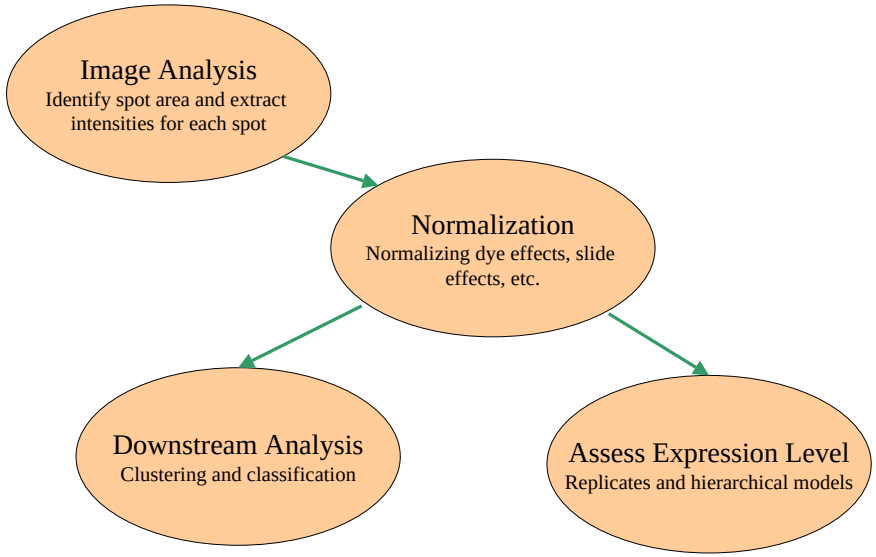
Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects

G.C. Tseng, M. Oh, L. Rohlin, J.C. Liao and W.H. Wong

Nucleic Acids Research Jun/15/2001

Statistical Issues in cDNA Microarray Analysis



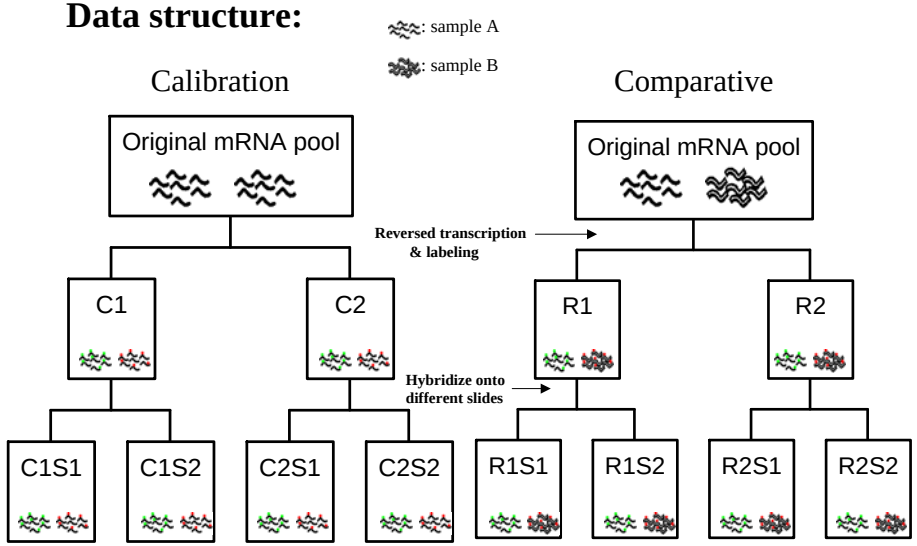
~ Outline ~

1. Normalization
- Quality filtering
- Rank-Invariant selection for non-differentially expressed genes
- Fit normalization curve

2. Assessment of gene expression level
- Hierarchical model

3. Program with interface on the IE browser
- Algorithms and methods developed in the paper

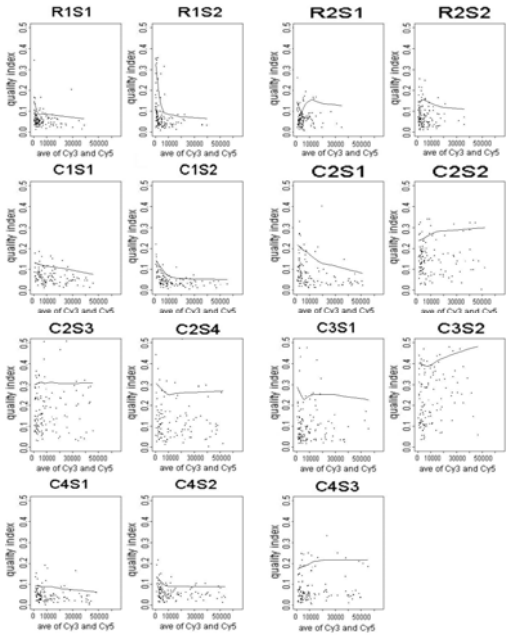
Data structure:



Samples: E. coli grown in acetate or glucose
125-gene project: each gene is spotted four times
4129-gene project: each gene is singly spotted

		slides in the experiment	samples in Cy3	samples in Cy5	Denhardt's solution
125	Calib	C1S1~C1S2	Acetate	Acetate	all slides
		C2S1~C2S4	Glucose	Glucose	None
		C3S1~C3S2	Glucose	Glucose	C3S1
		C4S1~C4S3	Glucose	Glucose	C4S1~C4S2
	Comp	R1S1~R1S2	Acetate	Glucose	all slides
		R2S1~R2S2	Acetate	Glucose	all slides
4129	Calib	C1S1~C1S2	Acetate	Acetate	all slides
		C2S1~C2S2	Glucose	Glucose	all slides
	Comp	R1S1~R1S2	Acetate	Glucose	all slides
		R2S1~R2S2	Acetate	Glucose	all slides

Quality filtering (125)

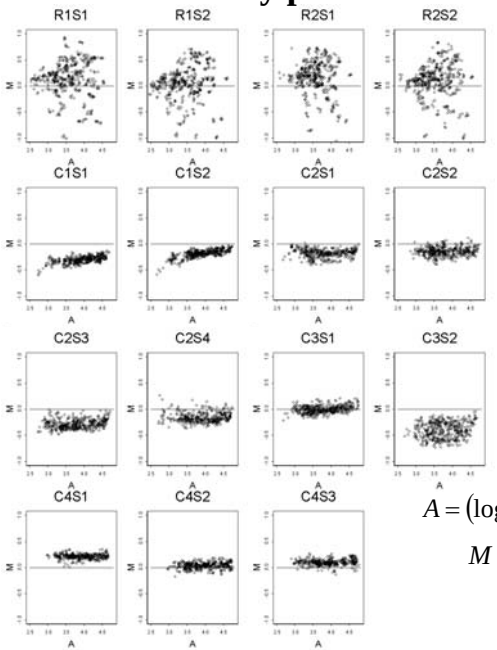


Quality Index:

$$m_i = Cy5_i / Cy3_i$$

$$CV = std(m_i)/mean(m_i)$$

Intensity plot

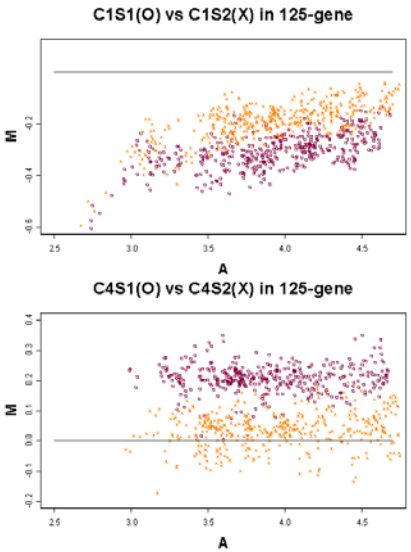


$$A = (\log(Cy5) + \log(Cy3))/2$$

$$M = \log(Cy5/Cy3)$$

Normalization: (why normalization needed?)

Calibration: apply the same samples on both dyes (E. Coli grown in glucose). Purple and orange represent two replicate slides.



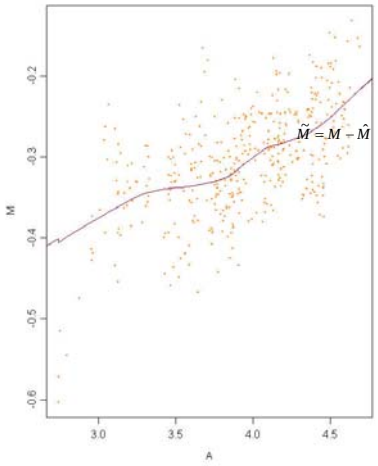
$$A = (\log(Cy5) + \log(Cy3))/2$$

$$M = \log(Cy5/Cy3)$$

Normalization in calibration experiment:

Fit $\hat{M} = \hat{f}(A)$

by 'Lowess' function in S-Plus



Normalized Log ratio:

$$\tilde{M} = M - \hat{M}$$

Normalization in comparative experiment:

Current popular methods:

- House-keeping genes : Select a set of non-differentially expressed genes according to experiences. Then use these genes to normalize.
- Constant normalization factor :
  - Use mean or median of each dye to normalize.
  - ANOVA model (Churchill's group)
- Average-intensity-dependent normalization:
  - Robust nonlinear regression(Lowess) applied on whole genome. (Speed's group)
  - Select invariant genes computationally (rank-invariant method). Then apply Lowess. (Wong's group)

Normalization:

Rank-invariant method (Schadt et al. 2001, Tseng et al. 2001**):**

$$G = \{g : abs(rank(Cy3_g) - rank(Cy5_g)) < 5\}$$

Iterative selection :

$$S_0 = \{ g : |Rank(Cy5_g) - Rank(Cy3_g)| < p*G \;\&\; l < Rank(((Cy5_g + Cy3_g)/2) < G - l \}$$

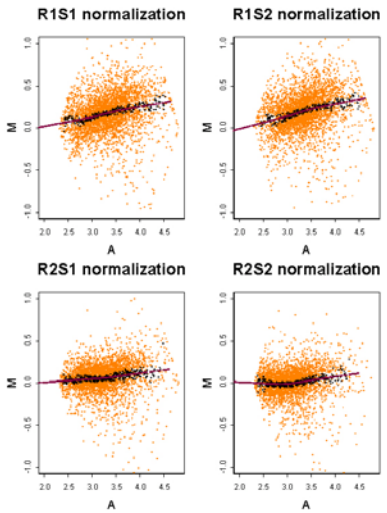
$$S_i = \{ g : g \in S_{i-1} \;\&\; |Rank_{g \in S_{i-1}}(Cy5_g) - Rank_{g \in S_{i-1}}(Cy3_g)| < p*|S_{i-1}| \}$$

Idea:

- If a particular gene is up- or down- regulated, then its Cy5 rank among whole genome will significantly different from Cy3 rank.
- Iterative selection helps to select a more conserved invariant set when number of genes is large.

Normalization: (Wong group)

Data: E. Coli. Chip, ~4000 genes, from Liao lab.



Blue points are invariant genes selected by rank-invariant method. Red curves are estimated by Lowess and extrapolation.

Hierarchical Model

A version of empirical Bayes

$$x_{seg} \sim N(\mu_{eg}, \tau_g^2) \qquad \mu_{eg} \sim N(\theta_g, \sigma_g^2)$$

$$\tau_g^2 \sim k\tilde{\tau}_g^2 / \chi_k^2 \qquad \sigma_g^2 \sim h\tilde{\sigma}_g^2 / \chi_h^2 \qquad p(\theta_g) \propto 1$$

x : normalized logratios, $\log(Cy5/Cy3)$.

e : experiment, s : slide, g : gene.

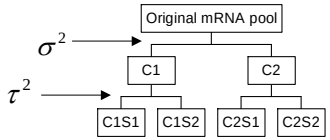
θ : underlying true expression level.

σ^2 : experimental(cultural) variation.

τ^2 : slide variation. h and k are adjustable parameters.

Note only x s are observed.

e.g. ((0.75, 0.67), (0.45, 0.51))



How to specify the prior?

Empirical Bayes:

Use empirical data to help specify hyperparameters $(\tilde{\tau}_g^2, \tilde{\sigma}_g^2)$.

A common version of EB is usually achieved by integrating out intermediate parameters and maximize the resulting marginal likelihood.

$$\max_{\tilde{\tau}^2, \tilde{\sigma}^2} p(\tilde{\tau}^2, \tilde{\sigma}^2 \mid \underset{\sim}{X}) = \int p(\tilde{\tau}^2, \tilde{\sigma}^2, \mu, \theta, \tau^2, \sigma^2 \mid \underset{\sim}{X})\, d\mu d\theta d\tau^2 d\sigma^2$$

It is hard to implement in this three - layer hierarchical model.

Another version of EB:

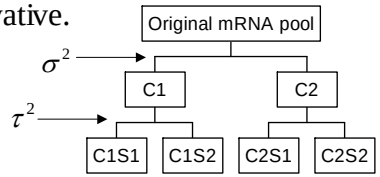
Estimate parameters in prior from empirical data.

$$\tilde{\tau}^2 = \sum_{g,s,e} (y_{gse} - y_{g\bullet e})^2 / G(S-1)E \quad \text{(between slides variation)}$$

$$\tilde{\sigma}^2 = \sum_{g,e} (y_{g\bullet e} - y_{g\bullet\bullet})^2 / G(E-1) \quad \text{(between experiment variation)}$$

Note :

- Since there are thousands of genes, the common problem of reusing data and getting over - confident prior in EB is alleviated.
- The estimation of $\tilde{\sigma}^2$ is more conservative.



MCMC for hierarchical model:

1. Compute $(\mu_{eg})^{(0)} = x_{\bullet eg}$

2. $$\sigma_g^2 \mid \mu_{eg} \sim \frac{\sum_e (\mu_{eg} - \mu_{\bullet g})^2 + h\tilde{\sigma}^2}{\chi^2_{E+h-1}}$$

3. $$\theta_g \mid \mu_{eg}, \sigma_g^2 \sim N\left(\mu_{\bullet g}, \frac{\sigma_g^2}{E}\right)$$

4. $$\tau_g^2 \mid \mu_{eg}, x_{seg} \sim \frac{\sum_{j=1}^{E}\sum_{s=1}^{s_e} (x_{seg} - \mu_{eg})^2 + k\tilde{\tau}_g^2}{\chi^2_{s_1+\ldots+s_E+k}}$$

5. $$\mu_{eg} \mid x_{seg}, \tau_g^2, \theta_g, \sigma_g^2 \sim N\left(\frac{s_e x_{\bullet eg}\sigma_g^2 + \tau_g^2\theta_g}{s_e\sigma_g^2 + \tau_g^2}, \frac{\tau_g^2\sigma_g^2}{s_e\sigma_g^2 + \tau_g^2}\right)$$

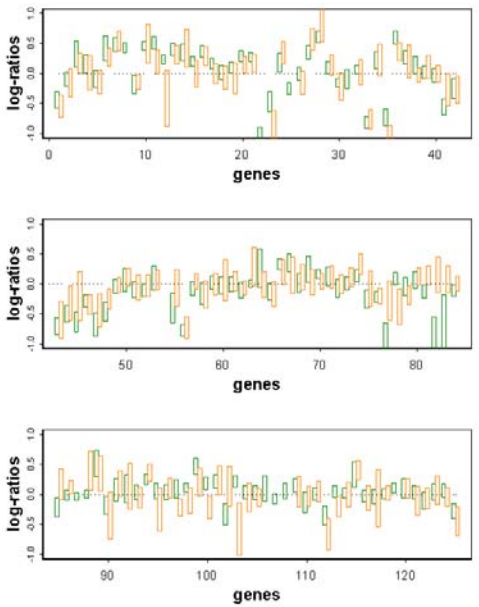
Assumptions in the model:

1. Uniform prior on theta.
2. Common slide variation in different experiments.
3. Normal distribution.

Computation concern when relaxing the model:

Now the simulation converges in 4000 times in contrast to 100,000 times in common hierarchical model. The fast convergence is due to the simple normal model and the conjugate prior.

Speed: implement ~4000 genes in around 20 minutes in C but up to hours in R.



Result:

The 95% probability intervals in the 125-gene and 4129-gene projects correspond to 0.73 to 1.4 and 0.61 to 1.6 fold change respectively .

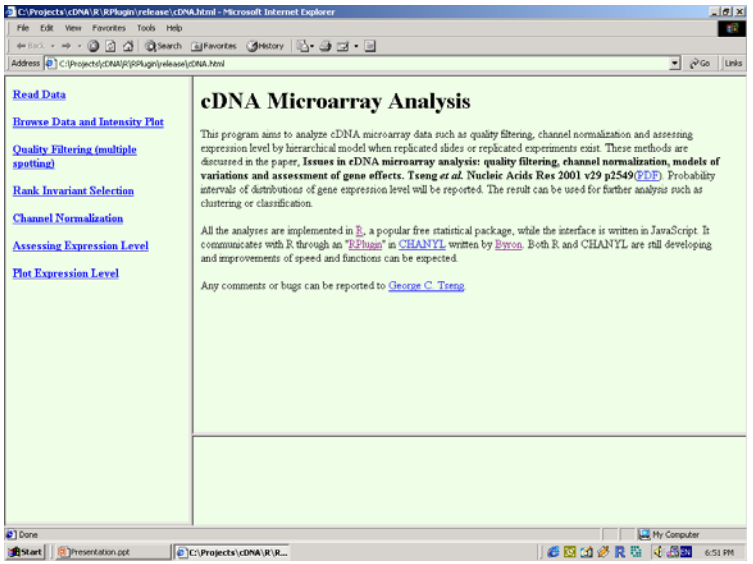
In the few strong disagreement genes of two projects, we found that most of them are grouped in some pathways, such as metE, metB, aroL, aroG, and aroF. This suggests that these strong disagreements may reflect real biological variation between the cultures used in the two different projects.

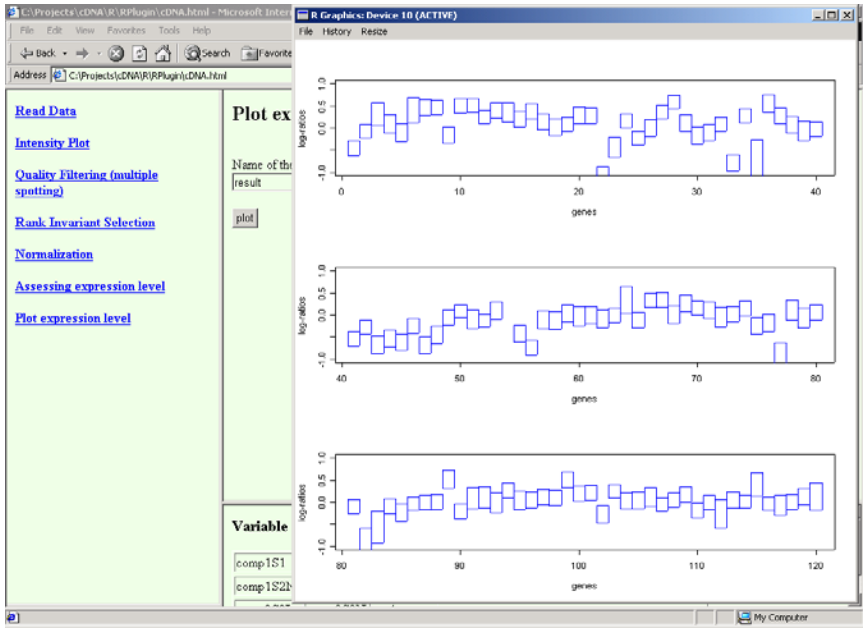
Multiple comparison:

We have not discussed how to account for multiple comparisons, i.e. selecting apparently differentially expressed genes from the large number of genes in the genome.

Download page

http://www.biostat.harvard.edu/~ctseng/download.html





Other issues in cDNA microarray:

1. Different choices of reference sample:

a) Normal patient or time 0 sample in time course study
b) Pooled samples
c) Embryonic cells
d) Commercial kit

2. Experimental design issues

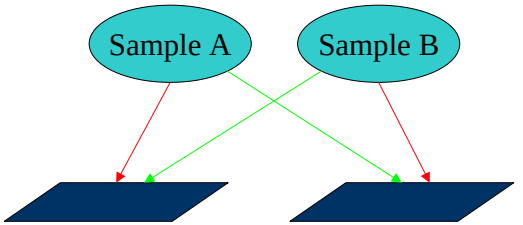
(i) Calibration:

- Use the same sample on both dyes for hybridization.
- Calibration experiments help to validate experiment quality and gene-specific variability.

(ii) Replicates: (replicate spots, slides)

- Multiple-spotting helps to identify local contaminated spots but will reduce number of genes in the study.
- Multi-stage strategy: Use single-spotting to include as many genes as possible for pilot study. Identify a subset of interesting genes and then use multiple-spotting.
- Replicate slides help to verify reproducibility on the slide level.

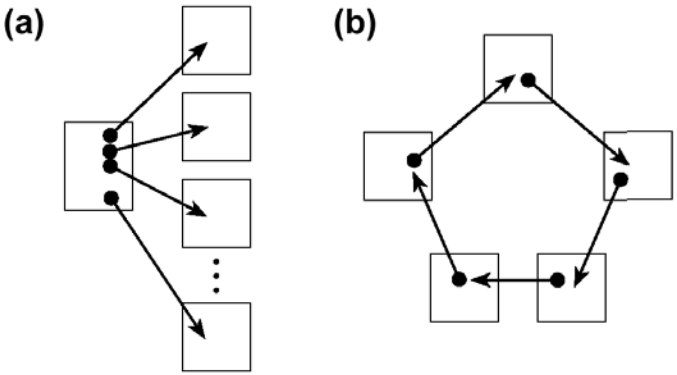
(iii) Reverse labelling:

- 

- Advantage:
  - Cancel out linear normalization scaling and simplifies the analysis. However, the linear assumption is often not true.
  - Help to cancel out gene-label interactions if it exists.

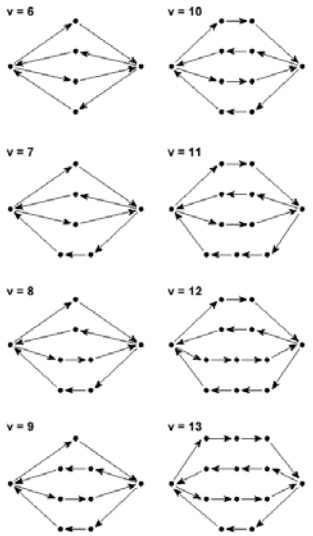
(iv) Design issues:

(a) Reference design
(b) Loop design
(c) Balance design



(c)

v samples with *v+2* experiments



v samples with *2v* experiments