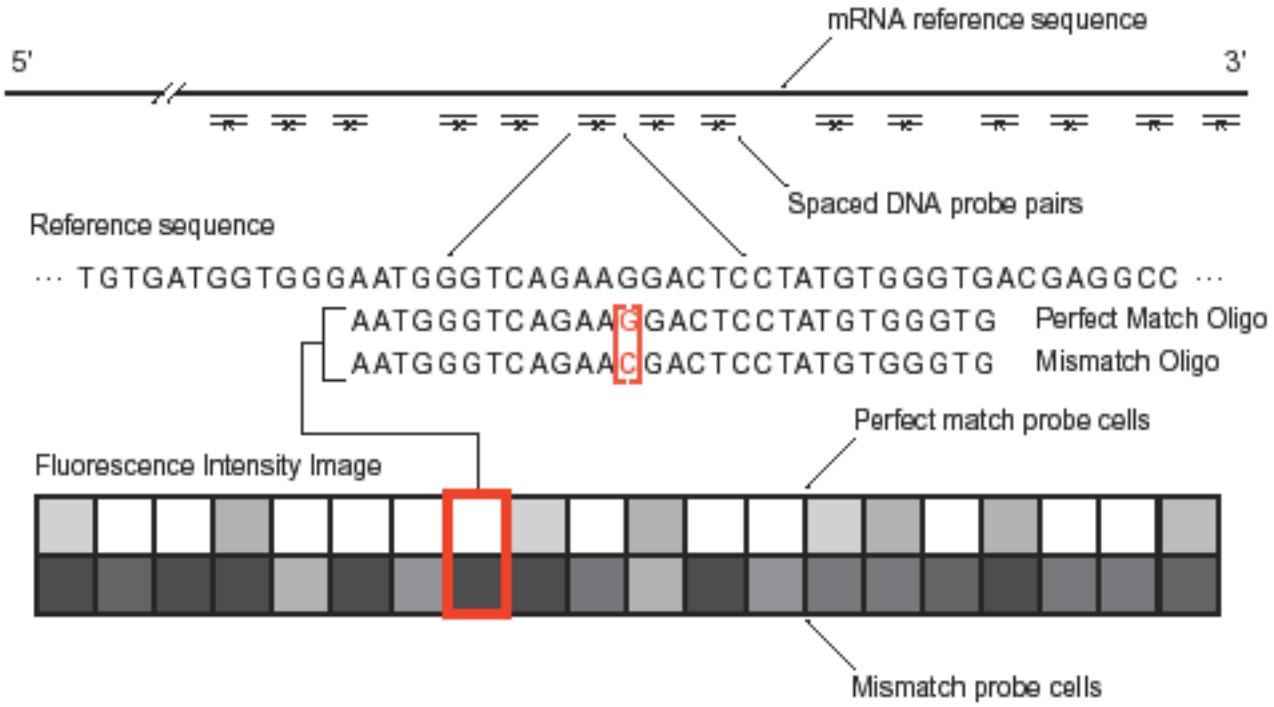




Gene Expression Arrays



Affymetrix 1 Color Arrays



Spotted vs Affymetrix vs Agilent

Spotted Arrays:

- Advantages: Long pieces.
- Disadvantages: Uncertainties in spot reading.

Affymetrix Arrays:

- Advantages: Probes in same place, can be read precisely.
- Disadvantages: Short pieces. Must assemble probe information.

Agilent:

- Advantages: Medium pieces. Advantages of Affymetrix and Agilent.
- Disadvantages: None in principle.

Concentrations from 1 Color Experiments

$$\frac{RNA_{\text{experiment}}}{RNA_{\text{control}}} = \frac{Intensity_{\text{experiment}}}{Intensity_{\text{control}}}$$

Probeset intensity as an average of probe intensities

$$I_{probeset} = \sum_{j=1,k} \frac{\log_2(PM_j - MM_j)}{k}$$

Problems with averaging probes

1. Var(probes within probeset) > Var(The same probe across slides)
2. MM>PM (40% of slides)

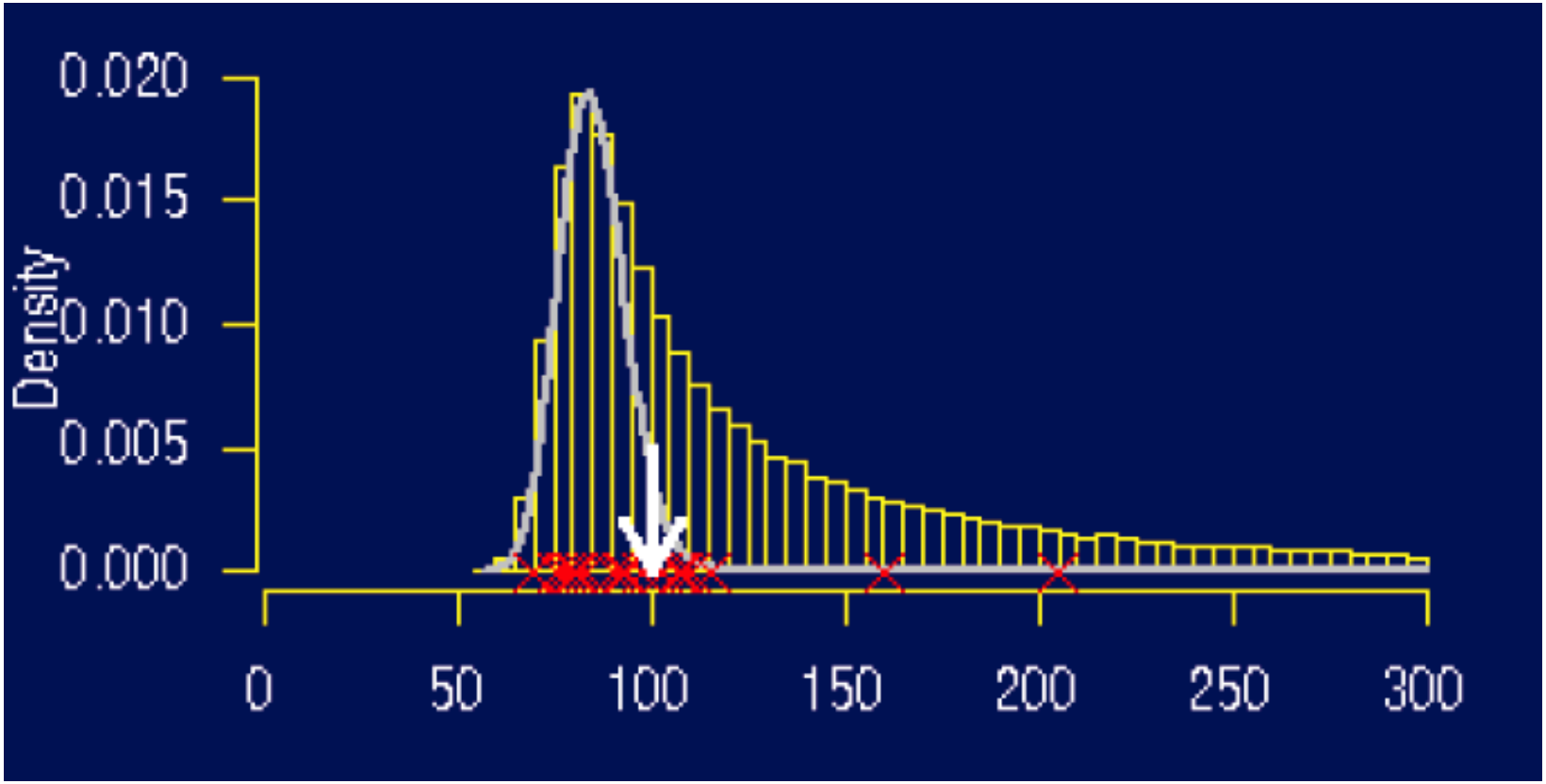
Problems to be solved in chip reading

1. Highly variable probe intensities compared to probes set intensity.
2. Correct for nonspecific binding realistically.
3. Correct for background within chips.
4. Correct for intensity variation between chips.

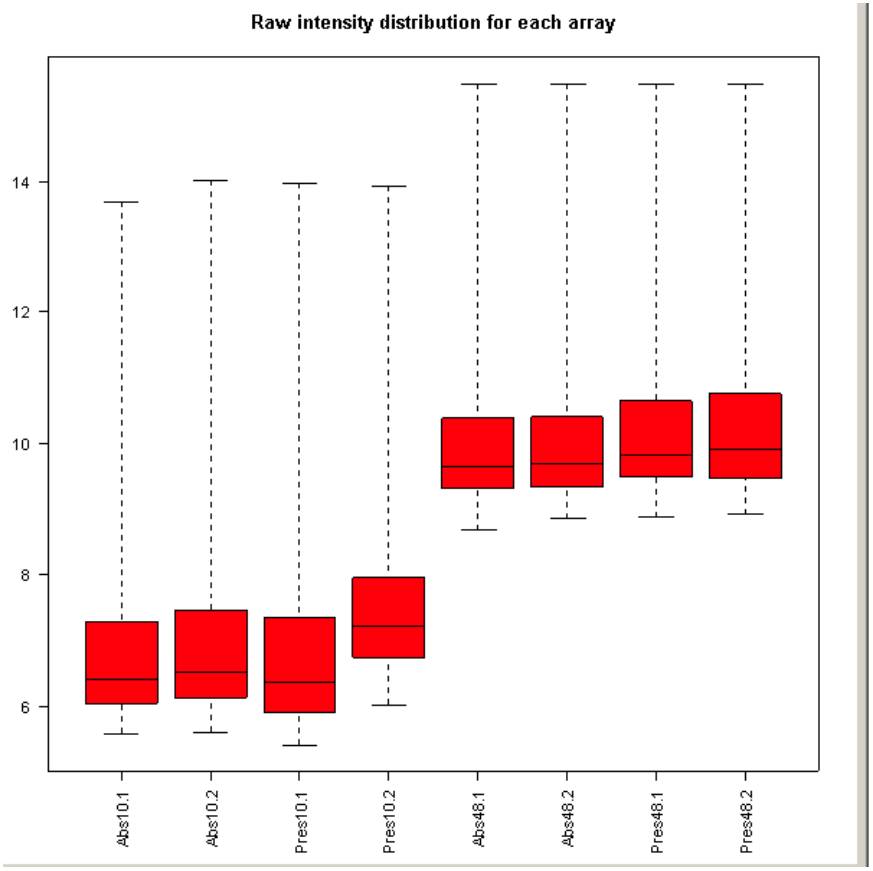
Steps in RMA

1. Background correction- in each chip.
2. Normalization - between chips.
3. Summarization of probes to probe sets.

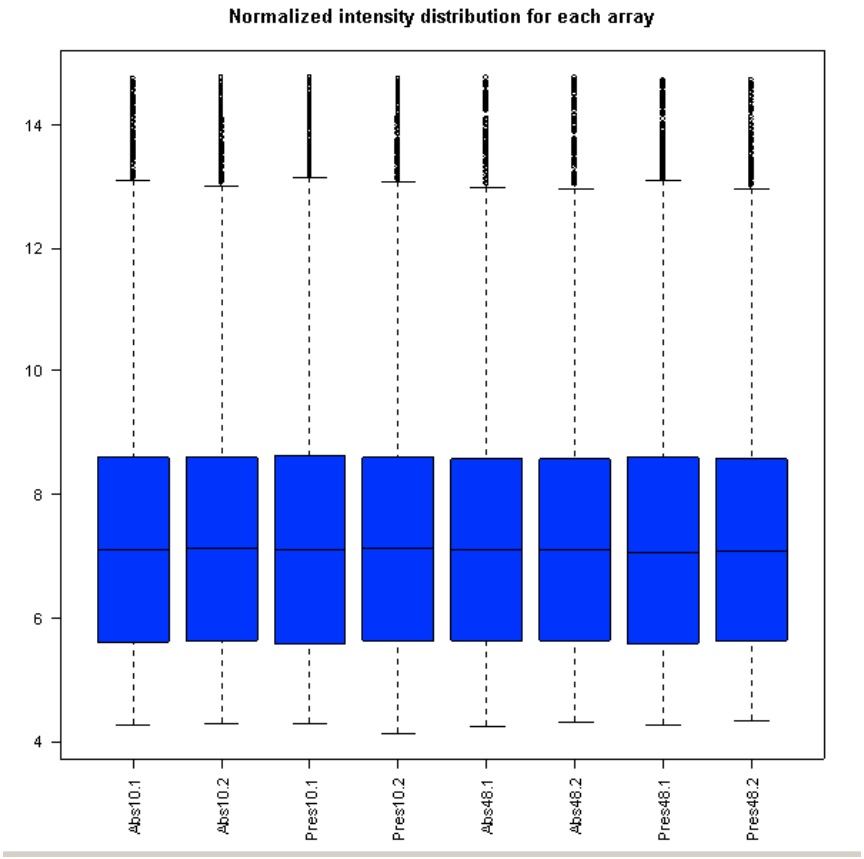
RMA Background Correction



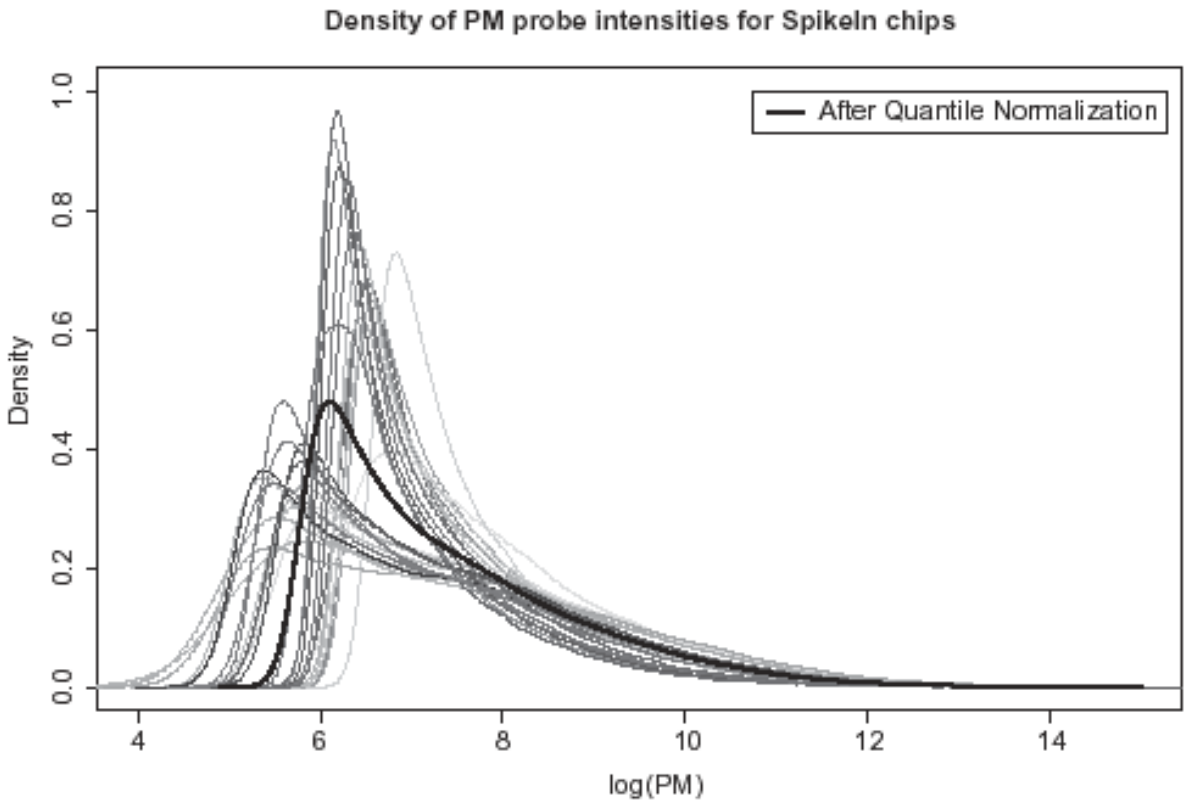
Boxplot of Unnormalized Chips



Quantile Normalization



Intensity Plots



Contributions to probe intensity

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

Constraint on probe-effects

$$\sum_{j=1,J} \alpha_{jn} \cong 0$$

Simple Example

1 probeset,

3 probes/probeset

2 chips

$Y_{11}=\mu_1 + \alpha_1$ $Y_{21}=\mu_2 + \alpha_1$

$Y_{12}=\mu_1 + \alpha_2$ $Y_{22}=\mu_2 + \alpha_2$

$Y_{13}=\mu_1 + \alpha_3$ $Y_{23}=\mu_2 + \alpha_3$

$\alpha_1 + \alpha_2 + \alpha_3 = 0$

5 unknowns, 7 equations

Median polish algorithm

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \varepsilon_{ijn}$$

GCRMA

- Similar to RMA but has a probe composition dependent background correction.
- GC base-pair 3 hydrogen bonds.
- AT base-pair 2 hydrogen bonds.
- Non-specific binding to GC higher than to AT.
- GCRMA implements RMA-type background correction dependent on GC content. Mismatch intensities of probes with the same GC content are pooled.

Expression ratios

$$\frac{x_2}{x_1} \geq 2, OR \frac{x_2}{x_1} \leq \frac{1}{2}$$

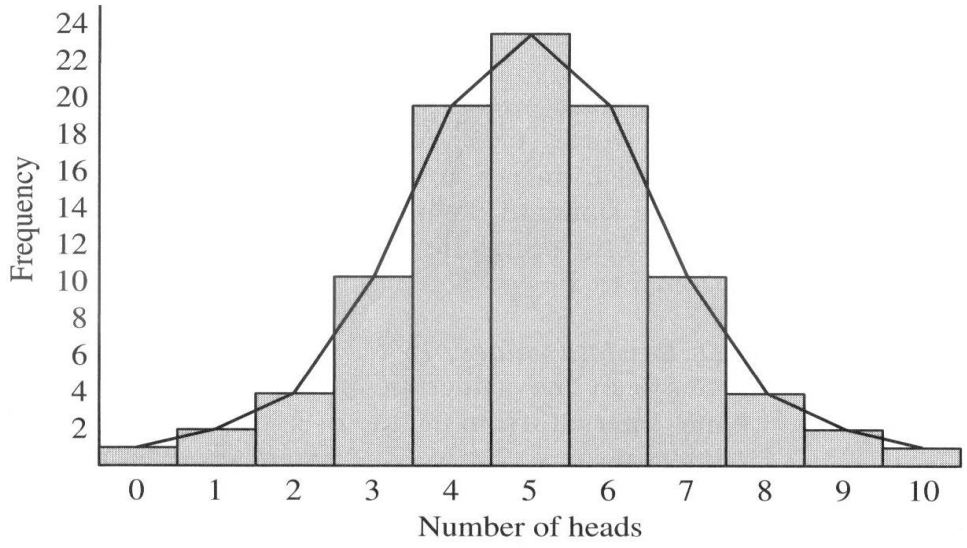
Need for a measure of variability

Experiment	Replicate A	Replicate B	Average
1	2	6	4
2	1	15	8

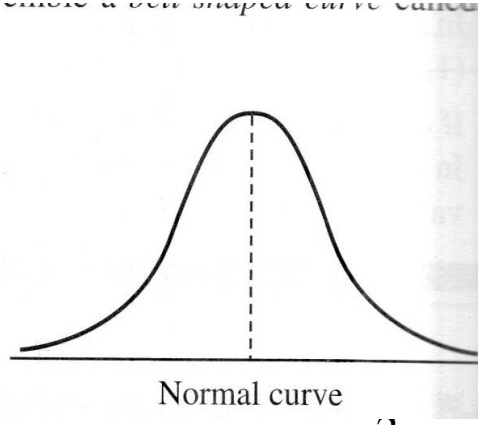
$$\frac{\bar{x}_2}{\bar{x}_1} = \frac{\left(\frac{1+15}{2}\right)}{\left(\frac{2+6}{2}\right)} = \frac{\left(\frac{16}{2}\right)}{\left(\frac{8}{2}\right)} = \frac{8}{4} = 2$$

Approximation of the normal distribution

Number of heads	0	1	2	3	4	5	6	7	8	9	10
Frequency	1	2	4	11	20	24	20	11	4	2	1



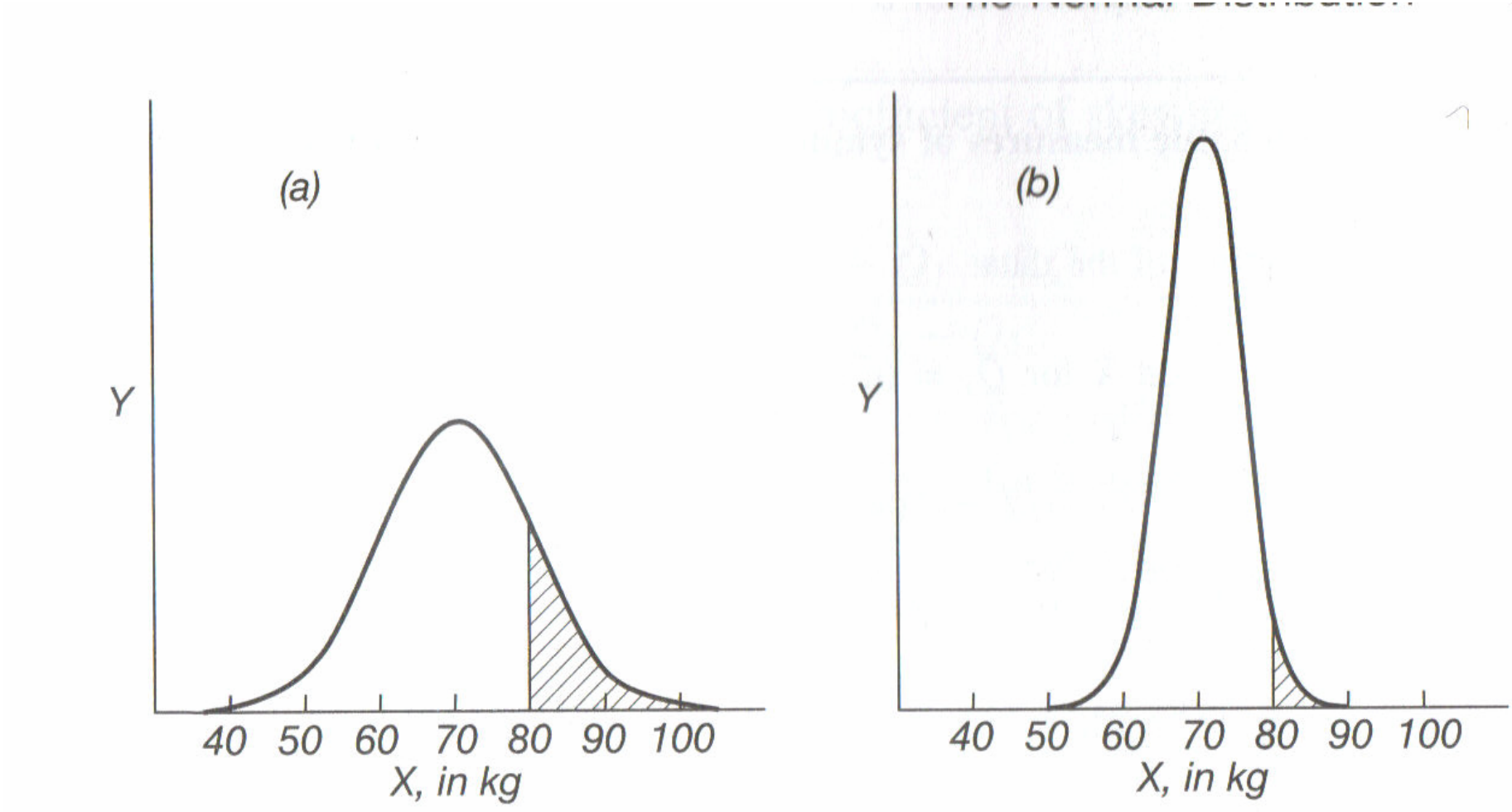
Equation of the normal distribution



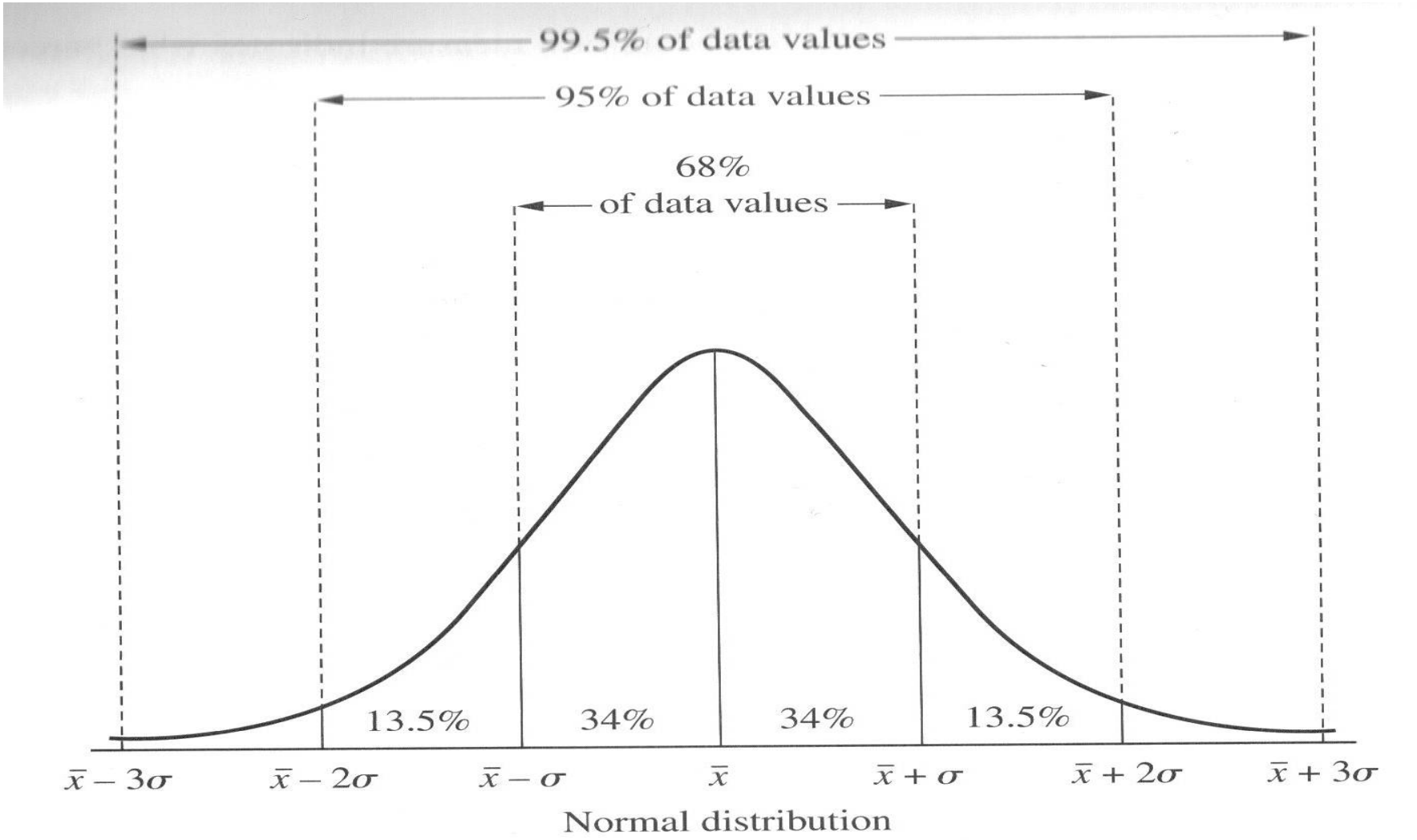
$$\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma}}$$

μ= mean(average) σ= standard deviation

Effect of the standard deviation



Standard deviation and percent



Estimates of the mean and standard deviation of the mean

$$\bar{x} = \frac{\sum_{i=1,N} x_i}{N} \qquad s_x = \sqrt{\frac{\sum_{i=1,N} (x_i - \bar{x})^2}{N(N-1)}}$$

Standard deviation of the mean

$$\sigma_x = \sqrt{\frac{\sigma}{n}}$$

68% of all of the means within 1 standard deviation of the mean.

95% of all of the means within 2 standard deviation of the mean.

The z distribution

$$z = \frac{\bar{x} - \mu}{s_x}$$

Does experimental CO_2=10.00mg/m^3

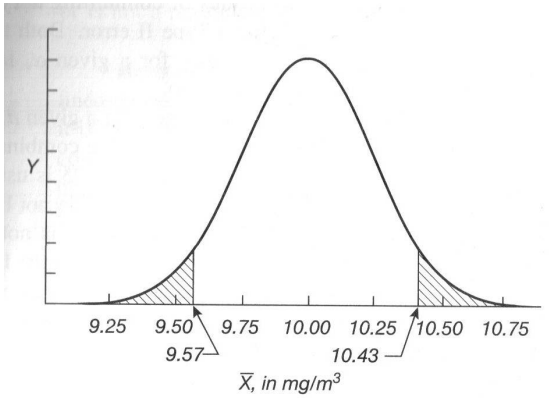
$$\bar{x} = 10.43(mg/m^3)$$

$$\mu = 10.00(mg/m^3)$$

$$s_x = 0.24(mg/m^3)$$

$$z = \frac{\bar{x} - \mu}{s_x}$$

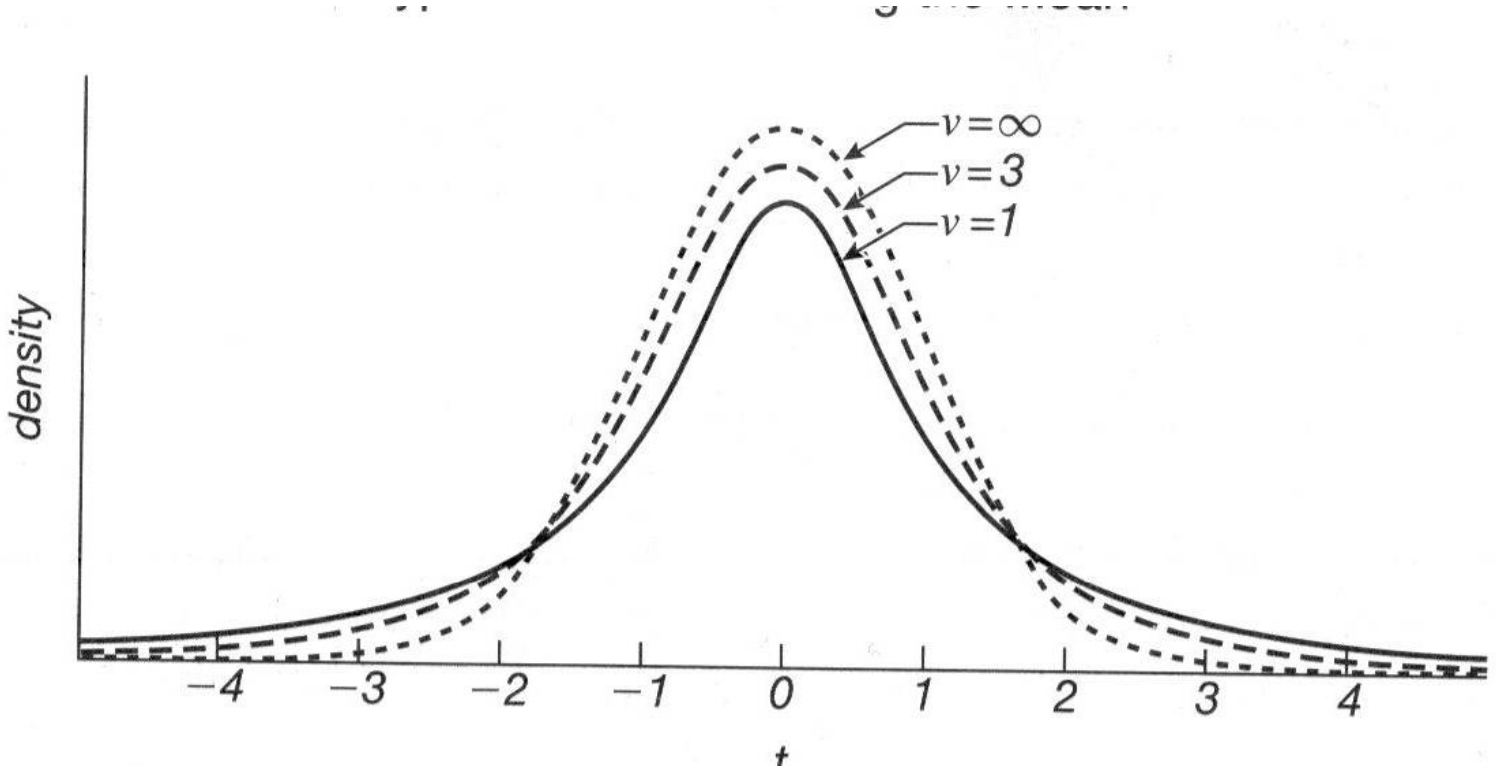
$$z = \frac{10.43 - 10.00}{0.24} = 1.79$$



$$p(z \geq 1.79 \text{ AND } z \leq -1.79) = 0.07$$

$$0.07 \geq 0.05$$

The t-distribution



$$t = \frac{\bar{x} - \mu}{s_x}$$

The t-distribution of the difference of 2 means

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}}$$

$$s_p^2 = \frac{\sum_{i=1,N_1} (x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2} (x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}$$

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\sum_{i=1,N_1} (x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2} (x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}\right)}}$$

Problems applying t-test to microarrays

1. Multiple tests - thousands of genes.
2. Multiple conditions- more than 2 conditions.

Solution: LIMMA

LInear Models for Microarray Analysis.

The log transformation of intensities

$$x\text{–} > \log_2(x)$$

$$m = \log_2 x_2 - \log_2 x_1 = \log_2\left(\frac{x_2}{x_1}\right)$$

The t-distribution of the difference of 2 means

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}}}$$

$$s_p^2 = \frac{\sum_{i=1,N_1} (x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2} (x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}$$

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\sum_{i=1,N_1} (x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2} (x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}\right)}}$$

Empirical Bayesian correction

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2} + s_0}}$$

Adding S_0

Denominatorincreases.

t decreases.

p increases.

#false positives decreases.

Empirical Bayesian Correction

Frequentist Statistics- Standard deviation estimated from data.

Bayesian Statistics- Prior estimate of standard deviation modifed

from data.

Empirical Bayesian Statistics- Prior estimate of standard deviation obtained from

from all data and modified for individual data.

Benjamini-Hochberg False Discovery Correction

Uncorrected p-value= rate of false discovery if only 1 test.

Corrected p-value= rate of false discovery if all of the genes above it on the p-value list were tested and accepted.

False Discovery vs Raw Pvalue

Raw p-value is the probability of getting the t-statistic or a larger one by chance if there is no difference.

False discovery rate is the proportion of differences that are accepted at or above a given p-value for which there is really no difference.

Multiple Conditions

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\sum_{i=1,N_1}(x_{i1} - \bar{x}_1)^2}{N_1 + N_2 - 2} + \frac{\sum_{i=1,N_2}(x_{i2} - \bar{x}_2)^2}{N_1 + N_2 - 2}\right)}}$$

Comparing 2 conditions out of 3

$$t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\left(\frac{1}{N_1} + \frac{1}{N_2}\right)\left(\frac{\sum_{i=1,N_1}(x_{i1} - \bar{x}_1)^2}{N_1 + N_2 + N_3 - 3} + \frac{\sum_{i=1,N_2}(x_{i2} - \bar{x}_2)^2}{N_1 + N_2 + N_3 - 3} + \frac{\sum_{i=1,N_3}(x_{i3} - \bar{x}_3)^2}{N_1 + N_2 + N_3 - 3}\right)}}$$

Cutoffs for differential expression

$B = \ln(\text{odds of differential expression})$

Cutoff1: $p_{fdr} \leq .05$

Cutoff2: $B = \ln(odds) \geq 0$

(Cutoff2: odds ≥ 1)

Cutoff3: $p_{raw} \leq .001$

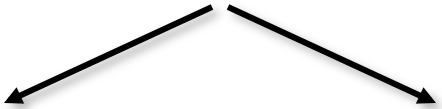
But take FDR into account!

AffylmGUI

R (Statistical Programming Language)

Bioconductor (R Programs for Biology)

LIMMA



AffylmGUI
1 Color

Limma GUI
2 Color