Microarray Data Analysis

Statistical methods to detect differentially expressed genes

Outline

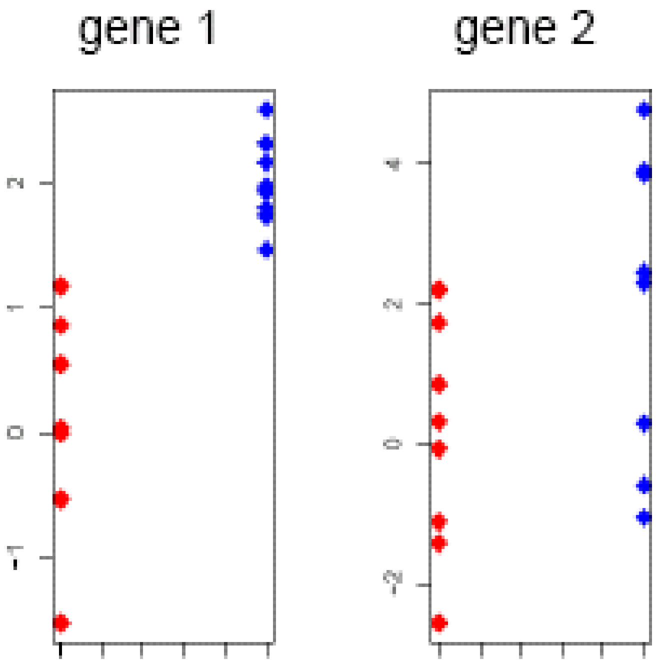
- The *class comparison* problem
- Statistical tests
  - Calculation of p-values
  - The volcano plot
- Multiple testing
- Extensions
- Examples

Class comparison

- Identifying *differentially expressed* genes means ->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 from 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 (1)

- *From now on we will use R as a measure of the (log) ratio, instead of M.*
- For **Direct comparisons** *in two colour* or *paired-one colour*.

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

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

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

“Natural” measures of discrepancy (2)

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

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

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

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 instability**
  - **Non-normality of the data**
- Other related to big number of variables
  - **Multiple testing**

Variance instability

- 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

Courtesy of Y.H. Yang

Variance unstability (1): outliers

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

Courtesy of Y.H. Yang

Variance unstability (2): tiny variances

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

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)

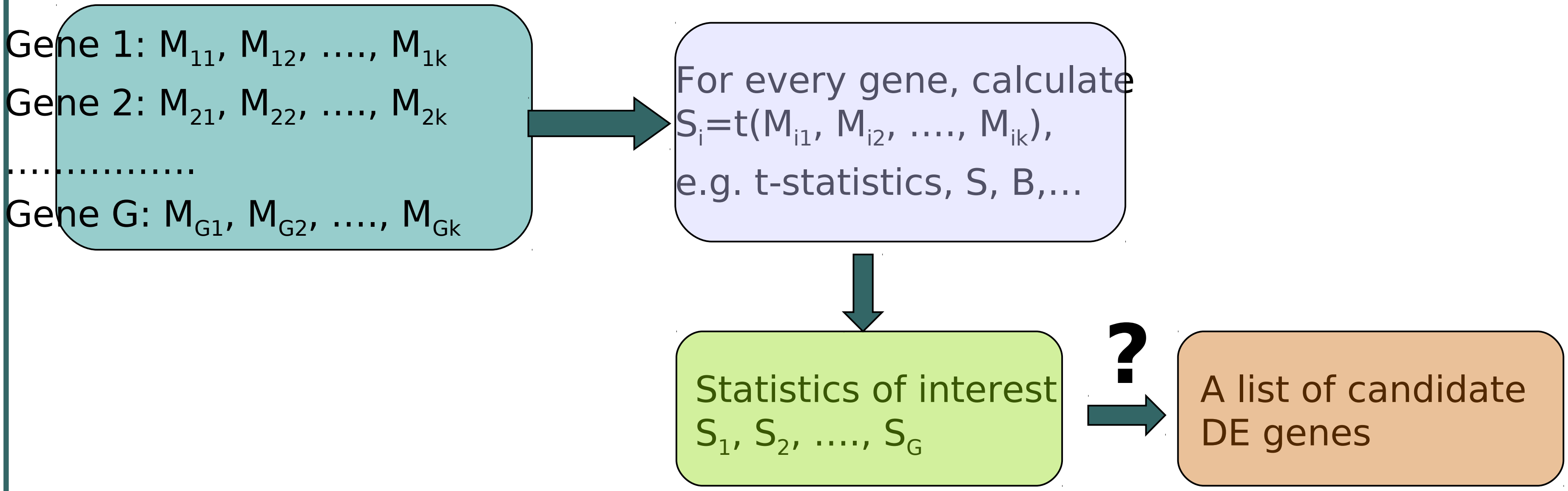
$$t = \frac{R_g}{\sqrt{\frac{v_0 SE^2 + (n-1)SE_g^2}{v_0 + n - 2}}}$$

EB-moderated t
(Smyth, 2003)

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

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:



Gene 1: M_{11}, M_{12}, …., M_{1k}

Gene 2: M_{21}, M_{22}, …., M_{2k}

……………..

Gene G: M_{G1}, M_{G2}, …., M_{Gk}

For every gene, calculate S_i=t(M_{i1}, M_{i2}, …., M_{ik}), e.g. t-statistics, S, B,…

Statistics of interest S_1, S_2, …., S_G

?

A list of candidate DE genes

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

Hypothesis testing overview for a single gene

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

Calculation of p-values

- Standard methods for calculating p-values:

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

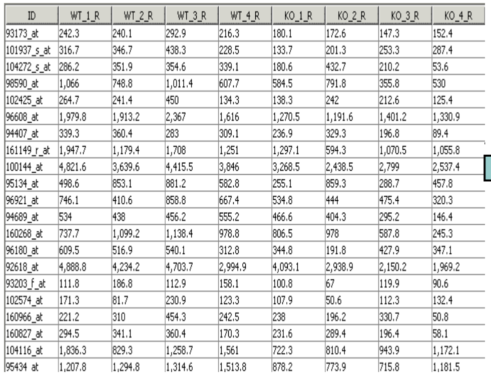
(ii) Perform a permutation analysis

(i) Tabulated p-values

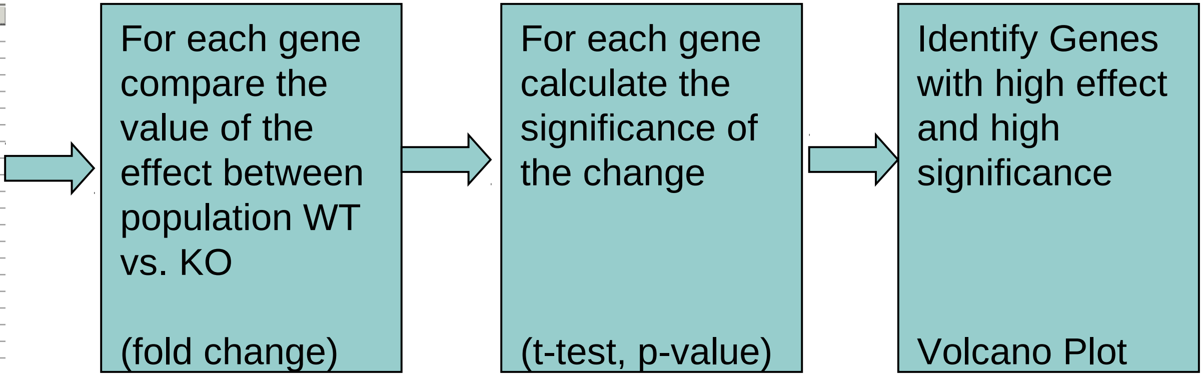
- Tabulated *p*-values can be obtained for standard test statistics (e.g.the *t*-test)
- They often rely on the assumption of normally distributed errors in the data
- This assumption can be checked (approximately) using a
  - Histogram
  - Q-Q plot

Volcano Plots

Volcano plots are a graphical means for visualising results of large numbers of t-tests allowing us to plot both the Effect and significance of each test in an easy to interpret way



ID	WT_1_R	WT_2_R	WT_3_R	WT_4_R	KO_1_R	KO_2_R	KO_3_R	KO_4_R
93173_at	242.3	240.1	292.9	216.3	180.1	172.6	147.3	152.4
101937_s_at	316.7	346.7	438.3	228.5	133.7	201.3	253.3	287.4
104272_s_at	286.2	351.9	354.6	339.1	180.6	432.7	210.2	53.6
98590_at	1,066	748.8	1,011.4	607.7	584.5	791.8	355.8	530
102425_at	264.7	241.4	450	134.3	138.3	242	212.6	125.4
96608_at	1,979.8	1,913.2	2,367	1,616	1,270.5	1,191.6	1,401.2	1,330.9
94407_at	339.3	360.4	283	309.1	236.9	329.3	196.8	89.4
161149_r_at	1,947.7	1,179.4	1,708	1,251	1,297.1	594.3	1,070.5	1,055.8
100144_at	4,821.6	3,639.6	4,415.5	3,846	3,268.5	2,438.5	2,799	2,537.4
95134_at	498.6	853.1	881.2	582.8	255.1	859.3	288.7	457.8
96921_at	746.1	410.6	858.8	667.4	534.8	444	475.4	320.3
94689_at	534	438	456.2	555.2	466.6	404.3	295.2	146.4
160268_at	737.7	1,099.2	1,138.4	978.8	806.5	978	587.8	245.3
96180_at	609.5	516.9	540.1	312.8	344.8	191.8	427.9	347.1
92618_at	4,888.8	4,234.2	4,703.7	2,994.9	4,093.1	2,938.9	2,150.2	1,969.2
93203_f_at	111.8	186.8	112.9	158.1	100.8	67	119.9	90.6
102574_at	171.3	81.7	230.9	123.3	107.9	50.6	112.3	132.4
160966_at	221.2	310	454.3	242.5	238	196.2	330.7	50.8
160827_at	294.5	341.1	360.4	170.3	231.6	289.4	196.4	58.1
104116_at	1,836.3	829.3	1,258.7	1,561	722.3	810.4	943.9	1,172.1
95434 at	1,207.8	1,294.8	1,314.6	1,513.8	878.2	773.9	715.8	1,181.5



For each gene compare the value of the effect between population WT vs. KO

(fold change)

For each gene calculate the significance of the change

(t-test, p-value)

Identify Genes with high effect and high significance

Volcano Plot

Volcano plots

- In a volcano plot:
- X-axis represents effect measured as fold change:

$$\text{Effect} = \log_2(\overline{WT}) - \log_2(\overline{KO}) \qquad = \log_2(\overline{WT} / \overline{KO})$$

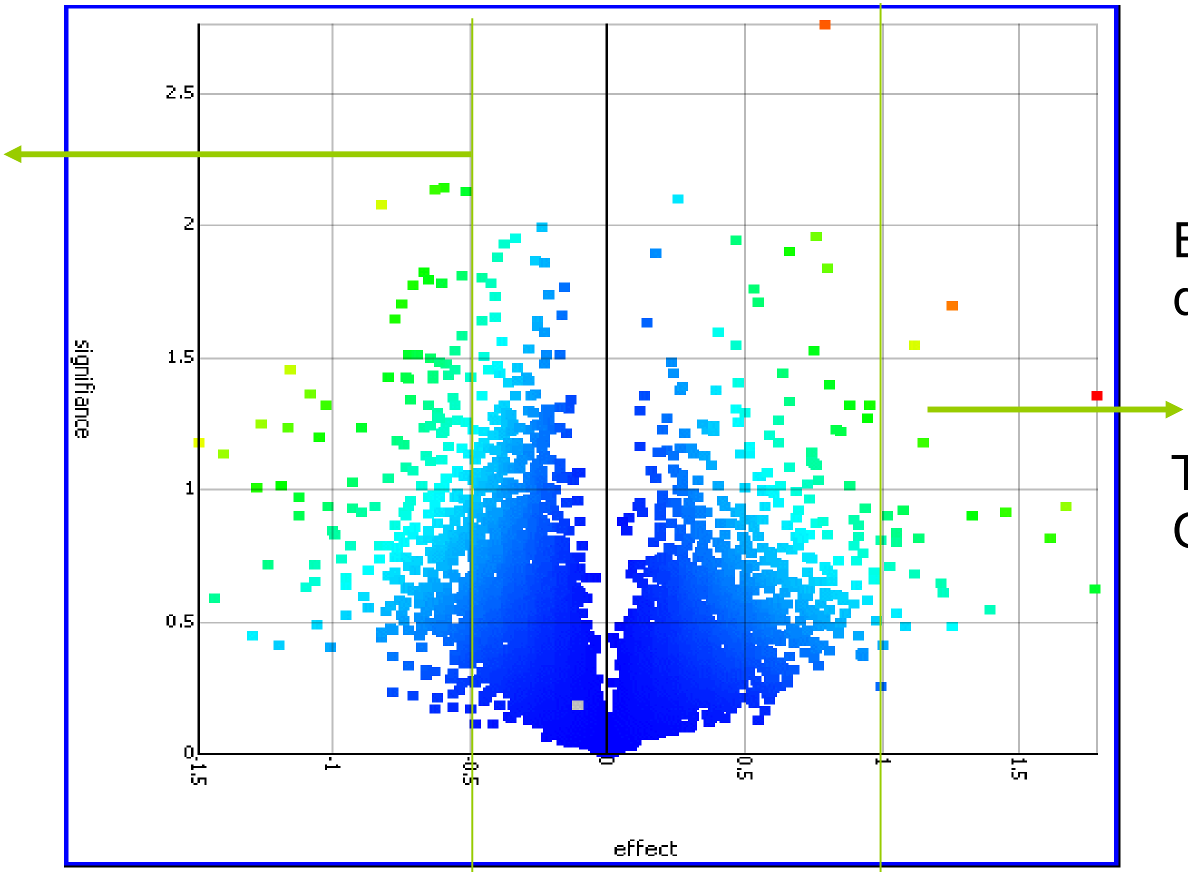
If WT = KO,   Effect Fold Change = 0 ,

If WT = 2 KO, Effect Fold Change = 1

...

Numerical Interpretation (Effect)

Effect has halved=$2^{0.5}$



Effect has doubled =2^1

Two Fold Change

Using $\log_2$ for X axis:

Volcano plots

- In a volcano plot:
- y-axis represents the number of zeroes in the p-value
  - (remember with a p-value of 0.0001, you are more confident than with a p-value of 0.01
  - This is just a trick so that higher values on the graph are more important

Calculate Significance as : $-\log_{10}$ (p_value)

If p = 0.1, -log(0.1) = 1 (1 decimal point)

If p = 0.01, -log (0.01) = 2 (2 decimal points)

...

Numerical Interpretation (Significance)

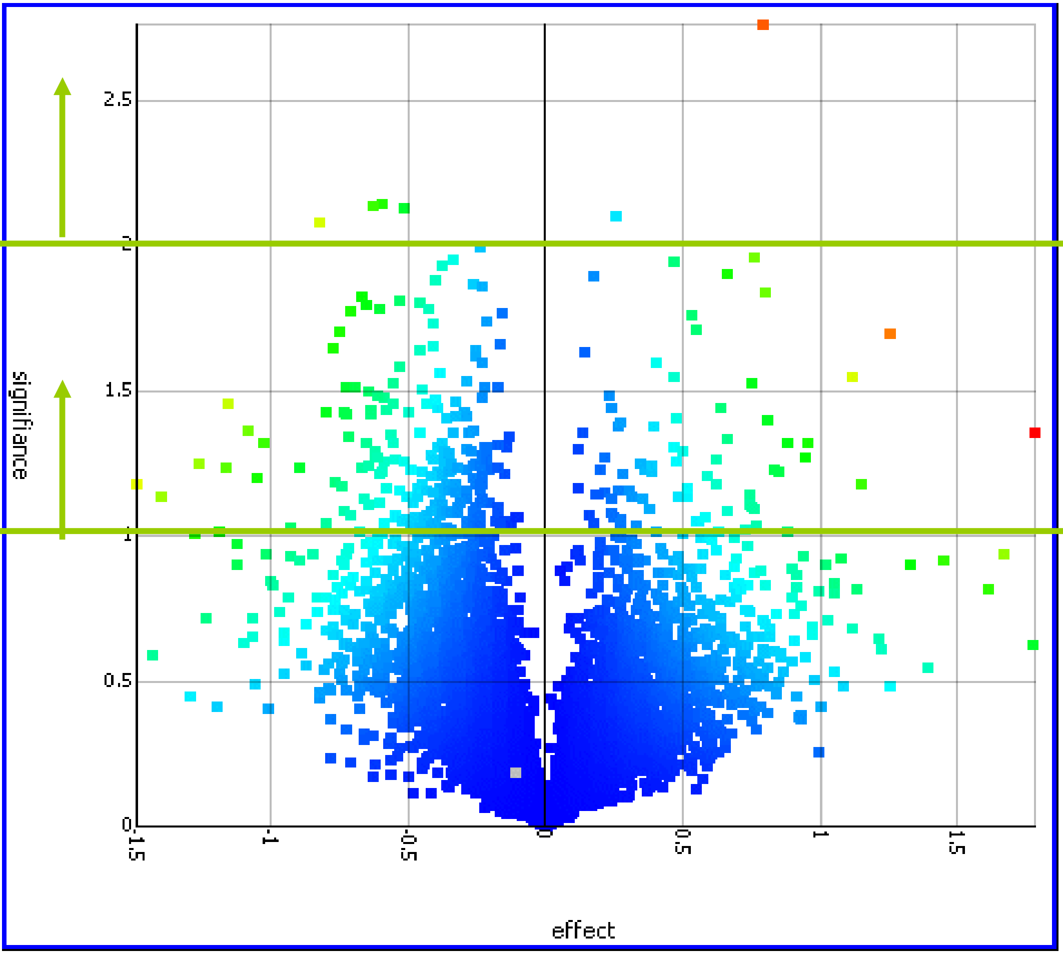
p< 0.01

(2 decimal places)

p< 0.1

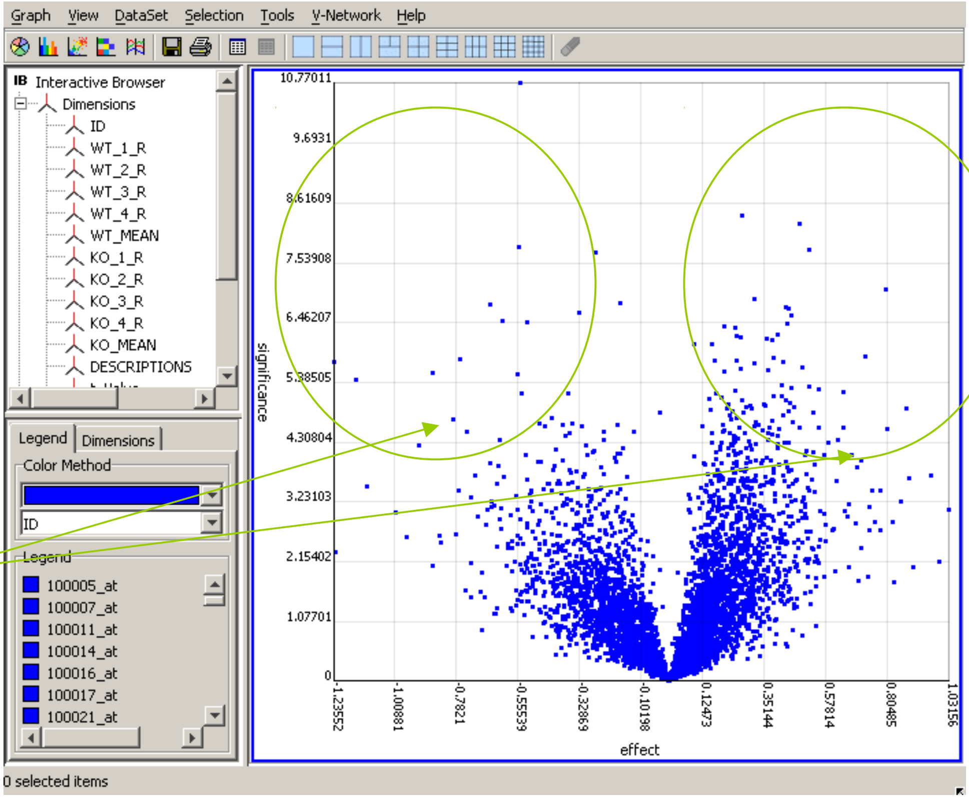
(1 decimal place)

Using $\log_{10}$ for Y axis:



Visualise the Result :Volcano Plot

- Effect vs. Significance
- Selections of items that have both a large effect and are highly significant can be identified easily.



High Effect & Significance

High Significance

Low Significance

-ve effect

+ve effect

Choosing log scales is a matter of convenience

Effect can be both +ve or -ve

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

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

Multiple testing: Counting errors

		Decision reported				
		H_0 is Rejected *(Genes Selected)*		H_0 is accepted *(Genes not Selected)*		Total
State of the nature ("Truth")	H_0 is false *(Affected)*	$m_\alpha - \alpha m_0$	(S)	$(m-m_o)-(m_\alpha - m_{0)}$	(T)	$m-m_o$
	H_0 is true *(Not Affected)*	αm_0	(V)	$m_o - \alpha m_0$	(U)	m_o
Total		m_α	(R)	$m-m_\alpha$	(m-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

Extension 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]

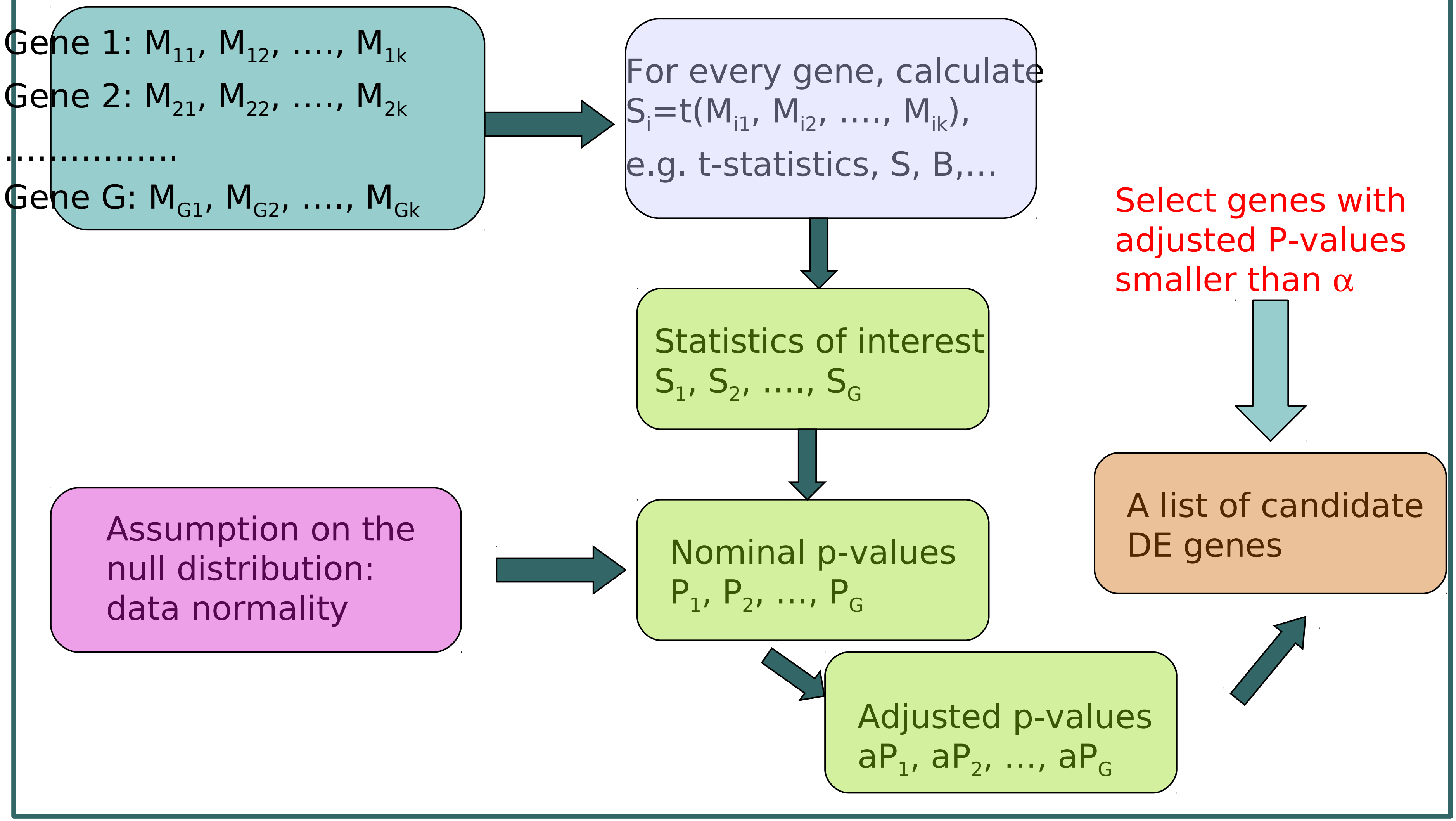
FDR and FWER controlling procedures

- 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)



Gene 1: $M_{11}, M_{12}, \ldots., M_{1k}$

Gene 2: $M_{21}, M_{22}, \ldots., M_{2k}$

…………….

Gene G: $M_{G1}, M_{G2}, \ldots., M_{Gk}$

For every gene, calculate $S_i=t(M_{i1}, M_{i2}, \ldots., M_{ik})$, e.g. t-statistics, S, B,…

Statistics of interest $S_1, S_2, \ldots., S_G$

Assumption on the null distribution: data normality

Nominal p-values $P_1, P_2, \ldots, P_G$

Adjusted p-values $aP_1, aP_2, \ldots, aP_G$

Select genes with adjusted P-values smaller than α

A list of candidate DE genes