



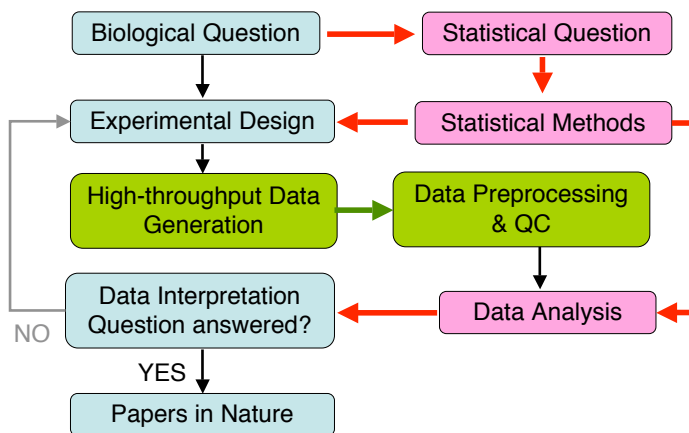
BMI 209, Fall 2006 Statistical Methods in Bioinformatics: Case Studies

Ru-Fang Yeh, Jane Fridlyand,
Yuanyuan Xiao and Mark Segal
Center for Bioinformatics & Molecular Biostatistics
UCSF Division of Biostatistics

Announcements

- Please sign up.
- Please **register for credit if you are eligible** (graduate students). We need a critical mass of registered students to keep the course. The course requirement will be minimum -- to attend and participate in class!
- All announcements & handouts, lecture slides will be available on the course website:
<http://www.biostat.ucsf.edu/cbmb/bmi209>
- **Next lecture 9/21 ONLY in GH-S202**
- **11/16 class rescheduled to 11/30 due to BMI retreat.**

Course Aims



Sequencer

SAGE

Spotted array

Mass Spectrometry

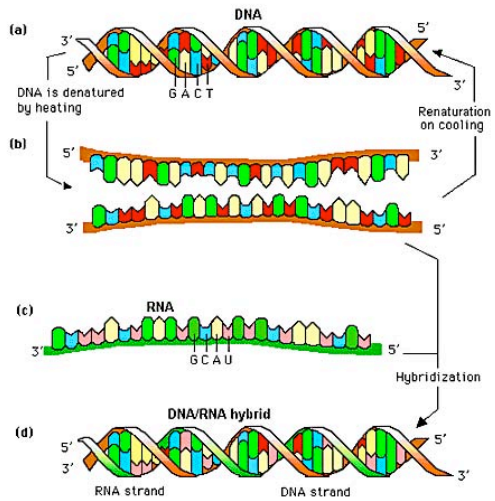
GeneChip Affymetrix

Illumina Bead Array

Agilent: Long oligo Ink Jet

Nylon membrane

Different High-throughput Technology



Nucleic Acid Hybridization

Microarray Applications

Array	Probes <i>on the array</i>	Targets <i>to be hybridized</i>	Large-scale Analysis of...
Gene Expression	DNA (cDNA, oligos; gene representatives)	mRNA/cDNA	transcriptional alterations
CGH	DNA (clones, oligos)	DNA	Genomic changes in cancers
SNP	DNA (oligos)	DNA	Genotyping; Genomic changes
Methylation	DNA (CpG island)	DNA (bisulfite-treated)	Methylation-status in genes
Promoter	DNA (promoter ~1kb)	DNA (ChIP-enriched)	Transcription factor binding sites; histone modifications
Tiling	DNA	All of the above	All of the above; sequencing; gene annotation
Protein	antibody	protein	Protein expression (ELISA)
Tissue	tissues	proteins	Histology; protein expression (immunohistochemistry)

Statistical Issues

Technology dependent:

- Preprocessing & QC

Biological question dependent:

- Hypothesis testing & Multiplicity
- Classification, prediction
- Clustering

Both technology & biological question dependent:

- Experimental design
- Meta-data integration

Roadmap

Week 1: Overview and expression array analysis

Week 2: Tiling array analysis

Week 3: Genotype calls from SNP arrays

Week 4: Copy number arrays and meta-analysis

Week 5: Classification

Week 6: Statistical validation for classification

Week 7: Mass spectrometry data analysis

Week 8: Gene network

Week 9: Phenotype array analysis

Week 10 (11/30): DNA sequence evolution

Case Study: Differential Expression Using Spotted Arrays

Question:

Identifying σ^E -dependent genes in *E. coli*
(Rhodius et al 2006 *PLoS Biol.*)

Experiment:

Transcript profiling using spotted arrays in wild type *E. coli* K-12 strand (low σ^E) versus over-expressing σ^E strand.

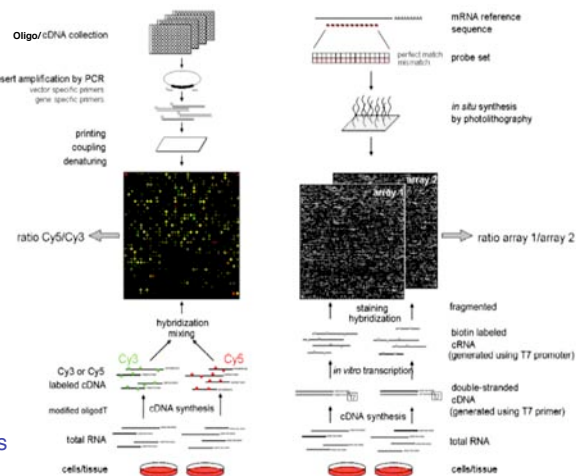
Statistical Issues

- **Experimental design:**
 - Which array? How many samples? Which to co-hybridize?
- **Data preprocessing:**
 - Quality assessment, image analysis, normalization
- **Combining replicates for differential expression:**
 - Effect estimate, statistical significance, multiple testing
- **Annotation:**
 - Shared characteristics (functional groups, motifs) among DE genes?
- **Data archive:**
 - Data submission to public database

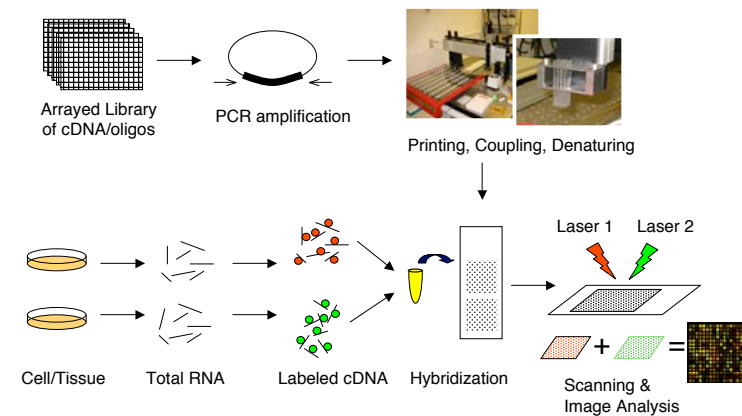
Experimental Design

Probe design:
which sequence to print on the array in which platform

Target design:
allocation of mRNA samples to the slides



Two-Color Spotted Arrays



Experimental Design

Considerations

- Scientific Aims
- Practical considerations
 - Types and amount of mRNA samples
 - Number of slides/chips available
- Other information: prior experiments, controls planned, etc.
- **Statistical principle: randomization, replication, local control.**

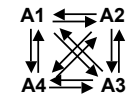
Two-color array specific design issues

- **Direct vs Indirect Comparison?**

$$\begin{array}{ccc}
 A \rightleftharpoons B & & \begin{array}{l} A \\ B \end{array} \rightarrow \text{Ref} \\
 \text{average}(\log(A/B)) & & \log(A/\text{Ref}) - \log(B/\text{Ref})
 \end{array}$$

- **K > 2 Conditions**

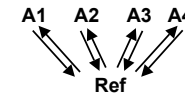
(i) All pairs



#arrays needed for the same precision for all pair comparisons

$$r \frac{K(K-1)}{2}$$

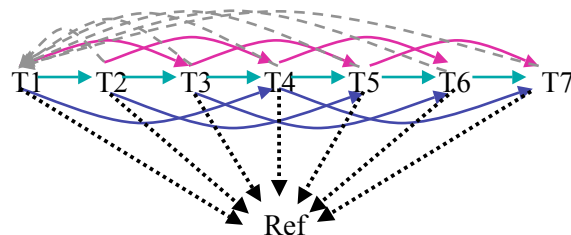
(ii) Common reference



$$r K$$

Two-color Array Design (cont.)

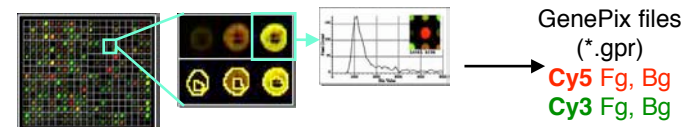
- **Time Course:**



Reference: Yang YH and Speed TP. 2002. *Nat. Rev. Genetics*.
Design Issues for cDNA Microarray Experiments.

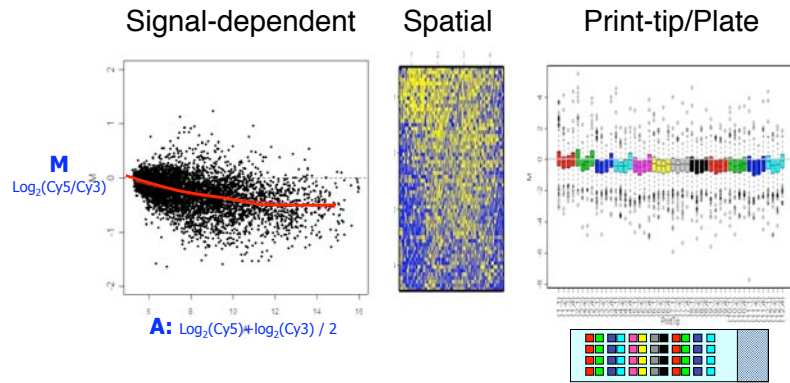
Preprocessing

- **Image Analysis:** To identify & extract signal & background

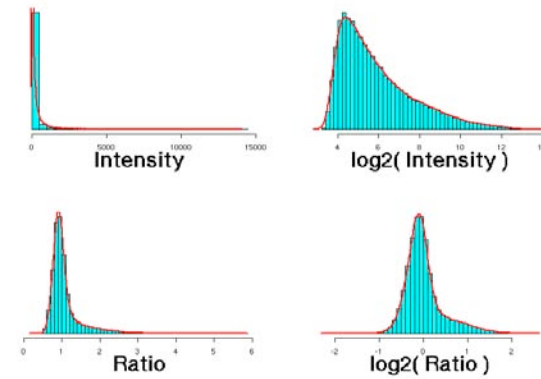


- **Within-Slide Normalization:** To identify and remove *systematic effects* not due to real biological signal.
- **Quality Assessment:** To identify bad quality arrays/spots.

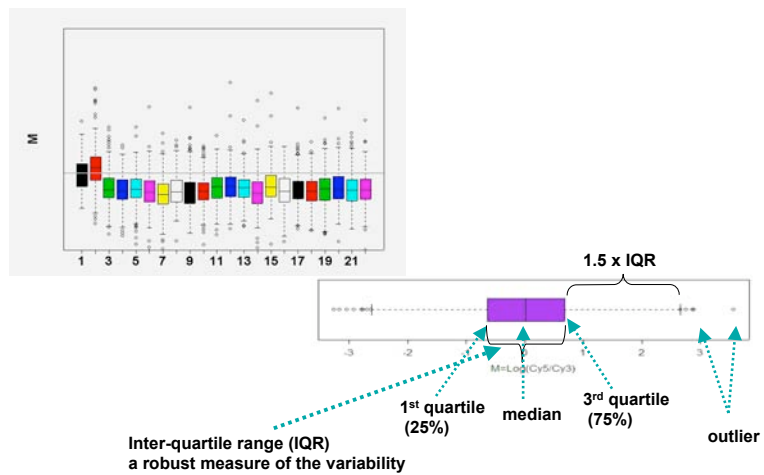
Type of "Bias"



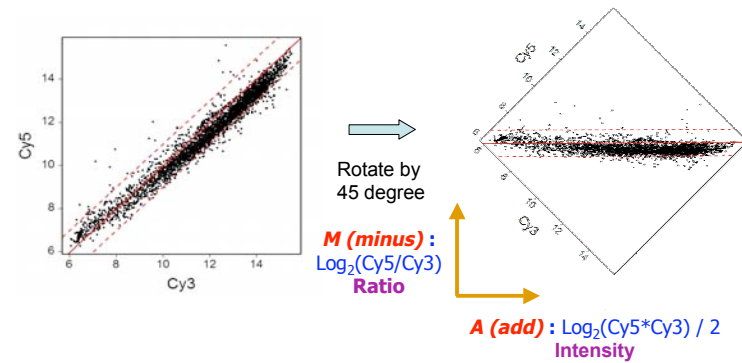
Use Log (base 2) Intensity for Analysis



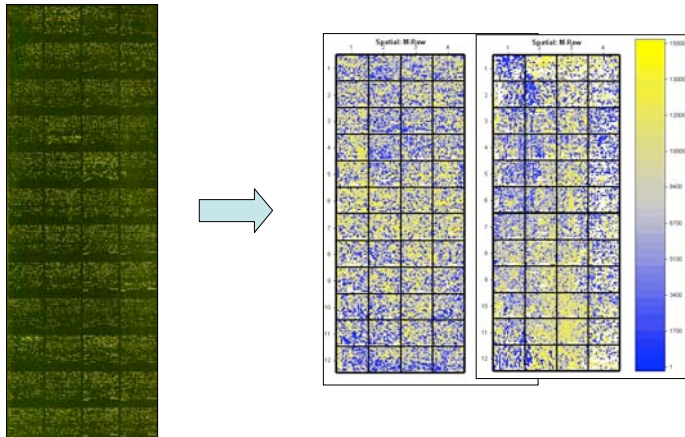
Boxplots



MA plot (log Ratio vs log Intensity)



Spatial plots/Heatmaps



Normalization

Which genes to use

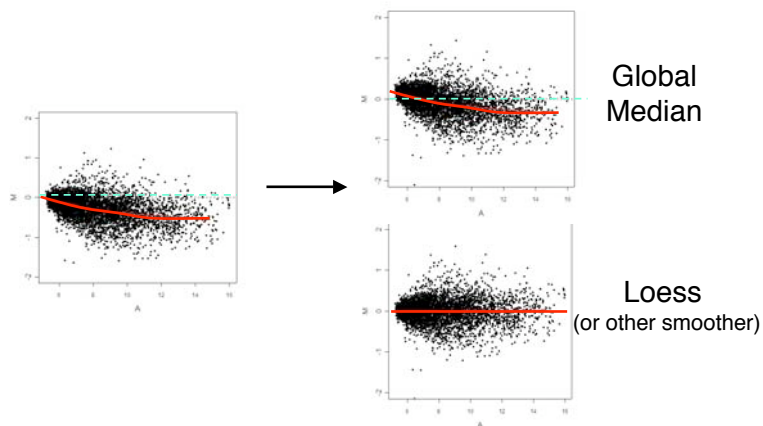
- **All genes on the array.**
- Constantly expressed genes.
- Spiked controls (e.g. plant genes).
- Genomic DNA titration series.
- Rank invariant set.

Normalization methods

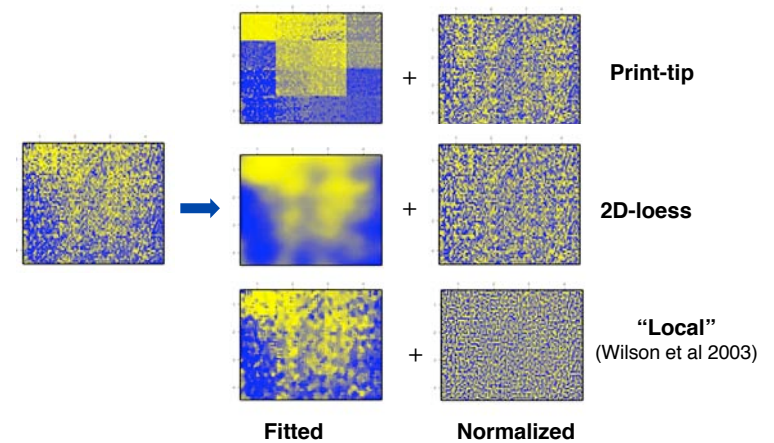
- **Ratios** [two channels], within-slide
 - Median
 - **Loess**
 - Print-tip / pins
- **Intensities** [separate channel], within- and between-slide
 - ANOVA
 - Quantile normalization
 - VSN

Assumption: No correlation between M and A, or M and layout.

Normalization (Cont.)



Normalization (Cont.)

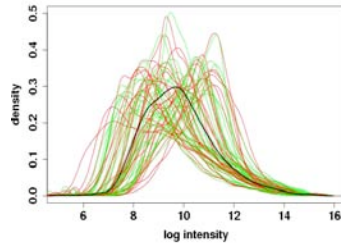


Single-channel Normalization

- Necessary for analysis methods that model log-intensities (eg: Wolfinger et al, Kerr et al)

Two stages:

- within** slide: use two-channel methods to remove spatial bias
- between** all single channels normalization: use quantile normalization or other Affy normalization methods



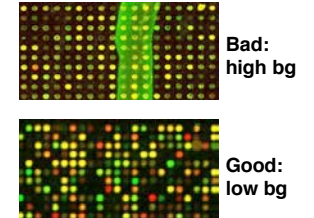
Quality Assessment

Tools:

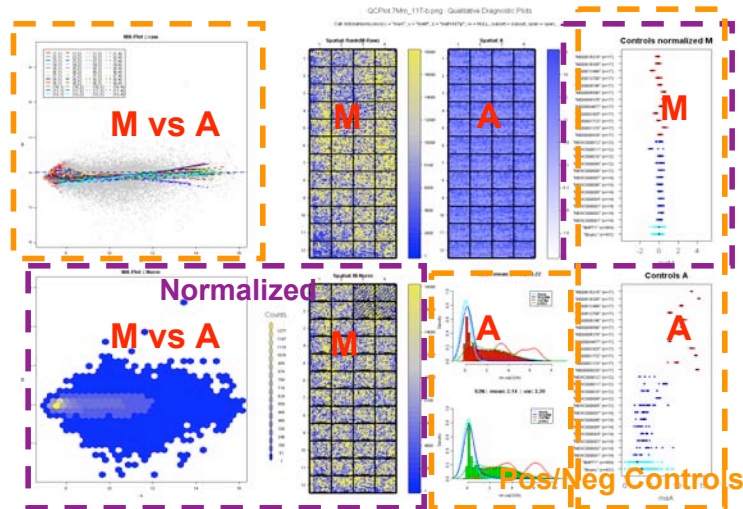
- Red/Green overlay images
- Diagnostic plots, especially on sets of positive and negative controls
- Quantitative statistics: signal to noise ratio, mean/median, variances...

Software: R/Bioconductor

- ```
> library(arrayQuality)
> gpQuality(organism="Mm")
```



## gpQuality() diagnostic plot



## Software for Preprocessing

- R / Bioconductor:
  - marray** maNorm
  - limma** normalizeWithinArrays, normalizeBetweenArrays
  - arrayQuality** gpQuality
  - vsn**
- GUI:
  - limmaGUI** <http://bioinfo.wehi.edu.au/limmaGUI/>
- Acuity and most gene expression analysis packages offer standard
  - Global normalization
  - (Print-tip) loess normalization.

## Preprocessing Reference

### Bioconductor book:

Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Edited by R Gentleman, V Carey et al. Springer-Verlag New York. 2005.



- Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, Vol. 30, No. 4, e15.
- W Huber, A von Heydebreck, H Sueltmann, A Poustka, M Vingron. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18, Suppl. 1, S96-S104.
- Y.H. Yang and N. Thorne (2003) Normalization for Two-color cDNA Microarray Data. *Science and Statistics: A Festschrift for Terry Speed, D. Goldstein (eds.)*, IMS Lecture Notes, Monograph Series, Vol 40, pp. 403-418.

## $\sigma^E$ target genes Example

### • Experimental design:

- Array: homemade E. coli array
- Targets: 4 pairs of  $\sigma^E$ -overexpressing vs wildtype E.coli

### • Data preprocessing:

- Background subtraction, Loess normalization

Data: a 4094 x 4 matrix of  $M$ , normalized  $\log_2(\text{Cy5/Cy3})$

|       | Array1 | Array2 | Array3 | Array4 |
|-------|--------|--------|--------|--------|
| Gene1 | 0.46   | 0.30   | 0.80   | 1.51   |
| Gene2 | -0.10  | 0.49   | 0.24   | 0.06   |
| Gene3 | 0.15   | 0.74   | 0.04   | 0.10   |
| Gene4 | -0.45  | -1.03  | -0.79  | -0.56  |
| Gene5 | -0.06  | 1.06   | 1.35   | 1.09   |
| ...   | ...    | ...    | ...    | ...    |

## $\sigma^E$ target genes Example (Cont.)

### • Experimental design:

- Array: homemade E. coli array
- Targets: 4 pairs of  $\sigma^E$ -overexpressing vs wildtype E.coli

### • Data preprocessing:

- Loess normalization

### • Combining replicates for differential expression:

- Effect estimate, statistical significance, multiple testing

### • Annotation:

- Shared characteristics (functional groups, motifs) among DE genes?

### • Archive and publish microarray data:

- Data submission to public databases

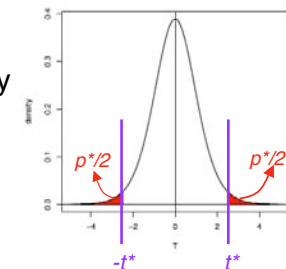
## For each gene, business as usual...

### • Effect size = *Average M* ( $\log_2$ fold-change)

### • Statistical significance by two-sample *t-statistic*:

$$T = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{M}}{se(\bar{M})}$$

- The *p-value*  $p^*$  is the probability that, under the null hypothesis  $H_0: M=0$ , the test statistic is at least as extreme as the observed value  $t^*$ .



## Problems with looking at ~4000 genes using n=4 replicates...

- (Fold-change) average  $\bar{M}$  and t-statistic  $\frac{\bar{M}}{se(\bar{M})}$  unstable when sample size (# arrays) is small.  
 ⇒ Use robust or specialized methods incorporating multi-gene information.
- p-values derived from single-gene tests are not that “statistically significant”! We will expect  $4094 \times 0.01 = 41$  genes with p-value  $< 0.01$  in 4094 random, non-DE genes!  
 ⇒ Multiple testing adjustment/consideration necessary to justify statistical significance.

## Specialized Methods: “Modified” t

- **Penalized-t (SAM, Tusher et al 2001, Efron et al 2000):**  

$$t^* = \frac{\bar{M}}{(s+a)/\sqrt{n}}$$
 Estimate penalty term  $a$  by 90th percentile of s.d. of all genes, or by minimizing the coefficient of variation of the absolute  $t$ .
- **Moderated-t (Limma, Smyth 2004):**  

$$t^* = \frac{\bar{M}}{\tilde{s}/\sqrt{n}}$$
 Use shrinkage s.d.  $\tilde{s}^2 = \frac{s^2 d + s_0^2 d_0}{d + d_0}$  estimated by an empirical Bayes method  
 $s_0$ : pooled s.d.,  $d_0$ : d.f. of prior
- **Regularized-t (Cyber-T, Baldi P & Long AD 2001):**  

$$t^* = \frac{\bar{M}}{\tilde{s}/\sqrt{n}}$$
 Use regularized s.d.  $\tilde{s}^2 = \frac{v_0 \sigma_0^2 + (n-1)s^2}{v_0 + n - 2}$   
 $v_0$ : prior strength  
 $\sigma_0^2$ : background s.d.

## Alternative Statistics

- **B-statistic (Lonnstedt & Speed 2002):** The log posterior odds ratio that a gene is DE vs not DE, estimated by the empirical Bayes method.

$$B = \log \frac{\Pr\{\text{DE}\}}{\Pr\{\text{not DE}\}} = \log \frac{p}{1-p} \left( \frac{v}{v+v_0} \right)^{1/2} \left( \frac{t^2 + d_0 + d}{t^2 \frac{v}{v+v_0} + d_0 + d} \right)^{(1+d+d_0)/2}$$

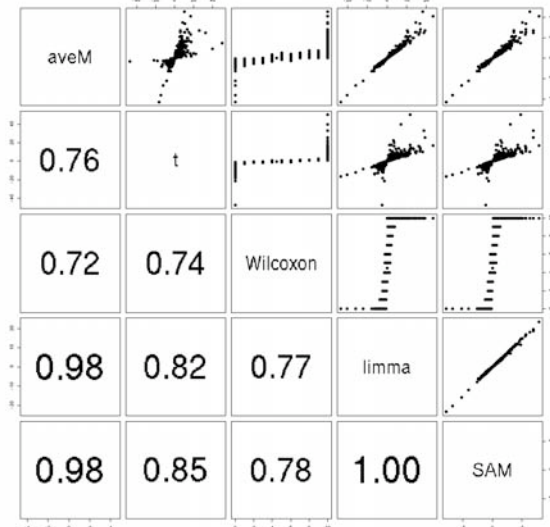
- Equivalent to **moderated-t** in terms of ranking genes.
- Dependent on **p = expected proportion of DE genes**

- **Distance Synthesis (DEDS, Yang et al 2004):** Define a *distance* statistic based on measures of choice, and estimate false discovery rates using appropriate null distribution.
- Single channel methods modelling absolute Cy5 & Cy3 expression (Newton et al 2001, Wolfinger et al 2001)

## DE Method Summary

| Method                                      | Error model                        | Multi-gene | Handles > two-sample? | Statistical significance |
|---------------------------------------------|------------------------------------|------------|-----------------------|--------------------------|
| Mean/Median                                 | NO                                 | NO         | NO                    | NO                       |
| t-statistic                                 | Log-normal                         | NO         | YES using ANOVA       | Single-gene p-value      |
| Wilcoxon statistic                          | Rank-based                         | NO         | NO                    | Single-gene p-value      |
| Moderated t-statistic / B-statistic (limma) | Log-normal hierarchical            | YES        | YES                   | Single-gene p-value      |
| Penalized t (SAM)                           | Log-normal                         | YES        | YES, limited          | Estimating FDR, q-value  |
| Regularized t (Cyber-T)                     | Log-normal                         | YES        | YES                   | Single-gene p-value      |
| EBarrays                                    | Gamma-gamma or log-normal mixtures | YES        | YES                   | Single-gene p-value      |
| maanova                                     | Log-normal                         | NO         | YES                   | Single-gene p-value      |
| DEDS                                        | Depending on choices of measures   |            |                       | Estimating FDR           |

WT vs rpoE+ E.coli K12 data from Rhodius et al 2006, 4 Replicates, 4094 genes



## Software Tools: DE (free, non-inclusive)

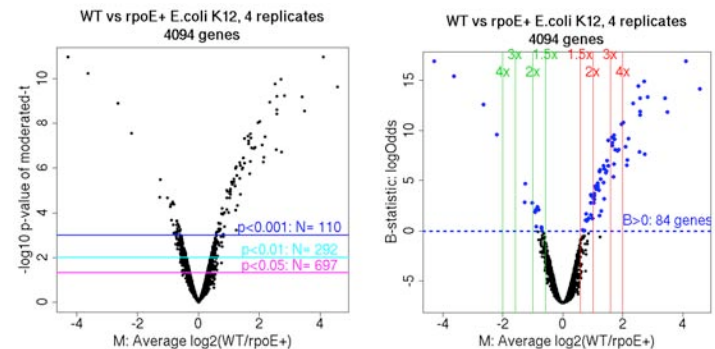
| Method              | What                                                            | URL                                                                                                                                                          |
|---------------------|-----------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>R</b>            | Statistical computing environment                               | <a href="http://www.r-project.org/">http://www.r-project.org/</a>                                                                                            |
| <b>Bioconductor</b> | A collection of R packages for genomic data                     | <a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a>                                                                                      |
| <b>Limma</b>        | Linear Models for MicroArray data; empirical Bayes moderated-t  | R/limma; R/limmaGUI; R/affymGUI<br><a href="http://bioinf.wehi.edu.au/limma/">http://bioinf.wehi.edu.au/limma/</a>                                           |
| <b>SAM</b>          | Significance Analysis of Microarrays; penalized-t; estimate FDR | R/samr, Excel macro (PC only)<br><a href="http://www-stat.stanford.edu/~tibs/SAM/">http://www-stat.stanford.edu/~tibs/SAM/</a>                               |
| <b>Cyber-T</b>      | Regularized-t                                                   | R/hdarray, R/bayesreg, R/bayesAnova<br><a href="http://visitor.ics.uci.edu/genex/cybert/index.shtml">http://visitor.ics.uci.edu/genex/cybert/index.shtml</a> |
| <b>EBarrays</b>     | Parametric empirical bayes                                      | R/EBarrays                                                                                                                                                   |
| <b>maanova</b>      | Fixed/mixed effect ANOVA                                        | R/maanova<br><a href="http://www.jax.org/staff/churchill/labsite/software/Rmaanova/">http://www.jax.org/staff/churchill/labsite/software/Rmaanova/</a>       |

## DE References

**Bioconductor book:** Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Edited by R Gentleman, V Carey et al. Springer-Verlag New York. 2005.

- Limma:** Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3, No. 1, Article 3.
- SAM:** Tusher, Tibshirani and Chu. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* 98: 5116-5121.
- Cyber-T:** P. Baldi and A.D. Long. (2001). A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes, *Bioinformatics* 17, 509-519.
- EBarrays:** Kendziorski, C.M., M.A. Newton, H. Lan, and M.N. Gould. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22:3899-3914, 2003.
- maanova:** Kerr, Martin and Churchill(2000), Analysis of variance for gene expression microarray data, *Journal of Computational Biology*, 7:819-837.
- Wolfinger RD, Gibson G, et al. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol.* 8(6):625-37.

## Assessing Statistical Significance How to set a cutoff for DE?



Expect  $0.05 * 4094 \approx 105$  genes with  $p < 0.05$  just by chance even if there is no DE!

## Multiple Testing Consideration

- Consider an “overall” error rate by either
- adjust single-gene p-values to control for, or
  - estimate the overall error (false positives).

“Overall” type-I error rate:

- Family-Wise Error Rate (FWER)**  
=  $P(\text{FP} > 0)$
- False Discovery Rate (FDR)**  
= Expected proportion of false positives  
=  $E(\text{FP}/nP)$  if  $nP > 0$ , 0 if  $nP=0$ .

|       |      |    |
|-------|------|----|
|       | test |    |
|       | -    | +  |
| truth | -    | +  |
| -     | TN   | FP |
| +     | FN   | TP |
|       | nN   | nP |

## Adjusted p-values controlling the FWER

Choosing all genes with adjusted p-value  $\tilde{p}_g \leq \alpha$  controls the FWER at level  $\alpha$

- The **Bonferroni** correction:  $\tilde{p}_g = mp_g$   
Most conservative adjustment; assume independence among genes.
- Alternatives:
  - Sidák**:  $\tilde{p}_g = 1 - (1 - p_g)^m$
  - minP** (Westfall & Young):  $\tilde{p}_g = \Pr(\min_{k=1, \dots, m} P_k \leq p_g | H_0)$   
estimated through permutation; allow dependency between genes.
  - maxT**: replace  $p_g$  by test statistics  $T_g$ , min by max. Less computationally intensive than minP.
  - Step-down; step-up; ...

## Adjust p-values Controlling the FDR

(Benjamini & Hochberg 1995)

- Order raw p-values:  $p_{r_1} \leq p_{r_2} \leq \dots \leq p_{r_m}$ .
- To control *FDR* at level  $\alpha$ , reject the hypothesis  $H_{r_j}$  for  $j = 1, \dots, j^*$ .  $j^* = \max\{j : p_{r_j} \leq (j/m)\alpha\}$ .
- Adjusted p-values:  $\tilde{p}_{r_j} = \min_{k=j, \dots, m} \left\{ \min\left(\frac{m}{k} p_{r_k}, 1\right) \right\}$ .

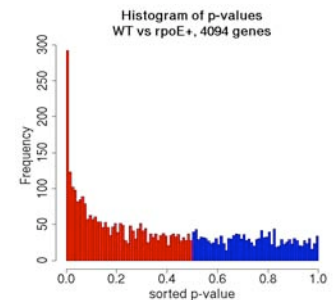
| j                     | 1      | 2      | 3      | 4      | 5    |
|-----------------------|--------|--------|--------|--------|------|
| rawP                  | 0.0003 | 0.0004 | 0.01   | 0.07   | 0.08 |
| Bonferroni-adjusted P | 0.0015 | 0.002  | 0.05   | 0.35   | 0.4  |
| FDR= rawP * 5 / j     | 0.0015 | 0.001  | 0.0167 | 0.0875 | 0.08 |
| Adjusted-P            | 0.001  | 0.001  | 0.0167 | 0.08   | 0.08 |

Interpretation: expect 5% false positives among genes with < 0.05 FDR-adjusted p-values.

## “Estimation” of the FDR

(SAM, Tusher et al 2001, Storey&Tibshirani 2003)

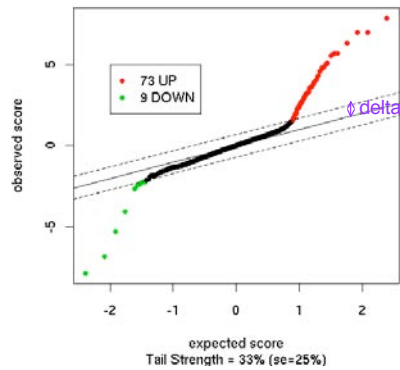
- Select a set of “null” genes  
Eg: genes with top 50% (large) p-values, which are unlikely to be DE.
- For the chosen cutoff value of the test statistic, estimate the expected proportion of false positives by permutations of the class labels of the null genes.
  - Any resulting test statistic exceeding the cutoff is a “false positive”.



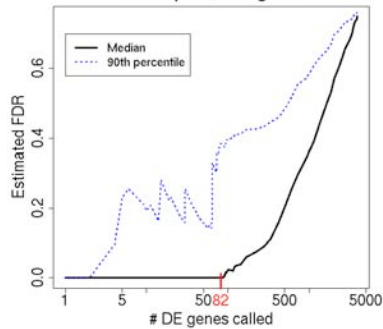
*q-value* = min estimated FDR at which it appears significant

## SAM-FDR: Example (cont)

QQ-plot of SAM Analysis: WT vs rpoE+  
delta=0.708, median # false positives = 0



Estimated False Discovery Rate  
WT vs rpoE+, 4094 genes



Multiple Testing Procedures  
Moderated-t p-values of WT vs rpoE+ from Rhodius et al 2006

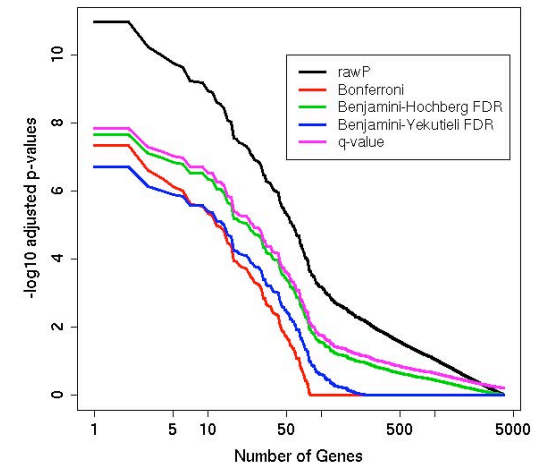


TABLE 2  
Properties of multiple testing procedures

| Procedure                              | Type I error rate              | Strong or weak control | Stepwise structure | Dependence structure           |
|----------------------------------------|--------------------------------|------------------------|--------------------|--------------------------------|
| Bonferroni                             | FWER                           | Strong                 | Single             | General/ignore                 |
| Šidák                                  | FWER                           | Strong                 | Single             | Positive orthant dependence    |
| min $P$                                | FWER                           | Strong                 | Single             | Subset pivotality              |
| max $T$                                | FWER                           | Strong                 | Single             | Subset pivotality              |
| Holm (1979)                            | FWER                           | Strong                 | Down               | General/ignore                 |
| Step-down Šidák                        | FWER                           | Strong                 | Down               | Positive orthant dependence    |
| Step-down min $P$                      | FWER                           | Strong                 | Down               | Subset pivotality              |
| Step-down max $T$                      | FWER                           | Strong                 | Down               | Subset pivotality              |
| Hochberg (1988)                        | FWER                           | Strong                 | Up                 | Some dependence (Simes)        |
| Troendle (1996)                        | FWER                           | Strong                 | Up                 | Some dependence                |
| Benjamini and Hochberg (1995)          | FDR                            | Strong                 | Up                 | Positive regression dependence |
| Benjamini and Yekutieli (2001)         | FDR                            | Strong                 | Up                 | General/ignore                 |
| Yekutieli and Benjamini (1999)         | FDR                            | Strong                 | Up                 | Some dependence                |
| Unadjusted $p$ -values                 | PCER                           | Strong                 | Single             | General/ignore                 |
| SAM, Tusher, Tibshirani and Chu (2001) | PFER (PCER)                    | Strong                 | Single             | General/hybrid                 |
| SAM, Efron et al. (2000)               | PFER (PCER)                    | Weak                   | Single             | General                        |
| Golub et al. (1999), step-down         | $\Pr(R \geq r   H_0^C)$ (FWER) | Weak                   | Down               | General                        |
| Golub et al. (1999), step-up           | $\Pr(R \geq r   H_0^C)$        | Weak                   | Up                 | General                        |

Dudoit et al 2003.

## Multiple Testing References

A comprehensive review:

S Dudoit, JP Shaffer & JC Boldrick (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, Vol. 18, 71-103.

Additional references:

- Benjamini, Y & Y Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289-300.
- Westfall, PH and SS Young (1993) Resampling-based multiple testing. Wiley.
- VG Tusher et al. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98, 5116-5121.
- Storey JD and Tibshirani R. (2003) Statistical significance for genome-wide experiments. *Proceedings of the National Academy of Sciences*, 100: 9440-9445.

R/Bioconductor package: [p.adjust](#), [multtest](#), [qvalue](#)

## FWER or FDR?

Depending on the study aims:

- Choose more conservative **FWER** if high confidence in all/most selected genes is desired: e.g., selecting a few candidate genes for experimental validation.
- Use more flexible **FDR** procedures if certain proportions of false positives are tolerable: e.g. gene discovery, selecting candidate co-regulated gene sets for GO/pathway analysis.
- In practice: effect size + statistical significance

some fold-change threshold, eg:  $M=4x, 2x, 1.5x$  ; and/or  
Bonferroni-adjusted  $p < 0.05$  or FDR-adjusted  $p < 0.05$  or  $B > 0$

More stringent

less stringent

## Alternative ways to cope with multiple testing

- Carefully interpret unadjusted p-values and comment on number of expected false positives.
- Reduce the number of tests by
  - **Filtering** genes: low intensity
  - **Grouping** genes: considering gene sets defined by clustering or functional annotation such as Gene Ontology categories, (KEGG, BioCarta) pathways, genes with common known TF binding sites.

## Functional Annotation (manually curated)

### • Functional category:

Gene Ontology: <http://www.geneontology.org>

### • Pathways:

GenMAPP: <http://www.genmapp.org/>

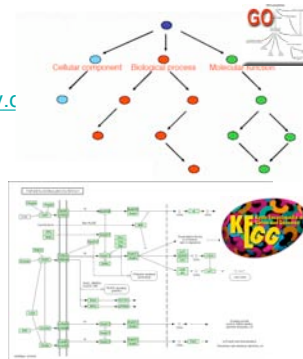
KEGG: <http://www.genome.jp/kegg/>

BioCarta: <http://www.biocarta.com/>

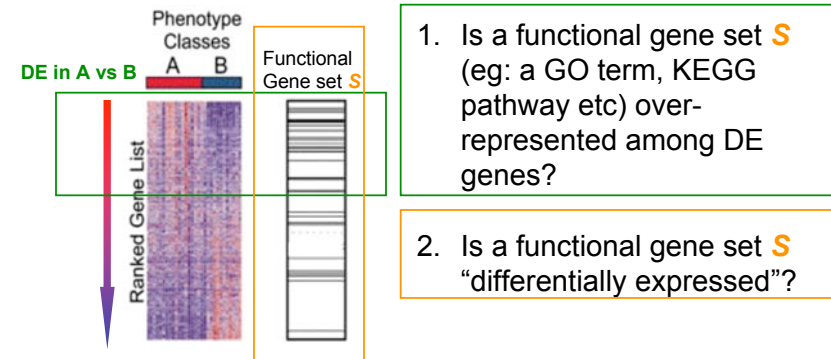
### • Protein Interactions:

BIND: <http://bind.ca/>

Human protein reference database: <http://www.hprd.org/>



## Inference: Functional Enrichment



## 1. Is a GO term over-represented among DE genes?

|                                                    | Contingency Table                                |      |      | P-value                                |
|----------------------------------------------------|--------------------------------------------------|------|------|----------------------------------------|
| count genes with GO term in set                    | 51                                               | 416  | 467  | 8x10 <sup>-52</sup>                    |
| count genes without GO term in set                 | 125                                              | 8588 | 8713 |                                        |
|                                                    | 173                                              | 9004 | 9177 |                                        |
| count in set (e.g. differentially expressed genes) | Count in reference set (e.g. all genes on array) |      |      | Fisher's exact test or chi-square test |

**Note: Multiple testing with complex dependence!**

## GO Tools

Listing: <http://www.geneontology.org/GO.tools.shtml>

### Finding over-representation among a list of genes:

- Gostat: <http://gostat.wehi.edu.au/>
- MAPPFinder: <http://www.genmapp.org/MAPPFinder.html>
- EASE: <http://david.niaid.nih.gov/david/ease.htm>
- FantiGO: <http://fatigo.bioinfo.cnio.es/>
- GoMiner: <http://discover.nci.nih.gov/gominer/>
- Bioconductor: GOSTats and goTools

and many more...

## 2. Is a GO term “differentially expressed”?

- Use **Wilcoxon Signed Rank test** or **Kolmogorov-Smirnov test** (Mootha et al 2003; Subramanian et al 2005. **Gene Set Enrichment Analysis**) to determine whether the ranks for a particular GO term are significantly higher or lower than usual.
- **Gostat**: <http://gostat.wehi.edu.au>  
**GSEA**: <http://www.broad.mit.edu/gsea/>  
 R/Bioconductor: **limma (geneSetTest)**

## Archiving & Searching Microarray Data

- Array Database:
  - NCBI, Gene Expression Omnibus (GEO).  
<http://www.ncbi.nlm.nih.gov/geo/>
  - EBI, ArrayExpress.  
<http://www.ebi.ac.uk/arrayexpress/>
- Data format: MIAME

 © 2001 Nature Publishing Group <http://genetics.com> *commentary*

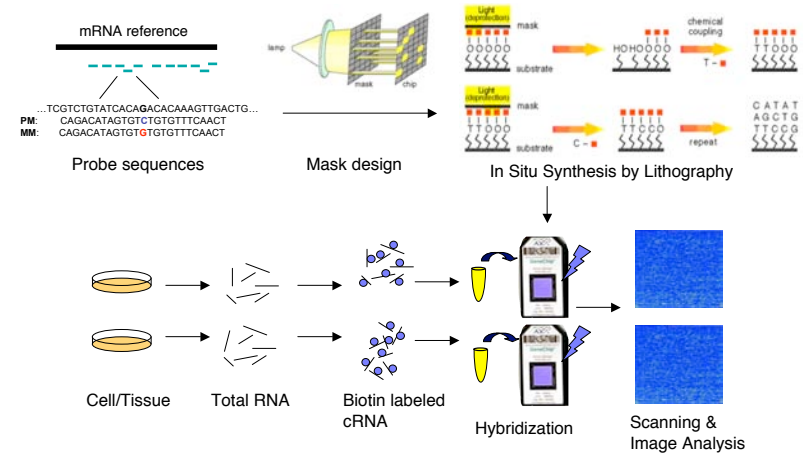
### Minimum information about a microarray experiment (MIAME)—toward standards for microarray data

Alvis Brazma<sup>1</sup>, Pascal Hingamp<sup>2</sup>, John Quackenbush<sup>3</sup>, Gavin Sherlock<sup>4</sup>, Paul Spellman<sup>5</sup>, Chris Stoeckert<sup>6</sup>, John Aach<sup>7</sup>, Wilhelm Ansorge<sup>8</sup>, Catherine A. Ball<sup>9</sup>, Helen C. Causton<sup>9</sup>, Terry Gaasterland<sup>10</sup>, Patrick Glenisson<sup>11</sup>, Frank C.P. Holstege<sup>12</sup>, Irene F. Kim<sup>5</sup>, Victor Markowitz<sup>13</sup>, John C. Matise<sup>4</sup>, Helen Parkinson<sup>1</sup>, Alan Robinson<sup>1</sup>, Ugis Sarkans<sup>1</sup>, Steffen Schulze-Kremer<sup>14</sup>, Jason Stewart<sup>15</sup>, Ronald Taylor<sup>16</sup>, Jaak Vilo<sup>1</sup> & Martin Vingron<sup>17</sup>

## Summary

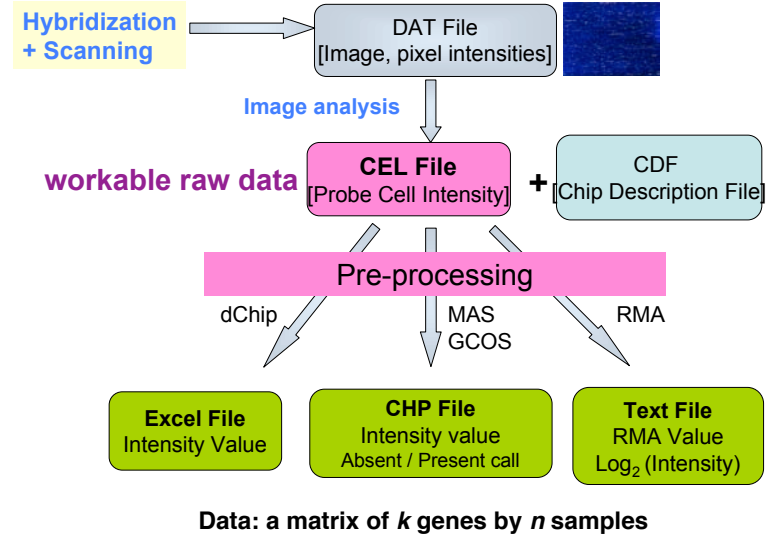
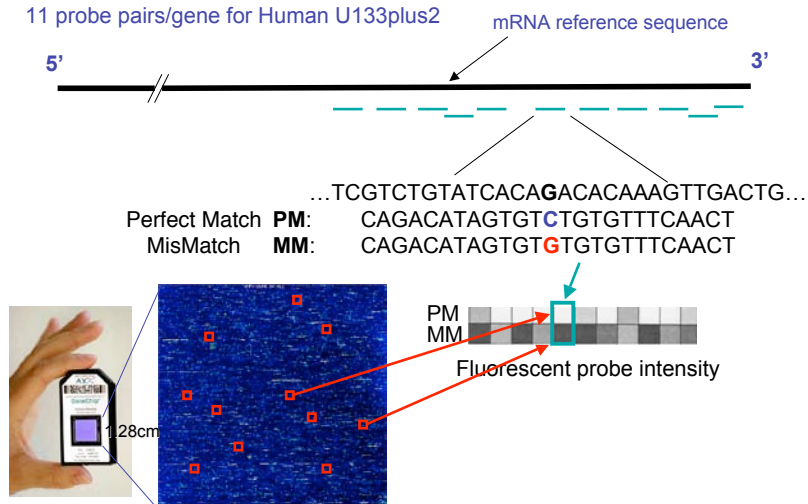
- Replication, randomization and proper controls are required for a good microarray experiment.
- Be aware of the assumptions of any normalization procedure. Many assume no substantial global changes, or approximate up/down symmetry. Note that the 'systematic bias' to be removed can contain real biological signals.
- Normalization can not fix bad quality data.
- Differential expression measures are continuous. Any threshold is ad hoc. But MT-adjusted p-values or log-odds can provide more rationalized cutoffs.

## Alternative DNA Microarray Platform: Affymetrix High Density Short Oligo Arrays



### For one gene (probe set):

11 probe pairs/gene for Human U133plus2



Data: a matrix of  $k$  genes by  $n$  samples

## Preprocessing of Affymetrix data

Computing expression measures as a three-step procedure:

- **Background subtraction (B)**
- **Normalization (N)** to facilitate *between-array* comparison
- **Summarization** of 11-20 probe pair (PM/MM) intensities to one probe set value (**S**).

Let X be CEL file data from multiple arrays then

$$\text{Expression values} = S(N(B(X)))$$

## Affymetrix Data Preprocessing Methods

- Affymetrix: **MAS v5.1**  
Signal = TukeyBiweight $\{\log(\text{PM}_j - \text{MM}_j^*)\}$
- Robust multi-array analysis (**RMA**, Irizarry et al 2003):  
use median polish or robust regression fit (IRLS) and estimate background from PMs  
 $\log_2(\text{PM-BG})_{ij}^* = \text{chip}_i + \text{probe}_j + \varepsilon_{ij}$

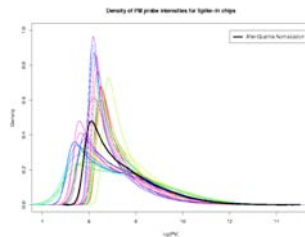
RMA log expression values

- **dChip** (Li, Schadt & Wong 2001): Maximum likelihood estimate.  
 $\text{PM}_{ij} - \text{MM}_{ij} = \text{chip}_i \text{probe}_j + \varepsilon_{ij}$

dChip expression values

## Between-Array Normalization

- **MAS5**: N/A. Single-chip method. Global scaling suggested for probe set data.
- **dChip**: Use piecewise linear normalization of rank-invariant sets (estimated non-DE genes) between each array and a baseline array.
- **RMA**: Quantile normalization (B Bolstad et al 2003).



## Affymetrix array Quality Assessment using weights from affyPLM

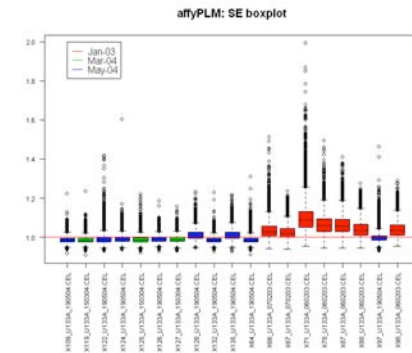
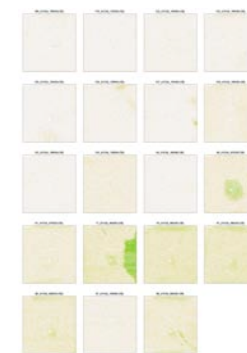


Image Gallery:

<http://stat-www.berkeley.edu/users/bolstad/PLMImageGallery/index.html>

## Affymetrix Preprocessing Softwares

- **RMA:**
  - R/Bioconductor package & functions
    - affy** *rma, justRMA*
    - affyPLM** *fitPLM, image, boxplot*
    - gcrma** *justGCRMA*
  - Standalone GUI:
    - affyImGUI:** <http://bioinf.wehi.edu.au/affyImGUI/>
    - RMAExpress:** <http://rmaexpress.bmbolstad.com/>
- **dChip:** <http://www.dchip.org>
- Affymetrix own software: **MAS5, PLIER**

Preprocessing Method Comparison: AffyComp  
<http://affycomp.biostat.ihsp.edu/>

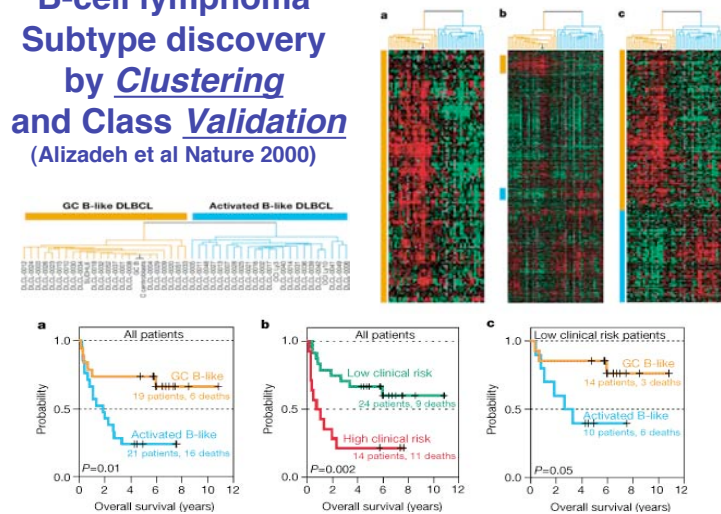
## Affymetrix Preprocessing Reference

Bioconductor book

- RMA:** Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics* 4(2): 249-64.
- GCRMA:** Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F. (2004). A Model Based Background Adjustment for Oligonucleotide Expression Arrays In the Journal of the American Statistical Association. 99, 909–917.
- Quantile Normalization:** Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* 19(2):185-193
- dChip:** Li, C. and W. H. Wong (2001). "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." *Proc Natl Acad Sci U S A* 98(1): 31-6.
- Method comparison**  
 Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*. 2006 Apr 1;22(7):789-94.  
 Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol*. 2005;6(2):R16.

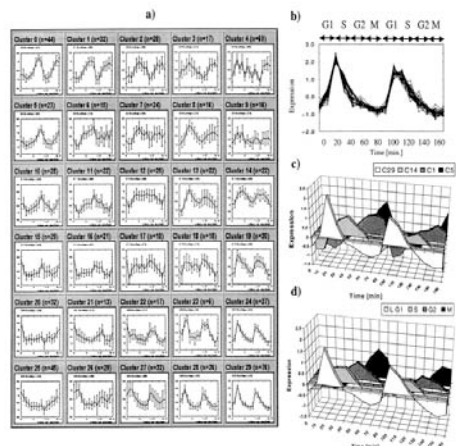
## Other Statistical Analysis Questions

### B-cell lymphoma Subtype discovery by Clustering and Class Validation (Alizadeh et al Nature 2000)



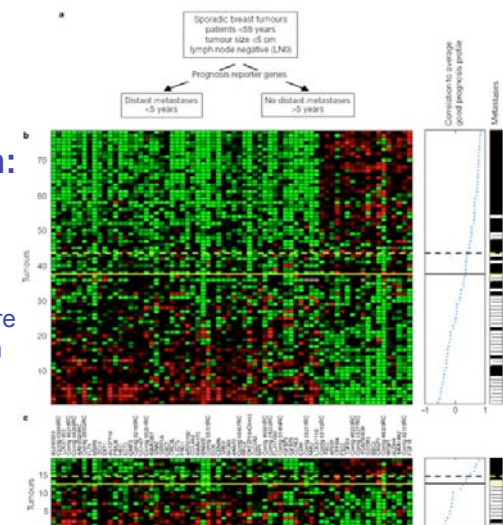
## Finding Groups of Co-Regulated Genes with similar time profile in Yeast Cell Cycle using SOM Clustering

(Tamayo *et al*, PNAS 1999, Data from Cho *et al*, Mol. Cell 1998)



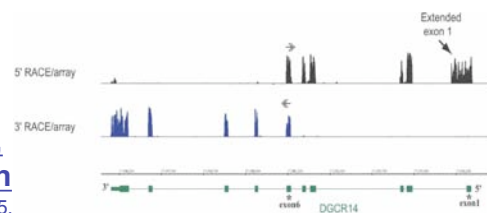
## Prognosis Prediction with Gene Expression: Classification & Validation

(van't Veer LJ *et al*, Nature 2002; Further validated in NEJM 2002).



## Transcript (Exon) Detection Using Tiling Arrays: Estimation, Testing, Meta-data Integration

(Kapranov *et al*. Genome Res 2005.)



## Chromosomal Aberrations by array CGH: Segmentation, Meta-data Integration

(Selzer RR *et al*. Genes Chromosomes Cancer. 2005)

