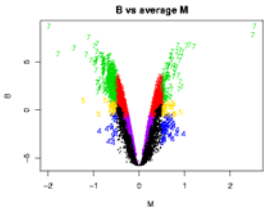
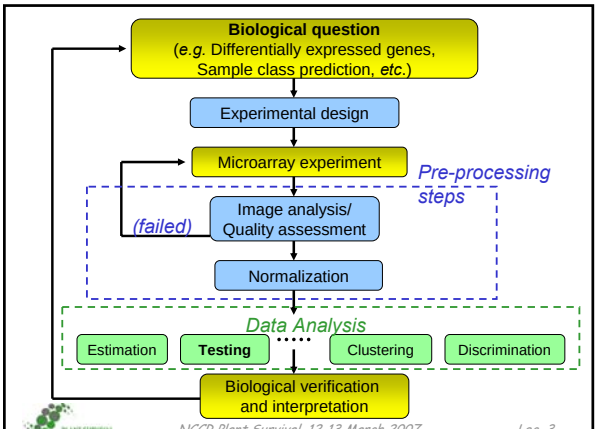
## *Statistics for cDNA microarrays*

*Multiple hypothesis testing, identifying differentially expressed genes*




http://www.isrec.isb-sib.ch/~darlene/NCCR-PS/








## Hypothesis testing

- 2 'competing theories' regarding a population parameter:
  - *NULL* hypothesis *H* ('straw man')
  - ALTERNATIVE hypothesis *A* ('claim', or theory you wish to test)
- *H:* NO DIFFERENCE
  - any observed deviation from what we expect to see is due to *chance variability*
- *A:* THE DIFFERENCE IS *REAL*



## Hypotheses

- Examples of NULL hypothesis:
  - The coin is fair (no preference for heads or tails)
  - This new drug is no better (or worse) than a placebo
  - There is no difference in weight between two given strains of mice
- Examples of Alternative hypothesis:
  - The coin is biased (either towards tail or head)
  - The drug is better than a placebo



## Test statistic

- Measure how far the observed data are from what is expected *assuming the NULL H* by computing the value of a *test statistic* (TS) from the data
- The particular TS computed depends on the parameter
- For example, to test the population mean, the TS is the *sample mean* (or standardized sample mean)



## Errors in hypothesis testing

| Truth \ Decision | not rejected | rejected |
|---|---|---|
| true H | ☺ specificity (True -) | X Type I error (False +) α |
| false H | X Type II error (False -) β | ☺ Power 1 - β; sensitivity (True +) |

## p-value

- Decide on whether or not to *reject* the NULL hypothesis *H* based on the chance of obtaining a TS *as or more extreme* (as far away from what we expected or even farther, in the direction of the ALT) than the one we got, *ASSUMING THE NULL IS TRUE*
- This chance is called the *observed significance level*, or *p-value*
- A TS with a p-value less than some pre-specified false positive *level* (or *size*) $\alpha$ is said to be 'statistically significant' at that level



## More on p-values

- Calculate p based on distribution of the TS
- Historically, compare observed to values in a table for pre-defined p-values (e.g. .05, .01)
- Tables specific for the TS you are using
- Computers can now calculate exact p-values, which are reported as output



## p-value interpretation

- The interpretation of a p-value is a little tricky
- The p-value represents the chance that we would see a difference as big as we saw (or bigger) *if* there were really nothing happening other than chance variability
- 'a single convenient number giving a measure of the degree of surprise which the experiment should cause a believer of the null hypothesis' (Hodges and Lehmann)



## A p-value is *NOT*

- Confusion about p-values:
  - It is ***not*** the chance that the NULL hypothesis is true
  - It is ***not*** the chance of making an error
  - It is ***not*** the chance that a followup experiment would give a different result



## Steps in hypothesis testing

1. Identify the population parameter being tested
2. Formulate the NULL and ALT hypotheses
3. Compute the TS
4. Compute the p-value
5. (Optional) *Decision Rule:* REJECT H if the p-value $\leq \alpha$
   (This is a type of argument by contradiction)



## One sample t-test

- Is the mean of a population equal to a given value ?
- *Example* :
  - Given a gene and several replicate microarray measurements (log ratios) $g_1, g_2, \ldots, g_n$. Is the gene differentially expressed?
- Hypotheses:
  - NULL: mean equals 0
  - ALT: could be for example
    - mean different from 0
    - mean larger than 0

## TS for the one sample t-test

- Test-statistic (Student's t-statistic):

$$T = \frac{\bar{x} - \mu_0}{\sqrt{S^2/n}}$$

- Here
  - $\bar{x}$ is the average of the observations
  - S is the (estimated) standard deviation
  - n is the number of observations
  - $\mu_0$ is the NULL value (= 0 for our example)



## t-distribution

- If the NULL is true, the value T follows a *known distribution* (*t*-distribution)
- The shape of the *t*-distribution depends on the number of observations
- If the average is made of n observations, we use the *t*-distribution with *n-1 degrees of freedom* ($t_{n-1}$)



## Example: Trees

- A study is conducted to investigate the growth of a certain type of tree at an elevation of 675 meters
- The variable of interest is the core measurement (in cm) for a 10 year period
- The theory is that the mean should be at least 1.75
- In a random sample of 10 measurements, the mean was 2 with an SD of .5



## Steps in hypothesis testing (I)

1. Identify the population parameter being tested

   *Here, the parameter being tested is the population mean core measurement* $\mu$

2. Formulate the NULL and ALT hypotheses

   *H:* $\mu = 1.75$ *(or* $\mu \leq 1.75$*)*

   *A:* $\mu > 1.75$

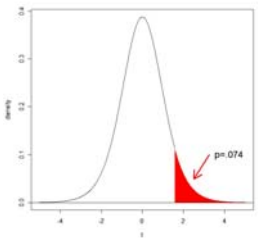
3. Compute the TS

   $t = (2 - 1.75)/(.5/\sqrt{10}) = 1.58$



## One sample t-test: example

- Since we have 10 observations, compare observed t to *t*-distribution with 9 df






## Steps in hypothesis testing (II)

4. Compute the p-value

   *Here,* $p = P(T_9 > 1.58) = .074$

5. (Optional) *Decision Rule:* REJECT H if the p-value $\leq \alpha$

   *If we use* $\alpha = .05$*, the decision here will be DO NOT REJECT H (but just barely!)*

## Comments

- The *t*-test assumes that the different observations are *independent* and that they follow a *normal distribution*
- The basic *t*-test is not popular for microarrays, because the estimation of S is unstable (small sample size)



## Pitfalls in hypotheses testing

- Even if a result is 'statistically significant', *it can still be due to chance*
- Conversely, if a result is *not* statistically significant, it may be only because you do not have enough data
- *Statistical significance* is not the same as *practical importance*
- A test of significance does not say how *important* the difference is, or *what caused it*
- A test does not check the *study design*
- *Data-snooping* makes p-values hard to interpret: the test should be fully defined **BEFORE** looking at the data



## Multiple testing problem

- Simultaneously test *G* null hypotheses, one for each gene *j*

  $H_j$: no association between expression measure of gene j and the response
- Because microarray experiments simultaneously monitor expression levels of thousands of genes, there is a *large multiplicity issue*
- Increased chance of *false positives*
- Would like some sense of how 'surprising' the observed results are



## Coin tossing

- Observation of 11 consecutive "tails"
- *Very improbable* with a fair coin ($p < 0.0004$)
- Does it mean that the coin is *biased*?
- p-values are valid if *only one test is done*
  - 11 consecutive tails in *11 tosses* would be significant
  - 11 consecutive tails in *1000 tosses* is not
- If *several tests are conducted*, the significance of each of them is *reduced*
  - If you try more often, you are more likely to succeed, even just by chance
  - If an event has probability 1/1000 every time you try, and you try 1000 times, it is likely to happen at least once



## Hypothesis truth vs. decision

| Truth \ Decision | # not rejected | # rejected | totals |
|---|---|---|---|
| # true H | U | V (F +) Type I error | $m_0$ |
| # non-true H | T Type II error | S | $m_1$ |
| totals | m - R | R | m |



## Type I (false positive) error rates

- *Per-family Error Rate*

  PFER = E(V)
- *Per-comparison Error Rate*

  PCER = E(V)/m
- *Family-wise Error Rate*

  FWER = $p(V \geq 1)$
- *False Discovery Rate*

  FDR = E(Q), where

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

## Strong vs. weak control

- All probabilities are *conditional* on which hypotheses are true
- *Strong control* refers to control of the Type I error rate under *any combination* of true and false nulls
- *Weak control* refers to control of the Type I error rate only under the *complete null hypothesis* (i.e. *all* nulls true)
- In general, weak control without other safeguards is unsatisfactory



## Adjusted p-values (p*)

- Statistical procedures can adapt the results in the case of multiple testing
  - Most well-known and conservative: Bonferroni
  - Divide significance threshold by number of repetitions
  - Example:
    - 1000 tests with threshold 0.01 ➔ corrected threshold = 0.01/1000 = 0.00001
  - Other procedures are less stringent
- Particularly relevant for microarrays:
  - Showing that 1 *preselected* gene is differentially expressed (p=0.01) may be interesting
  - Showing that 1 gene *out of 10,000* is differentially expressed (p=0.01) is probably *not* interesting



## Some multiple testing procedures

- What follows is given for completeness only
- You don't need to follow all of the details ☺
- There are *several possibilities* providing strong control of the FWER (not just Bonferroni!)
- In the microarray context, it is generally more useful to consider control of the *FDR*



## A Little Notation

- For hypothesis $H_j$, j = 1, …, m
  observed test statistic: $t_j$
  observed unadjusted p-value: $p_j$
- Ordering of observed (absolute) $t_j$: $\{r_j\}$
  such that $|t_{r_1}| \geq |t_{r_2}| \geq \ldots \geq |t_{r_G}|$
- Ordering of observed $p_j$: $\{r_j\}$
  such that $|p_{r_1}| \leq |p_{r_2}| \leq \ldots \leq |p_{r_G}|$
- Denote corresponding RVs by upper case letters (T, P)



## Control of the FWER

- *Bonferroni single-step* adjusted p-values
  $p_j^* = \min(mp_j, 1)$
- *Sidak single-step (SS)* adjusted p-values
  $p_j^* = 1 - (1 - p_j)^m$
- *Sidak free step-down (SD)* adjusted p-values
  $p_j^* = 1 - (1 - p_{(j)})^{(m - j + 1)}$



## Control of the FWER

- *Holm (1979) step-down* adjusted p-values
  $p_{r_j}^* = \max_{k = 1\ldots j} \{\min((m-k+1)p_{r_k}, 1)\}$
  - *Intuitive explanation*: once $H_{(1)}$ rejected by Bonferroni, there are only m-1 remaining hyps that might still be true (then another Bonferroni, *etc.*)
- *Hochberg (1988) step-up* adjusted p-values (Simes inequality)
  $p_{r_j}^* = \min_{k = j\ldots m} \{\min((m-k+1)p_{r_k}, 1)\}$

## Control of the FDR

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

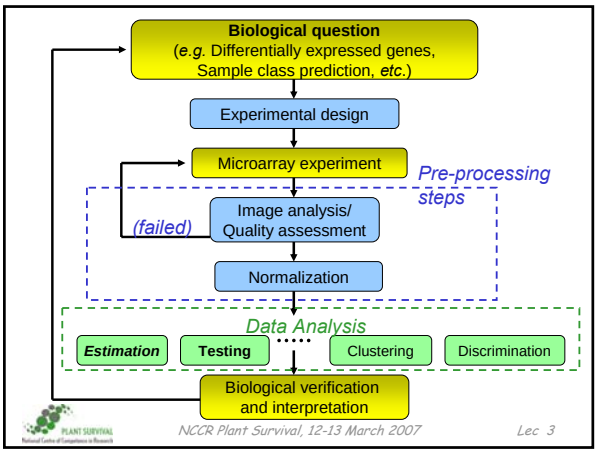
$$p_{r_j}^* = \min_{k = j \ldots m} \{ \min ([m/k]\, p_{r_k}, 1) \}$$

- *Benjamini & Yuketieli (2001): conservative step-up* procedure which controls the FDR under general dependency structures

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

- *Yuketieli & Benjamini (1999):* resampling based adjusted p-values for controlling the FDR under certain types of dependency structures








## cDNA gene expression data

Data on *G* genes for *n* samples:

mRNA samples

| Genes | | sample1 | sample2 | sample3 | sample4 | sample5 | ... |
|---|---|---|---|---|---|---|---|
| | 1 | 0.46 | 0.30 | 0.80 | 1.51 | 0.90 | ... |
| | 2 | -0.10 | 0.49 | 0.24 | 0.06 | 0.46 | ... |
| | 3 | 0.15 | 0.74 | 0.04 | 0.10 | 0.20 | ... |
| | 4 | -0.45 | -1.03 | -0.79 | -0.56 | -0.32 | ... |
| | 5 | -0.06 | 1.06 | 1.35 | 1.09 | -1.09 | ... |

*Gene expression level* of gene *i* in mRNA sample *j*

= (normalized) $\text{Log}_2$( Red intensity / Green intensity)



## Identifying Differentially Expressed Genes

- Goal: Identify genes associated with covariate or response of interest
- Examples:
  - Qualitative covariates or factors: treatment, cell type, tumor class
  - Quantitative covariate: dose, time
  - Responses: survival, cholesterol level
  - Any combination of these!



## Informal methods

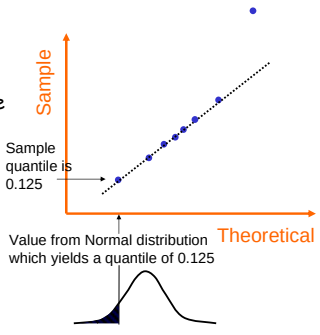
- If no replication (*i.e.* only have a single array), there are not many options
- Common methods include:
  - (log) Fold change exceeding some threshold, *e.g.* more than 2 (or less than -2)
  - Graphical assessment, *e.g.* QQ plot
- However, the threshold is pretty arbitrary



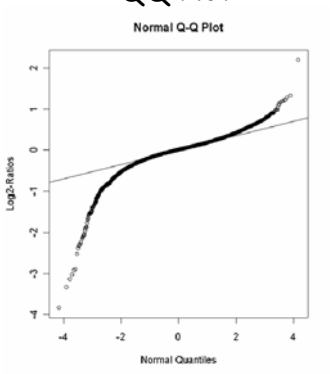
## QQ-Plots

Used to assess whether a sample follows a particular (e.g. normal) distribution (or to compare two samples)

A method for looking for outliers when data are mostly normal



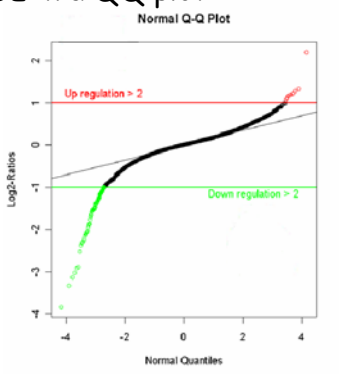
## QQ Plot






## DE in a QQ plot

In this case, the ratios are from a *self-self hyb* – *i.e.* NO genes are truly DE!






## Single-slide methods

- Model-dependent rules for deciding whether (R,G) corresponds to a differentially expressed gene
- Amounts to drawing two curves in the (R,G)-plane; call a gene differentially expressed if it falls outside the region between the two curves
- Difficult to justify these strong modeling assumptions
- n = 1 slide may not be enough (!)



## Replicated experiments

- Have *n* replicates
- For each gene, have *n* values of M = $\log_2$ fold change, one from each array
- *Summarize* $M_1, \ldots, M_n$ for each gene by
  - M = average $(M_1, \ldots, M_n)$
  - s = SD$(M_1, \ldots, M_n)$
- *Rank* genes in order of strength of evidence in favor of DE
- How might we do this?



## Which genes are DE?

- Difficult to judge significance
  - massive multiple testing problem
  - genes dependent
  - don't know null distribution of M
- Strategy
  - aim to *rank* genes
  - assume most genes are not DE (depending on type of experiment and array)
  - find genes *separated* from the majority



## Ranking criteria

- Genes $i = 1, \ldots, p$
- $M_i$ = average $\log_2$ fold change for gene i
  - *Problem*: genes with large variability likely to be selected, even if not DE
- Fix that by taking variability into account: use $t_i = M_i / (s_i/\sqrt{n})$
  - *Problem*: genes with extremely small variances make very large *t*
  - When the number of replicates is small, the smallest $s_i$ are likely to be underestimates

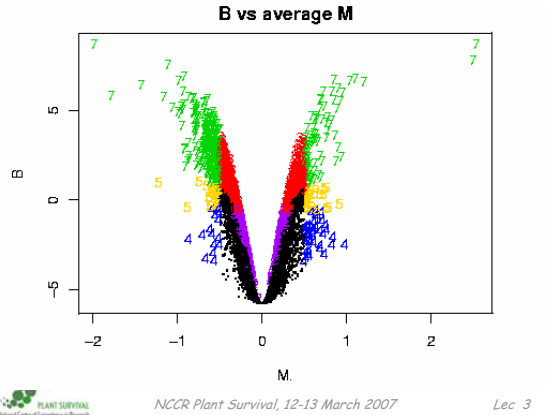
## Shrinkage estimators

- *Idea:* borrow information across genes
- Here, we 'shrink' the $t_i$ towards zero by modifying the $s_i$ in some way (get $s_i^*$)
- mod $t_i = t_i^* = M_i/(s_i^*/\sqrt{n})$

$t_i$ ⟷ $t_i^*$ ⟷ $M_i$

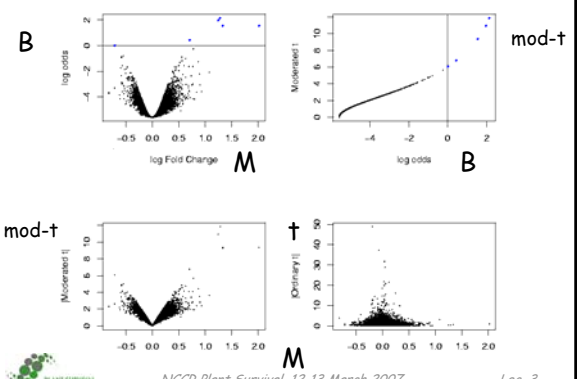
- Many ways to get a value for $s_i^*$
- We will use the version implemented in the BioConductor package **limma**
- Similar to *B-statistic* [log(P(DE)/P(not DE)]





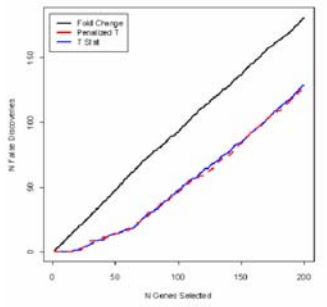



## M, B, mod t, t



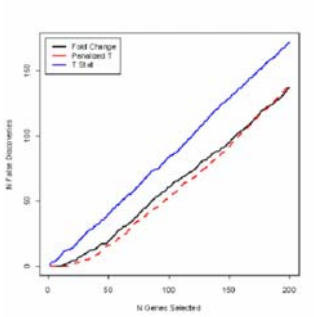



## Simulation study: genes with different SDs



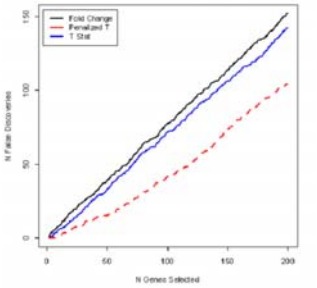



## Simulation study: genes with similar SDs






## Simulation study: genes with different SDs, small number of arrays



## Significance of results

- Assessing significance is difficult, due to complicated (and unknown) dependence structure between genes and unknown distribution for log ratios
- B statistic does not yield absolute cutoff values, because $p$ is not estimated ($p$ is necessary for the calibration)
- Possible to compute approximate adjusted $p$-values by resampling methods
- *Conclusion*: use mod $t$ (or B) statistic for ranking genes, don't believe associated $p$-value



(*BREAK*)