The essentials of microarray data analysis

(from a complete novice)

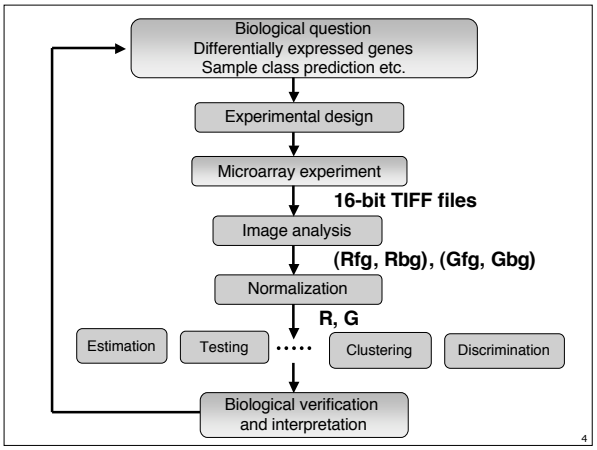
Thanks to Rafael Irizarry for the slides!

Outline

- Experimental design
- Take logs!
- Pre-processing: affy chips and 2-color arrays
- Clustering
- Identifying differentially expressed genes

Resources

- Expressionists working group
  (Search google for "hopkins expressionists")
  "Top ten things to know…" held monthly
- "Stats for gene expression" course (140.688)
  http://www.biostat.jhsph.edu/~ririzarr/Teaching/688



Experimental design

- Choice of platform
- Array design
  - Creation of probes
  - Location on the array
  - Controls
- Target samples

5

Experimental design

Proper experimental design is needed to ensure that questions of interest ***can*** be answered and that this can be done **accurately and precisely**, given experimental constraints, such as cost of reagents and availability of mRNA.

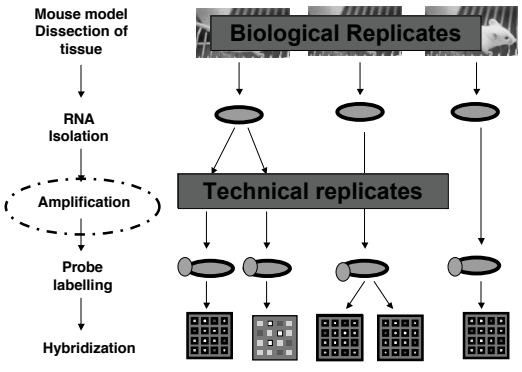
6

Avoidance of bias

- Conditions of an experiment (mRNA extraction and processing, the reagents, the operators, the scanners and so on) can leave a "global signature" in the resulting expression data.
  - Balance
  - Randomization

7

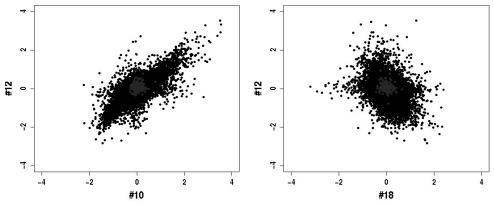
Preparing mRNA samples:



8

Technical replication - amplification

- 3 sets of two different samples performed on different days
- #10 and #12 were from the same RNA isolation and amplification
- #12 and #18 were from different dissections and amplifications
- All 3 data sets were labeled separately before hybridization



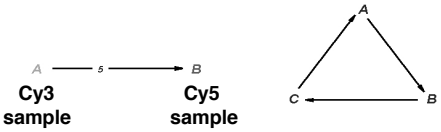
Data provided by
Dave Lin (Cornell)

9

Two-color layouts

For two color platforms it is assumed that the size of the spot/probe effect is too big to trust the absolute intensites. Thus we always use relative measurements

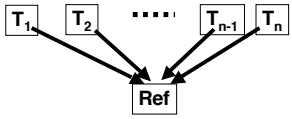
***Vertices*: mRNA samples; *Edges:* hybridization;**
***Direction:* dye assignment.**



(a) *(b)*

10

Common reference design



Experiment for which the common reference design is appropriate

Meaningful biological control (C) Identify genes that responded differently / similarly across two or more treatments relative to control.

Large scale comparison. To discover tumor subtypes when you have many different tumor samples.

Advantages:

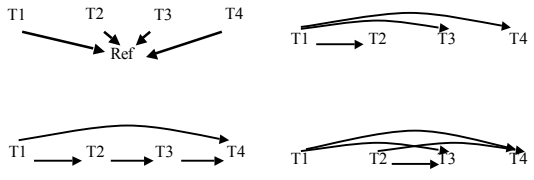
Ease of interpretation.

Extensibility - extend current study or to compare the results from current study to other array projects.

11

Experiment for which a number of designs are suitable for use

Time Series



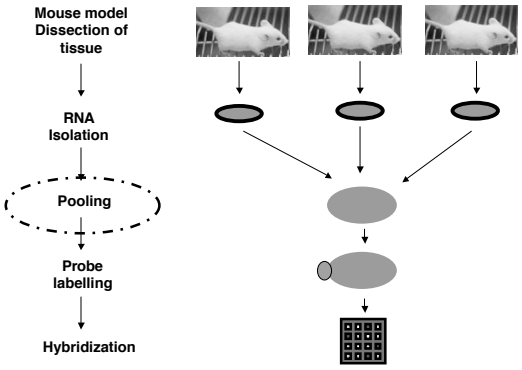
12

Pooling

- **Should I pool mRNA samples across subjects in an effort to reduce the effect of biological variability?**
- **Notice, in many cases, samples are cheap but arrays are expensive**

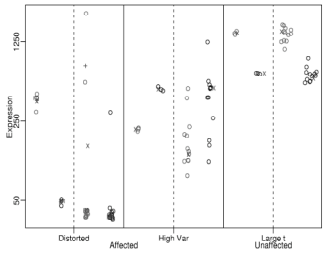
13

Pooling: looking at very small amount of tissues



14

Problem with pooling everything



- You cannot measure variability
- You cannot take logs before "averaging"

15

Alternative pooling strategy

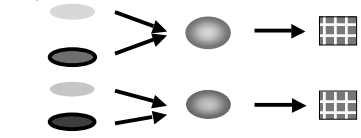
- **Instead of pooling everything, how about pooling groups?**
- **For example, will I obtain the same results with 12 individuals on 12 chips as with 12 individuals on 4 chips ?**

16

Design 1

Pooled vs Individual samples

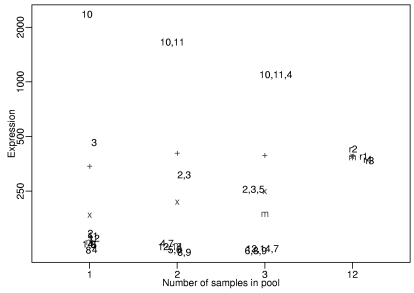
Design 2



Taken from Kendziorski etl al (2003)

17

More on Pooling and outliers



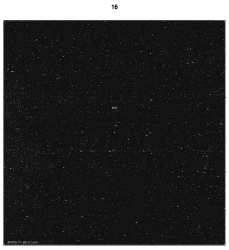
18

Outline

- Experimental design
- Take logs!
- Pre-processing: affy chips and 2-color arrays
- Clustering
- Identifying differentially expressed genes

19

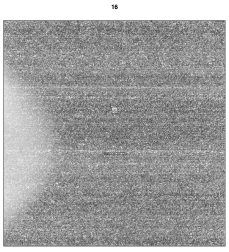
Images of probe level data



This is the raw data

20

Images of probe level data



Log scale version much more informative

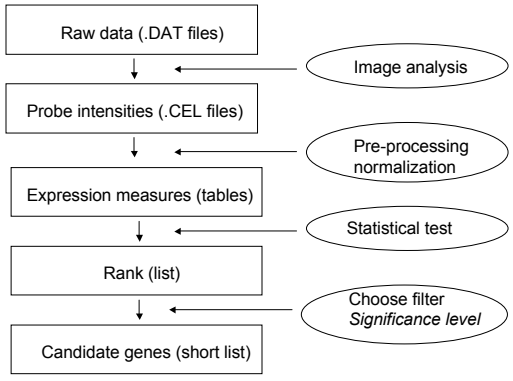
21

Outline

- Experimental design
- Take logs!
- Pre-processing: affy chips and 2-color arrays
- Clustering
- Identifying differentially expressed genes

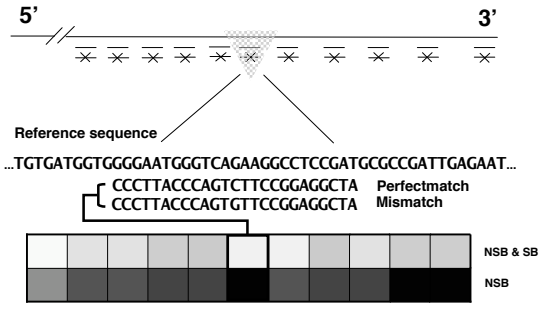
22

Work flow



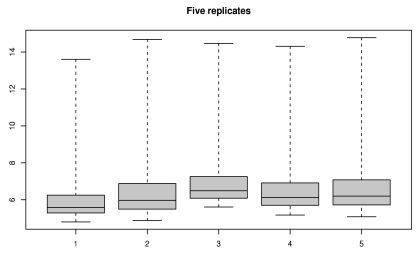
23

Affymetrix GeneChip Design



24

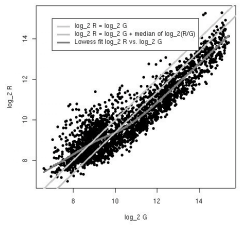
Affy Technical Replicates Boxplot



Different scanners were used.

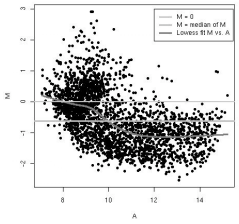
Self-self hybridization

$\log_2$ R vs. $\log_2$ G



Self-self hybridization

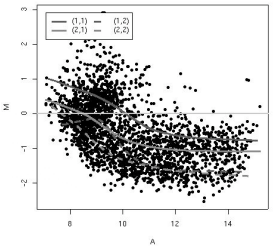
M vs. A



$M = \log_2 R - \log_2 G$, $A = (\log_2 R + \log_2 G)/2$

Self-self hybridization

M vs. A

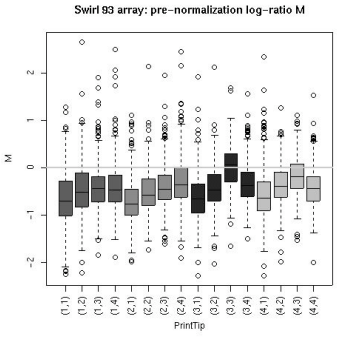


Robust local regression
within sectors
(print-tip-groups)
of intensity log-ratio M
on average log-intensity A.

$M = \log_2 R - \log_2 G$, $A = (\log_2 R + \log_2 G)/2$

Boxplots by print-tip-group

Intensity log-ratio, M



29

What can we do?

- Throw away the data and start again? Maybe.
- Statistics offers hope:
  - Use control genes to adjust
  - Assume most genes are not differentially expressed
  - Assume distribution of expression are the same

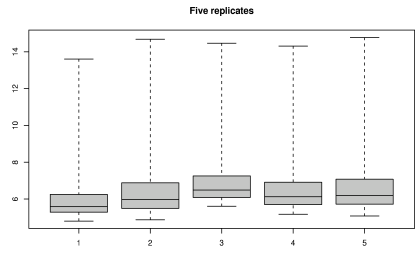
30

Simplest Idea

- Assume all arrays have the same median log expression or relative log expression
- Subtract median from each array
- In cDNA this makes the median log ratio 0 (we assume there are as many over-expressed as under-expressed)
- In Affy we usually add a constant that takes us back to the original range

31

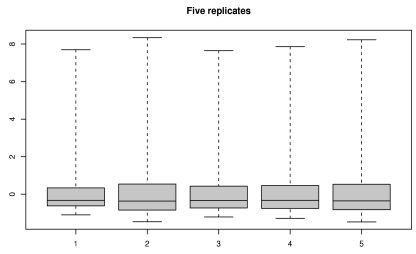
How does it work



Notice subtracting in the original scale won't work well

32

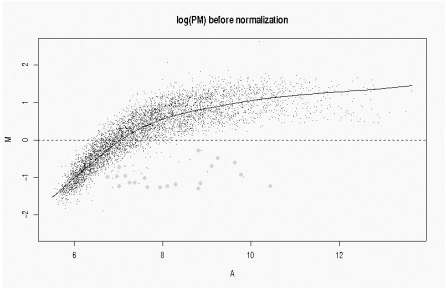
How does it work



The medians match but there are some discrepancies

33

What are the consequences



These are two technical replicates with spike-ins.

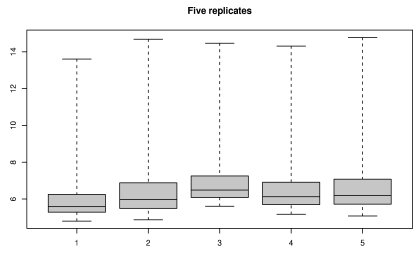
34

Quantile normalization

- All these non-linear methods perform similarly
- Quantiles is my favorite because its fast
- Basic idea:
  - order value in each array
  - take average across probes
  - Substitute probe intensity with average
  - Put in original order

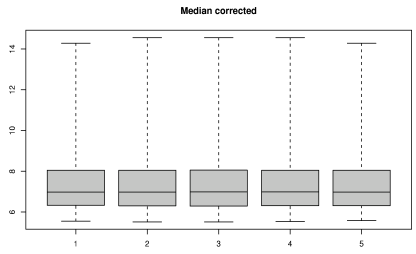
35

How does it work



36

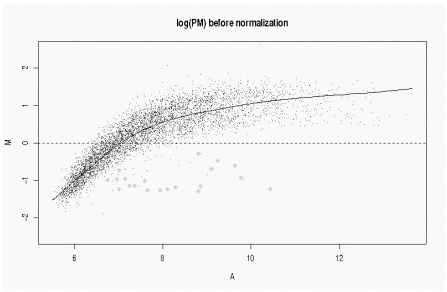
How does it work



Does it wash away real differential expression?

37

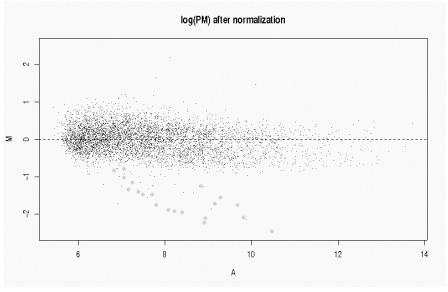
Before and



These are two technical replicates. Only the red are differentially expressed

38

After quantile normalization



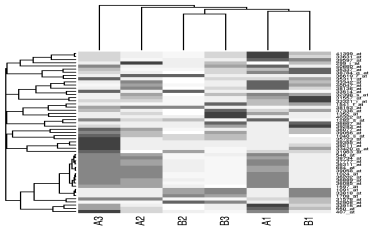
39

Outline

- Experimental design
- Take logs!
- Pre-processing: affy chips and 2-color arrays
- Clustering
- Identifying differentially expressed genes

40

Clustering is not a universally good tool



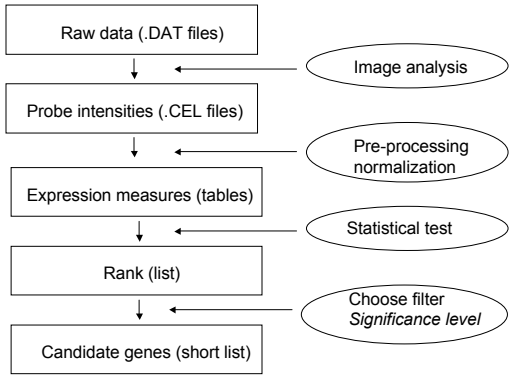
41

Outline

- Experimental design
- Take logs!
- Pre-processing: affy chips and 2-color arrays
- Clustering
- Identifying differentially expressed genes

42

Work flow



43

Differential gene expression

- Identify genes whose expression levels are associate with a response or covariate of interest
  - clinical outcome (e.g. survival, response to treatment, tumor class)
  - covariate such as treatment, dose, time.
- Estimation: In a statistical framework, assigning a score can be viewed as estimating an effects of interest (e.g. difference in means, slope, interaction). We can also take the variability of these estimates into account.
- Testing: In a statistical framework, deciding on a cut-off can be viewed as an assessment of the statistical significance of the observed associations.

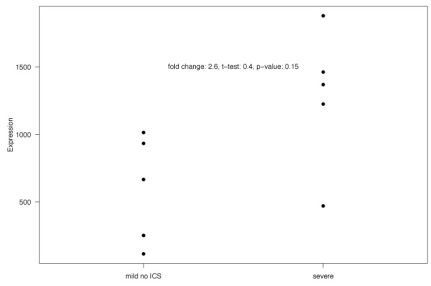
44

Example: Two populations

- A common problem is to find genes that are differentially expressed in two populations.
- Many method papers appear in both statistical and molecular biology literature.
- The proposed scores range from:
  - ad-hoc summaries of fold-change
  - variantes on the t-test
  - and posterior means obtained from Bayesian or empircal Bayes methods.
- What's the difference? Mainly the way in which the variation within population is incorporated

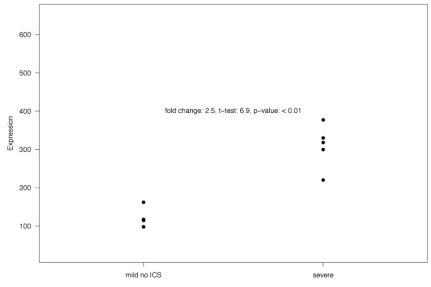
45

Should we consider variability?



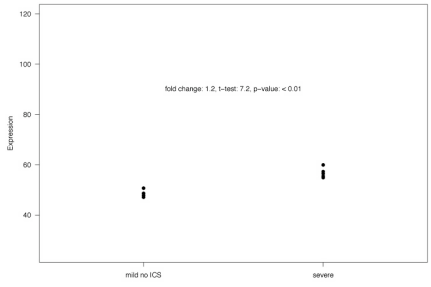
46

Should we consider variability?



47

Should we consider variability?



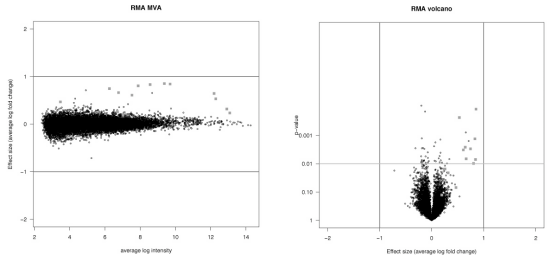
48

Useful Plots

- The MA plot shows M=log fold change, plotted against A=average log intensity
- If we have various replicates in each population we can plot M=difference in log averages in two populations
- The volcano plot shows, for a particular test, - log p-value against the effect size (M)

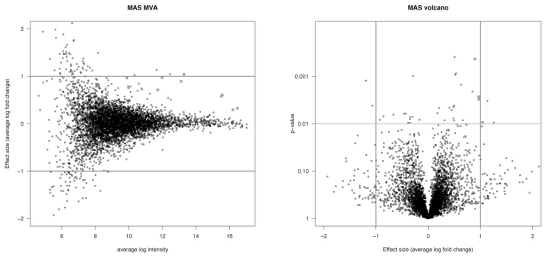
49

Example



50

If you insist on using MAS 5.0 it really helps



51

Summary

- Experimental design is critical
  - Biological vs technical replicates
- Pre-processing (calibration) can have an enormous effect on the end results
- Analysis should be guided not by what others have done, but by what you want to learn.
- Regarding multiple testing: view this process as exploratory, and don't get hung up on adjustments for multiple testing.

52