Microarray Data Analysis

Statistical methods to detect differentially expressed genes

Outline

- The *class comparison* problem
- Statistical tests
  - Calculation of p-values
  - Permutations tests
  - The volcano plot
- Multiple testing
- Extensions
- Examples

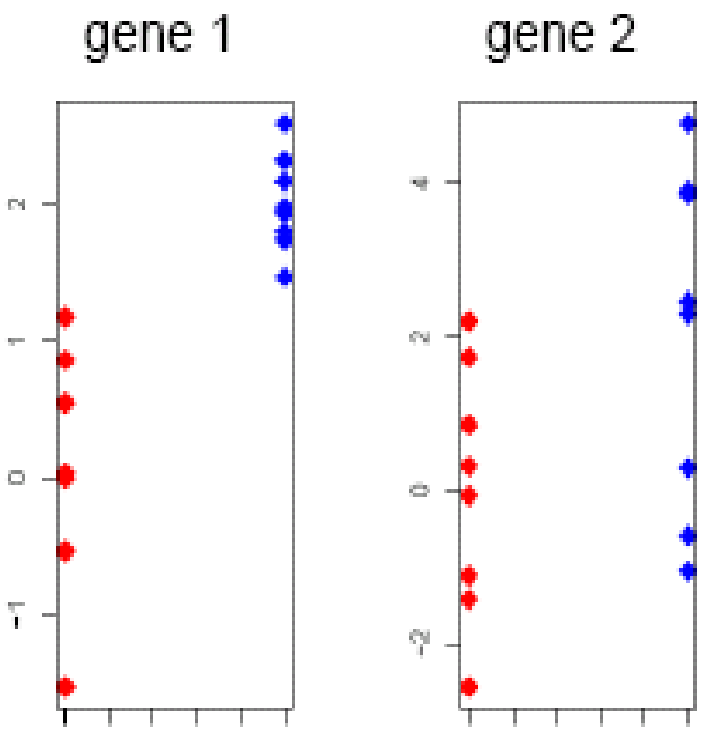
Class comparison: Identifying *differentially expressed* genes

- Identify genes whose expression is *significantly* associated with different conditions
  - Treatment, cell type,... (qualitative covariates)
  - Dose, time, ... (quantitative covariate)
  - Survival, infection time,... !
- Estimate effects/differences between groups probably using log-ratios, i.e. the difference on log scale:

  log(X)-log(Y) [=log(X/Y)]

What is a “significant change”?

- Depends on the variability within groups, which may be different from gene to gene.
- To assess the statistical significance of differences, conduct a statistical test for each gene.



Different settings for statistical tests

- Indirect comparisons: 2 groups, 2 samples, unpaired
  - E.g. 10 individuals: 5 suffer diabetes, 5 healthy
  - One sample fro each individual
  - Typically: Two sample t-test or similar
- Direct comparisons: Two groups, two samples, paired
  - E.g. 6 individuals with brain stroke.
  - Two samples from each: one from healthy (region 1) and one from affected (region 2).
  - Typically: One sample t-test (also called paired t-test) or similar, based on the individual differences between conditions.

Different ways to do the experiment

- An experiment use cDNA arrays (“two-colour”) or affy (“one-colour).
- Depending on the technology used allocation of conditions to slides changes.

Type of chip / Experiment	cDNA (2-col)	Affy (1-col)
10 indiv. Diab (5) Heal (5)	*Reference design.* (5) Diab/Ref (5) Heal/Ref	*Comparison design.* (5) Diab vs (5) Heal
6 indiv. Region 1 Region 2	*6 slides* *1 individual per slide* (6) reg1/reg2	*12 slides* (6) Paired differences

“Natural” measures of discrepancy

For **Direct comparisons** *in two colour or paired-one colour.*

Mean (log) ratio $= \frac{1}{n_T}\sum_{i=1}^{n_T} R_i$, (R or M used indistinctly)

Classical t-test = $t = (\bar{R})/SE$, (SE estimates standard error of $\bar{R}$)

Robust t-test = Use robust estimates of location &scale

For **Indirect comparisons** *in two colour or*
Direct comparisons *in one colour.*

Mean difference $= \frac{1}{n_T}\sum_{i=1}^{n_T} T_i - \frac{1}{n_C}\sum_{i=1}^{n_C} C_i = \bar{T} - \bar{C}$

Classical t-test = $t = (\bar{T} - \bar{C})/s_p\sqrt{1/n_T + 1/n_C}$

Robust t-test = Use robust estimates of location &scale

Some issues in gene selection

- Gene expression values have peculiarities that have to be dealt with.
- Some related with small sample sizes
  - **Variance unstability**
  - **Non-normality of the data**
- Other related to big number of variables
  - **Multiple testing**

Variance unstability

- Can we trust average *effect sizes* (average difference of means) alone?
- Can we trust the *t-statistic* alone?
- Here is evidence that the answer is no.

Gene	M1	M2	M3	M4	M5	M6	Mean	SD	t
A	2.5	2.7	2.5	2.8	3.2	2	2.61	0.40	16.10
B	0.01	0.05	-0.05	0.01	0	0	0.003	0.03	0.25
C	2.5	2.7	2.5	1.8	20	1	5.08	7.34	1.69
D	0.5	0	0.2	0.1	-0.3	0.3	0.13	0.27	1.19
E	0.1	0.11	0.1	0.1	0.11	0.09	0.10	0.01	33.09

***M** may be assumed to be the log-Fold change in a paired experiment*

Courtesy of Y.H. Yang

Variance unstability (1): outliers

- Can we trust average effect sizes (average difference of means) alone?
- Can we trust the t statistic alone?
- Here is evidence that the answer is no.

Gene	M1	M2	M3	M4	M5	M6	Mean	SD	t
A	2.5	2.7	2.5	2.8	3.2	2	2.61	0.40	16.10
B	0.01	0.05	-0.05	0.01	0	0	0.003	0.03	0.25
C	2.5	2.7	2.5	1.8	20	1	5.08	7.34	1.69
D	0.5	0	0.2	0.1	-0.3	0.3	0.13	0.27	1.19
E	0.1	0.11	0.1	0.1	0.11	0.09	0.10	0.01	33.09

•Averages can be driven by outliers.

Courtesy of Y.H. Yang

Variance unstability (2): tiny variances

- Can we trust average effect sizes (average difference of means) alone?
- Can we trust the t statistic alone?
- Here is evidence that the answer is no.

Gene	M1	M2	M3	M4	M5	M6	Mean	SD	t
A	2.5	2.7	2.5	2.8	3.2	2	2.61	0.40	16.10
B	0.01	0.05	-0.05	0.01	0	0	0.003	0.03	0.25
C	2.5	2.7	2.5	1.8	20	1	5.08	7.34	1.69
D	0.5	0	0.2	0.1	-0.3	0.3	0.13	0.27	1.19
E	0.1	0.11	0.1	0.1	0.11	0.09	0.10	0.01	33.09

•t's can be driven by tiny variances.

Courtesy of Y.H. Yang

Solutions: Adapt t-tests

- Let
  - R_g mean observed log ratio
  - SE_g standard error of R_g estimated from data on gene *g*.
  - SE standard error of R_g estimated from data across all genes.
- Global t-test: $t=R_g/SE$
- Gene-specific t-test $t=R_g/SE_g$

Some pro's and con's of t-test

Test	Pro's	Con's
Global t-test: $t=R_g/SE$	Yields stable variance estimate	Assumes variance homogeneity → biased if false
Gene-specific: $t=R_g/SE_g$	Robust to variance heterogeneity	■ Low power ■ Yields unstable variance estimates (due to few data)

T-tests extensions

SAM
(Tibshirani, 2001)

$$S = \frac{R_g}{c + SE_g}$$

Regularized-t
(Baldi, 2001)

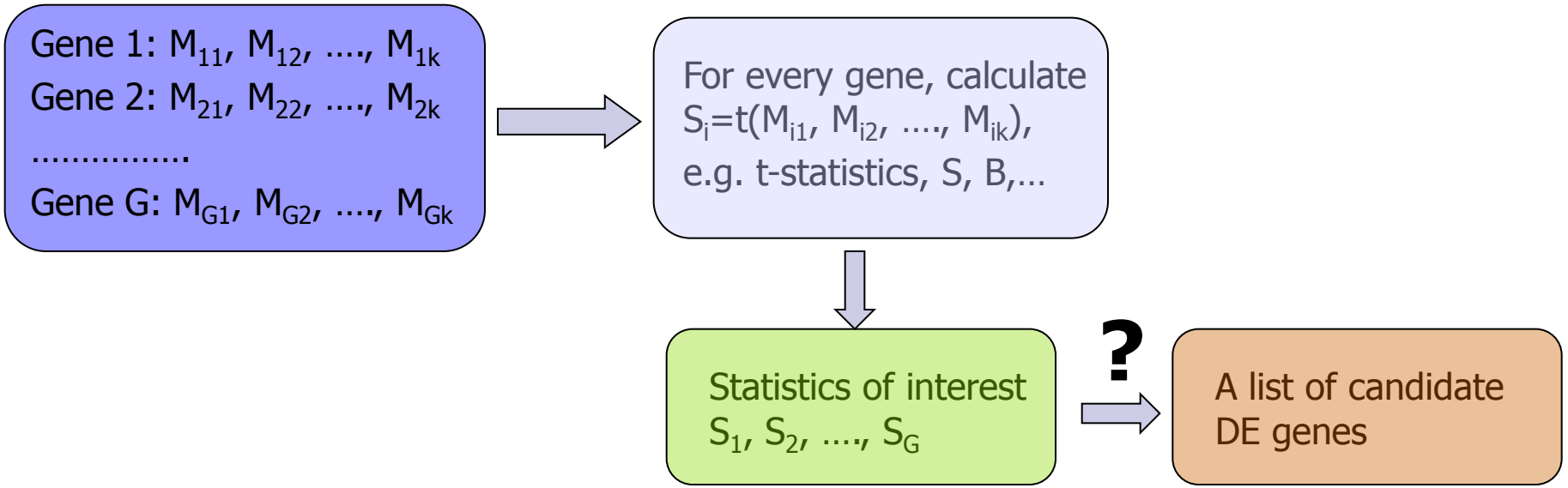
$$t = \frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1)SE_g^2}{v_0 + n - 2}}}$$

EB-moderated t
(Smyth, 2003)

$$t = \frac{R_g}{\sqrt{\frac{d_0 \cdot SE_0^2 + d \cdot SE_g^2}{d_0 + d}}}$$

Up to here…: Can we generate a list of candidate genes?

With the tools we have, the reasonable steps to generate a list of candidate genes may be:



We need an idea of how significant are these values
→We'd like to assign them *p-values*

Nominal p-values

- After a test statistic is computed, it is convenient to convert it to a *p*-value:

  The probability that a test statistic, say S(X), takes values equal or greater than the observed value, say $X^{0,}$ under the assumption that the null hypothesis is true

$$p=P\{S(X)>=S(X^0)|H_0 \ true\}$$

Significance testing

- Test of significance at the α level:
  - *Reject the null hypothesis if your p-value is smaller than the significance level*
  - It has advantages but not free from criticisms
- Genes with *p*-values falling below a prescribed level may be regarded as significant

Hypothesis testing overview for a single gene

		Reported decision		
		H_0 is Accepted *(gene not Selected)*	H_0 is Rejected *(gene is Selected)*	
State of the nature ("Truth")	H_0 is true *(Not Affected)*	TN , prob: β	FP, P[Rej H_0|H_0]<= α Type I error	Specificity TN/[TN+FP]
	H_0 is false *(Affected)*	FN, prob: 1-β Type II error	TP, prob: 1-α	Sensitiviy TP/[TP+FN]
		Negative predictive value TN/[TN+FN]	Positive predictive value TP/[TP+FP]	

Calculation of p-values

- Standard methods for calculating p-values:

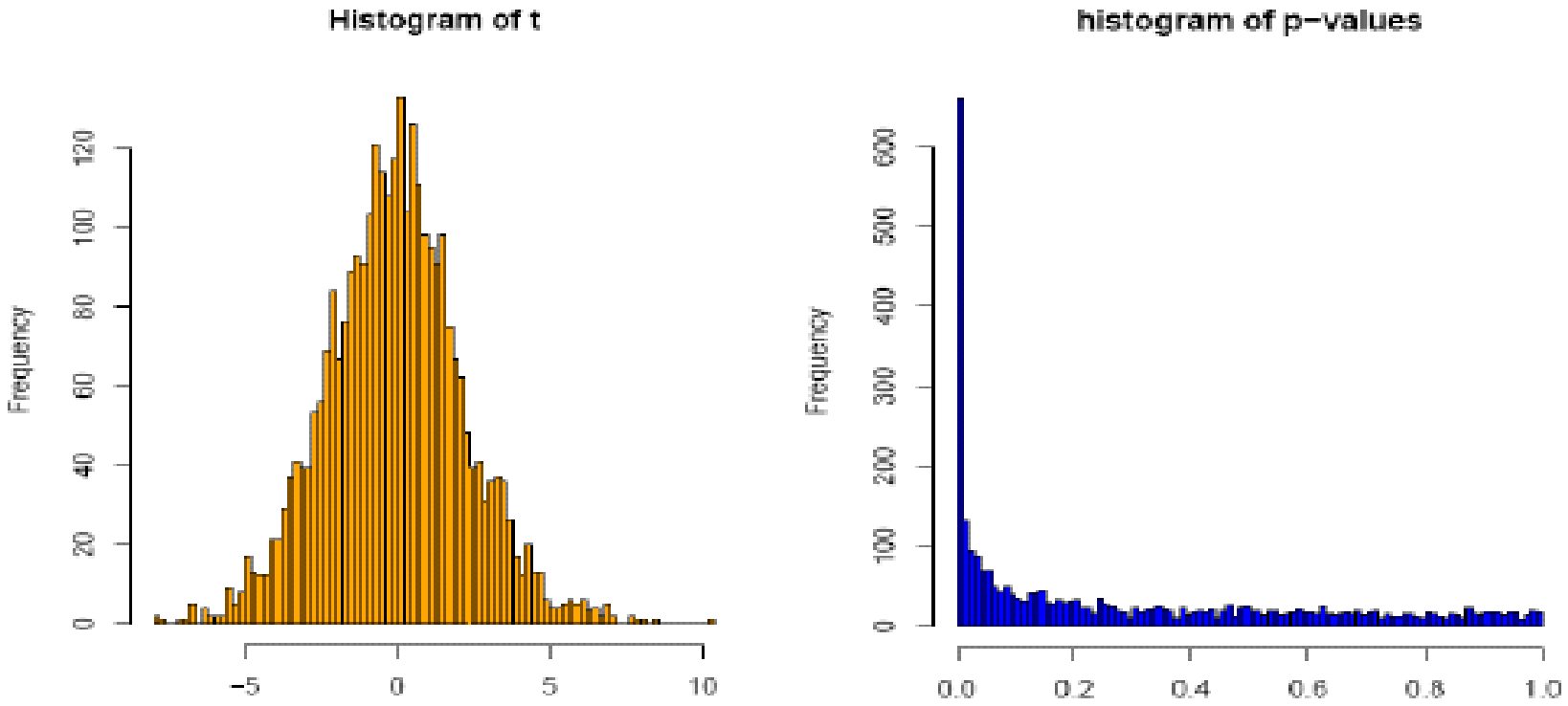
  (i) Refer to a statistical distribution table (*Normal, t, F*, …) or

  (ii) Perform a permutation analysis

(i) Tabulated p-values

- Tabulated *p*-values can be obtained for standard test statistics (e.g.the *t*-test)
- They often rely on the assumption of normally distributed errors in the data
- This assumption can be checked (approximately) using a
  - Histogram
  - Q-Q plot

Example (1)



Golub data, 27 ALL vs 11 AML samples, 3051 genes

A *t*-test yields 1045 genes with $p < 0.05$

(ii) Permutations tests

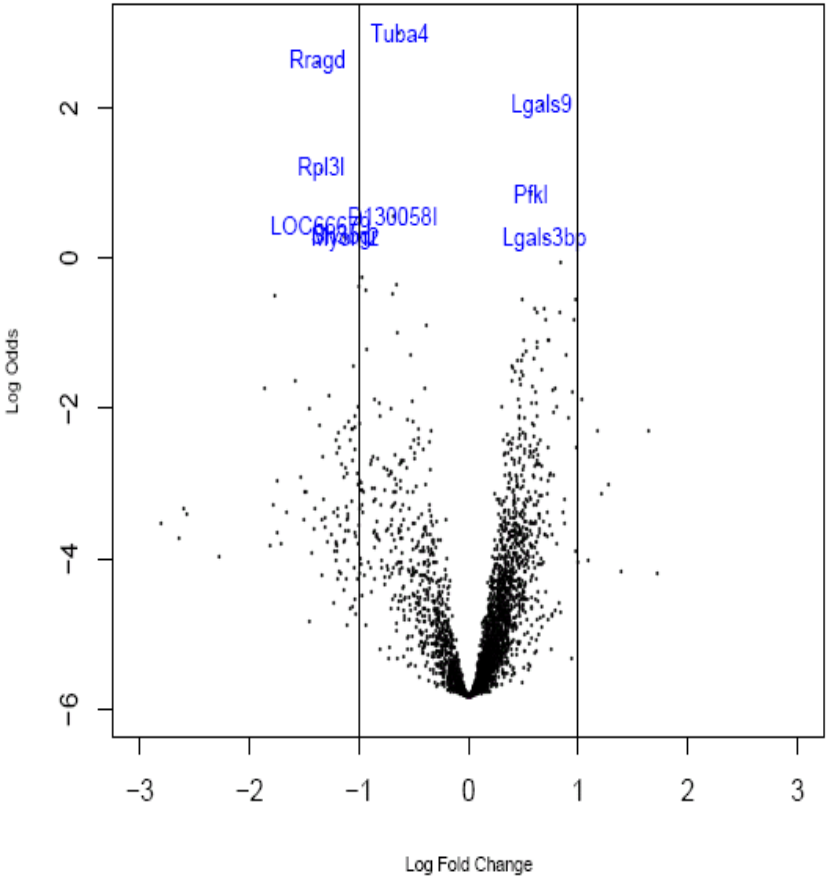
- Based on data shuffling. No assumptions
  - Random interchange of labels between samples
  - Estimate p-values for each comparison (gene) by using the permutation distribution of the *t*-statistics
- Repeat for every possible permutation, b=1…B
  - Permute the *n* data points for the gene (*x*). The first *n*1 are referred to as "treatments", the second *n*2 as "controls"
  - For each gene, calculate the corresponding two sample t-statistic, *tb*
- After all the *B* permutations are done put
  $p = \#\{b: |tb| \geq |tobserved|\}/B$

Permutation tests (2)

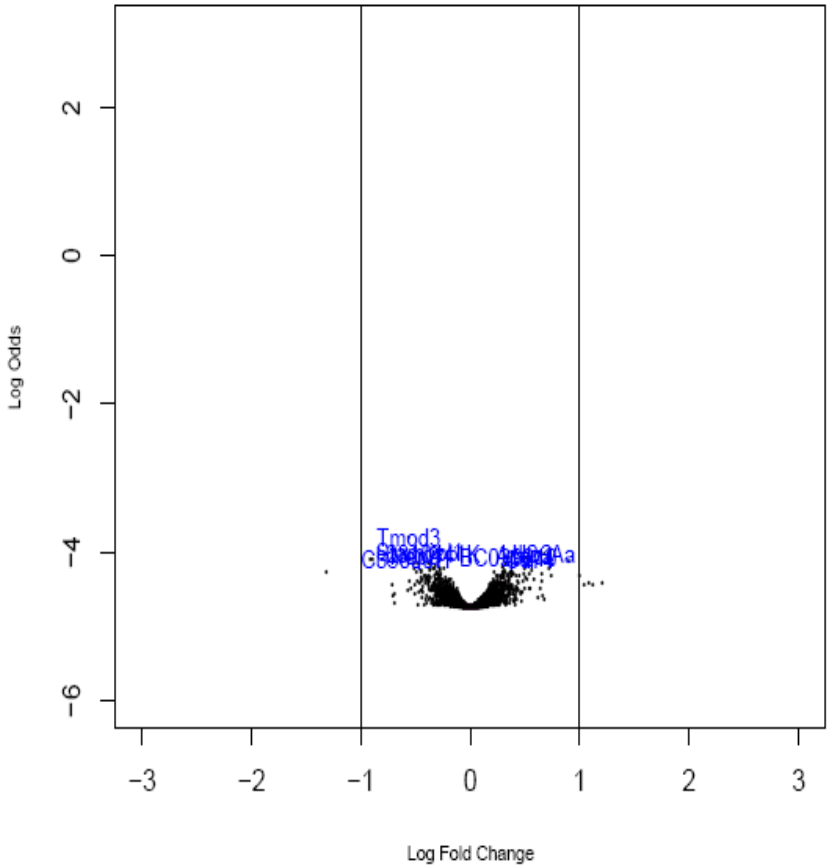
	Class 1 data values					Class 2 data values				Parametric *t*-statistic
original data:	−0.18	−0.10	−0.13	0.30	−0.14	0.15	0.84	0.66	0.52	$t = 3.64$
data permutation 1:	−0.18	−0.10	−0.13	0.30	−0.14	0.15	0.84	0.66	0.52	$t^* = \mathbf{3.64}$
data permutation 2:	−0.18	−0.10	−0.13	0.30	0.15	−0.14	0.84	0.66	0.52	$t^* = 2.15$
data permutation 3:	−0.18	−0.10	−0.13	0.15	0.84	0.30	−0.14	0.66	0.52	$t^* = 0.83$
data permutation 4:	−0.18	−0.10	−0.13	−0.14	0.15	0.30	0.84	0.66	0.52	$t^* = \mathbf{5.48}$
⋮			⋮				⋮			⋮
data permutation 124:	0.30	−0.14	0.84	0.66	0.52	−0.18	−0.10	−0.13	0.15	$t^* = -2.48$
data permutation 125:	0.30	0.15	0.84	0.66	0.52	−0.18	−0.10	−0.13	−0.14	$t^* = \mathbf{-4.49}$
data permutation 126:	−0.14	0.15	0.84	0.66	0.52	−0.18	−0.10	−0.13	0.30	$t^* = -2.48$

$$\text{permutation } p\text{-value} = \frac{\#\text{ of the 126 permuations where } |t^*| \geq |t|}{126} = \frac{3}{126}$$

The volcano plot: fold change vs log(odds)[1]



Significant change detected



No change detected

1: log(odds) is proportional to "-log (p-value)"

Multiple testing

How far can we trust the decision?

- The test: "*Reject H_0 if p-val ≤ α*"
  - is said to *control* the type I error because, under a certain set of assumptions, the probability of falsely rejecting H_0 is less than a fixed small threshold

    $$P[\text{Reject } H_0 | H_0 \text{ true}] = P[FP] \leq \alpha$$

  - Nothing is warranted about P[FN]→
    - "Optimal" tests are built trying to minimize this probability
    - In practical situations it is often high

What if we wish to test more than one gene at once? (1)

- Consider more than one test at once
  - Two tests each at 5% level. Now probability of getting a false positive is $1 – 0.95*0.95 = 0.0975$
  - Three tests → $1 – 0.95^3 = 0.1426$
  - *n* tests → $1 – 0.95^n$
  - Converge towards 1 as n increases
- Small p-values don't necessarily imply significance!!! → We are not controlling the probability of type I error anymore

What if we wish to test more than one gene at once? (2): a simulation

- Simulation of this process for 6,000 genes with 8 treatments and 8 controls
- ***All*** the gene expression values were simulated *i.i.d* from a *N* (0,1) distribution, i.e. ***NOTHING*** is differentially expressed in our simulation
- The number of genes falsely rejected will be on the average of $(6000 \cdot \alpha)$, i.e. if we wanted to reject all genes with a p-value of less than 1% we would falsely reject around 60 genes

See example

Multiple testing: Counting errors

		Decision reported		
		H_0 is accepted *(Genes not Selected)*	H_0 is Rejected *(Genes Selected)*	Total
State of the nature ("Truth")	H_0 is true *(Not Affected)*	m_o-$\alpha m0$ (U)	αm_0 (V)	m_o
	H_0 is false *(Affected)*	$(m-m_o)$-$(m_\alpha - \alpha m_0)$ (T)	$m_\alpha - \alpha m_0$ (S)	$m-m_o$
Total		$m-m_\alpha$ (m-R)	M_α (R)	m

V = # Type I errors [false positives]

T = # Type II errors [false negatives]

All these quantities could be known if m_0 was known

How does type I error control extend to multiple testing situations?

- Selecting genes with a p-value less than α doesn't control for *P[FP]* anymore
- What can be done?
  - Extend the idea of type I error
    - FWER and FDR are two such extensions
  - Look for procedures that control the probability for these extended error types
    - Mainly adjust raw p-values

Two main error rate extensions

- Family Wise Error Rate (FWER)
  - FWER is probability of at least one false positive

    FWER= Pr(# of false discoveries >0) = Pr(V>0)
- False Discovery Rate (FDR)
  - FDR is expected value of proportion of false positives among rejected null hypotheses

    FDR = E[V/R; R>0] = E[V/R | R>0]·P[R>0]

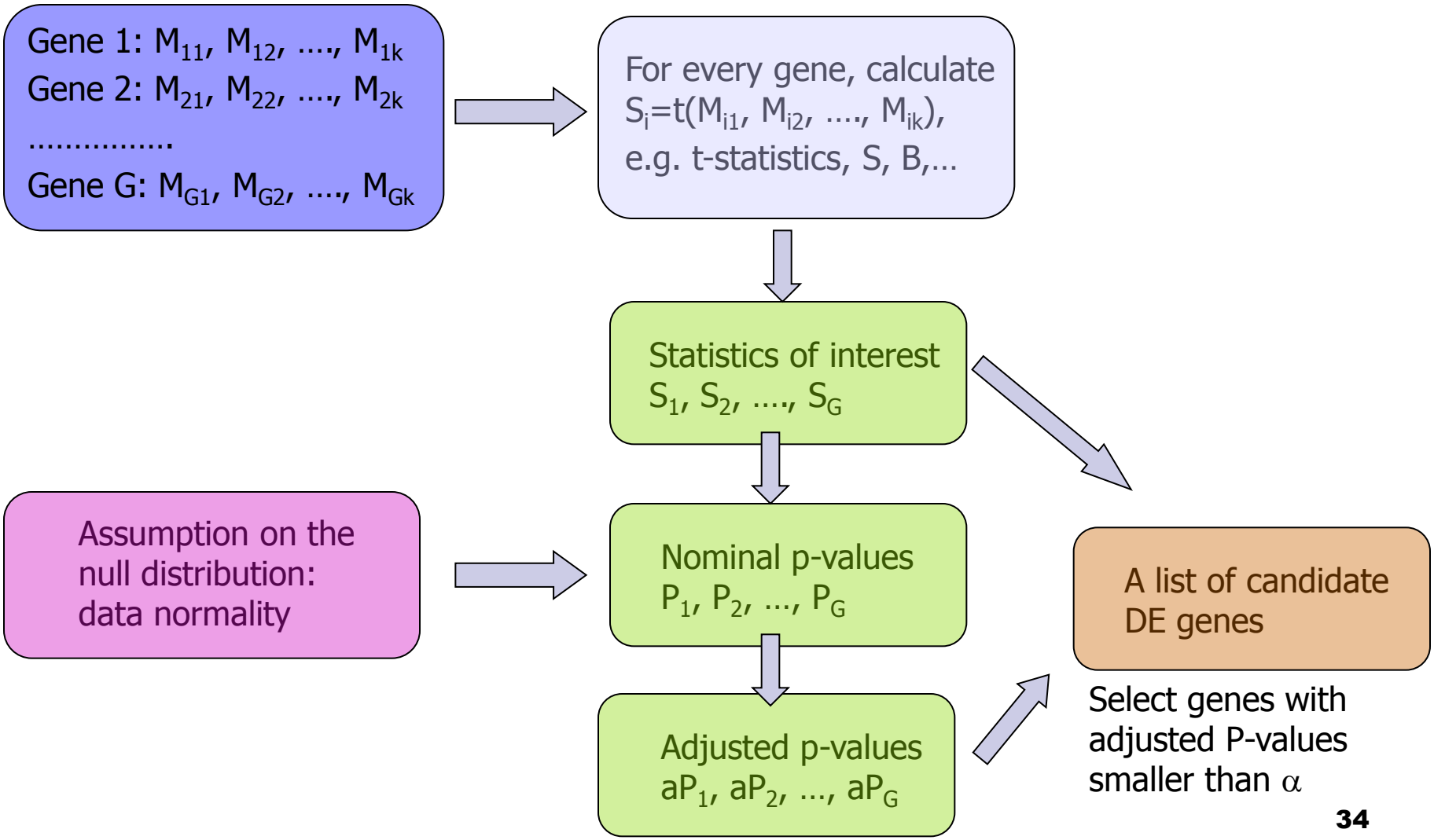
FDR and FWER controlling procedures

- FWER
  - Bonferroni (adj Pvalue = min{n*Pvalue,1})
  - Holm (1979)
  - Hochberg (1986)
  - Westfall & Young (1993) maxT and minP
- FDR
  - Benjamini & Hochberg (1995)
  - Benjamini & Yekutieli (2001)

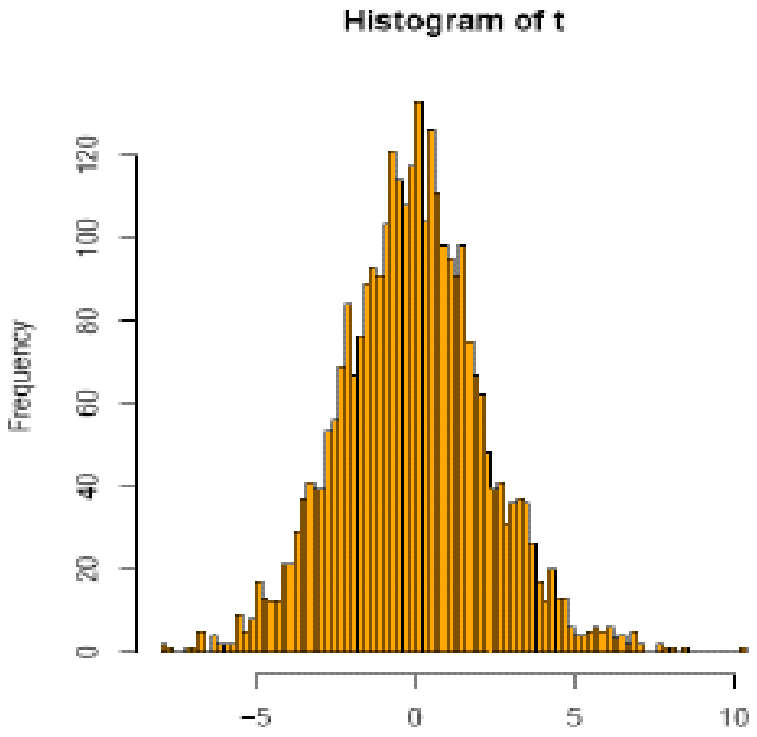
Difference between controlling FWER or FDR

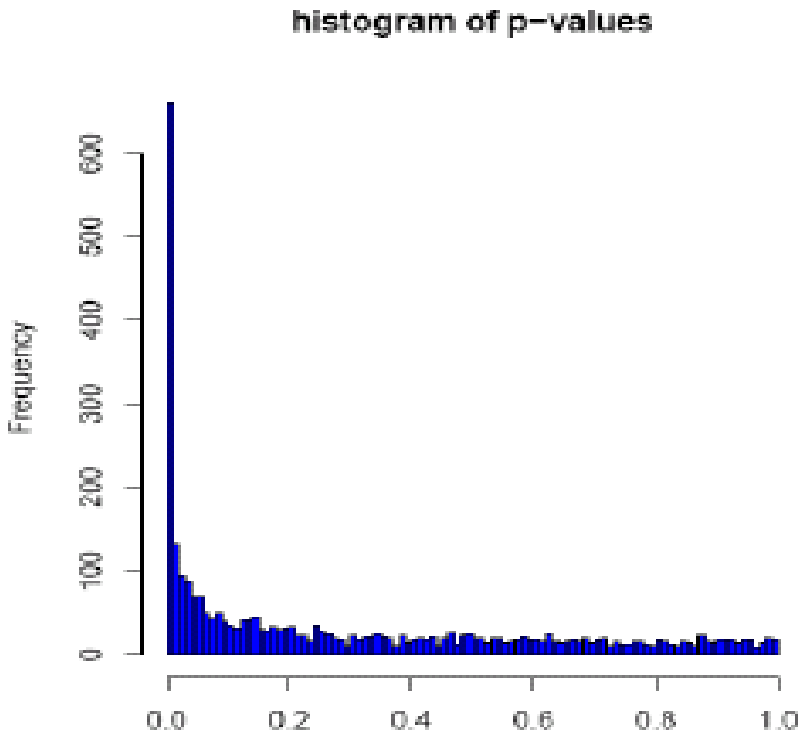
- FWER→ *Controls for no (0) false positives*
  - gives many fewer genes (false positives),
  - but you are likely to miss many
  - adequate if goal is to identify few genes that differ between two groups
- FDR→ *Controls the proportion of false positives*
  - if you can tolerate more false positives
  - you will get many fewer false negatives
  - adequate if goal is to pursue the study e.g. to determine functional relationships among genes

Steps to generate a list of candidate genes revisited (2)



Example (1b)





Golub data, 27 ALL vs 11 AML samples, 3051 genes

Bonferroni adjustment: 98 genes with $p_{adj} < 0.05$ ($p_{raw} < 0.000016$)

Example (2)

- Se the examples of testing in the case study found in this link

  http://www.ub.edu/stat/docencia/bioinformatica/microarrays/ADM/labs/Virtaneva2002/Ejemplo_AML8.R

Extensions

- Some issues we have not dealt with
  - Replicates within and between slides
  - Several effects: use a linear model
  - ANOVA: are the effects equal?
  - Time series: selecting genes for trends
- Different solutions have been suggested for each problem
- Still many open questions