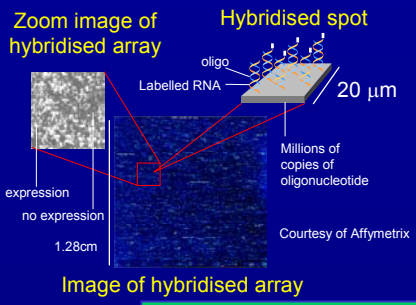
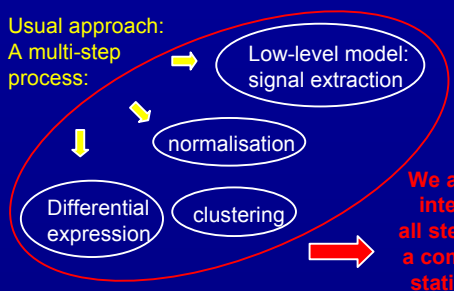
Bayesian Hierarchical Modelling of Gene Expression Data

Anne-Mette K. Hein, Alex Lewin, Clare Marshall and Sylvia Richardson, Imperial College, London



- Gene expression arrays: hybridise fluorescently labelled mRNA – intensity related to gene expression
- Thousands of genes represented on an array, low number of replicate arrays (costly!)
- Goal of microarray experiments: compare patterns of gene expression under different conditions (e.g. types of tumor, disease vs. control)
- Various sources and levels of variability



Within an array

a.hein@imperial.ac.uk

Affymetrix chip: each gene g is represented by a probe set: J probe pairs. A probe pair:

perfect match (PM_{gj}) and mismatch (MM_{gj})

Extracting gene expression measures from probes

$PM_{gj} \mid m_{gj}^1 \sim N(m_{gj}^1, \eta^2)$
$MM_{gj} \mid m_{gj}^2 \sim N(m_{gj}^2, \eta^2)$ → Background noise, additive

$m_{gj}^1 = S_{gj} + H_{gj}$
$m_{gj}^2 = \Phi S_{gj} + H_{gj}$ → Signal + cross-hybridization

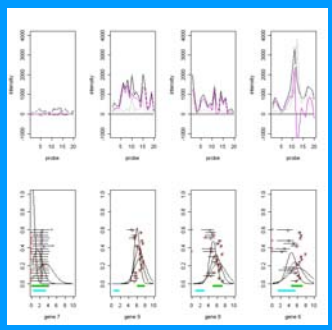
$\log(S_{gj} + 1) \sim TN(\mu_g, \tau_g^2)$ → Gene specific error term
$\log(H_{gj}) \sim N(\lambda, \sigma^2)$
$\log(\tau_g^2) \sim N(a, b)$ → Gene expression index: "Pools" information from the probes

Variances exchangeable

Data from one array: 4 genes, each at conc. 2.0: probe response and posterior distributions

PM:
MM:
PM-MM:
Log(PM-MM):

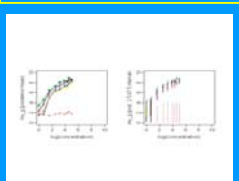
Probe response:



S_{gj}: μ_g: τ_g^2:
log(Sgj+1):
TN(μg,τg2)
(2.5,mean and 97.5)

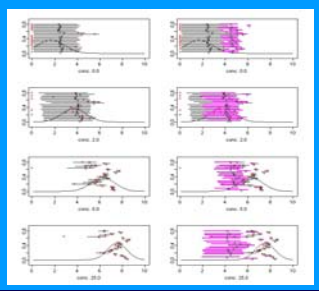
Mean posterior estimates:

Signals increase with concentration, cross doesn't



10 arrays: 11 genes spiked-in all at conc ci on array i (left):

One gene spiked in at 4 different cncentrations (right) :



Between arrays

a.m.lewin@imperial.ac.uk

Each gene g has one estimate of signal y_{gr} on each array r. Our data set has 3 arrays for each of 2 experimental conditions. The model estimates gene effects α_g, gene-condition effects δ_{gs}, array effects $\beta_{r(g)}$ and gene-specific variances σ_g^2.

Bayesian Hierarchical Model for Differential Expression

- 1st level

$y_{gsr} \sim N(\alpha_g + \delta_{gs} + \beta_{sr(g)}, \sigma_{gs}^2)$, $\Sigma_r \beta_{sr(g)} = 0$

$\beta_{sr(g)}$ = piece-wise quadratic in α_g with unknown locations of break-points, parameters {**a**} and {**b**}

- 2nd level

Priors for α_g, δ_{gs}, coefficients {**a**} and {**b**}

$\sigma_{gs}^2 \sim$ lognormal (μ_s, τ_s)

- 3rd level

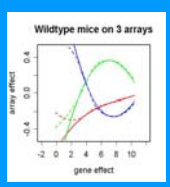
$\mu_s \sim N(c, d)$ $\tau_s \sim$ lognormal (e, f)

ARRAY EFFECTS ARE FUNCTIONS OF GENE EFFECT

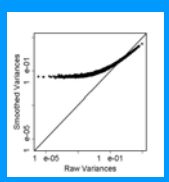
EXCHANGEABE VARIANCES

POSTERIOR DISTRIBUTIONS SIMULATED BY MCMC USING WINBUGS

Array effect as function of gene effect (cubic and loess smoothing)



Exchangeable variances (share information between genes) versus raw variances (3 measurements per gene)

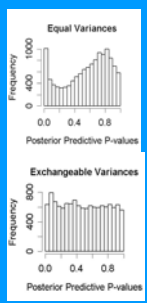


Bayesian Model Checking for Gene-Specific Variances

Predict sample variance $S_g^{2\,new}$ for each gene from the model, compare with observed sample variance $S_g^{2\,obs}$.

Bayesian p-value Prob($S_g^{2\,new} > S_g^{2\,obs}$)

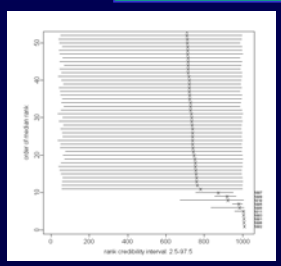
P-values Uniform if model true.



Integrated:

Integration of Within and Between Array Models

We obtain the joint distribution of posterior ranks of differential expression.



GeneLogic spike-ins: truth known: 11 of 1011 genes are differentially expressed. These are genes number 1001-1011. The ranks of their log ratios under the two conditions (three replicates under each) considered are:

Ranks:

True:	2	1	3/4	9	3/4	5	7	8	6	10	11
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