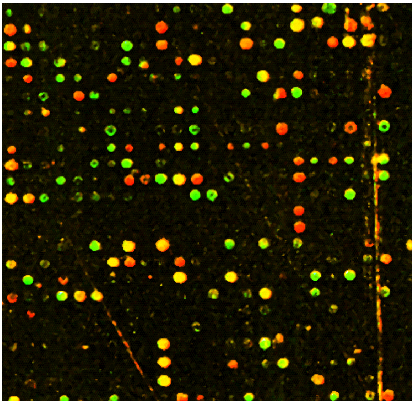


# Statistical Analysis of cDNA microarrays I

**Terry Speed**

WEHI Bioinformatics & UCB Statistics

# Image



## Spot – output

Spot_ID	Gmean	GIQR	Rmean	RIQR	Gvalley	Rvalley
1	13177	0.62	10327	0.47	463	1282
2	8000	0.14	5070	0.13	463	1185
3	3138	0.27	3211	0.19	481	1213
4	8433	0.25	8635	0.40	481	1265
5	4118	0.35	3776	0.22	463	1265
6	15473	0.90	5603	0.65	456	1231
7	2399	0.21	2995	0.16	481	1253
8	1245	0.27	2107	0.11	483	1265
9	35959	0.81	31807	0.73	483	1247

Gmorph	Rmorph	area	circularity	Gsn	Rsn	lratio
260	1208	31	.88	5.62	2.92	.50
261	1174	32	.83	4.88	1.73	.99
261	1146	56	.64	3.46	0.84	.48
260	1208	11	1.38	4.96	2.62	.13
262	1163	16	1.03	3.88	1.14	.57
262	1144	40	.74	5.86	1.95	1.77
250	1185	30	.94	3.03	0.57	.26
262	1157	20	1.12	1.91	-0.42	.12
262	1164	61	.85	7.09	4.68	.22

# Statistical problems involving microarray data

<http://www.stat.Berkeley.EDU/users/terry/zarray/Html/list.html>

Image analysis ones already mentioned.

6. Use of housekeeping genes.

7. Quality, II: Spots.

8. Normalization within an experiment:

- \* when few genes change.

- \* when many genes change.

- \* use of red-green and green-red pairs.

9. Normalization between experiments: location and scale effects.

10. Noise.

11. Variability.

12. Bias : Use of "truth" .

13. Quality, III: Ratios.
14. Who is up/down?
15. P-values.
16. Planning of experiments:
  1. design.
  2. sample sizes.
17. Analysis of factorial experiments.
18. Discrimination and allocation.
19. Clustering:
  1. of samples.
  2. of genes.
20. Time course experiments.
21. Gene networks.
22. Special problems.
  - \* Mixture analyses.
  - \* Pooled cDNA vs amplified DNA.

We start with 9 and go on to 14.

## Experiments

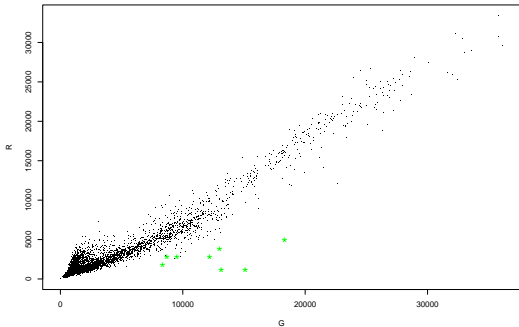
**Goal.** Identify genes with altered expression in the livers of two lines of mice with very low HDL cholesterol levels compared to inbred control mice.

Two experiments: (1) Apo AI knock-out mouse model and (2) SR-BI transgenic mouse model. In each experiment:

- 8 treatment (trt) mice (apo AI ko or SR-BI tg) and 8 control (ctl) mice (C57Bl /6 or FVB).
- 16 hybridizations: mRNA from each of the 16 mice is labeled with Cy5, pooled mRNA from control mice is labeled with Cy3.
- Probes: ~ 6,000 cDNAs, including 200 related to lipid metabolism.

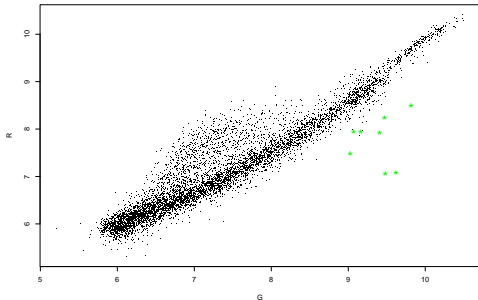
# Single Slide plots

R vs G (intensity scale)



# Plotting transformed intensities

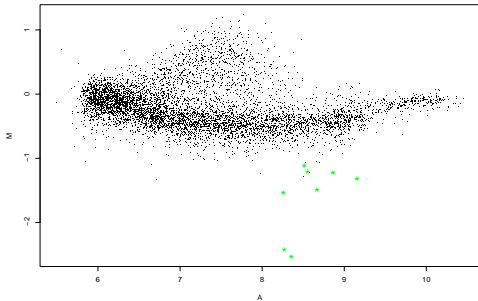
Log R vs Log G (base 2)



More informative.



$$M = \log_2 R/G \text{ vs } A = \frac{1}{2}(\log_2 G + \log_2 R)$$



More informative still.

## Within-slide normalization

Normalization balances red and green intensities. Imbalance may be caused by differential incorporation of dyes, different amounts of the two species of RNA, differential scanning, etc. In practice, we usually need to bump up the red intensity a bit to balance the green.

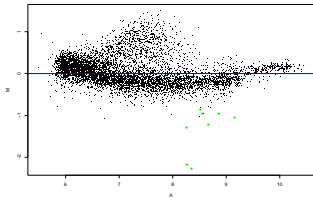
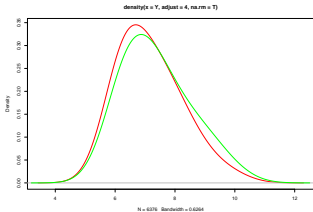
$$\log R/G \rightarrow \log R/G + c = \log kR/G$$

A standard choice is to arrange that normalized log ratios have zero mean or median. Our preference is to do this in an A-dependent way: we choose  $c = c(A)$  using lowess.

A proof that this is better than using a constant is currently lacking. It certainly changes things, and we are pretty sure it helps.

# Normalization - Median

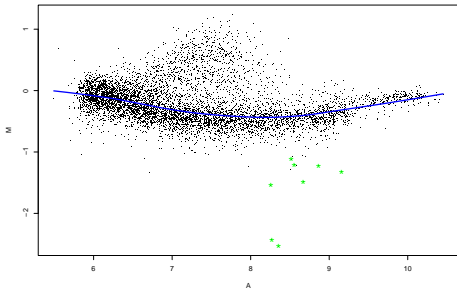
Assumption: Changes roughly symmetric



First panel: smoothed densities of  $\log_2 G$  and  $\log_2 R$ .  
Second panel:  $M$  vs  $A$  plot with median  $M$  put to 0.

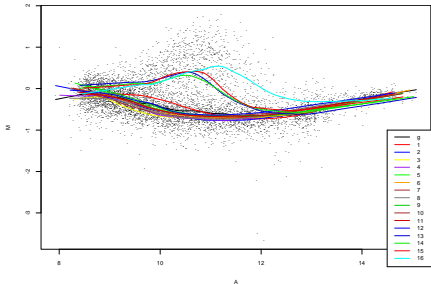
## Normalization - Lowess

Global lowess. Assumption: changes roughly symmetric at all intensities

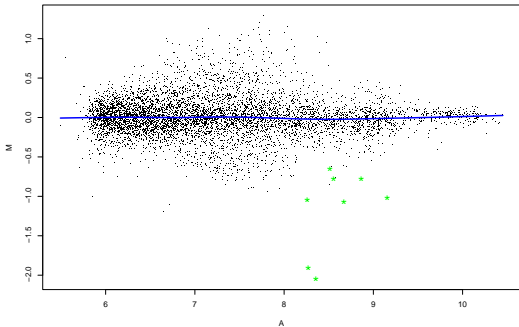


# Normalization - Print Tip Lowess

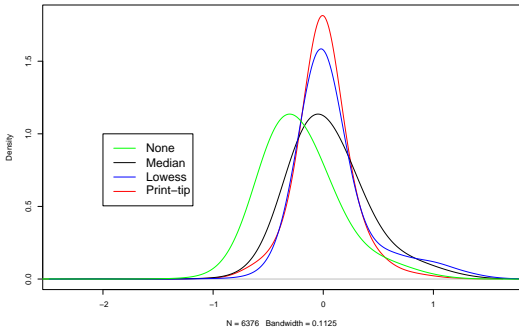
Print-tip lowess normalization. Need stronger assumption.



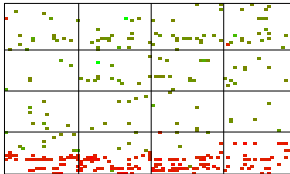
# M vs A – after print-tip normalization



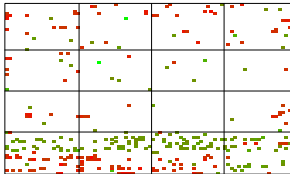
# Effects of normalization I



## Effects of normalization II



Before normalization

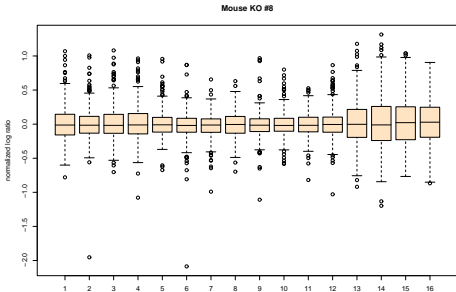


After normalization



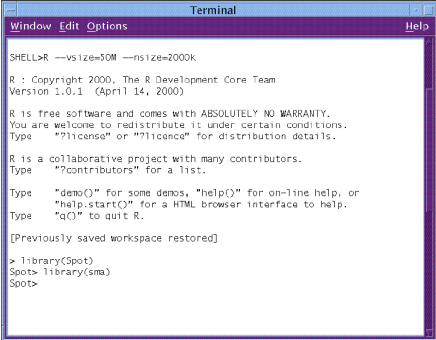
# Within print-tip box plots of print-tip normalized M

Print-tip scale effects remain: last four more variable.



# Statistical Software

Spplus or R (freeware)

A screenshot of a terminal window titled "Terminal". The window has a menu bar with "Window", "Edit", "Options", and "Help". The terminal text shows the R startup process: SHELL>R --vsize=50M --nsize=2000K, followed by copyright information for R 1.0.1 (April 14, 2000). It then displays the R license and contributor information. The user enters commands to load the Spot and sma libraries.

```
SHELL>R --vsize=50M --nsize=2000K

R : Copyright 2000, The R Development Core Team
Version 1.0.1 (April 14, 2000)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type  "?license" or "?licence" for distribution details.

R is a collaborative project with many contributors.
Type  "?contributors" for a list.

Type  "demo()" for some demos, "help()" for on-line help, or
      "help.start()" for a HTML browser interface to help.
Type  "q()" to quit R.

[Previously saved workspace restored]

> library(Spot)
Spot> library(sma)
Spot>
```

# Which genes have changed expression levels?

## Single-slide methods

*Existing 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 and calling a gene differentially expressed if its  $(R, G)$  falls outside the region between the two curves.

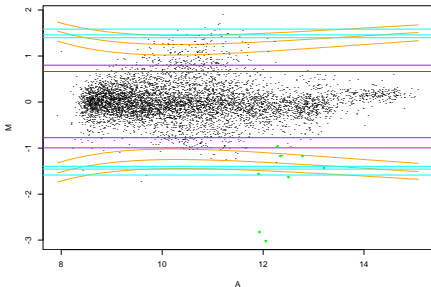
We probably do not know enough about the systematic and random effects within a microarray experiment to justify strong modeling assumptions or theory-based predictions.  
Conclusion  $n = 1$  slide may not be enough.

## Single-slide methods, cont'd

Existing methods differ in the distributional assumptions they make regarding  $(R, G)$ .

1. Chen *et al.* Each  $(R, G)$  is assumed to be normally and independently distributed with constant CV. Decision based on  $R/G$  only. (purple)
2. Newton *et al.* Gamma-Gamma-Bernoulli hierarchical model for each  $(R, G)$ . (yellow)
3. Roberts *et al.* Each  $(R, G)$  is assumed to be normally and independently distributed with variance depending linearly on the mean.
4. Sapir & Churchill. Each  $\log R/G$  is assumed to be distributed according to a mixture of normal and uniform distributions. Decision based on  $R/G$  only. (turquoise)

# Which genes have changed expression levels?



## A Bayesian approach for replicated slides

*Motivation* To combine information in  $M$  values, taking into account their variability within and between slides. Our Bayesian approach is meant to be a vehicle for doing this, but we do not take the probabilities implicit in it seriously. Mainly, we want to avoid being misled by means involving outliers, while taking care not to be too impressed with unusually small variances.

Sampling model Let  $M_{ij} = \log(R_{ij}/G_{ij})$  be the log ratio of our green ( $G_{ij}$ ) and red ( $R_{ij}$ ) intensities for a gene,  $i = 1 \dots m$  refers to the slides and  $j = 1 \dots n$  to replicates within slides. We suppose

$$EM_{ij} = \begin{cases} \mu_i & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

We also suppose

$$M_{ij} \sim pN(\mu_i, \sigma^2) + (1 - p)N(0, \sigma^2)$$

but with variances  $\sigma_b^2$  *between* slides and  $\sigma_w^2$  *within* slides,  $\sigma^2 = \sigma_b^2 + \sigma_w^2$ . If  $i = j = 2$

$$\text{Cov}(M_{ij}) = \begin{pmatrix} \sigma_w^2 + \sigma_b^2 & \sigma_w^2 & 0 & 0 \\ \sigma_w^2 & \sigma_w^2 + \sigma_b^2 & 0 & 0 \\ 0 & 0 & \sigma_w^2 + \sigma_b^2 & \sigma_w^2 \\ 0 & 0 & \sigma_w^2 & \sigma_w^2 + \sigma_b^2 \end{pmatrix}.$$

## Priors for $\mu$ and $\tau$

For an integer  $\nu$  and  $a > 0, c > 0$

$$a\nu\tau \sim \chi_\nu^2$$
$$N\left(0, (c\tau)^{-1}\right)^I \cdot \delta(0)^{1-I} \quad \text{where} \quad \delta(0) = \begin{cases} 1 & \text{if } \mu = 0 \\ 0 & \text{if } \mu \neq 0 \end{cases} .$$



## Technical point

It will be noted that we have chosen the standard conjugate prior for our normal means and variances. This was with the aim of getting a simple formula to use for the posterior odds ratio, see below. However, there appear to be no closed form expressions (simple or otherwise) when there are two components of variance, even in this balanced case. MCMC methods might work here, but we have 5,000 small samples, and have yet to try it out.

For these technical reasons, we therefore suppose that

$$\sigma_w^2 = k_1 \sigma_b^2, \text{ with } k_1 \text{ known.}$$

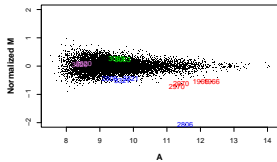
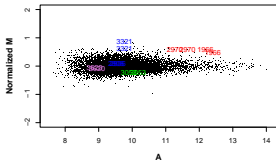
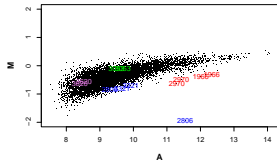
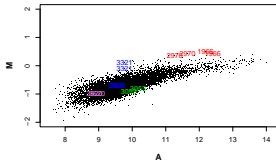
In fact we use the parametrization

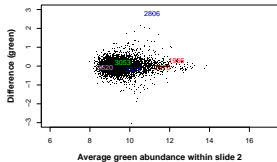
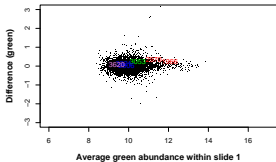
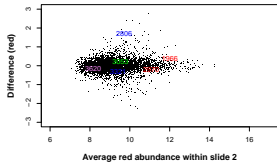
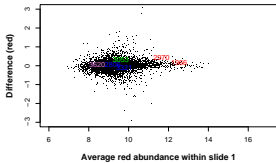
$$\tau^{-1} = \sigma_w^2 + n\sigma_b^2 \text{ and } \sigma_w^2 = k\tau.$$

## The log odds ratio R

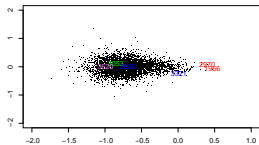
We are interested in whether or not the gene is differentially expressed, i.e. whether  $\mu \neq 0$  or  $\mu = 0$ , so we calculate the log posterior odds ratio

$$\begin{aligned} R &= \log \frac{\Pr(\mu \neq 0 | (M))}{\Pr(\mu = 0 | (M))} \\ &= \frac{p}{1-p} \left( \frac{c}{c+mn} \right)^{1/2} \left( \frac{va + mnM_{..}^2 + SSB + kSSW}{va + mnM_{..}^2 + SSB + kSSW - \frac{(mnM_{..})^2}{c+mn}} \right)^{\frac{\nu+mn}{2}-1} \end{aligned}$$



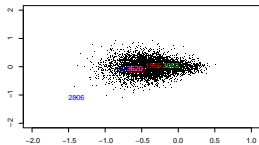


Difference between M's within slide 1



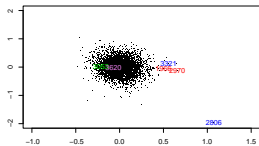
Average M within slide 1

Difference between M's within slide 2



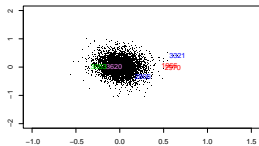
Average M within slide 2

Difference between M's between slides



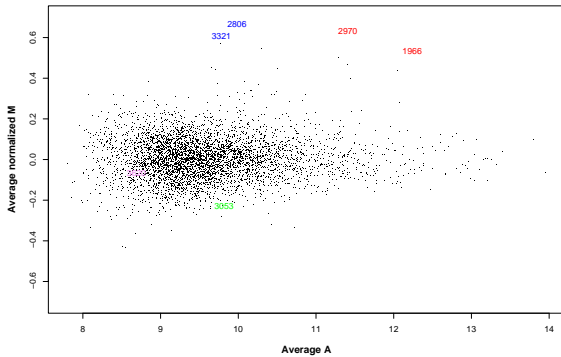
Average M between slides

Difference between M's between slides

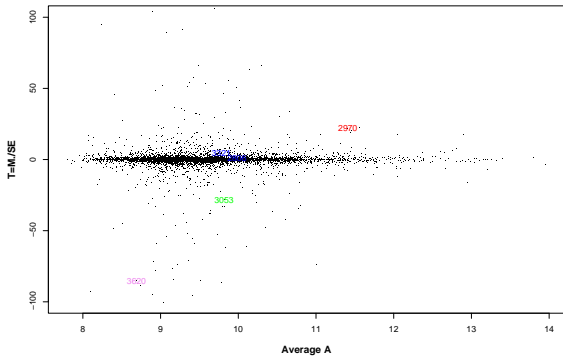


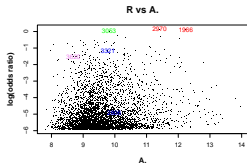
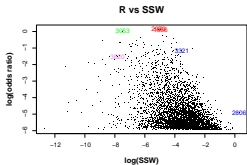
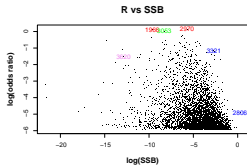
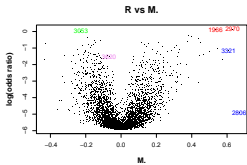
Average M' between slides

Normalized M. vs A.



T vs A.







# ACKNOWLEDGEMENTS

Yee-Hwa Yang  
Ingrid Lonnstedt  
Natalie Roberts  
Sandrine Dudoit  
Suzie Grant