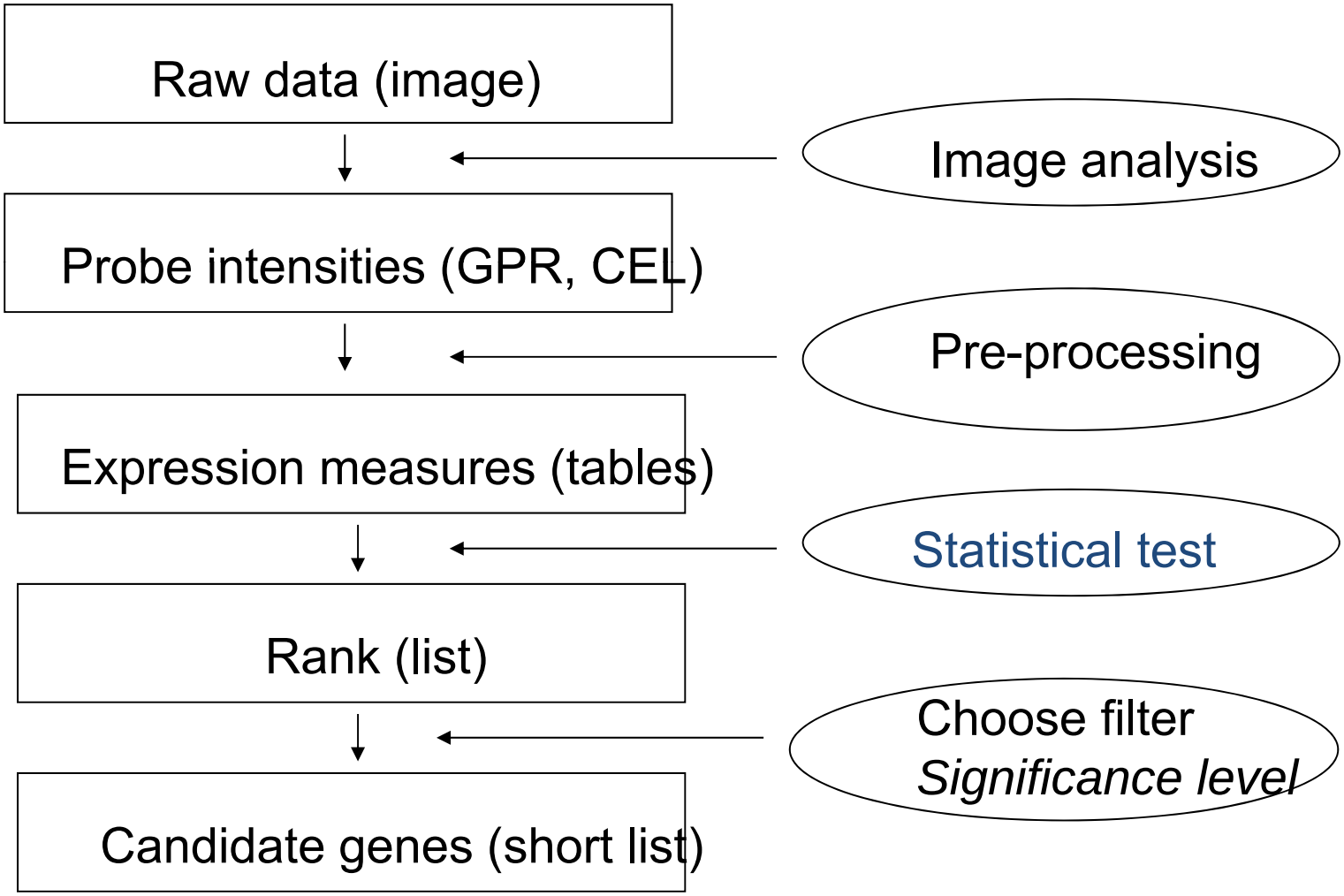
Differential Expression

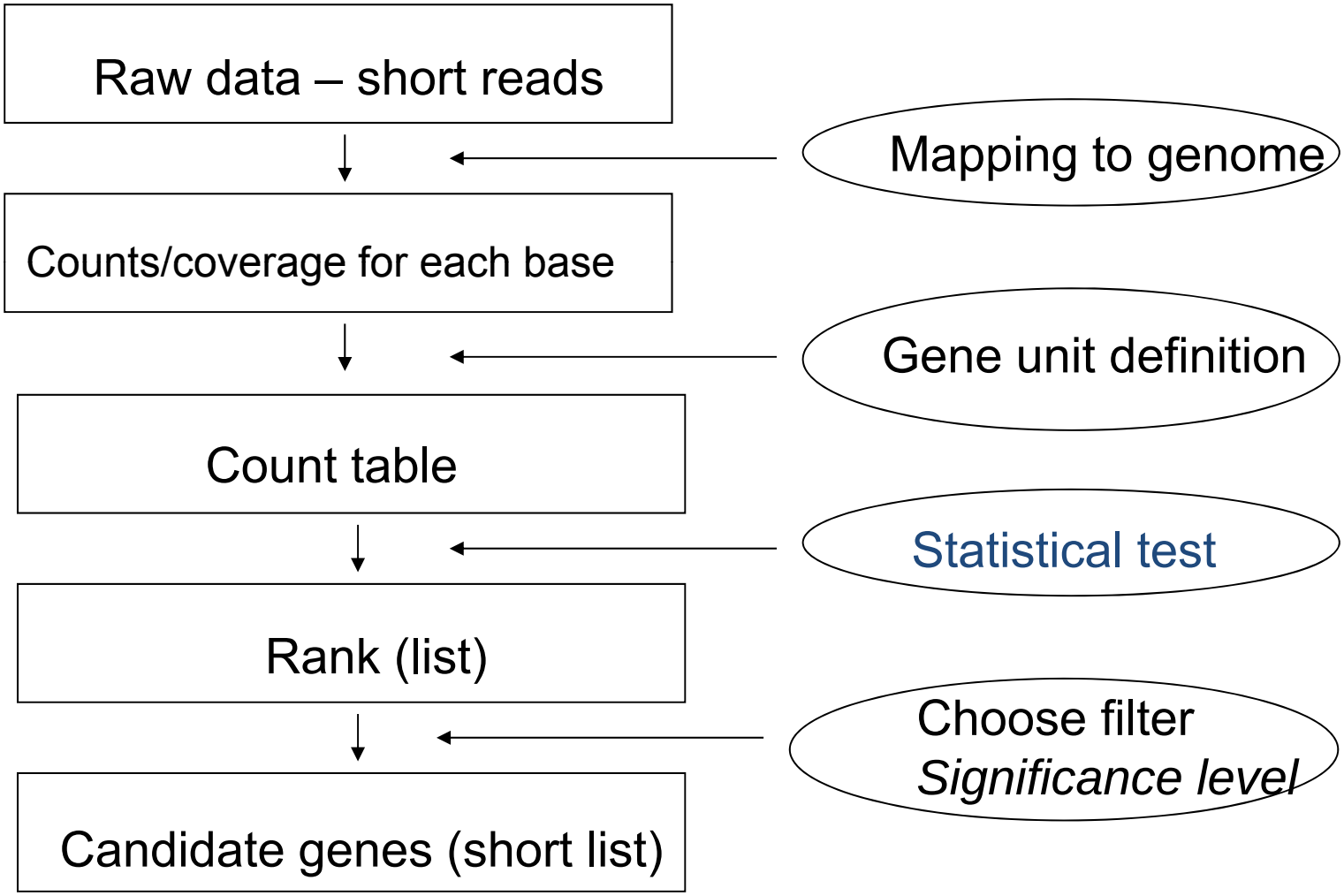
Acknowledgement: Rafael Irizarry
Sandrine Dudoit

- We use gene expression as an example but many of the issues and solutions apply, in principle, in other types of data as well

Microarray data



RNAseq data


Raw data – short reads
Mapping to genome
Counts/coverage for each base
Gene unit definition
Count table
Statistical test
Rank (list)
Choose filter
Significance level
Candidate genes (short list)

Differential gene expression

- **Goal:** Identify genes whose expression levels are associated 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.

Example: Two populations

- Finding genes that are differentially expressed in two populations, eg Cancer/normal, treated/control
  - Let X_g , Y_g represent the random variables of gene expression for gene *g* in the two populations
  - If a gene is not associated with the treatment, its mean expression levels in the two populations are expected to be equal (non-differentially expressed)

    $E[X_g] = E[Y_g]$ for these *g* s
  - It is likely that a number of genes could have differential expression in the two populations.

    $E[X_g] \neq E[Y_g]$

Which genes have the "most differential expression"?

- If we knew $E[X_g]$ and $E[Y_g]$, we may simply rank $abs(E[X_g] - E[Y_g])$
- Now we only have a few replicates
  - $x_{g1}, x_{g2}, \ldots, x_{gk}$ as k realizations of the random variable X_g
  - $y_{g1}, y_{g2}, \ldots, y_{gJ}$ as J realizations of the random variable Y_g

  We can estimate $E[X_g]$ and $E[Y_g]$ by the sample means.

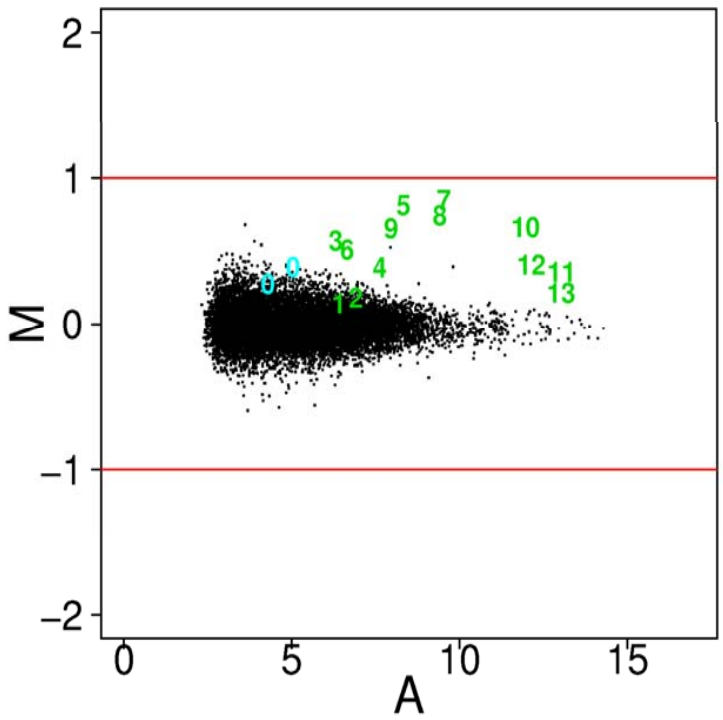
  ??? Rank by $abs(\overline{X} - \overline{Y})$???

Back to Basics

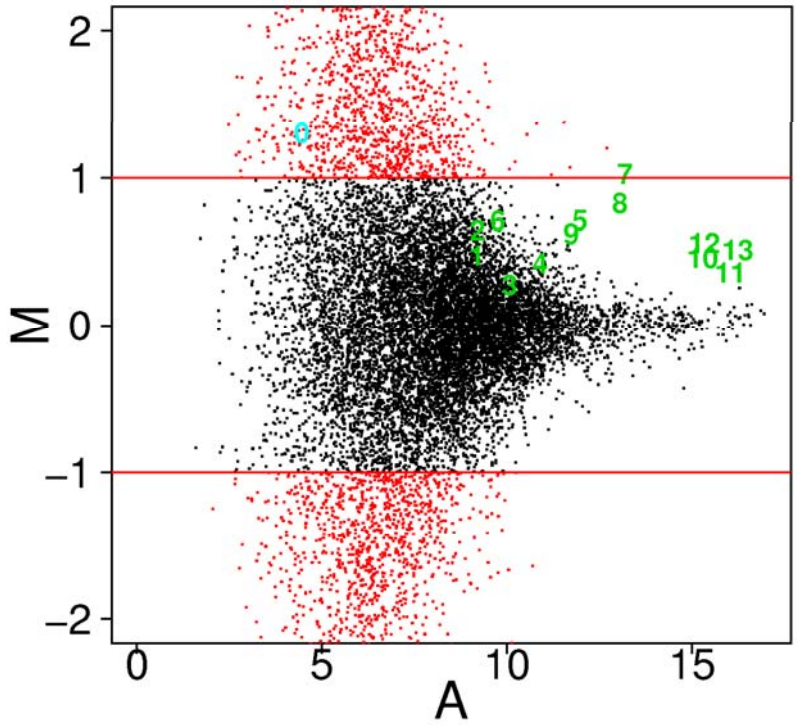
- If we are interested in genes with over-all large fold changes why not look at average log ratios?
- We can make MA plots:
  - M = difference in average log intensities and
  - A = average of average log intensities

Comparison of two arrays

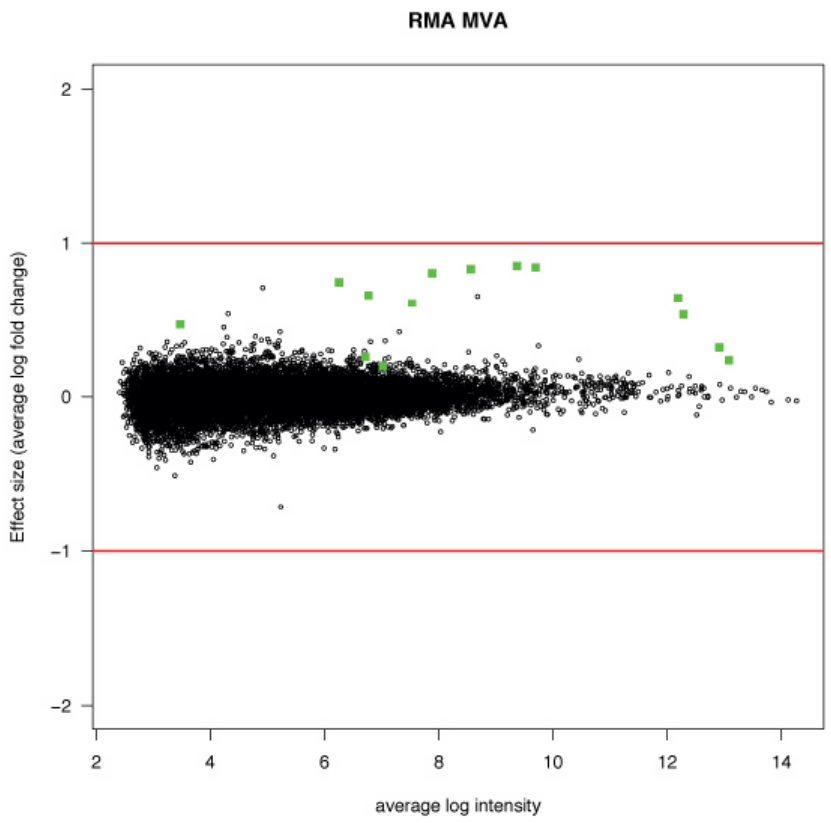
RMA

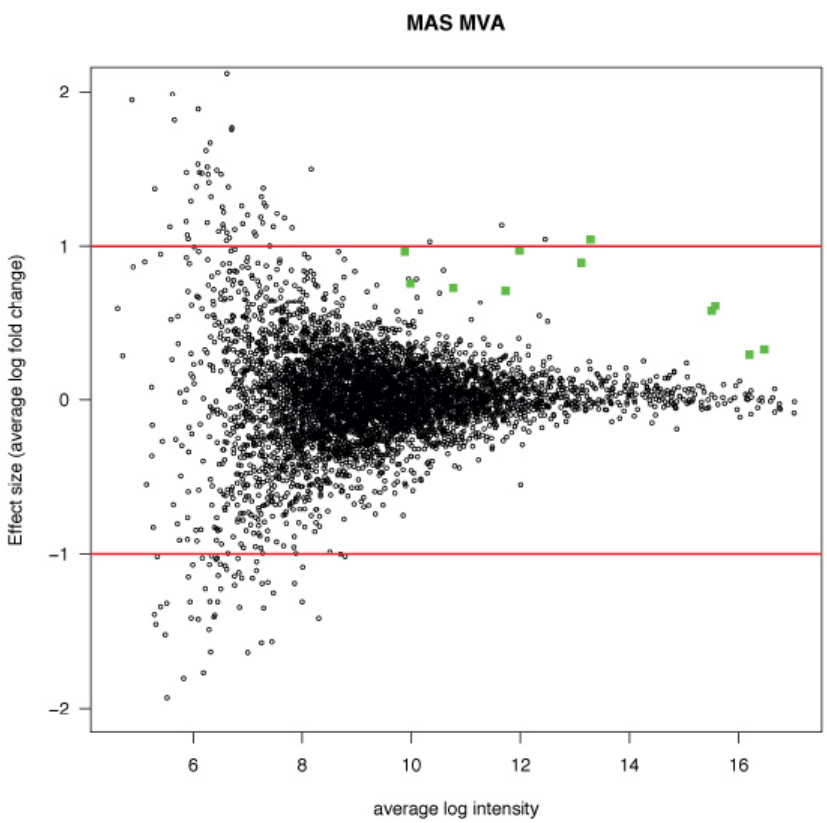


MAS 5.0

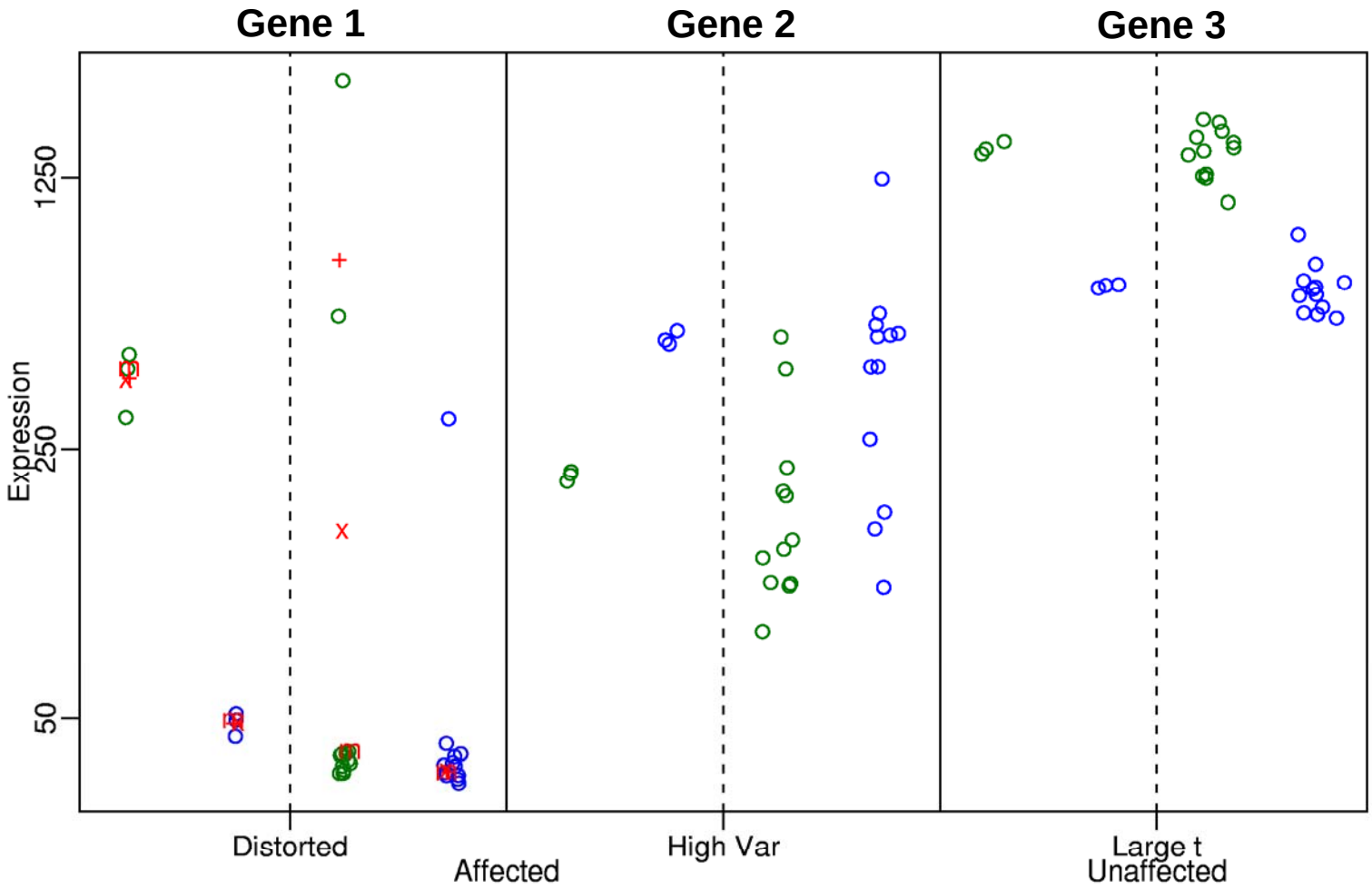


MA plot of average log ratios much better





Should we consider gene-specific variance?



Left panels are 3 replicate measurements from pooled RNA (12 mice).
Right panels have measurements for each of the 12. The colors represent two populations

- Importance of replicates

  Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations

  MT Lee et al, PNAS 2000 | vol. 97 | no. 18 | 9834-9839

- proposed scores range from:
  - ad-hoc summaries of fold-change
  - variants 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

Back to basics: one gene at a time

Observations: $X_1,\ldots,X_M \qquad Y_1,\ldots,Y_N$

Averages: $\bar{X} = \frac{1}{M}\sum_{i=1}^{M} X_i \qquad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

SD² or variances:

$$s_X^2 = \frac{1}{M-1}\sum_{i=1}^{M}(X_i - \bar{X})^2 \quad s_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

What you have learned in intro biostat courses

- If there is one gene of interest,
  - $X \sim N(\mu_1, \sigma^2)$ $Y \sim N(\mu_2, \sigma^2)$: Student t-test

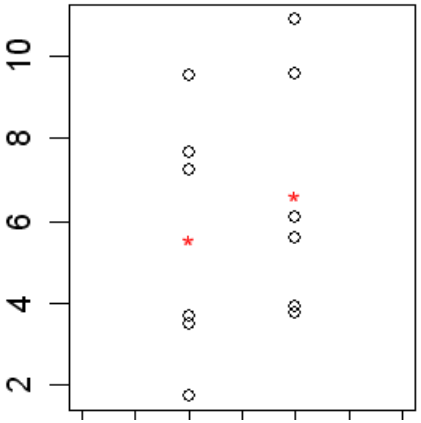
$$t = \frac{X_1 - X_2}{S_{X_1X_2} \cdot \sqrt{\frac{2}{n}}} \qquad S_{X_1X_2} = \sqrt{\frac{1}{2}(S_{X_1}^2 + S_{X_2}^2)}$$

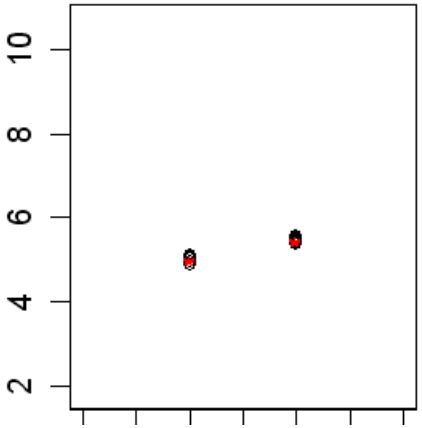
  - $X \sim N(\mu_1, \sigma_1^2)$ $Y \sim N(\mu_2, \sigma_2^2)$: Welch t-test

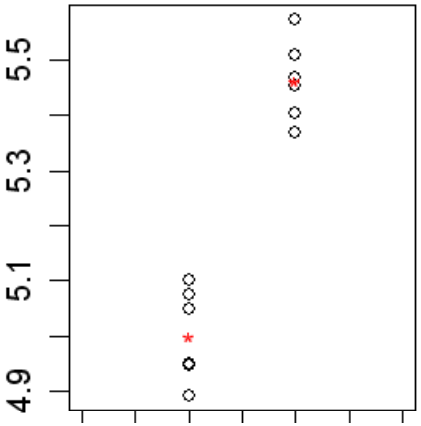
  - Not sure if its normal?
  - What happens to the t-statistic?

  - Non-parametric: Wilcoxon test

Back to Basics







Mean difference 1.1

t = -0.6269, df = 9.998,
p-value = 0.5448

Mean difference 0.46

t = -10.2293, df = 9.808,
p-value = 1.508e-06

Back to Basics

t - statistic:

T follows a t-distribution with degree of freedom v if

$$T = \frac{Z}{\sqrt{V/\nu}} = \frac{Z\sigma}{\sqrt{\sigma^2 V/\nu}}$$

Where Z is N(0,1), V ~ X^2(df=v)

These are satisfied if data are Normal and equal variance.

Back to Basics

t - statistic squared if N=M:

$$N \times \frac{(\overline{Y}_g - \overline{X}_g)^2}{s_{Yg}^2 + s_{Xg}^2}$$

Back to Basics

- If N and M are big then the t-statistic is normally distributed with mean 0 and and SD of 1
- If the observed data is normally distributed then the t-statistic follow a t distribution regardless of N,M
- Regardless, the square of the t-test is proportional to the ratio of across group variance to within group variance

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
  - For n paired comparisons,

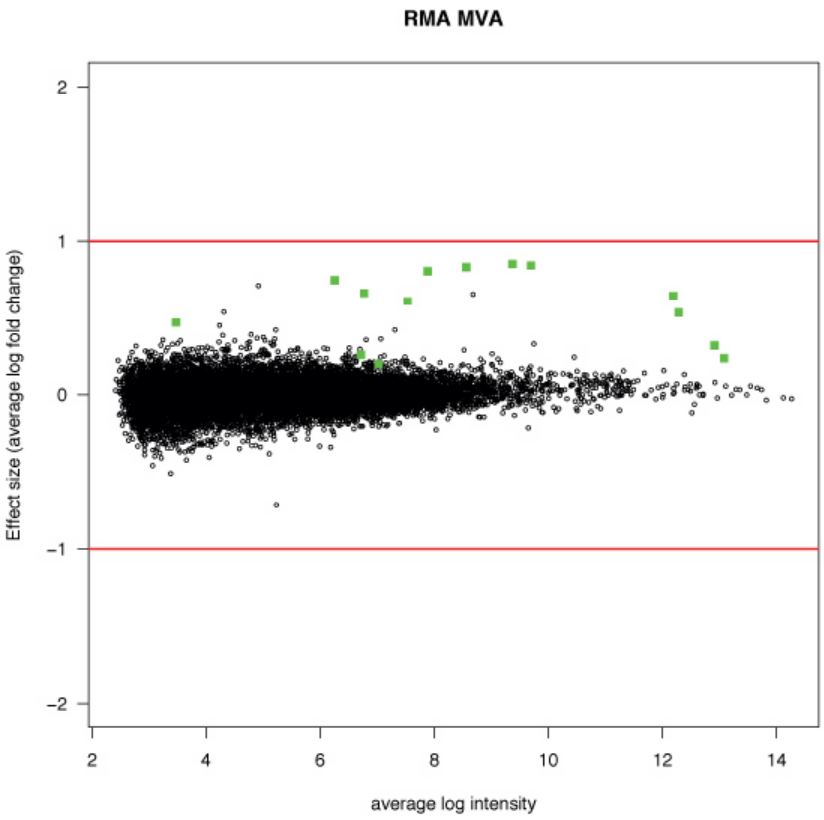
$$M = \frac{1}{n}\sum_{i=1}^{n} \log(y_i / x_i) = \frac{1}{n}\sum_{i=1}^{n} \log(y_i) - \frac{1}{n}\sum_{i=1}^{n} \log(x_i)$$

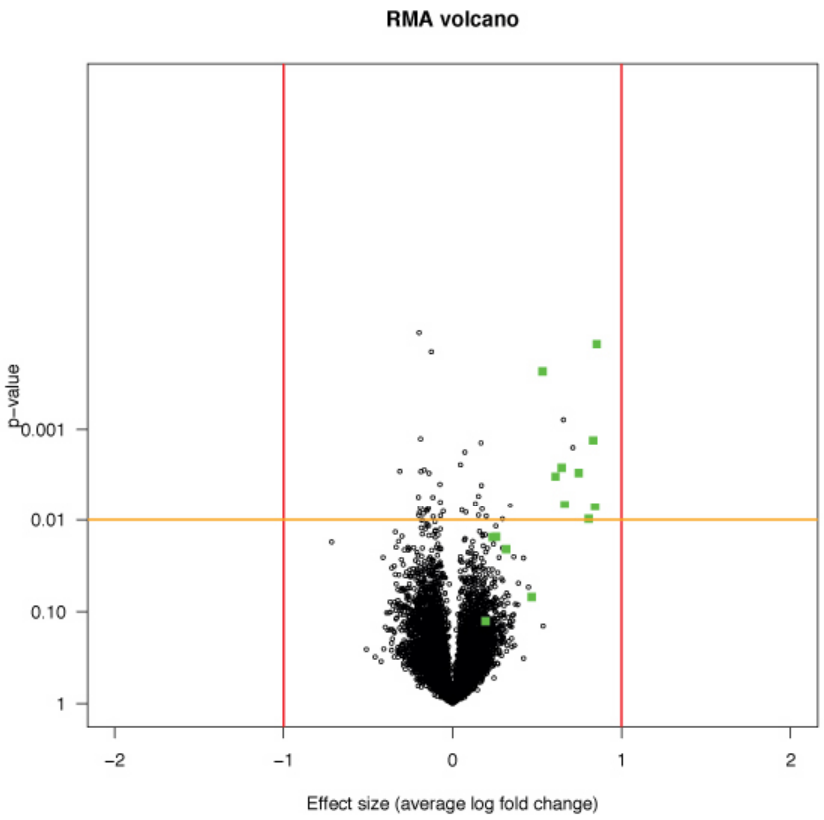
  - For unpaired comparisons,

$$M = \frac{1}{n}\sum_{i=1}^{n} \log(y_i) - \frac{1}{m}\sum_{i=1}^{m} \log(x_i)$$

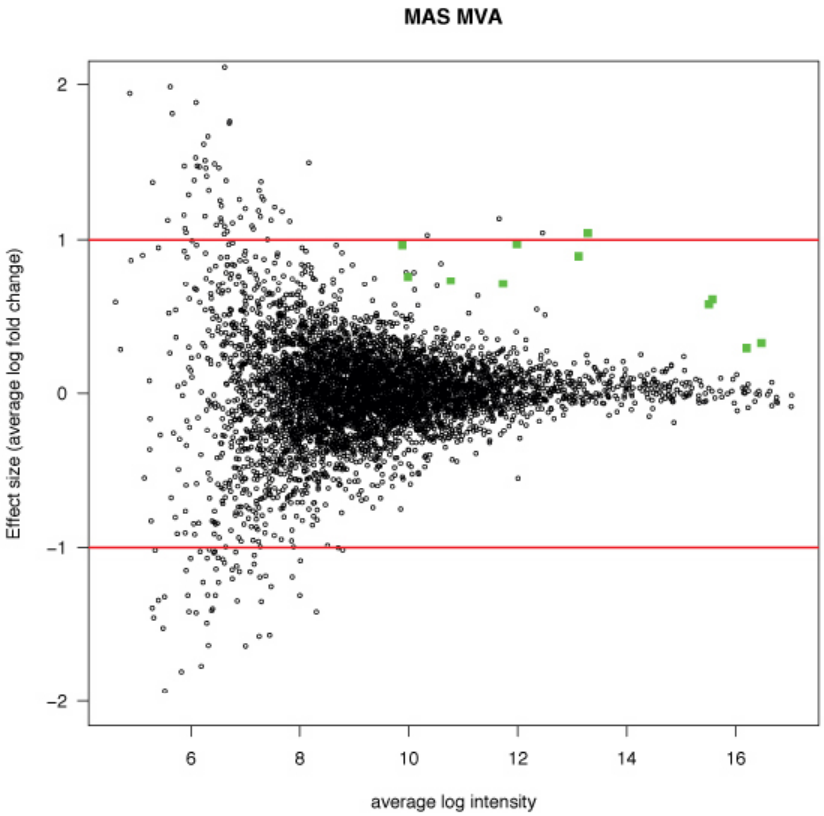
The volcano plot shows, for a particular test, negative log p-value against the effect size (M)
How do we get p-values? Are they really p-values?

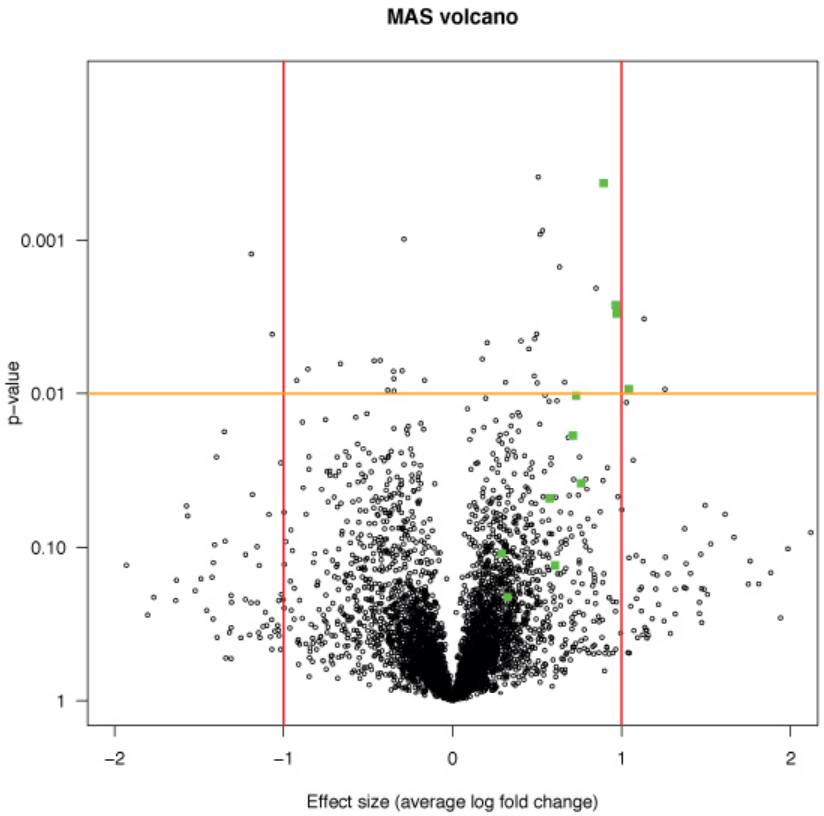
With RMA t-test is not powerful





If you insist on using MAS 5.0
it really helps





Estimating the variance

- If different genes (or probes) have different variation then it is not a good idea to simply use average log ratios
- Under a random model we need to estimate the SD
- The t-test divides by SD
- But with few replicates, estimates of SD are not stable
- This explains why t-test is not powerful
- There are many proposals for estimating variation
- Many "*borrow strength*" across genes

Some Examples of Tests

Notation:

- T is average log expression in Tx
- C is average log expression in Control
- S is SD

- Note taking log before average is important
- Tests:
  - Average log fold-change: (T-C)
  - t-statistic: (T-C) / S
  - SAM shrunken t-statistic: (T-C) / ($S + S_0$)
  - Bayesian posteriors: (T-C) / √(S^2+K^2)
  - Wilcoxon Rank test
  - Ad-hoc pairwise comparison No formula

More on some of these later

Sensitivity and specificity

- Sensitivity

  P(test + | D)

  For Pvalue, P(p< cutoff | D)

- Specificity

  P(test - | nonD)=1-P(test+ | nonD)

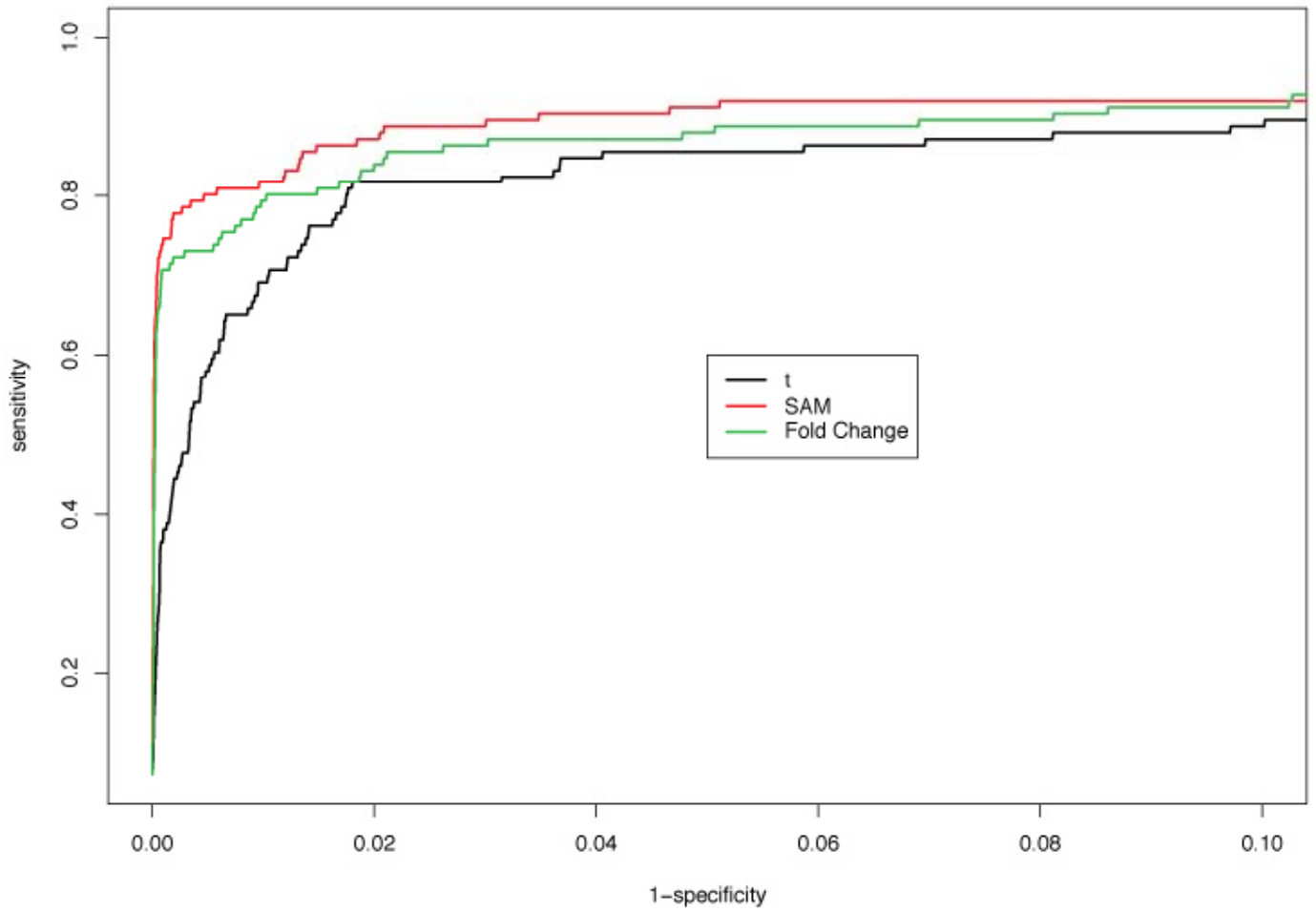
  P(p>=cutoff | non D)

		Truth	
		D	Non-D
Test	+		α
	-	β	

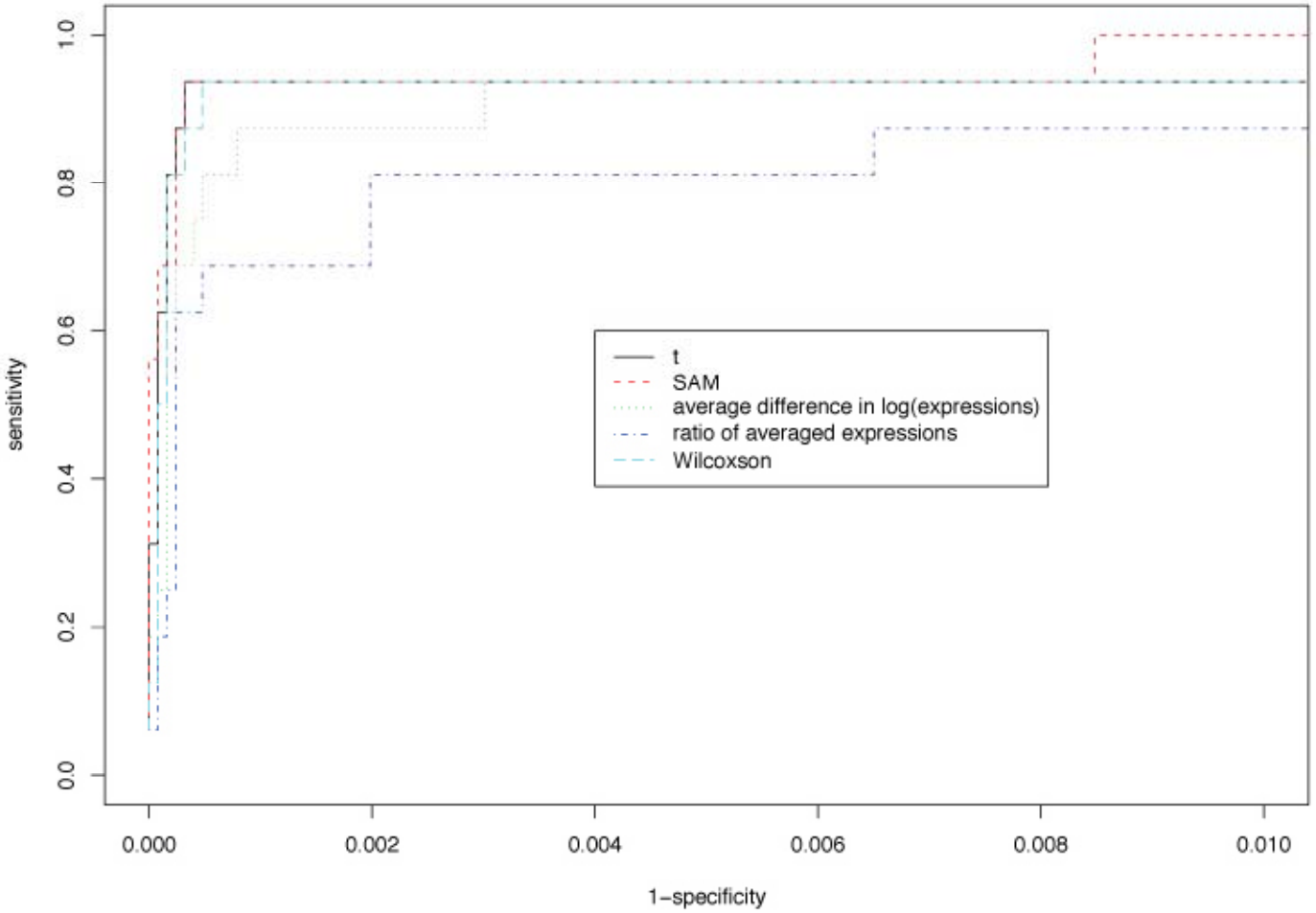
ROC curves

- ROC = Receiver Operator Characteristic
- To compare tests it is important to look at both specificity and sensitivity
- For every cut-off value there will be some true positives and some false positives
- We can make a curve that plots true positives versus false positives as we move the cut-off

Does it make a difference (N=3)?



Does it make a difference (N=12)?



Hypothesis testing

- Once you have a score for each gene, how do you decide on a cut-off?
  p-values are popular. Are they appropriate?
- Test for each gene null hypothesis: no differential expression.
- Notice that if you have look at 10,000 genes for which the null is true you expect to see 500 attain p-values of 0.05
- This is called the multiple comparison problem. Statisticians fight about it. But not about the above.
- Main message: p-values can’t be interpreted in the usual way

What do we do?

- Bonferoni correction (too conservative)
  - Adjusted p: $p^* = p \times n$ or $\alpha^* = \alpha/n$
- Give list of genes and report:
  - Family-Wise Error Rate: probability of including at least one non-differentially expressed gene
  - False discovery rate (FDR): expected proportion of Type I errors among the rejected hypotheses
  - pFDR: Expected proportion of false discoveries among the genes in your list conditioning on at least one gene is included in the differential list.
- Bayesian inference
- Forget about inference: use EDA and rank
- We will talk about this in another lecture

Moderated t-test:

Significance analysis of microarrays (SAM)

- A clever adaptation of the t-ratio to borrow information across genes
- A user-friendly EXCEL add-on to compute FDR's for sets of genes screened using the statistic above
- In Bioconductor, two packages are available: *siggenes* and *samr*
- *http://www-stat.stanford.edu/~tibs/SAM/*

SAM-statistic

- For gene i

$$d_i = \frac{\bar{y}_i - \bar{x}_i}{s_i + s_0}$$

$\bar{y}_i =$ mean of Irradiated samples

$\bar{x}_i =$ mean of Unirradiated samples

$s_i =$ Standard deviation of residuals for gene *i* *aassuming same variance*

$s_0 =$ Exchangeability factor estimated using all genes

The exchangeability factor

- Chosen to make signal-to-noise ratios independent of signal
- Computation
  - Let s^{α} be the α percentile of the s_i values. Let

$$d_i^{\alpha} = r_i / (s_i + s^{\alpha})$$

  - Compute the 100 quantiles of the s_i values, denoted by

$$q_1 < q_2 < \cdots < q_{100}$$

  - For $\alpha \in (0, 0.05, 0.10, \ldots, 1.0)$
    - Compute $v_j = mad(d_i^{\alpha} \mid s_i \in [q_j, q_{j+1})), j = 1, 2, \ldots, 99,$ where mad is the median absolute deviation from the median, divided by 0.64
    - Compute $cv(\alpha)$ = coefficient of variation of the v_j
    - Choose $\hat{\alpha} = \arg\min[cv(\alpha)].$ and $\hat{s}_0 = s^{\hat{\alpha}}.$

The reference distribution

- Order the values of d_i (could be any stat)

$$d_{(1)} \le d_{(2)} \le \cdots \le d_{(p)}$$

- Permute the treatment labels, and compute a new set of ordered values

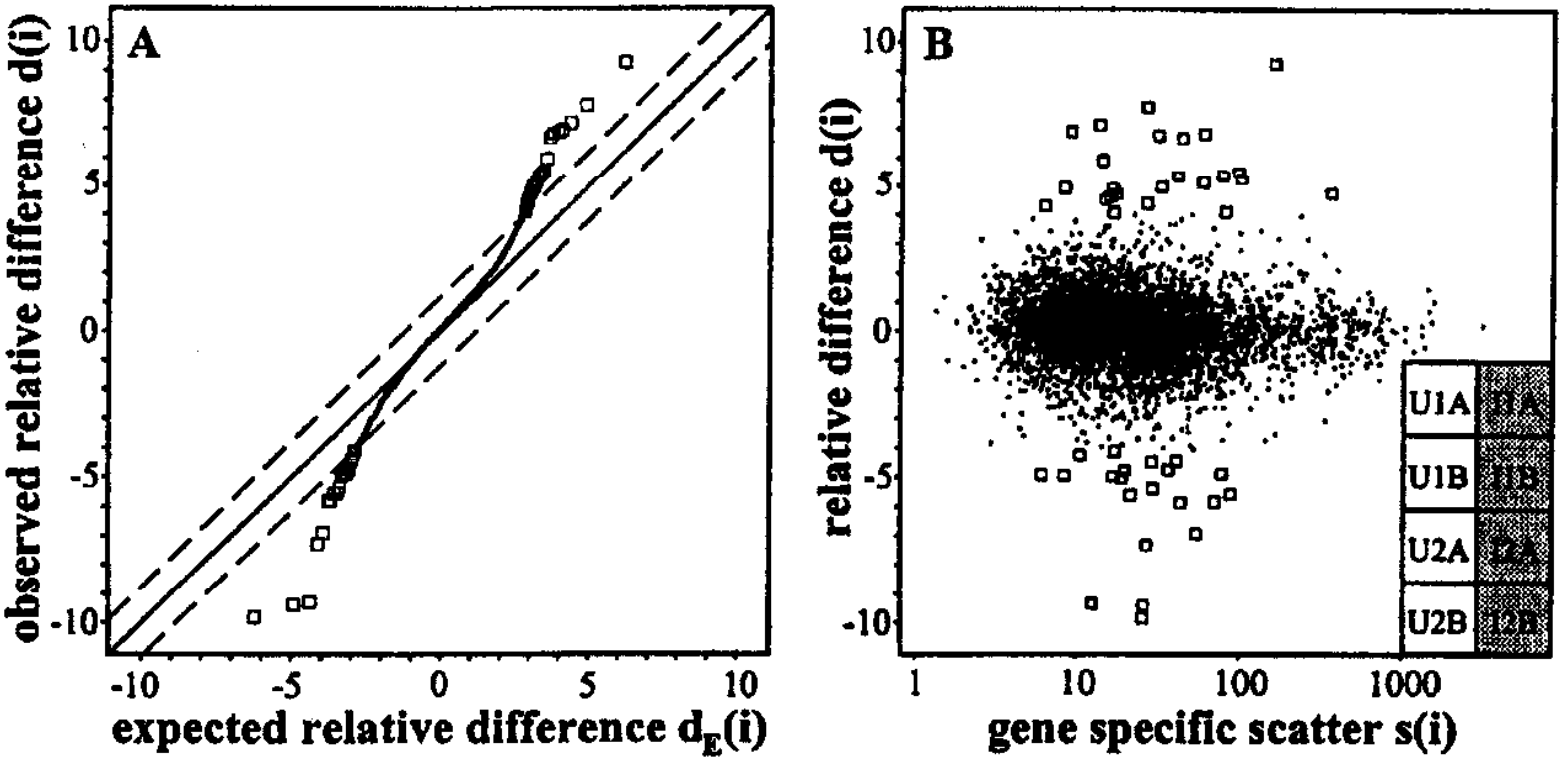
$$d^*_{(1)} \le d^*_{(2)} \le \cdots \le d^*_{(p)}$$

- Repeat step 2 for, say, 100 permutations:

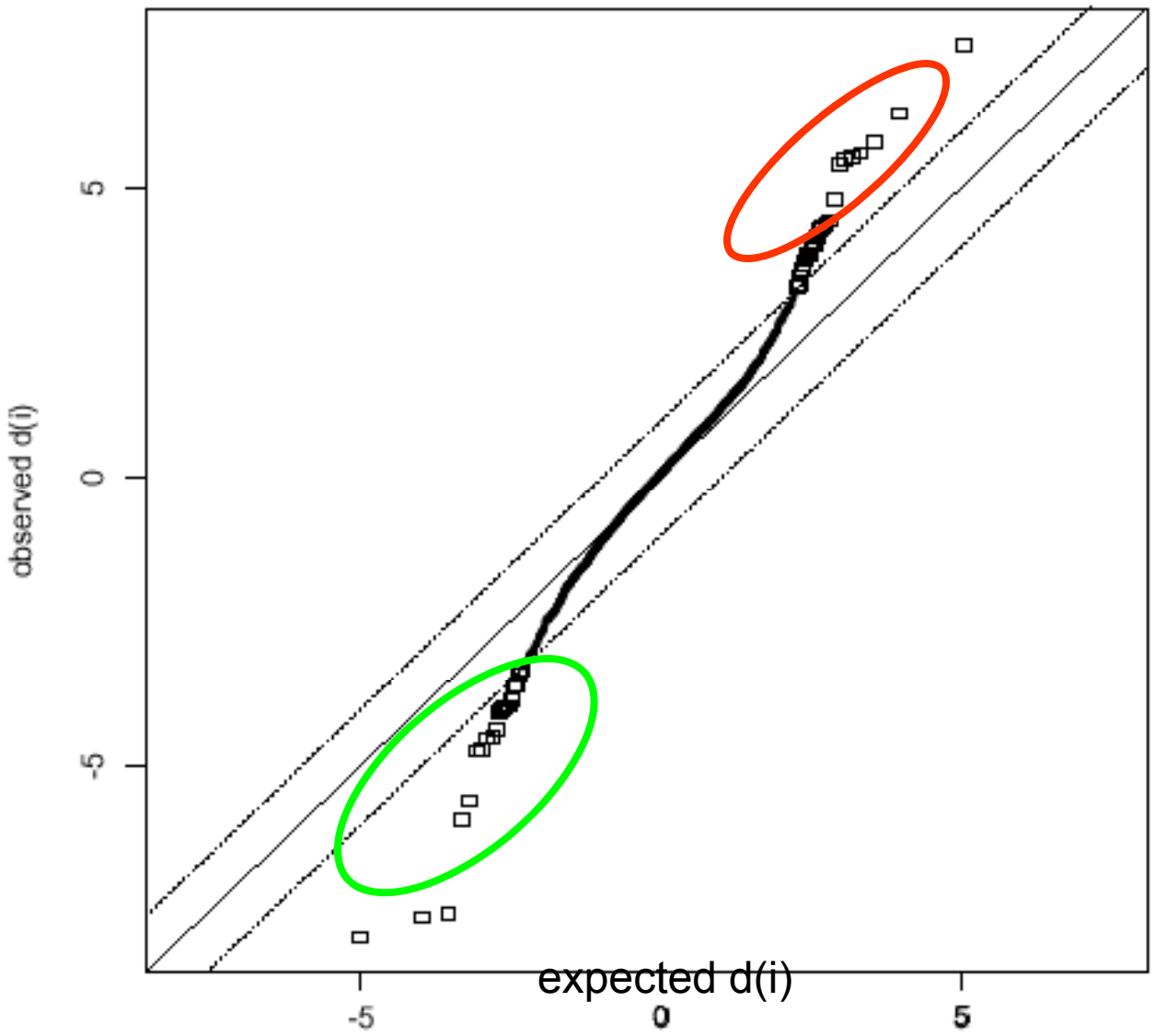
$$\begin{array}{ccccc} d^{*1}_{(1)} \le & d^{*1}_{(2)} \le & \cdots & \le d^{*1}_{(p)} \\ d^{*2}_{(1)} \le & d^{*2}_{(2)} \le & \cdots & \le d^{*2}_{(p)} \\ & \vdots & \vdots & \\ d^{*100}_{(1)} \le & d^{*100}_{(2)} \le & \cdots & \le d^{*100}_{(p)} \end{array}$$

- From these, compute the average largest, average second largest etc.

Selected genes



SAM plot



Delta	Ave # falsely significant	# called significant	Estimated False discovery rate
0.3	75.1	294	0.255
0.4	33.6	196	0.171
0.5	19.8	160	0.123
0.7	10.1	94	0.107
1.0	4.0	46	0.086

Delta is the half-width of the bar around the 45-degree line.

Other Empirical Bayes Method: Limma

- Design matrix X

$$E(\mathbf{y}_g) = X\boldsymbol{\alpha}_g \qquad \text{var}(\mathbf{y}_g) = W_g\sigma_g^2$$

$$\boldsymbol{\beta}_g = C^T\boldsymbol{\alpha}_g.$$

$$\text{var}(\hat{\boldsymbol{\alpha}}_g) = V_g s_g^2$$

$$\text{var}(\hat{\boldsymbol{\beta}}_g) = C^T V_g C s_g^2.$$

Let v_{gj} be the jth diagonal element of $C^T V_g C$

$$\hat{\beta}_{gj} \mid \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj}\sigma_g^2)$$

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}}$$

$$s_g^2 \mid \sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g}^2$$

Hierarchical model

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$

$$P(\beta_{gj} \neq 0) = p_j.$$

$$\beta_{gj} \,|\, \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$$

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}$$

(where $\tilde{s}_g^{-2}$ is the posterior mean of σ_g^{-2}

Inference

- Posterior odds

$$O_{gj} = \frac{p(\beta_{gj} \neq 0 \,|\, \tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0 \,|\, \tilde{t}_{gj}, s_g^2)} = \frac{p(\beta_{gj} \neq 0, \tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0, \tilde{t}_{gj}, s_g^2)} = \frac{p_j}{1 - p_j} \frac{p(\tilde{t}_{gj} \,|\, \beta_{gj} \neq 0)}{p(\tilde{t}_{gj} \,|\, \beta_{gj} = 0)}$$

- OR just moderated t- or F-statistic
  - Replace the sample variance with the shrinkage estimate

Estimation of hyper parameters

- Prior degree of freedom and prior variance scale d0 and s0
- s_g^2 follows a scaled F-distribution

$$s^2 \sim s_0^2 F_{d,d_0}$$

$$\tilde{t} \,|\, \beta = 0 \sim t_{d_0+d}$$

$z_g = \log s_g^2$ is approximately normal with finite moments

$$E(z_g) = \log s_0^2 + \psi(d_g/2) - \psi(d_0/2) + \log(d_0/d_g)$$

$$\text{var}(z_g) = \psi'(d_g/2) + \psi'(d_0/2)$$

where $\psi()$ and $\psi'()$ are the digamma and trigamma functions respectively.

Use method of moments: solve for $d_0,\ s_0^2$

References

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(1), 3.

Lönnstedt, I., and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

"Significance analysis of microarrays applied to the ionizing radiation response" (ps file). (pdf version). PNAS 2001 98: 5116-5121,