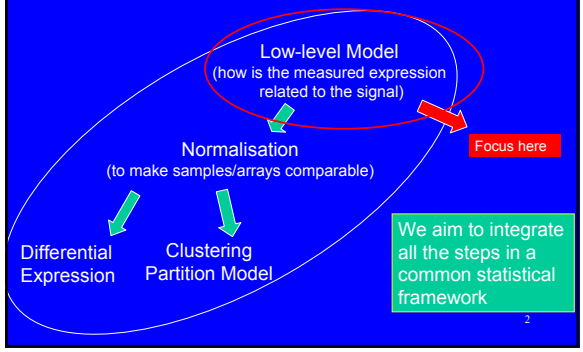
Bayesian Hierarchical models for analysing Affymetrix gene expression arrays using probe level data

Anne-Mette K. Hein

Department of Epidemiology and Public Health
Imperial College, London

Sylvia Richardson, Imperial College, St. Mary's
Helen Causton, Imperial College, Hammersmith

Gene expression analysis is a multi-step process



Overview:

- Model for extracting measure of gene expression from a single array
  - explore the model with respect to 'true signal' and cross hybridisation estimation (Data: GeneLogic varying concentration series)
- Extended model to encompass situations with different conditions and/or replicates
  - explore the model with respect to differential expression (Data: GeneLogic spike-in series)

Data and aim:

Data: Affymetrix chip:

- each gene g is represented by a probe set, consisting of a number of probe pairs (reporters) j
- Probe pair: consists of two probes: perfect match (PM) and mismatch (MM)

→ Aim: Formulate a model to combine PM and MM values into an expression value for the gene

- Base the model on biological assumptions
- Combine good features of Li and Wong (dChip) and RMA (Robust Multichip Analysis, Irizarry et al)

in a flexible Bayesian framework that will allow to integrate further components of the experimental design

Biological assumptions

- The intensity for the PM measurement for probe (reporter) j and gene g is due to binding
  - → of labelled fragments that perfectly match the oligos in the spot
    The true Signal S_{gj}
  - → of labelled fragments that do not perfectly match these oligos
    The cross-hybridisation H_{gj}
- The intensity of the corresponding MM measurement is caused
  - → by a binding fraction Φ of the true signal S_{gj}
  - → by cross-hybridisation H_{gj}

Model assumptions

- There is a background noise that affects PM and MM measurements
  - → additive error term for PM and MM, same variance
    (This explains why we sometimes observe PM < MM)
- S_{gj} and H_{gj} are positive and will be modelled on the log scale
- Will assume gene specific additive error on log scale signals
- Use hierarchical modelling of gene variances (to stabilise)

Model for one array:

g=1,…,G (thousands), j=1,…,J (11-20)

$PM_{gj} \mid m_{gj}^1 \sim N(m_{gj}^1, \eta^2)$

$MM_{gj} \mid m_{gj}^2 \sim N(m_{gj}^2, \eta^2)$

Background noise, additive

$m_{gj}^1 = S_{gj} + H_{gj}$ (signal + cross-hybridisation)

$m_{gj}^2 = \Phi S_{gj} + H_{gj}$

$\log(S_{gj} + 1) \sim TN(\mu_g, \tau_g^2)$

$\log(H_{gj} + 1) \sim N(\lambda, \sigma^2)$

$\log(1/\tau_g^2) \sim N(a, b)$

Gene expression index "Pools" information from the probes

Gene specific error term

Exchangeable model on the variance

Priors: "vague" $\Phi \sim B(1,1)$, $\eta^2 \sim \Gamma(0.001,0.001)$, cross: $\lambda \sim N(3,0.001)$, $\sigma^2 \sim \Gamma(0.001,0.001)$, $\mu_g \sim U(0,15)$, hyper prior on variances: $a \sim N(1,0.001)$, $b \sim \Gamma(0.01,0.01)$

Implementation

- in WinBugs
- Joint estimation of parameters in full Bayesian framework
- Base inference on posterior distribution of all unknown quantities, S_{gj}, H_{gj} , μ_g , …. (summarised by the posterior mean, variance, interval of credibility ….)

Data set 1: part of GeneLogic varying concentration series:

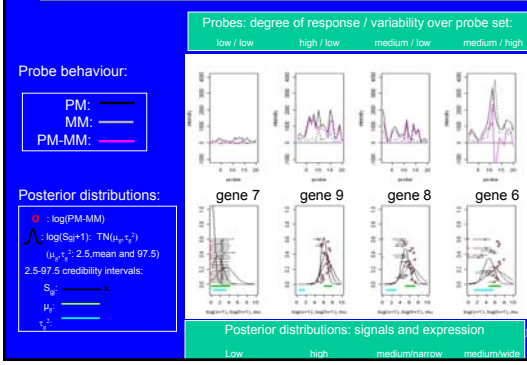
- cRNA from acute myeloid leukemia (AML) tumor cell line
- 10 samples
- In sample k: 11 genes spiked in, all at concentration c_k.
- Concentrations c_k:

sample:	1	2	3	4	5	6	7	8	9	10
concentration (pM):	0.0	1.0	2.0	5.0	12.5	25.0	50.0	75.0	100.0	150.0

- Each sample hybridised to an array
- Arrays analysed separately using single array model

Data from one array: All genes spiked in at conc. 2.0:

Shown: probe response and posterior distributions for 4 genes



Data from eight arrays analysed separately: A gene spiked in at 8 concentrations.

`true signal`/ expression index increases with concentration

As previous slide:

o : log(PM-MM)

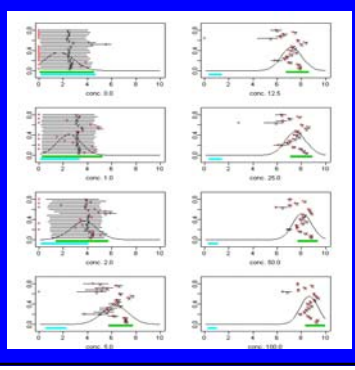
: distribution of log(Sgj+1): $TN(\mu_g,\tau_g^2)$ (using mean post estimates)

2.5-97.5 credibility interval for:

S_{gj}:

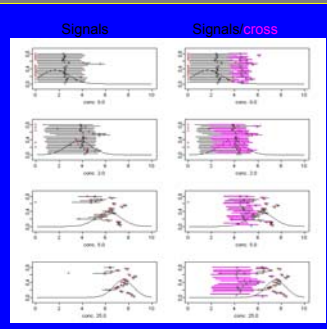
μ_g:

τ_g^2:



Data from 4 arrays analysed separately:

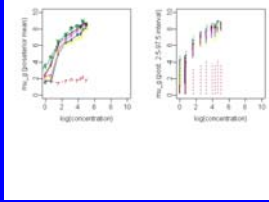
Gene 1 spiked in at different concentrations.



Data from 10 arrays analysed separately: 11 genes spiked in at different concentrations.

Expression index μ_g increases with concentration .....

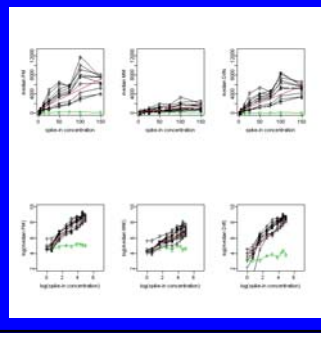
… except for gene 7 (which was the gene that had no detectable expression at conc. 2 (see slide 11.))



Data from 10 arrays analysed separately: 11 genes spiked in at different concentrations.

Median PM, MM and PM-MM values calculated for each of the 11 spike-in probe sets, for each array (that is, spike-in concentration)

Raw data for Gene 7 look suspicious…



Data set 2: Subset of GeneLogic spike-in: 6 arrays

Background mRNA: **AML tumor cell line**

Six samples/arrays: s_1 s_2 s_3 s_4 s_5 s_6

C_A C_A C_A C_B C_B C_B

Spiked in 11 genes at two sets of concentrations C_A and C_B:

Gene:	BioB-3	BioB-5	BioB-M	BioC-3	BioC-5	BioDn-3	CreX-3	CreX-5	DapX-3	DapX-5	DapX-M
C_A:	1	200	1	1	1	1	10	6.25	12.5	3.3	3
C_B:	50	1	37.5	4	37.5	33	1	1	1	1	1

➡ We analyse 1011 genes, including the 11 spiked-in genes

Subset of GeneLogic Spike-in data set: 6 arrays
Differential expression

The truth:

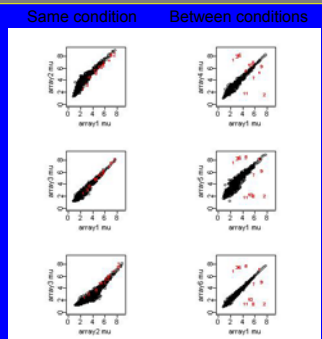
- Only the 11 spiked-in genes are differentially expressed
- some of the 1000 genes are expressed – but not differentially!
- We know the true ranks of the 11 spiked-ins

Gene	number	c1/c2 Ratios	Ranks of differential expression
BioB-3	1001	1:50	2
BioB-5	1002	200:1	1
BioB-M	1003	1:37.5	3/4
BioC-3	1004	1:4	9
BioC-5	1005	1:37.5	3/4
BioDn-3	1006	1:33	5
CreX-3	1007	10:1	7
CreX-5	1008	6.25:1	8
DapX-3	1009	12.5:1	6
DapX-5	1010	3.33:1	10
DapX-M	1011	3:1	11

6 Arrays analysed independently: Plot of (posterior mean) expression measures μ_g. Spiked in genes shown in red.

Same condition Between conditions

- Expression measures for all genes on replicate arrays correlate
- Expression measures of spike-in genes differ between conditions – others correlate
- Spikes 4, 5, 8 and 10 on array 4 differ from 5 and 6



Model for multiple arrays:
g: genes, j: probes c: conditions, r: replicates

$$\begin{cases} PM_{gjcr} \mid m_{gjcr}^{1} \sim N(m_{gjcr}^{1}, \eta_{cr}^{2}) \\ MM_{gjcr} \mid m_{gjcr}^{2} \sim N(m_{gjcr}^{2}, \eta_{cr}^{2}) \end{cases}$$

$$\begin{cases} m_{gjcr}^{1} = S_{gjcr} + H_{gjcr} + b_{cr} \\ m_{gjcr}^{2} = \Phi S_{gjcr} + H_{gjcr} + b_{cr} \end{cases} \qquad (b_{1,1} = 0)$$

$$\begin{cases} \log(S_{gjcr} + 1) \sim TN(\mu_{gc}, \tau_{gc}^{2}) \\ \log(H_{gjcr}) \sim N(\lambda_{cr}, \sigma_{cr}^{2}) \\ \log(\tau_{gc}^{2}) \sim N(a, b) \end{cases}$$

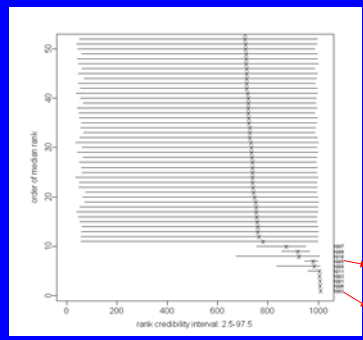
There is a choice of hierarchical structures !

6 Arrays analysed using multiple array model:
Comparison of ranks of differential expression: truth vs RMA, MBEI and Bayesian

Calculate gene expression measures (Bayesian: mean posterior estimates) for each gene under each condition, calculate fold changes between conditions and rank them (**Largest fold change: rank 1**)

Gene name:	BioB-3	BioB-5	BioB-M	BioC-3	BioC-5	BioDn-3	CreX-3	CreX-5	DapX-3	DapX-5	DapX-M
Ratios of conc:	1:50	200:1	1:37.5	1:4	1:37.5	1:33	10:1	6.25:1	12.5:1	3.3:1	3:1
Number:	1001	1002	1003	1004	1005	1006	1007	1008	1009	1010	1011
True ranks:	2	1	3/4	9	3/4	5	7	8	6	10	11
Ranks obtained											
RMA:	3	1	2	478	5	4	9	8	7	10	6
dChip:	4	1	2	65	7	6	13	232	11	860	10
Bayesian:	3	1	4	729	5	2	10	8	7	9	6

Distributions of ranks of differential expression: All spiked in genes except 1004 are in the top

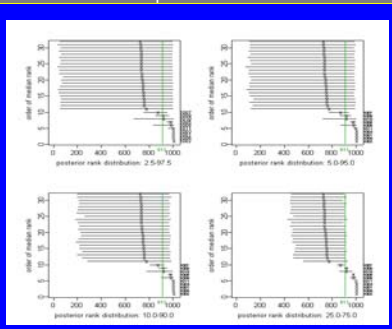


- Obtain distributions of the ranks of differential expression
  $rank_g$ (abs($\mu_g{}^1$- $\mu_g{}^2$))
- order genes by median rank
- Plot (in order of median ranks) median rank (x) and 2.5-97.5 credibility interval

Gene 1005 has 7th highest median rank (983): 2.5-97.5: 946-1001

Gene 1002 has highest median rank (1011): 2.5-97.5: 1010-1011

Distributions of ranks of differential expression: Different posterior intervals



Summary

- Developed a Bayesian hierarchical model for extracting gene expression measures
- Use PM and MM measurements – model 'true signal' and cross
- Include additive and multiplicative errors
- Extended model to multiple conditions/replicates
- Address differential expression within the same framework – errors of various steps propagated

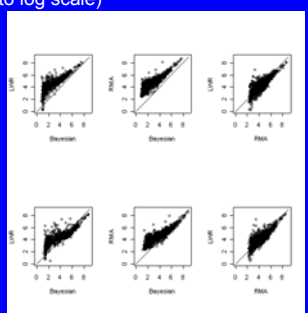
Future work:

- Posterior predictive model checks
- Extending/modifying model:
  - Φ GC content/trend specific
  - b-terms?
  - better normalization: array effects
  - explore other hierarchical structures
  - gene expression measure: mean, mode, median of TN or μ

Comparison of our gene index with those produced by Li and Wong (dChip) and RMA (Affy R package) (all transformed to log scale)

Array 1:

Array 2:



Alternative models for multiple arrays:
g: genes, j: probes c: conditions, r: replicates

$$PM_{gjcr} \mid m_{gjcr}{}^1 \sim N(m_{gjcr}{}^1, \eta_{cr}{}^2)$$
$$MM_{gjcr} \mid m_{gjcr}{}^2 \sim N(m_{gjcr}{}^2, \eta_{cr}{}^2)$$

Max Model:

$m_{gjcr}{}^1 = S_{gjcr} + H_{gjcr} + b_{cr}$
$m_{gjcr}{}^2 = \Phi S_{gjcr} + H_{gjcr} + b_{cr}$
$\log (S_{gjcr} + 1) \sim TN(\mu_{gc}, \tau_{gc}{}^2)$
$\log (H_{gjcr}) \sim N(\lambda_{cr}, \sigma_{cr}{}^2)$
$\log (\tau_{gc}{}^2) \sim N(a, b)$

Model 1:

$m_{gjcr}{}^1 = S_{gjc} + H_{gj} + b_{cr}$
$m_{gjcr}{}^2 = \Phi S_{gjc} + H_{gj} + b_{cr}$
$\log (S_{gjc} + 1) \sim TN(\mu_{gc}, \tau_{gc}{}^2)$
$\log (H_{gj}) \sim N(\lambda, \sigma^2)$
$\log (\tau_{gc}{}^2) \sim N(a, b)$

- TT (truncated t-distribtution)?
- Φ: GC content specific ?
  signal dependent (hybridisation/scanner saturation)?

Single array model analysis vs: max model and model 1 – density plots

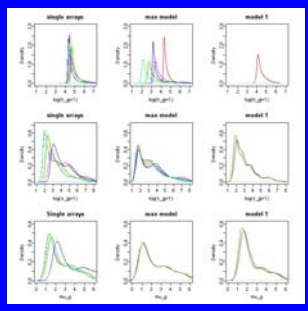
- single array analysis: 6 signal distributions (one per array), 6 cross distributions (one per array)
- max model: six cross distributions (one per array), two signal distributions (one per condition)
- additive mean background effects:

	$b_{1,1}$	$b_{1,2}$	$b_{1,3}$	$b_{2,1}$	$b_{2,2}$	$b_{2,3}$	phi
Max:	0	117	115	111	123	78	0.17
M1:	0	269	83	111	195	10	0.23

eta2 (*e+04)

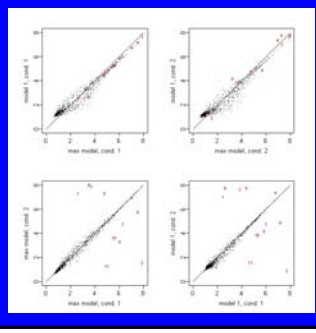
Max:	5.3	1.7	0.4	0.2	2.2	5.9
M1:	4.0	0.03	0.7	0.2	0.03	4.8

- Max model assumes 6 different cross hybridisation distributions: under this model the b-terms are not very different
- model1 assumes a common cross distribution for all six arrays: under this model the b-terms are very different



25

max model vs model 1: mu scatter plots



26

Subset of GeneLogic Spike-in data set: 6 arrays

The truth:

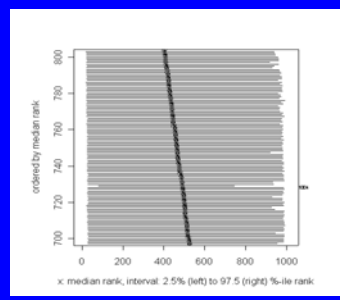
- Only the 11 spiked-in genes are differentially expressed
- some of the 1000 genes are expressed – but not differentially!
- We know the true ranks of the 11 spiked-ins

		S_1-S_3	S_4-S_6		
Gene	number	conc.	conc.	Ratios	Ranks#
BioB-3	1001	0.5	25.0	1:50	2
BioB-5	1002	100.0	0.5	200:1	1
BioB-M	1003	1.0	37.5	1:37.5	3/4
BioC-3	1004	25.0	100.0	1:4	9
BioC-5	1005	2.0	75.0	1:37.5	3/4
BioDn-3	1006	1.5	50.0	1:33	5
CreX-3	1007	50.0	5.0	10:1	7
CreX-5	1008	12.5	2.0	6.25:1	8
DapX-3	1009	37.5	3.0	12.5:1	6
DapX-5	1010	5.0	1.5	3.33:1	10
DapX-M	1011	3.0	1.0	3:1	11

27

#: of differential expression

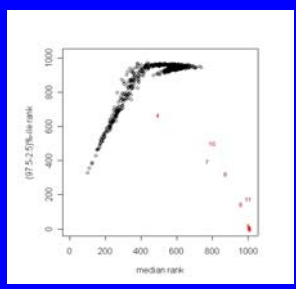
Model1: (not max Model!)



28

Plot of median rank versus length of 95% credibility interval for the ranks (Model 1!)

All spiked-in genes have high median ranks and small interval except gene 1004



29

Expression measures vs. spike-in concentration: 10 arrays

- **Bayesian**: obtained using single array model independently on 10 arrays (data not pre-normalised)

Fitted line (black):

mu = 2.89 + 1.10 * log(c)

- **MBEI**: invariant set normalised, PM-MM model. Obtained from dChip package).

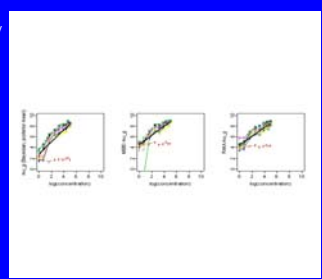
Fitted line (black):

mu = 4.55 + 0.87 * log(c)

- **RMA**: quantile normalised. Obtained using function 'rma' in AffyR package

Fitted line (black):

mu = 4.49 + 0.80 * log(c)



30

Max model vs. Model 1: mu_g