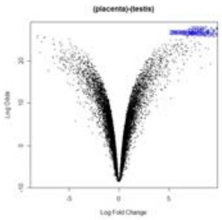
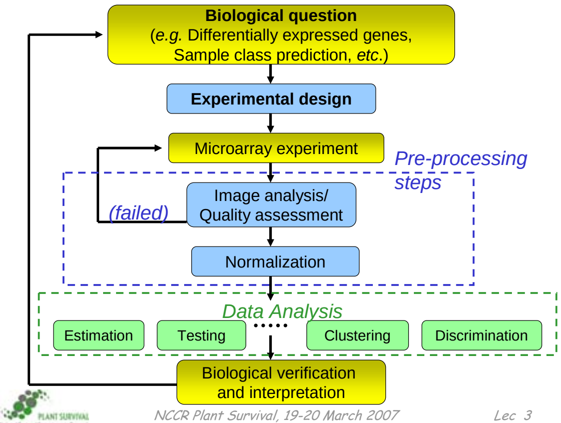
Statistics for Affymetrix GeneChips

Experimental design; Comparing two groups



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



Experimental Design – why do we care?

- Poor design *costs* :
  - time, money, ethical considerations
- To ensure *relevant* data are collected, and can be analyzed to test the scientific hypothesis/ question of interest
  - Decide *in advance* how data will be analyzed
  - 'Designing the experiment' = 'Planning the analysis'
- *The design is about the biology* – but requires knowledge of *statistics*

Three main ED principles: *Replication, Randomization, Blocking*

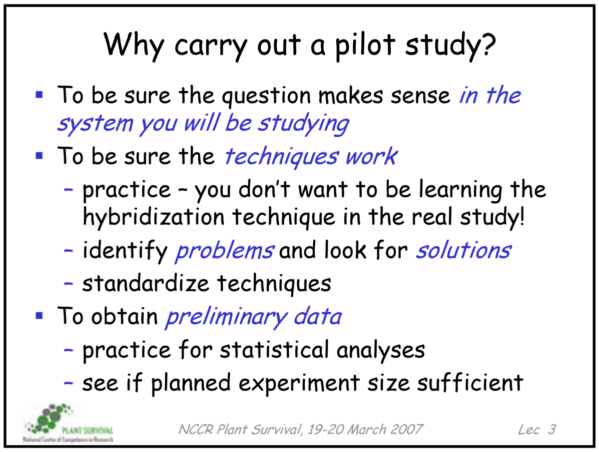
- *Replication* – to reduce random variation of the test statistic, increases generalizability
- *Randomization* – to remove bias
- *Blocking* – to reduce unwanted variation
  - Idea here is that units *within* a block are similar to each other, but different *between* blocks
- 'Block what you can, randomize what you cannot'

Three (biological) decisions

- What *measurements* to make (*response*)
  - In a microarray experiment, we measure *gene expression* (fluorescence intensity)
- What *conditions* to study (*treatments*)
- What experimental *material* to use (*units*)

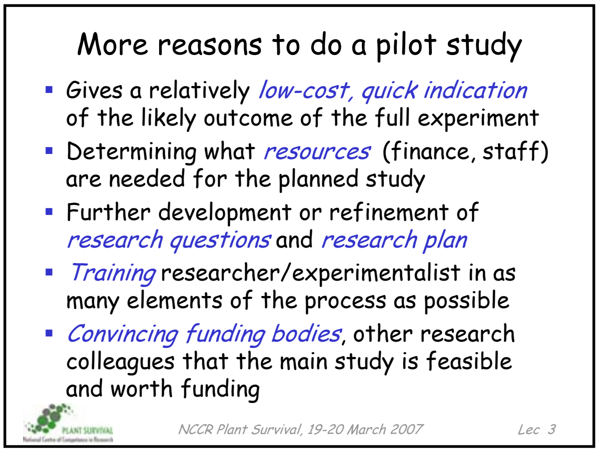
What is a pilot study?

- A pilot study is a *small scale version* of a full, larger experiment
- *'Mini-experiment'*
- Normally, a pilot is carried out as part of a larger experiment (or research program)
- Usually, the *pilot sample size is much smaller* than for the full experiment
- Carried out *before* the full experiment



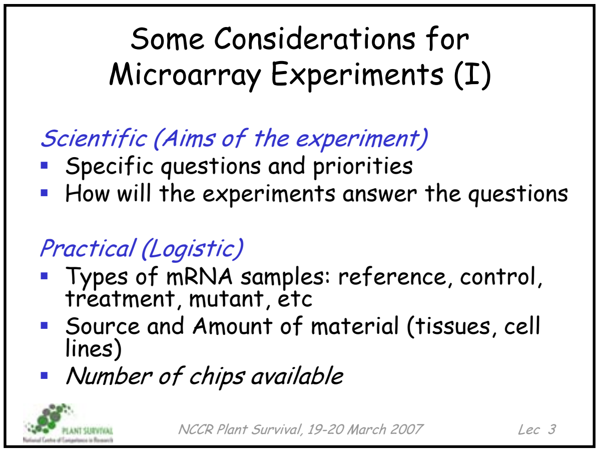
Why carry out a pilot study?

- To be sure the question makes sense *in the system you will be studying*
- To be sure the *techniques work*
  - practice – you don't want to be learning the hybridization technique in the real study!
  - identify *problems* and look for *solutions*
  - standardize techniques
- To obtain *preliminary data*
  - practice for statistical analyses
  - see if planned experiment size sufficient



More reasons to do a pilot study

- Gives a relatively *low-cost, quick indication* of the likely outcome of the full experiment
- Determining what *resources* (finance, staff) are needed for the planned study
- Further development or refinement of *research questions* and *research plan*
- *Training* researcher/experimentalist in as many elements of the process as possible
- *Convincing funding bodies*, other research colleagues that the main study is feasible and worth funding



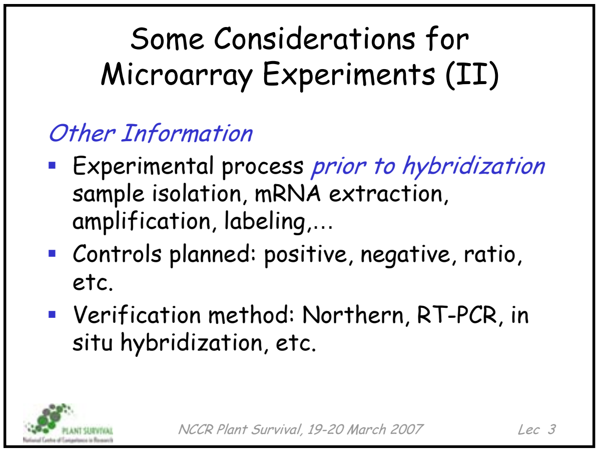
Some Considerations for Microarray Experiments (I)

Scientific (Aims of the experiment)

- Specific questions and priorities
- How will the experiments answer the questions

Practical (Logistic)

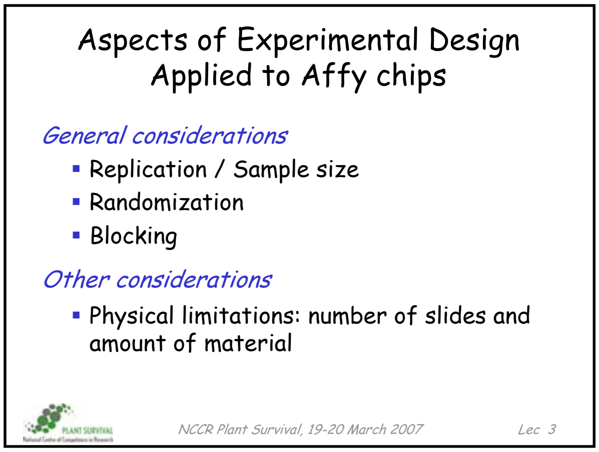
- Types of mRNA samples: reference, control, treatment, mutant, etc
- Source and Amount of material (tissues, cell lines)
- *Number of chips available*



Some Considerations for Microarray Experiments (II)

Other Information

- Experimental process *prior to hybridization* sample isolation, mRNA extraction, amplification, labeling,…
- Controls planned: positive, negative, ratio, etc.
- Verification method: Northern, RT-PCR, in situ hybridization, etc.



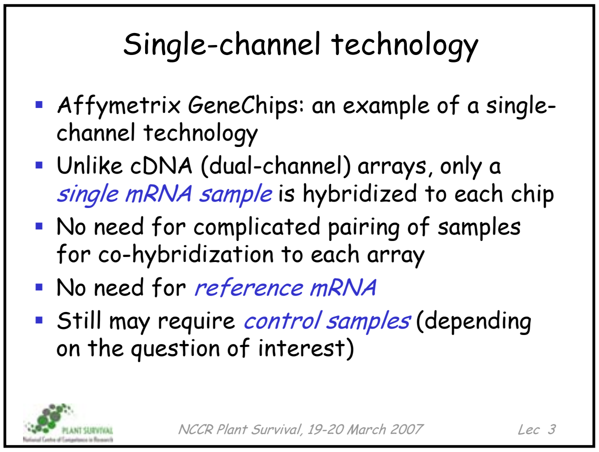
Aspects of Experimental Design Applied to Affy chips

General considerations

- Replication / Sample size
- Randomization
- Blocking

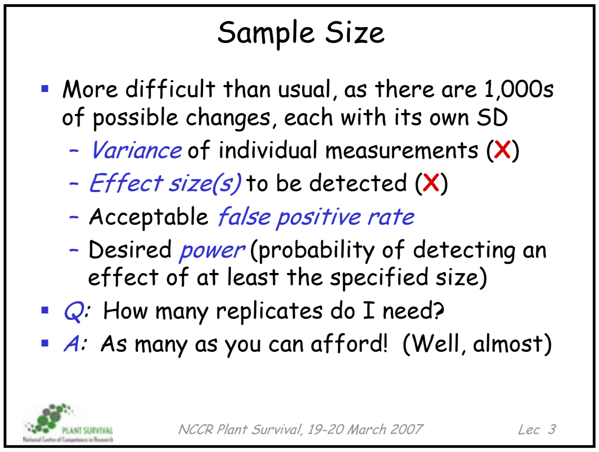
Other considerations

- Physical limitations: number of slides and amount of material



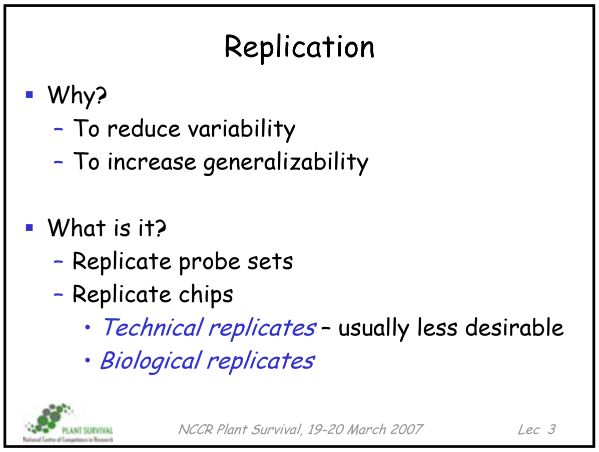
Single-channel technology

- Affymetrix GeneChips: an example of a single-channel technology
- Unlike cDNA (dual-channel) arrays, only a *single mRNA sample* is hybridized to each chip
- No need for complicated pairing of samples for co-hybridization to each array
- No need for *reference mRNA*
- Still may require *control samples* (depending on the question of interest)



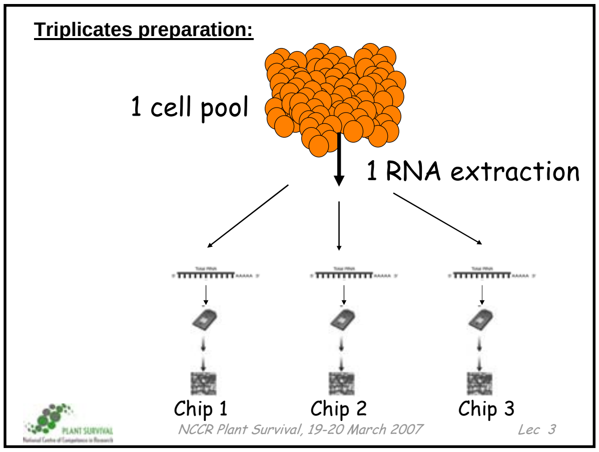
Sample Size

- More difficult than usual, as there are 1,000s of possible changes, each with its own SD
  - *Variance* of individual measurements (X)
  - *Effect size(s)* to be detected (X)
  - Acceptable *false positive rate*
  - Desired *power* (probability of detecting an effect of at least the specified size)
- *Q:* How many replicates do I need?
- *A:* As many as you can afford! (Well, almost)



Replication

- Why?
  - To reduce variability
  - To increase generalizability
- What is it?
  - Replicate probe sets
  - Replicate chips
    - *Technical replicates* – usually less desirable
    - *Biological replicates*

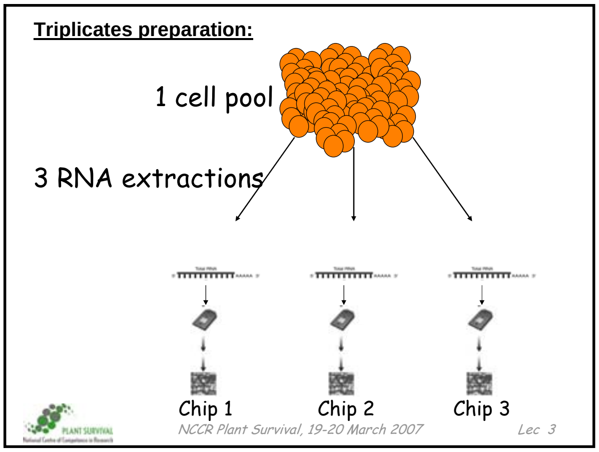


Triplicates preparation:

1 cell pool

1 RNA extraction

Chip 1 Chip 2 Chip 3

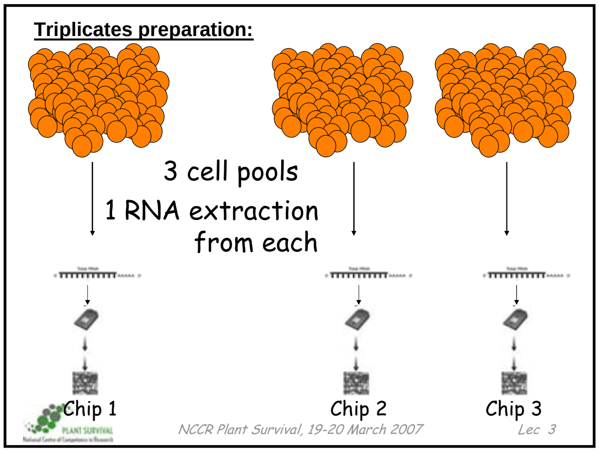


Triplicates preparation:

1 cell pool

3 RNA extractions

Chip 1 Chip 2 Chip 3

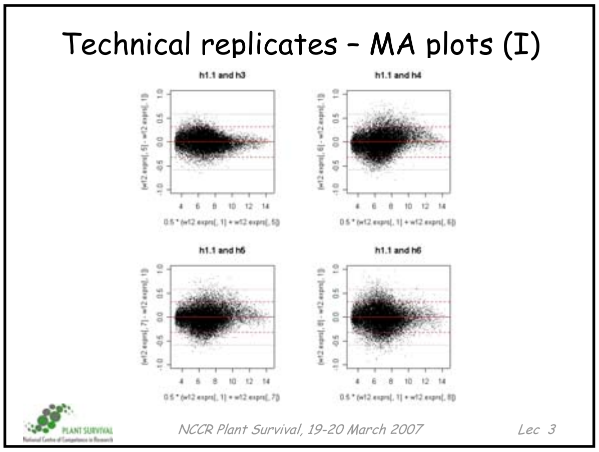


Triplicates preparation:

3 cell pools

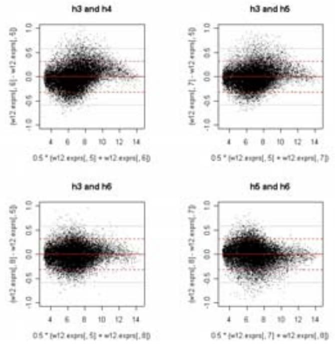
1 RNA extraction from each

Chip 1 Chip 2 Chip 3



Technical replicates – MA plots (I)

Technical replicates – MA plots (II)



Pooling samples

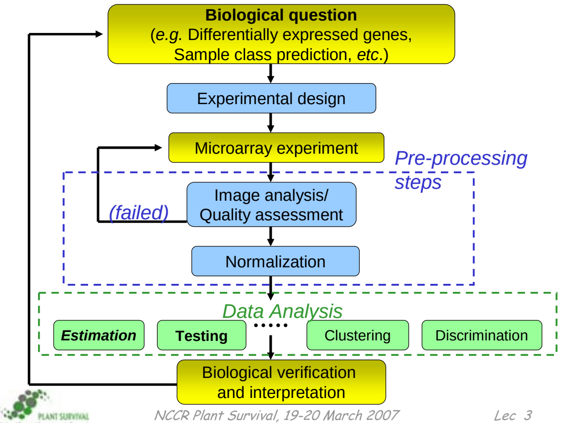
- To economize on the number of arrays, *pooling of samples* has been suggested (when samples are inexpensive)
- Sometimes pooling is *necessary* (*e.g.* when an insufficient amount of material is obtained from single individuals, such as ants)
- Pooling *should not be done* when individual-specific information is of interest, or when the goal is to identify unknown sub-groups

Results of pooling experiment

- Inference on differential expression for most genes not adversely affected by pooling
- For larger designs, pooling may be *beneficial when many subjects* are pooled, provided that independent samples contribute to *multiple pools*
- Pooling *only a few samples not advised* – the gain is small compared to the loss of individual specific information

Summary

- Balance of *direct* and *indirect* comparisons
- Optimize precision of the estimates among comparisons of interest
- Must satisfy *scientific and physical constraints* of the experiment
- It can save you a lot of *time*, *money* and *heart-ache* to consult with an experienced analyst on design issues ***before any steps of the experiment have been carried out***



Biological question
(*e.g.* Differentially expressed genes, Sample class prediction, *etc.*)

Experimental design

Microarray experiment

Pre-processing steps

(failed)

Image analysis/ Quality assessment

Normalization

Data Analysis

Estimation

Testing

Clustering

Discrimination

Biological verification and interpretation

Affymetrix gene expression data

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

mRNA samples

		sample1	sample2	sample3	sample4	sample5	...
	1	10.24	10.29	10.28	10.32	10.19	...
	2	6.83	6.62	6.61	6.83	6.67	...
Genes	3	7.97	8.25	8.41	7.90	7.92	...
	4	9.05	8.78	8.79	8.93	8.99	...
	5	5.49	5.18	5.24	5.28	5.27	...

Gene expression level (RMA value) of gene *i* in mRNA sample *j*

RMA = estimated chip effect for quantile normalized $\log_2$(PM – BG)

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!

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?

Ranking criteria

- Genes *i* = 1, ..., *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

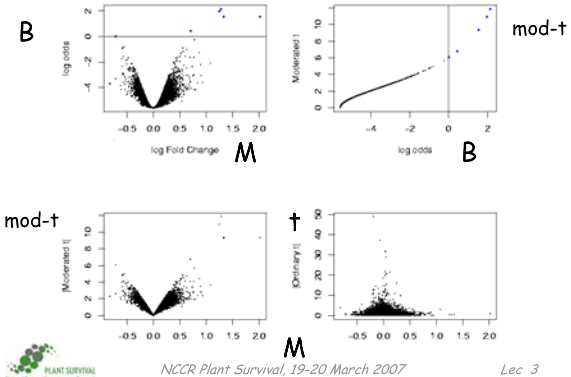
Shrinkage estimators

- *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 $t_i = t_i^* = M_i/(s_i^*/\sqrt{n})$

$$t_i \longleftrightarrow t_i^* \longleftrightarrow M_i$$

- 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



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

Some common experiments

- Comparison of *2 conditions*/types ('treatment vs. control')
  - mutant vs. wild type plants
  - liver vs. heart in mouse
- Comparison of *many treatments* to a control
- *Clinical studies* (*e.g.* cancer patients)
- *Time course* – measurements at different times
- *Factorial study* – multiple conditions varied and studied *simultaneously*

Experiments to compare 2 groups

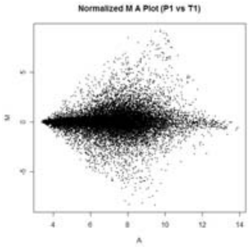
- Examples:
  - mutant – wild type
  - `treated' – control
- Generally want to compare the *same tissues in the same organism*
  - data more reliable
- *Minimum* number of chips: 3-5 *per group*

How many chips?

- The answer to this question depends on (among other things) the *degree of differential expression* you wish to detect
- *Minimum* number of chips: 3-5 *per group*
- These should be *biological replicates*
- More chips needed to distinguish *small* differences

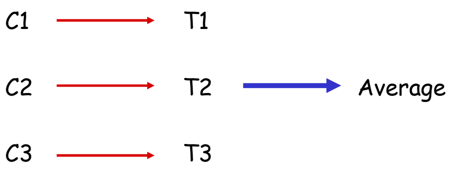
How to analyze?

- If you had only 1 chip from each group, you could look at the log fold change between the two conditions



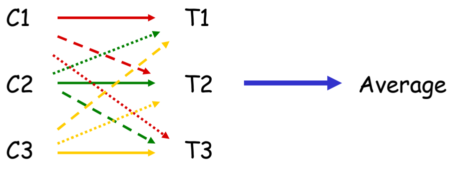
Combining replicate chips – how **NOT** to (I)

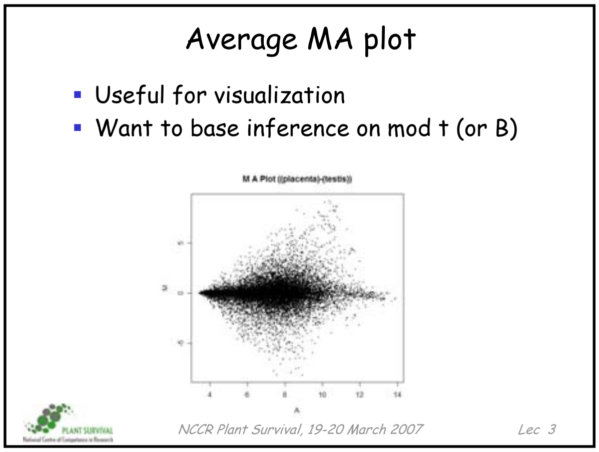
- *BUT:* You followed statistical advice and made replicate chips
- Something you **SHOULD NOT DO:**



Combining replicate chips – how **NOT** to (II)

- Something else you **SHOULD NOT DO:**

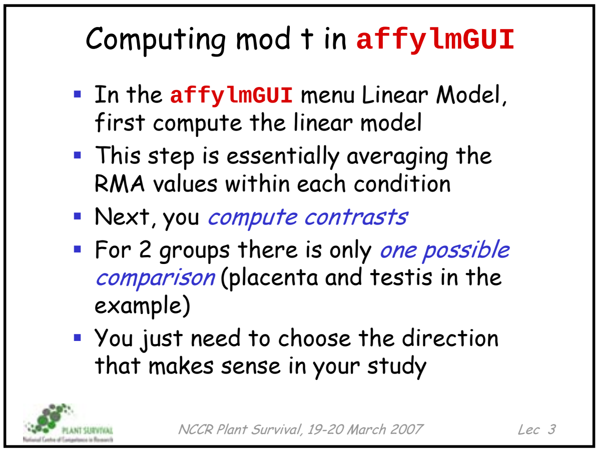




Average MA plot

- Useful for visualization
- Want to base inference on mod t (or B)

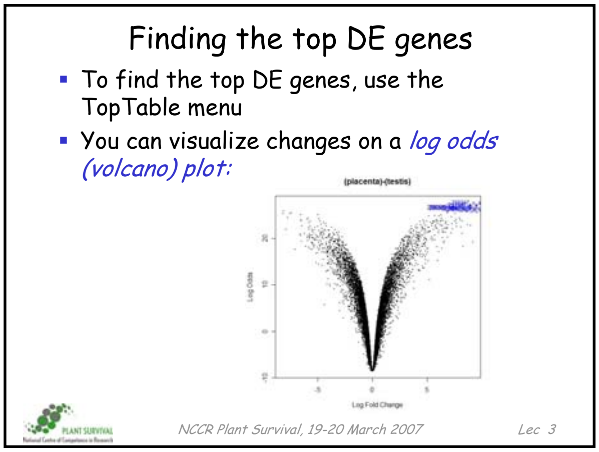
NCCR Plant Survival, 19-20 March 2007 Lec 3



Computing mod t in **affylmGUI**

- In the **affylmGUI** menu Linear Model, first compute the linear model
- This step is essentially averaging the RMA values within each condition
- Next, you *compute contrasts*
- For 2 groups there is only *one possible comparison* (placenta and testis in the example)
- You just need to choose the direction that makes sense in your study

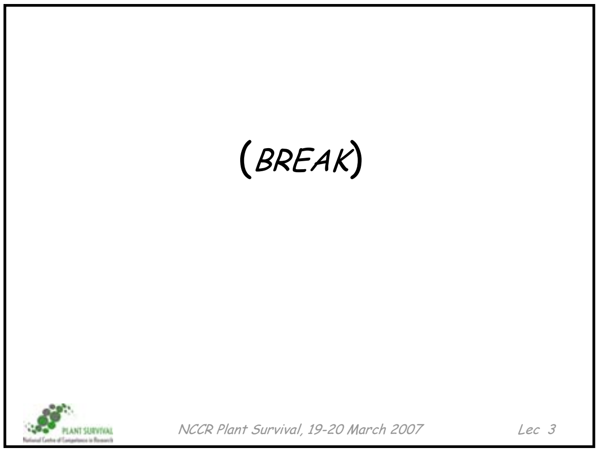
NCCR Plant Survival, 19-20 March 2007 Lec 3



Finding the top DE genes

- To find the top DE genes, use the TopTable menu
- You can visualize changes on a *log odds (volcano) plot:*

NCCR Plant Survival, 19-20 March 2007 Lec 3



(*BREAK*)

NCCR Plant Survival, 19-20 March 2007 Lec 3