Gene expression

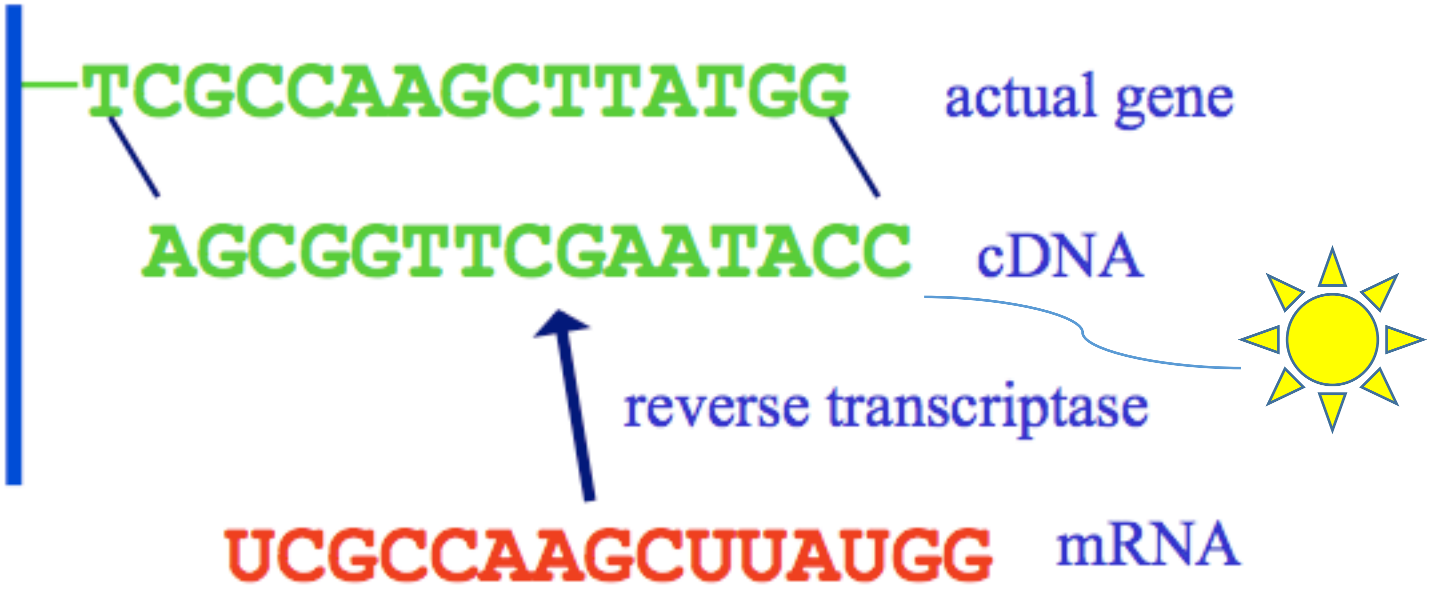
Microarrays

- Each cell has the same genome but doesn’t use it in the same way
- We can predict structural protein features from sequence and assign molecular function but some questions are difficult to answer looking at sequence
  - Sequence: this is a kinase with a an SH2 domain
  - Functional genomics: What tissue/organ/condition is this gene expressed in?
- Given the sequence of genes in the genome we can measure their simultaneous activity in a sample of interest
- Possible questions
  - What genes are different between cancer and normal tissue
  - What genes are required for response to ionizing radiation

How it works

Complementary hybridization:

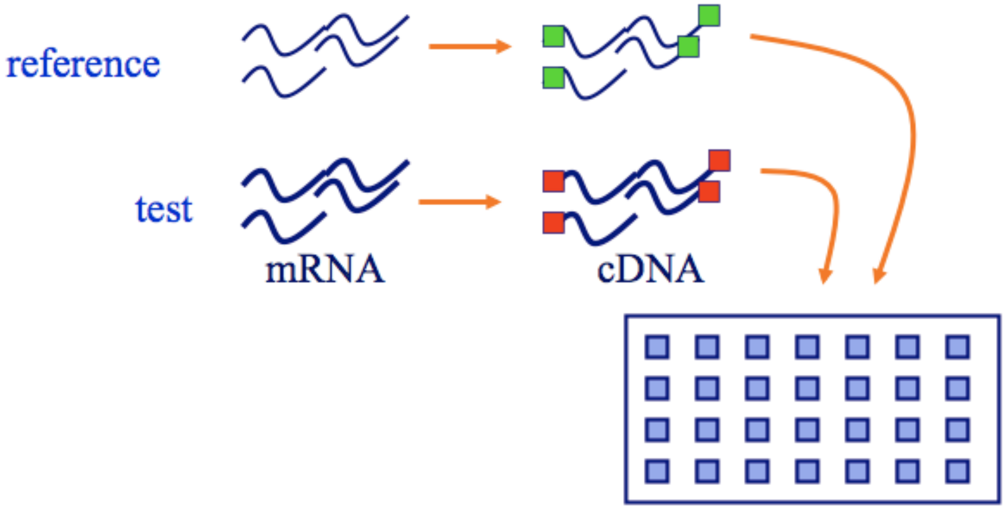
- Put a part of the gene sequence on the array
- convert mRNA to cDNA using reverse transcriptase



Fluorescent probe for quantification

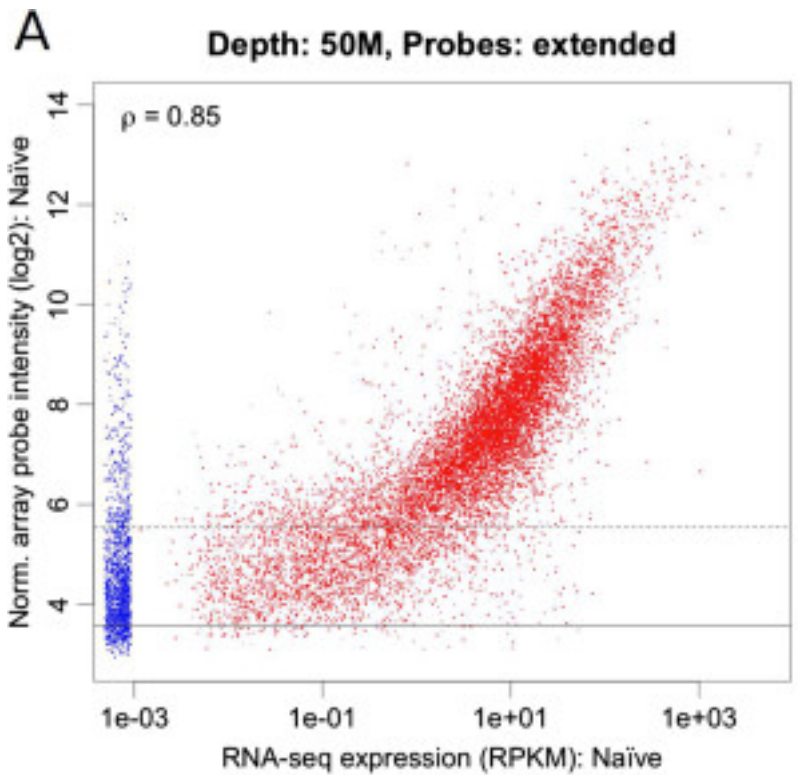
Type of arrays

- Spotted (old) –probes are synthesized and then deposited
- Oligonucleotide – probes are synthesized in place
- 2-channel
  - Two mRNA samples (reference, test) are labeled with fluorescent dyes (Cy3, Cy5) and allowed to hybridize to array
  - No comparisons across probes
- Single channel
  - One sample is hybridized
  - Intensity is related to total abundance
- Most arrays today are single channel oligonucleotide



RNAseq vs microarray

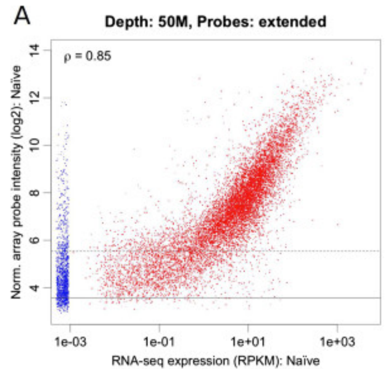
- RNAseq-direct sequencing of mRNA
  - Don't need to know what you are looking for
  - No probes
  - More certainty that you are detecting specific genes
  - Not based on fluorescent read out-better dynamic range
- Microarray vs RNAseq
  - Transcript misidentification
  - Saturation – low and high end



A comparison of RNA-seq and exon genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat arrays for whole

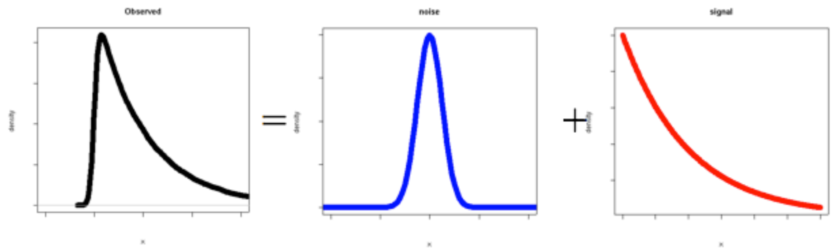
Microarray background estimation

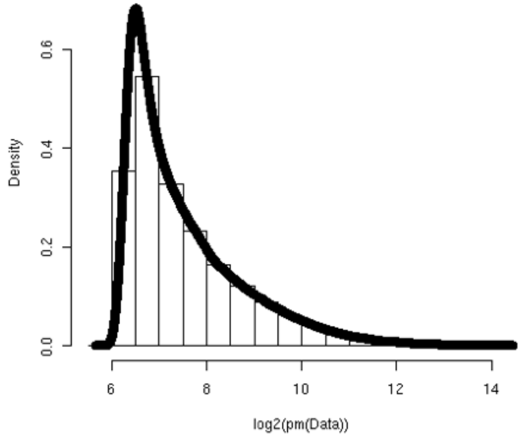
- Past some intensity point detection is not reliable
- General rule: for mammals about half of all possible genes are expressed in any given tissue/organ
- Use distribution characteristics to filter out unreliable measurements
- Signal intensity is modeled as a convolution of a normal and exponential distribution
- Illumina beadArray chips provide a detection p-value
    - These are often very close to a distribution based estimate



Intensity = Background + Signal

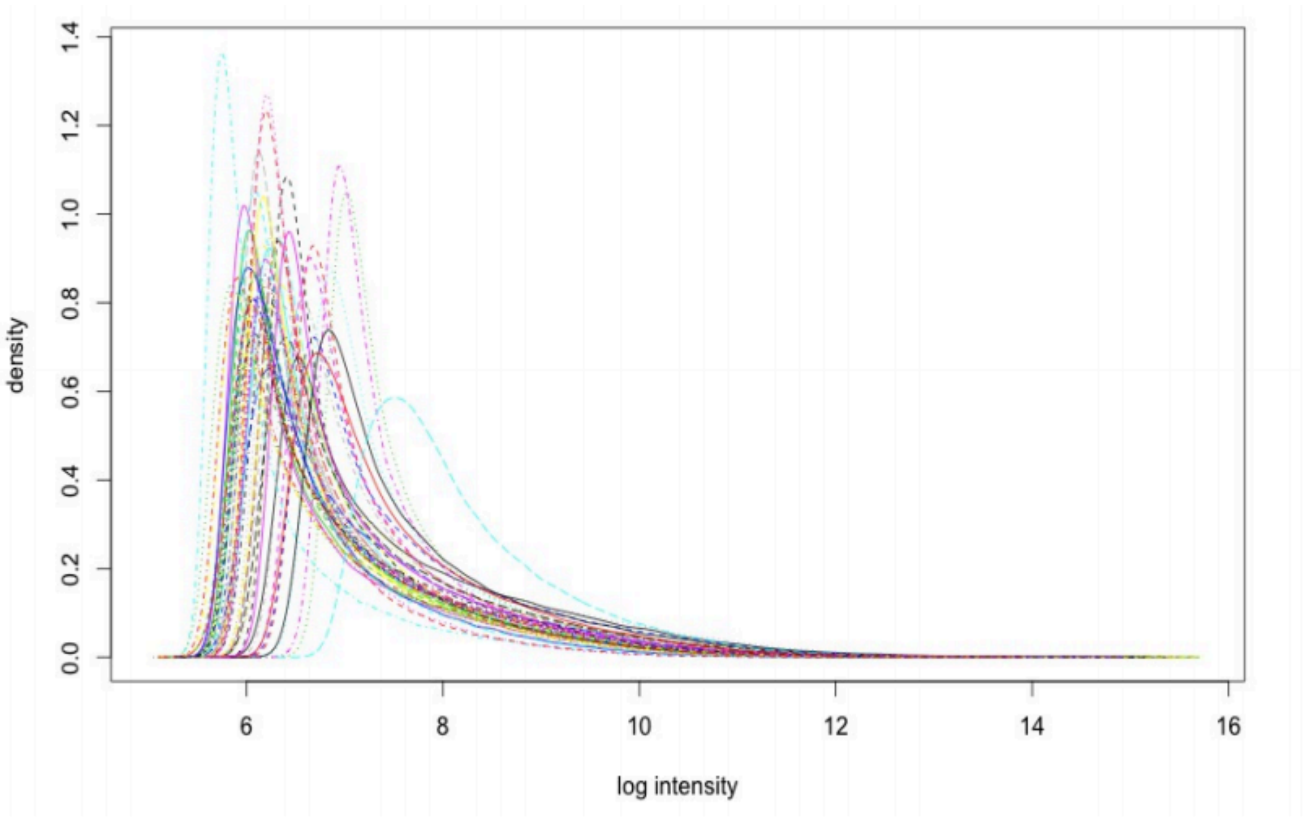
$N(\mu,\sigma^2)$ Exponential(α)





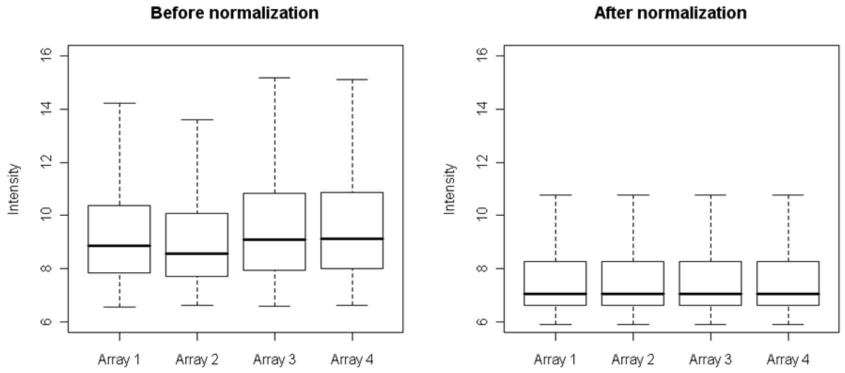
Normalization

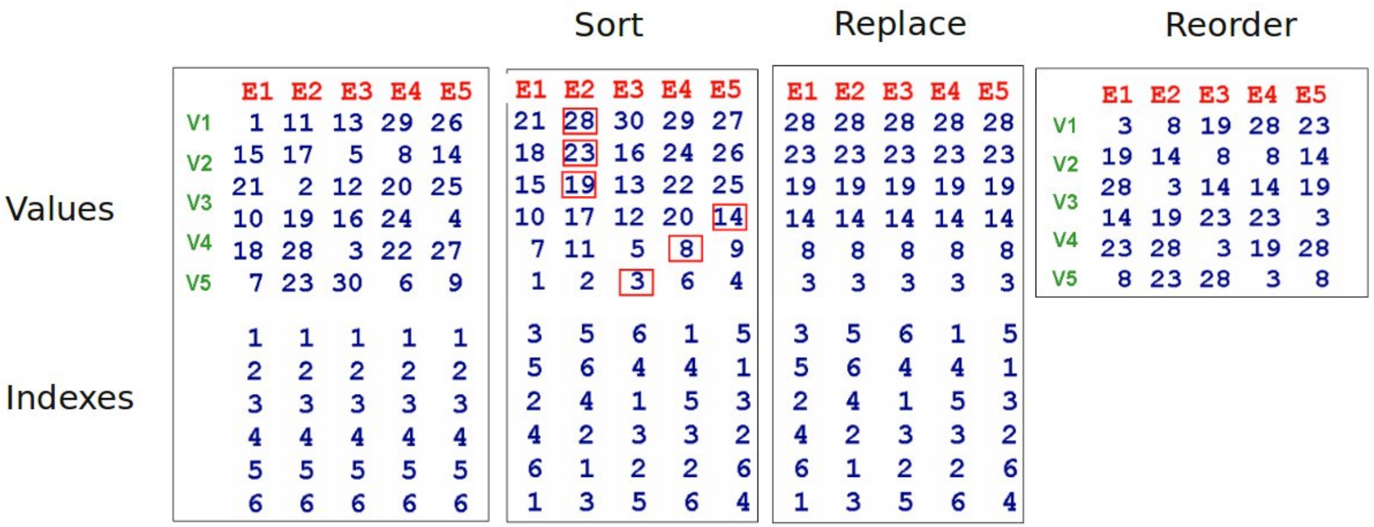
- We want to measure mRNA abundance but we measure fluorescence intensity
- Intensity is related through abundance through some arbitrary function
- This function depends on many experimental parameters and is different for different samples
- We have: $A_i = F_i(I_i)$
- Ideally we would like: $A_i = \mathbf{F}(I_i)$—all the functions are the same—though they still don't report true intensity



Normalization

- Many methods have been proposed
- Most widely used is **quantile normalization**
- Force the distributions to be the same by assigning the gene in each rank to the median value in that rank (not necessarily the same gene) across samples





Values

	E1	E2	E3	E4	E5
V1	1	11	13	29	26
V2	15	17	5	8	14
V3	21	2	12	20	25
	10	19	16	24	4
V4	18	28	3	22	27
V5	7	23	30	6	9

Indexes

E1	E2	E3	E4	E5
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6

Sort

E1	E2	E3	E4	E5
21	28	30	29	27
18	23	16	24	26
15	19	13	22	25
10	17	12	20	14
7	11	5	8	9
1	2	3	6	4

3	5	6	1	5
5	6	4	4	1
2	4	1	5	3
4	2	3	3	2
6	1	2	2	6
1	3	5	6	4

Replace

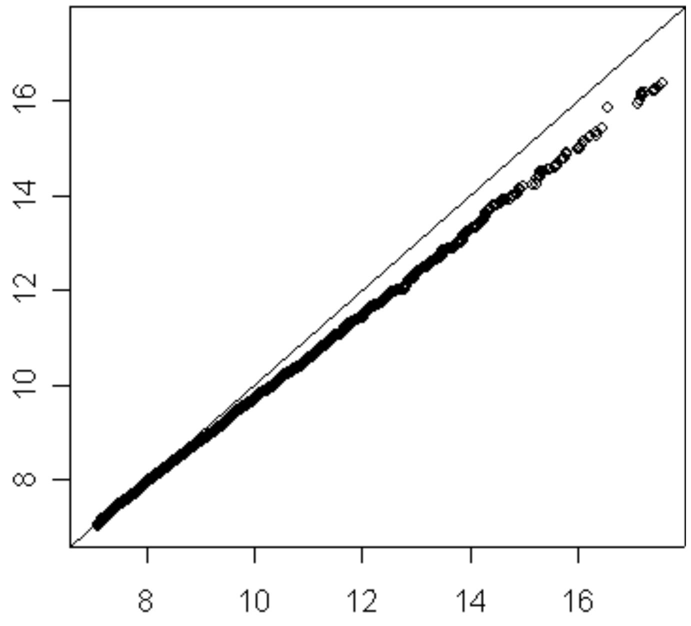
E1	E2	E3	E4	E5
28	28	28	28	28
23	23	23	23	23
19	19	19	19	19
14	14	14	14	14
8	8	8	8	8
3	3	3	3	3

3	5	6	1	5
5	6	4	4	1
2	4	1	5	3
4	2	3	3	2
6	1	2	2	6
1	3	5	6	4

Reorder

	E1	E2	E3	E4	E5
V1	3	8	19	28	23
V2	19	14	8	8	14
V3	28	3	14	14	19
V4	14	19	23	23	3
	23	28	3	19	28
V5	8	23	28	3	8

Quantile-quantile (QQ plot)



Quantile normalization

- Assumption—abundance distributions are the same
    - May be very far from true for different tissues!
    - Not true in general
- Sophisticated methods can normalize just a subset truly equivalent genes
- Can be applied to other datatypes
- Works best when you set of measurements is large and complete—not preselected to test a specific hypothesis

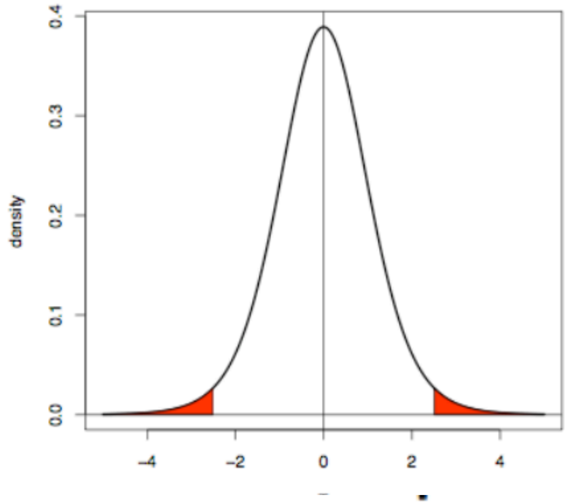
Quantile normalization—biggest assumption

- Mapping of abundance to intensity is monotone! –intensity value depends only on the true abundance
- This is not true and is sample dependent
- One important factor: GC content of sequence which affects:
  - cDNA synthesis
  - Hybridization kinetics
- Non monotonicity factors must be known and modeled explicitly
- Many methods mode GC content

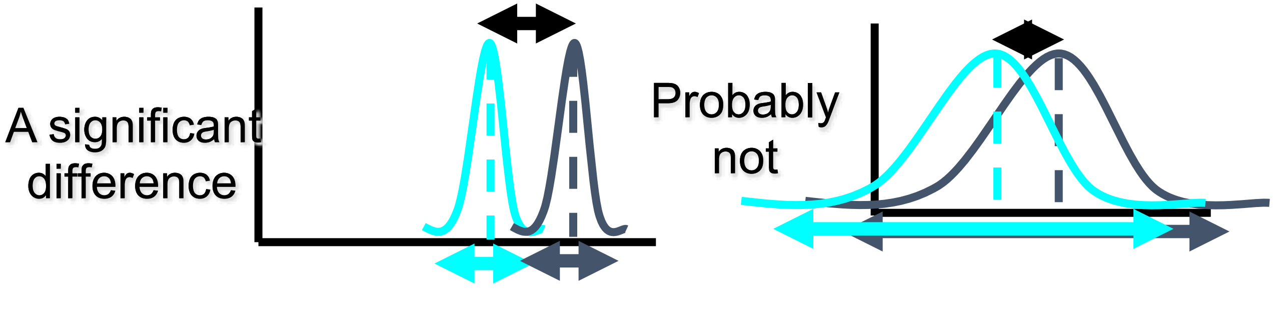
Statistical inference

- Now that the data looks good what can we say about the biology
- Simplest experimental design 2 groups
- T-test—signal to noise ratio
- Has T distribution when the two means are actually equal
- Assign p-value –small p-values means the T statistic was very unlikely for equal means
- We did an experiment and found 150 genes are differentially expressed with a p-value<0.005
- Is this a good result?

We measured 30,000 genes



$$T_g = \frac{\overline{X}_{g1} - \overline{X}_{g2}}{s_g \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$


A significant difference

Probably not

Type I/ Type II Error

Your Statistical Decision	True state of null hypothesis	
	H_0 True No difference between means	H_0 False Difference between means
Reject H_0 Conclude that samples are different	*Type I error (α)* *BIG MISTAKE*	*Correct*
Do not reject H_0 (ex: you conclude that there is insufficient evidence that the samples are different))	*Correct*	*Type II Error (β)*

Testing many hypothesis at once: Error rates

- Per-family Error Rate

  $\mathrm{PFER} = E(V)$

- Per-comparison Error Rate

  $\mathrm{PCER} = E(V)/m$

- Family-wise Error Rate

  $\mathrm{FWER} = p(V \geq 1)$

- False Discovery Rate

  $\mathrm{FDR} = E(Q)$, where

  $Q = V/R$ if $R > 0$;

  $Q = 0$ if $R = 0$

Truth Decision	# true H	# non-true H	totals
# rejected	V (Type I/big mistake)	S	R
# not rejected	U	T	$m - R$
totals	m_0	m_1	m

Adjusted p-values

- If interest is in controlling, e.g., the FWER, the adjusted p-value for hypothesis H_j is:

$$p_j^* = \inf \{\alpha: H_j \text{ is rejected at FWER } \alpha\}$$

- Hypothesis H_j is rejected at FWER α if $p_j^* \leq \alpha$

Correction procedures

- m is the total number of tests

- Bonferroni single-step adjusted p-values –controls FWER—probability of making at least one Type I error

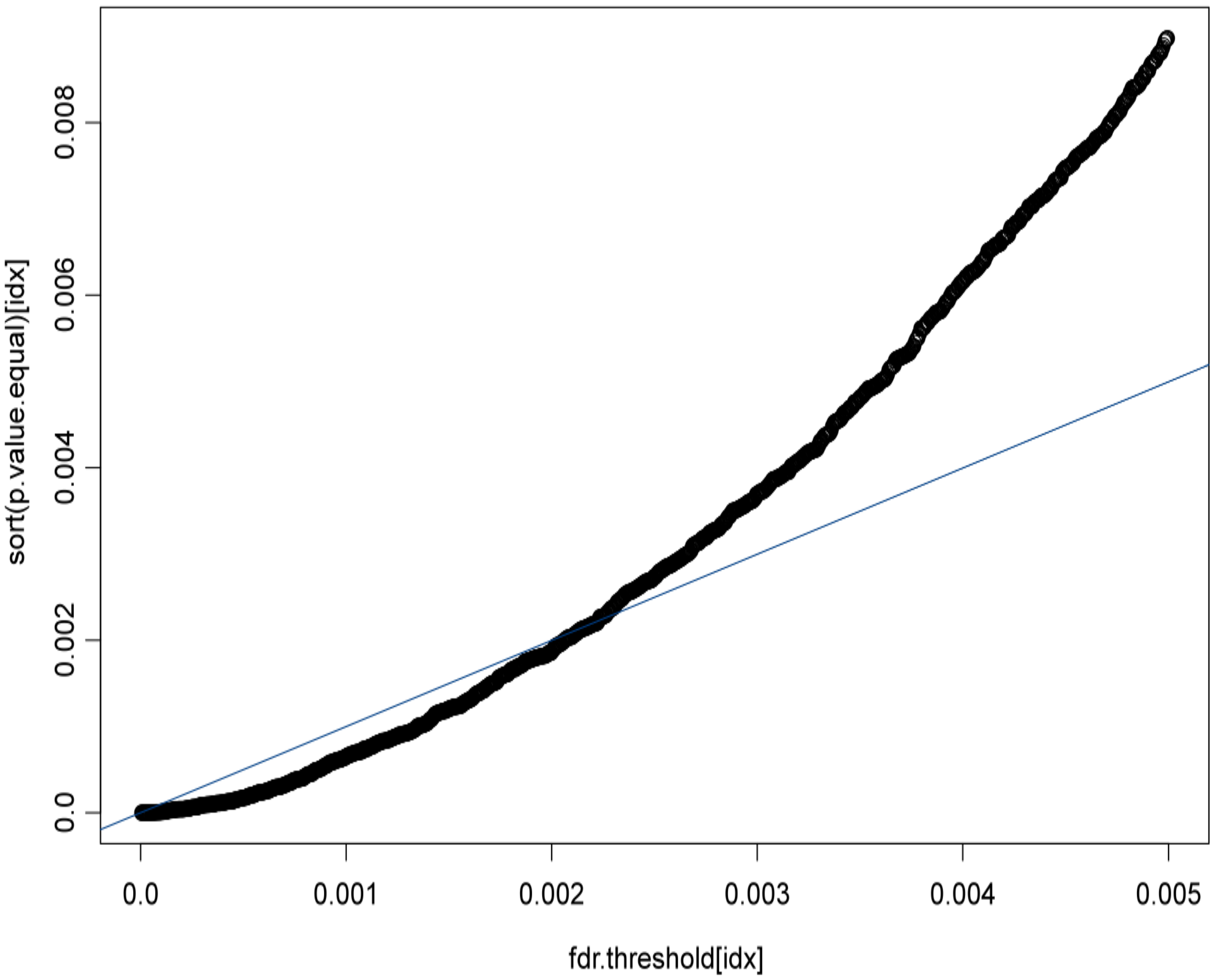
$$p_j^* = \min(mp_j, 1)$$

- Benjamini & Hochberg (1995): step-up procedure which controls the FDR under some dependency structures

$$p_{r_j}^* = \min_{k = j \ldots m} \{ \min([m/k]\, p_{r_k}, 1) \}$$

  - In practice
  - Sort p-values from smallest to largest
    - Multiply the first by m, the second by m/2, the third by m/3 ….
    - Each p_j^* is at most the minim of the ones after it --monotonicity

Visual interpretation

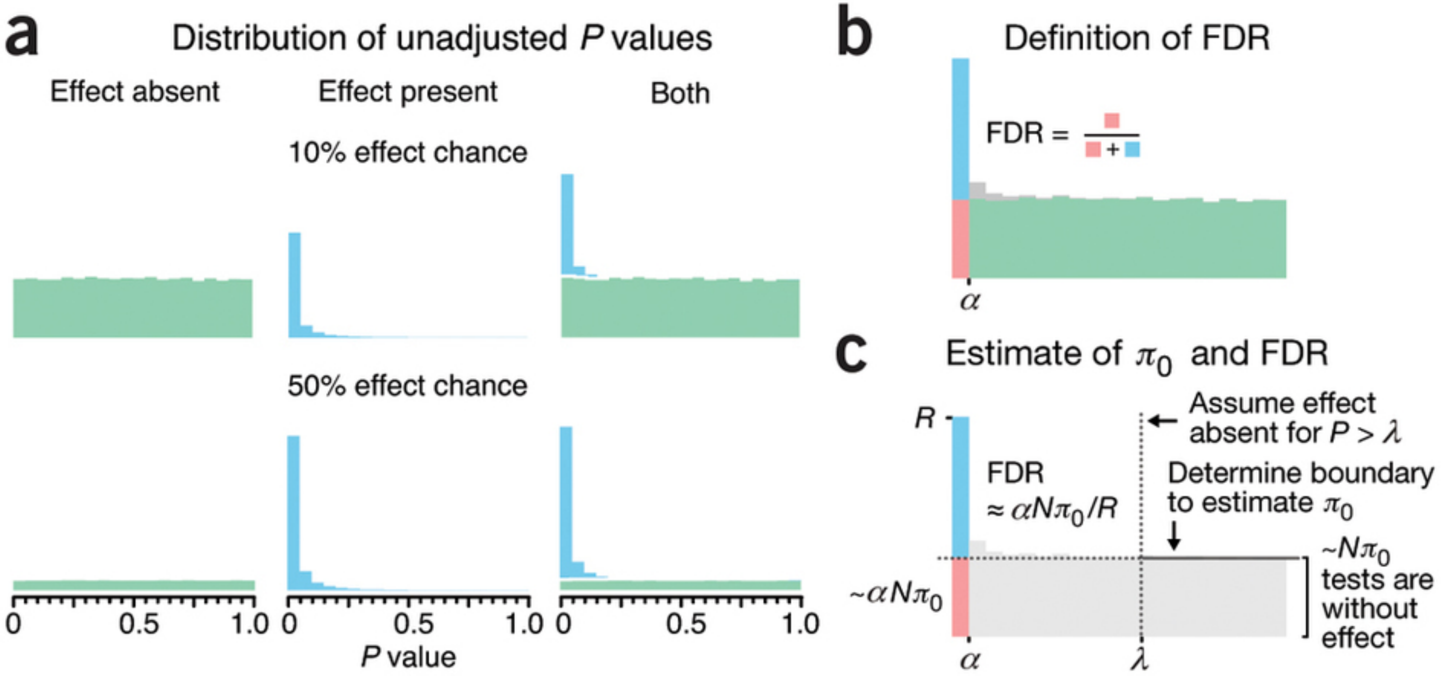


$mp/k < \alpha$

$p < k\alpha/m$

Q-value

- Storey & Tibshirani, *PNAS*, 2003
- Empirically derived – uses the p-value distribution
- Storey's method first estimates the fraction of comparisons for which the null is true, π_0,
- counting the number of *P* values larger than a cutoff λ (such as 0.5) relative to $(1 - \lambda)N$ (such as $N/2$), the count expected when the distribution is uniform
- Multiply the Benjamini & Hochberg FDR by π_0, strictly less conservative

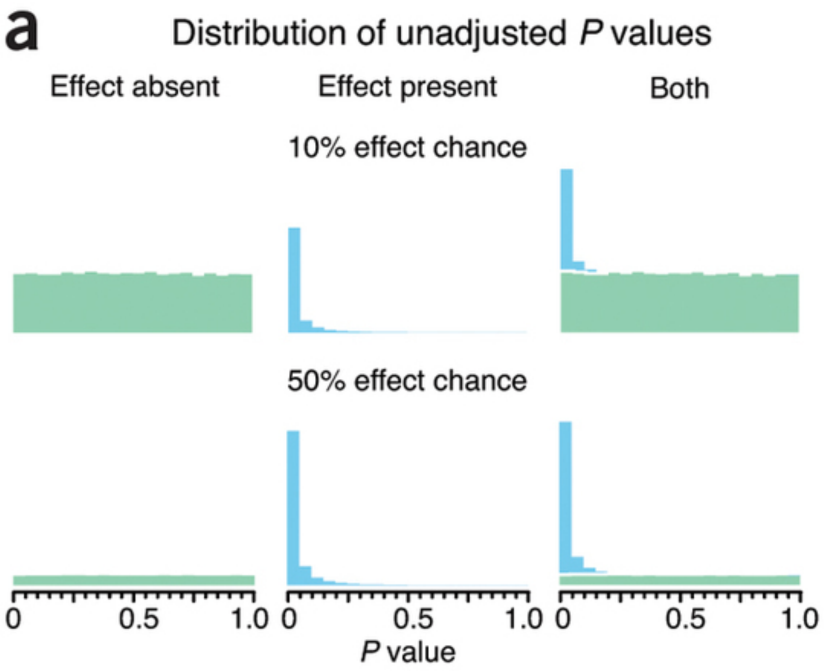


Form Points of significance: Comparing samples—part II *Nature Methods*

P-value summary

- P-value histogram can tell you there is an effect overall
    - Expect it to be uniform when there is no effect—even though individual test can return very small p-value

- π_0<1 can be used to argue that there is a difference even when no single gene is significant
    - Propose further testing such as aggregating across genes—pathway analysis (discussed later)



Beyond T-test: Significance analysis of microarrays (SAM)

- Significance analysis of microarrays applied to the ionizing radiation response Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu
- 2001
- With small sample sizes low and high variance can occur by chance
- Variance depends on expression level
- Choose S_0 so that variance is independent of expression level

Difference between the means of the two conditions

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0}$$

Fudge Factor

Estimate of the standard deviation of the numerator

Assigning significance by permutation

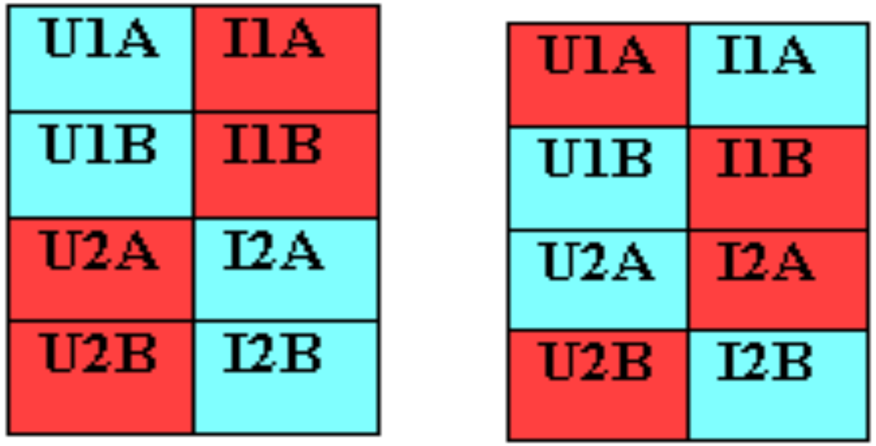
- We have calculated a new statistics and we don’t have a parametric description of the null distribution
- Solution: generate an empirical null distribution form a set of experiments where all hypotheses should be null
- Generate permutations of data labels so no difference is expected
- For each permutation p, calculate $d_p(i)$.
- Define FDR
  - Pick a threshold d_p for calling genes significant
  - Calculate the number of genes above the threshold X
  - Calculate the number of expected falsely differentially expressed genes at that threshold Y from the permuted sample analysis
  - Compute Y/X
  - 46 real DE genes , 8.4 average across permutation—FDR=.18 (8.4/46)

$$d_p(i) = \frac{\bar{x}_{G1}(i) - \bar{x}_{G2}(i)}{s(i) + s_0}$$

More on permutations

- Very small experiment-random permutations may create unbalanced groups
  - Solution: restrict to balanced permutations-each permutation should split the real groups equally
- Can be applied to up/down regulated genes separately
- Permutation analysis can be applied to any complicated statistical procedure!

Balanced permutations
Number of red and cyan groups is equal



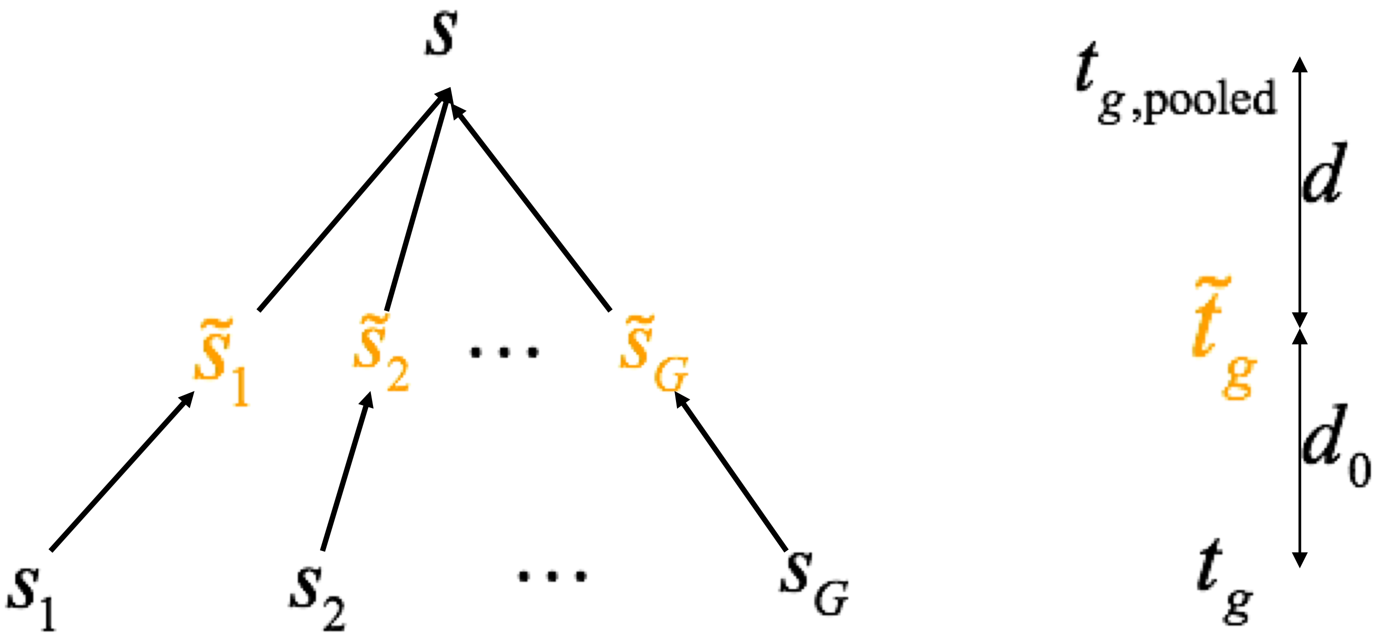
🟦 U1A	🟥 I1A
🟦 U1B	🟥 I1B
🟥 U2A	🟦 I2A
🟥 U2B	🟦 I2B

🟥 U1A	🟦 I1A
🟦 U1B	🟥 I1B
🟦 U2A	🟥 I2A
🟥 U2B	🟦 I2B

Why does SAM work

$$d_p(i) = \frac{\bar{x}_{G1}(i) - \bar{x}_{G2}(i)}{s(i) + s_0}$$

- Sample variance in not an accurate assessment of the true variance
- What would the per gene variance be we had an infinite number of samples?
- SAM is in example of **moderated T statistic**
- Many current methods use a more principled Bayesian method



Bayesian reasoning: short intro

- Synthesize prior knowledge and evidence
- Main theorem
- Simple derivation

P(A and B)=

P(A|B)P(B)=p(B|A)p(A)

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

Classical example

- Duchenne Muscular Dystrophy (DMD) can be regarded as a simple recessive sex-linked disease caused by a mutated X chromosome (X).
    - An XY male expresses the disease, whereas an XX female is a carrier but does not express the disease
- Suppose neither of a woman's parents expresses the disease, but her brother does. Then the woman's mother must be a carrier, and the woman herself therefore may be a carrier
- P(C)=1/2
- What if she has a healthy son?

$$p(C|\text{h.s.}) = \frac{p(\text{h.s.}|C)p(C)}{p(\text{h.s.})}$$

$$\frac{p(\text{h.s.}|C)p(C)}{p(\text{h.s.}|C)p(C) + p(\text{h.s.}|\overline{C})p(\overline{C})} =$$

$$\frac{(1/2)\cdot(1/2)}{(1/2)\cdot(1/2) + 1\cdot(1/2)} = \frac{1}{3}$$

Bayesian approach to statistics

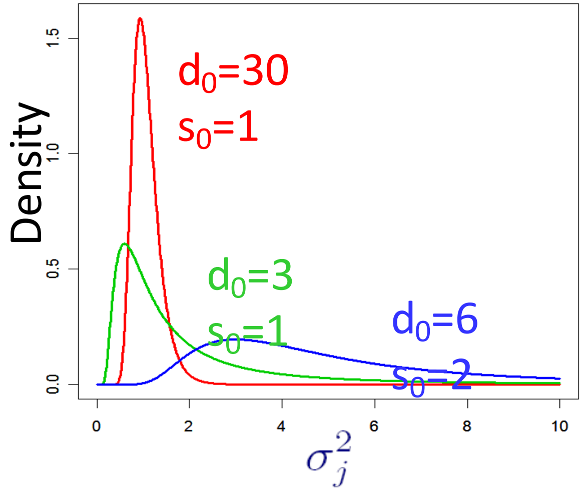
- Last example: incorporate evidence into strong prior belief
- Statistics
  - Naïve approach: is estimate the parameters from observation only
  - Bayesian approach: have some prior expectation
  - Prior expectation: Variance should not be too big or too small
- Bayesian statistical analyses:
  - begin with ‘prior’ distributions describing beliefs about the values of parameters in statistical models prior to analysis of the data at hand
  - require specification of these parameters
  - ‘Empirical Bayes’ methods use the data at hand to guide prior parameter specification
  - Then given the data, these prior distributions are updated to give posterior results

A few more details

- Gene specific variance is sampled from a distribution of variance parameters
  - True variance is unknown
  - Only sample variance is known
- Scaled inverse chi-squared distribution

$$\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2 \sim G(\theta)$$

$$\frac{s_0^2}{\sigma_1^2}, \frac{s_0^2}{\sigma_2^2}, \ldots, \frac{s_0^2}{\sigma_J^2} \sim \frac{\chi_{d_0}^2}{d_0}$$



Moderated estimate

Sample estimate

$$\tilde{s}_j^2 = \frac{d s_j^2 + d_0 s_0^2}{d + d_0}$$

Distribution parameters

Software: Limma

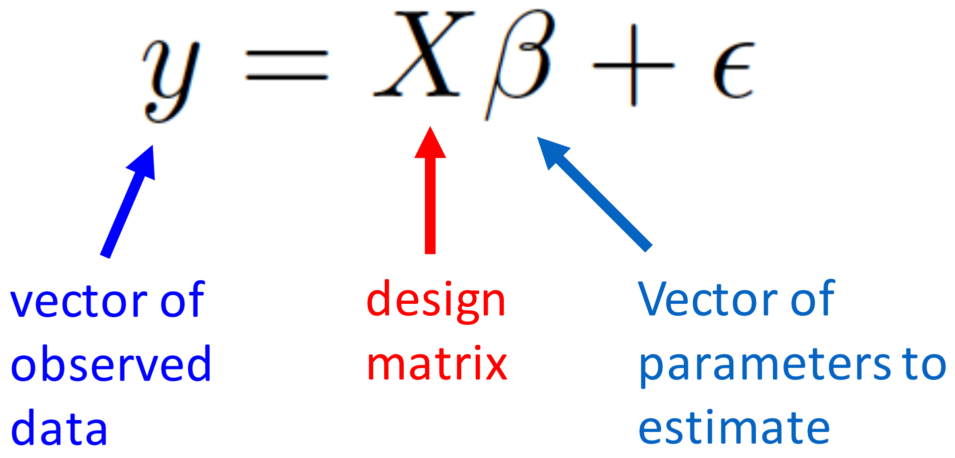
Baldi & Long 2001, Wright & Simon 2003, Smyth 2004

More complicated models

- So far we only consider 2 group experiments
- Many other possibilities
  - Factorial: two groups each has two treatments--Are treatment effects different across groups?
  - Continuous variables: dosage of a drug
  - Continuous discrete variables
    - 2 groups, 3 drug doses—do the drugs affect the groups differently?

General framework for differential expression

- Linear models
- Model the expression of each gene as a linear function of explanatory variables
  - Groups
  - Treatments
  - Combinations of groups and treatments
  - Etc…



$$y = X\beta + \epsilon$$

vector of observed data

design matrix

Vector of parameters to estimate

Example of a design matrix

Normal sample x 2

Cancer Sample x 2

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

β_1 = normal log-expression

β_2 = cancer – wt

$E[y_1]=E[y_2]=\beta_1$

$E[y_3]=E[y_4]= \beta_1+ \beta_2$

More examples

$$y = X\beta + \epsilon$$

- 6 samples
- 2 groups + drug treatment
- Group and treatment effect are additive

Global mean	Group 2	Drug dose
1	0	0.25
1	0	1
1	0	4
1	1	0.25
1	1	1
1	1	4

3 coefficients to estimate

More examples

$$y = X\beta + \epsilon$$

Global mean	Group 2	Drug dose	Drug dose + Group 2
1	0	0.25	0
1	0	1	0
1	0	4	0
1	1	0.25	0.25
1	1	1	1
1	1	4	4

- 6 samples
- 2 groups + drug treatment
- Treatments affect groups differently

4 coefficients to estimate

Linear model parameter estimation

Model is specified –how do we find the coefficients

$$y = X\beta + \epsilon$$

- Minimize squared error

$$\epsilon'\epsilon = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)$$

- Take derivative

$$\frac{d}{d\beta}((\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta)) = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta)$$

- Set to 0

$$-2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}$$

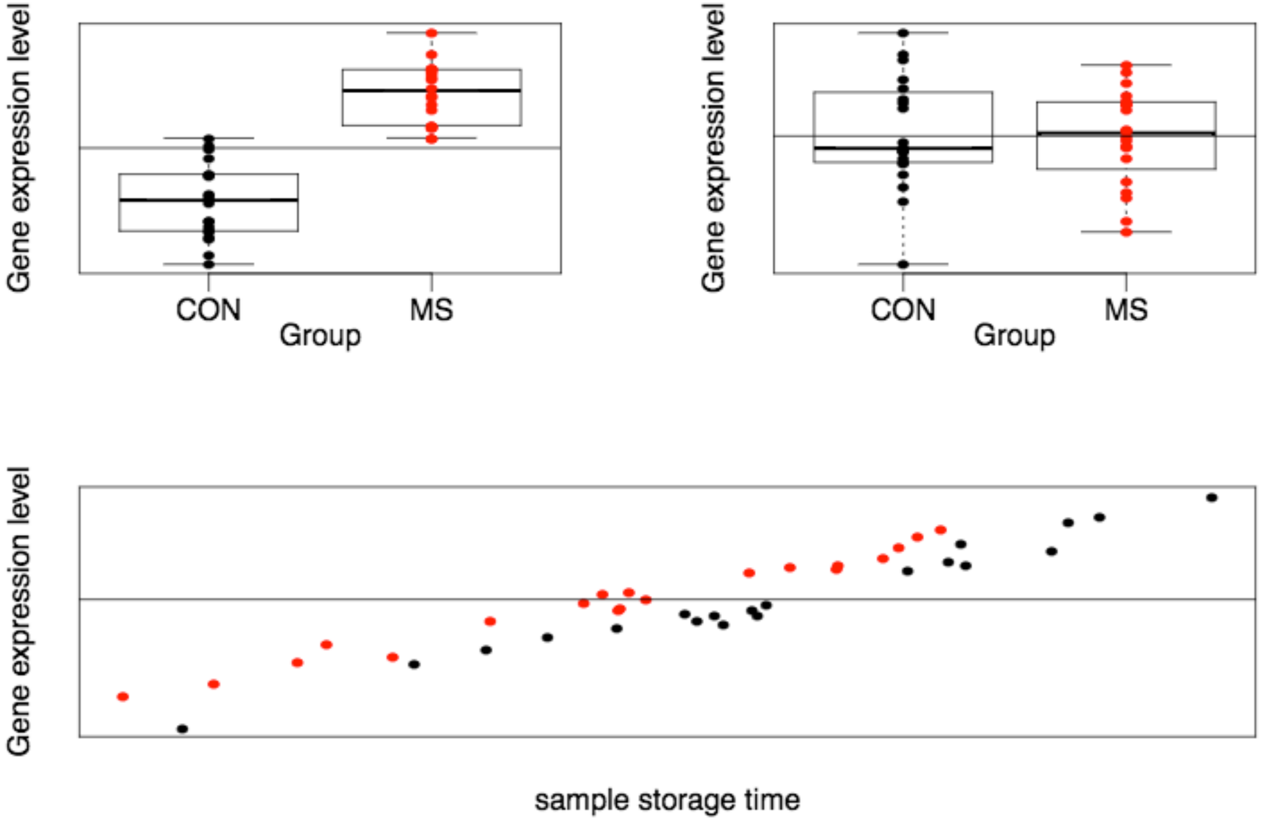
- Solve
  - Significance of coefficients is tested with a T-test

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})\beta$$

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

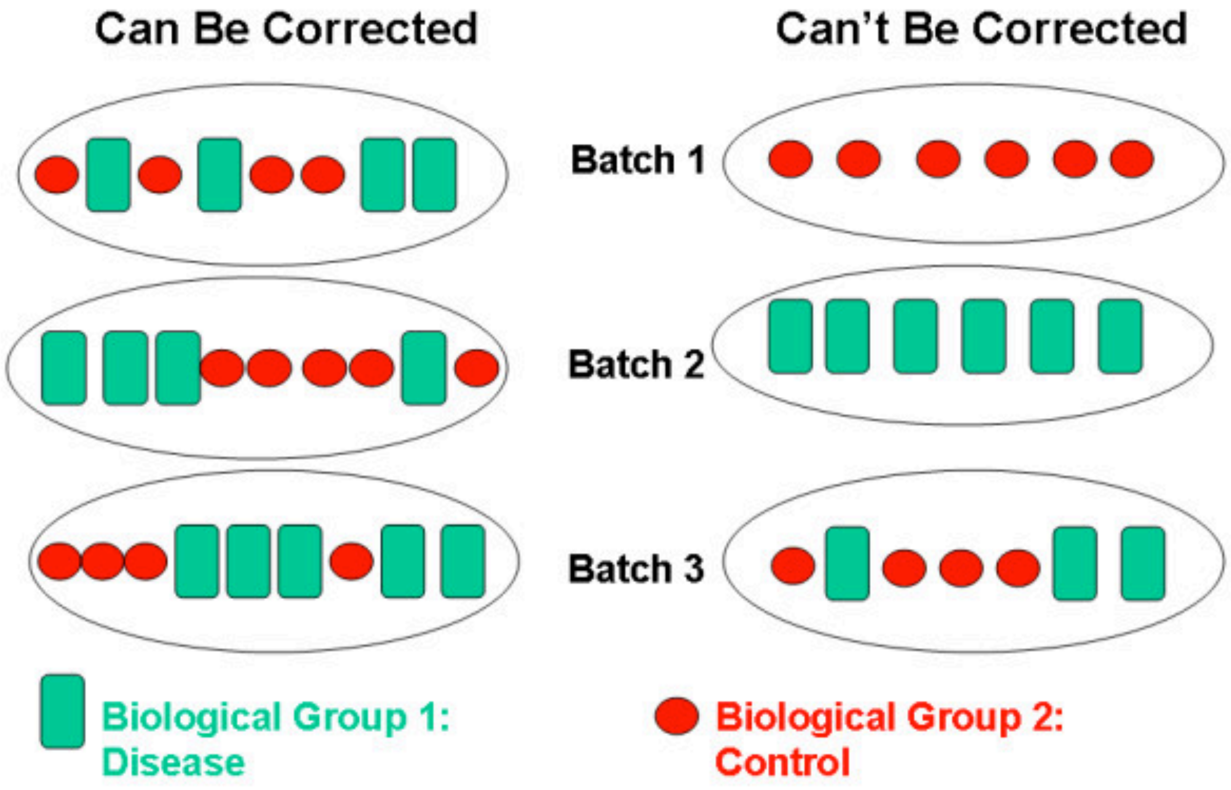
Linear models for data clean up

- Linear models are useful for including nuisance variables--Technical factors
- Variables that have an effect on measurements but are not themselves of interest
- 2 group design
- Control vs MS
- Variable sample storage time



Batch effects

- Technical variables that effect gene expression
- Often occur when samples were processed in “batches”
  - At different times
  - Different locations
  - Different technician
  - Different protocol
- Batch variables are often discrete but can be continuous (such as: storage time)
  - With RNAseq we can model sample specific GC bias
- Batch effects can be corrected if they don’t align with a variable of interest
- Experimental design is important!



General approaches to multi dimensional data

- Many more measurements than samples
- Use measurement distributions for normalization and filtering
- Borrow information across measurements for hypothesis testing

Comparisons

- In general, for a given multiple testing procedure,

$$\text{PCER} \leq \text{FWER} \leq \text{PFER},$$

and

$$\text{FDR} \leq \text{FWER},$$

with $\text{FDR} = \text{FWER}$ under the complete null

Cluster analyses:

1) Usually outside the normal framework of statistical inference;

2) less appropriate when only a few genes are likely to change.

3) Needs lots of experiments

Single gene tests:

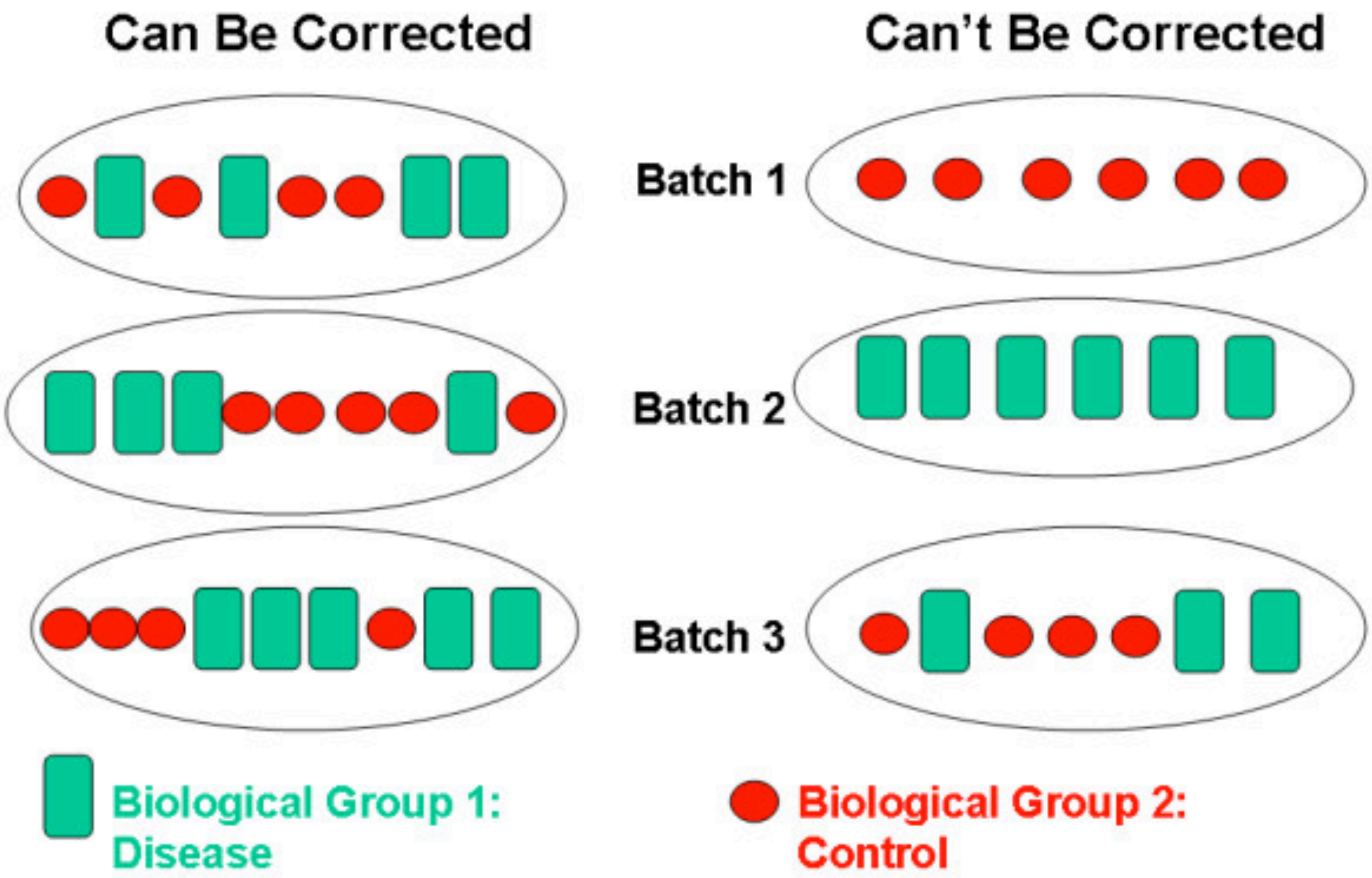
1) may be too noisy in general to show much

2) may not reveal coordinated effects of positively correlated genes.

3) hard to relate to pathways.

Example

- 150 genes were difference with a T test p-value of < 0.05
- Is this a good result?



Can Be Corrected

Can't Be Corrected

Batch 1

Batch 2

Batch 3

🟩 Biological Group 1: Disease

● Biological Group 2: Control