BGX: a Bioconductor R package for the analysis of GeneChip data

$\sim$

BGMix: an R package for differential expression

Ernest Turro
Alex Lewin

Department of Epidemiology & Public Health
Imperial College London

BGX: Bayesian Gene eXpression

Integrated modelling of Affymetrix GeneChip data

Hein et al 2005, Biostatistics
Hein & Richardson 2006, BMC Bioinformatics
Turro et al 2007, under review

GeneChips

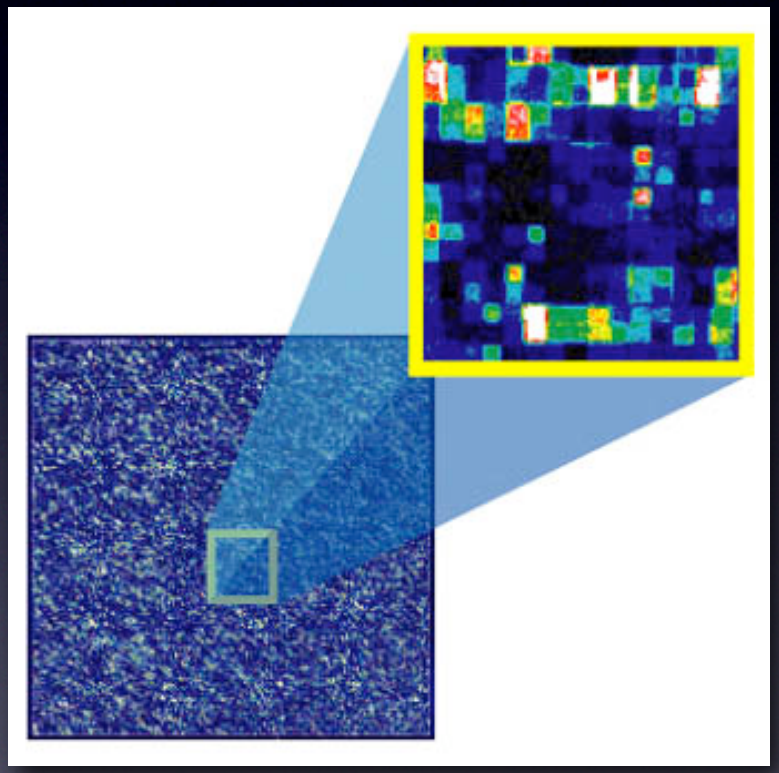


Image courtesy of Affymetrix

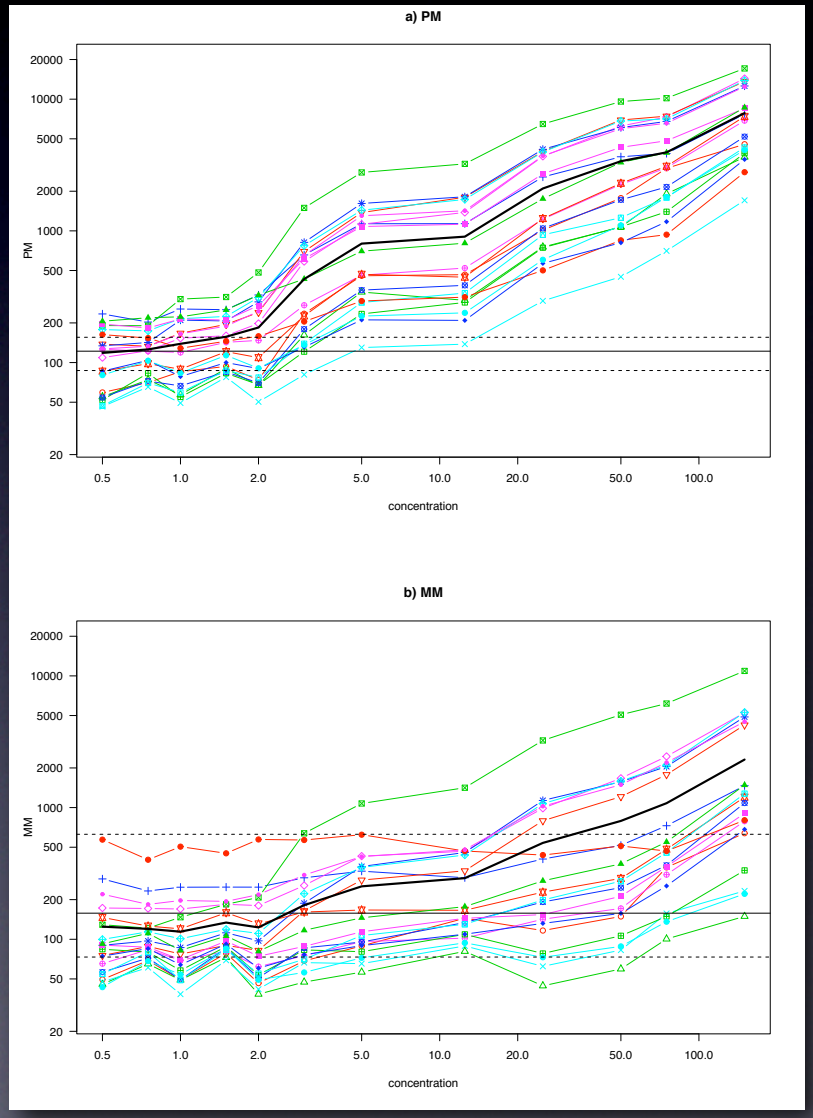
Gene g is represented by g probe sets of j probe pairs:

Perfect match: PM_{g1},..., PM_{gJ}

Mismatch: MM_{g1},..., MM_{gJ}

Noise

- Background Noise: both PM and MM bind to target (MM less so than PM)
- Biological and technical variability: both PM and MM hybridise non-specifically



Irizarry et al 2003

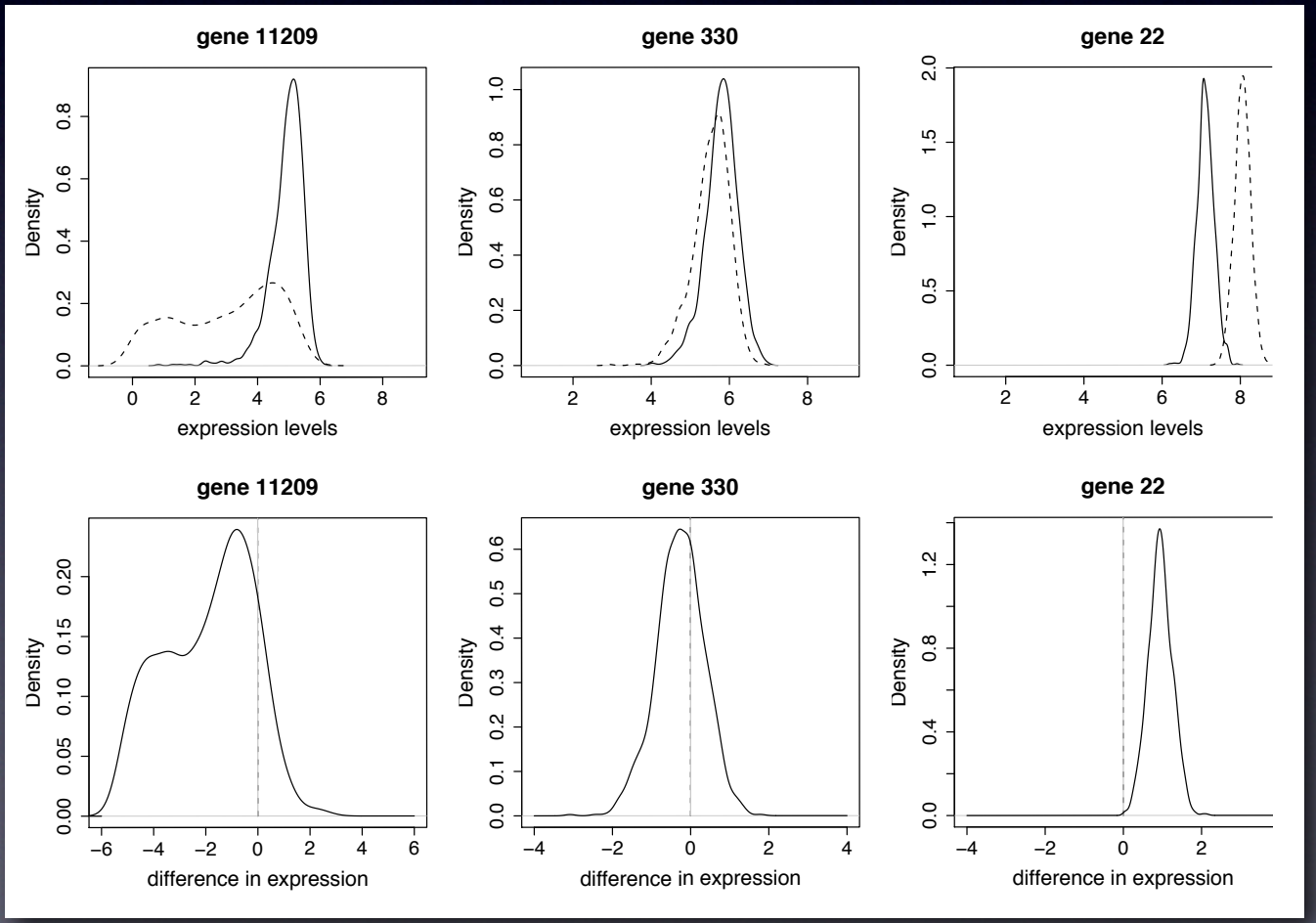
BGX model

Simultaneous modelling of all levels of errors

PMs, MMs ➡ Signal, Non-specific hybridisation ➡ Distribution of log gene expression

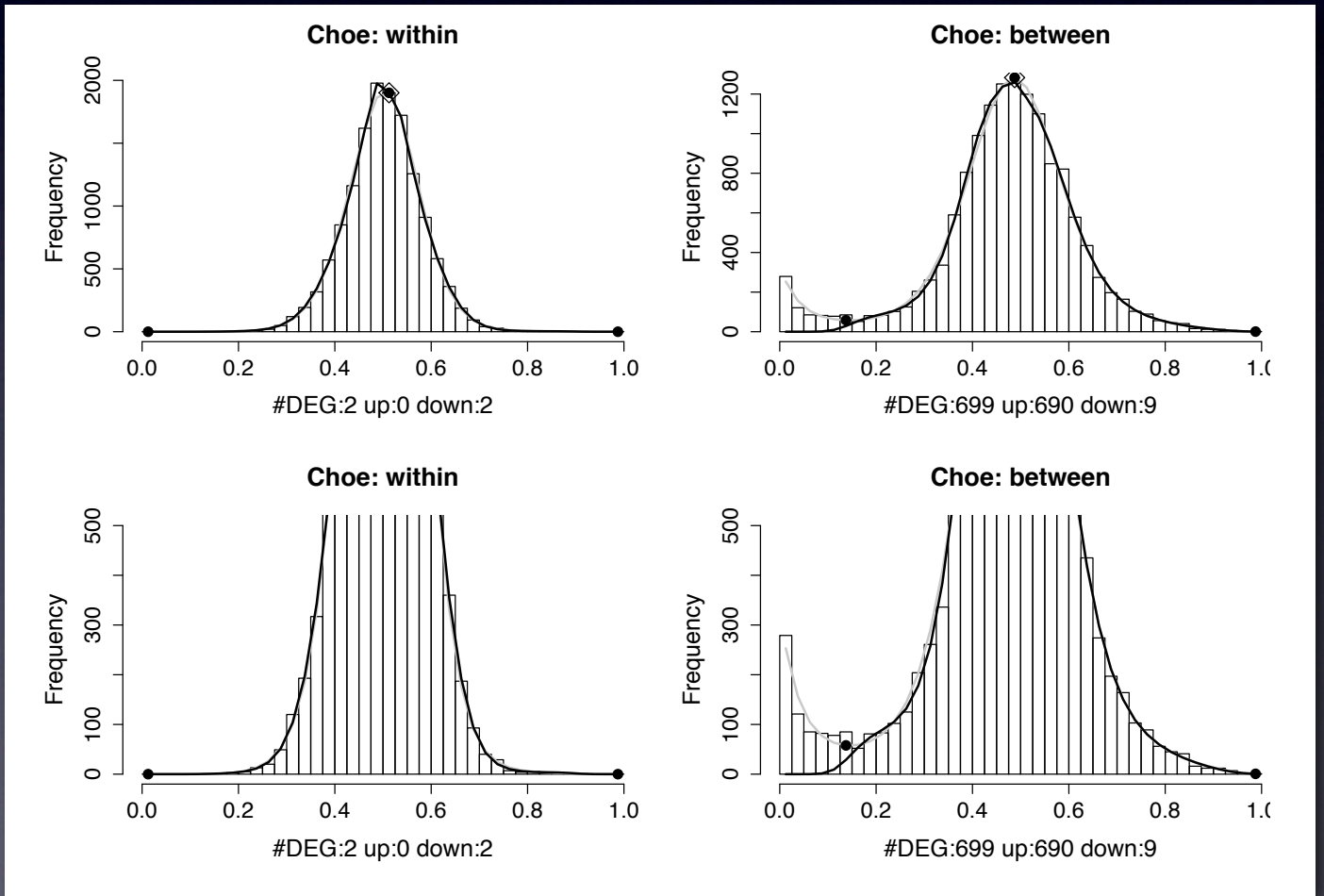
Conditions 1 and 2

Condition 2 - Condition 1



BGX model

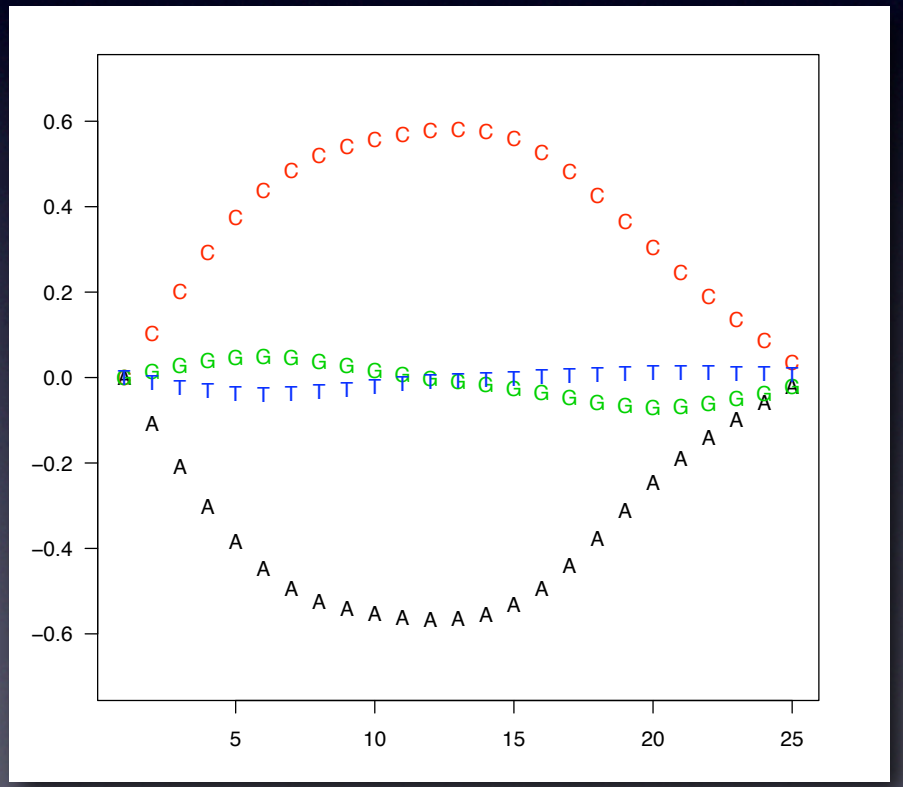
Estimate number of differentially expressed genes



GCBGX extension

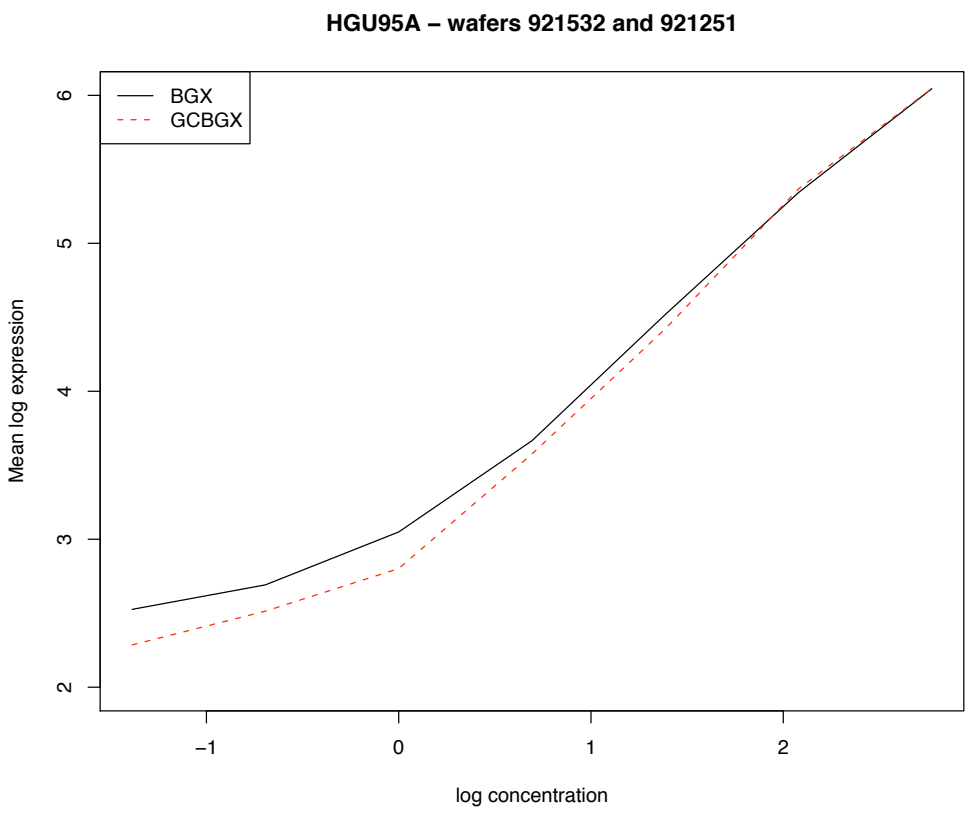
We take into account probe GC content (like GCRMA)

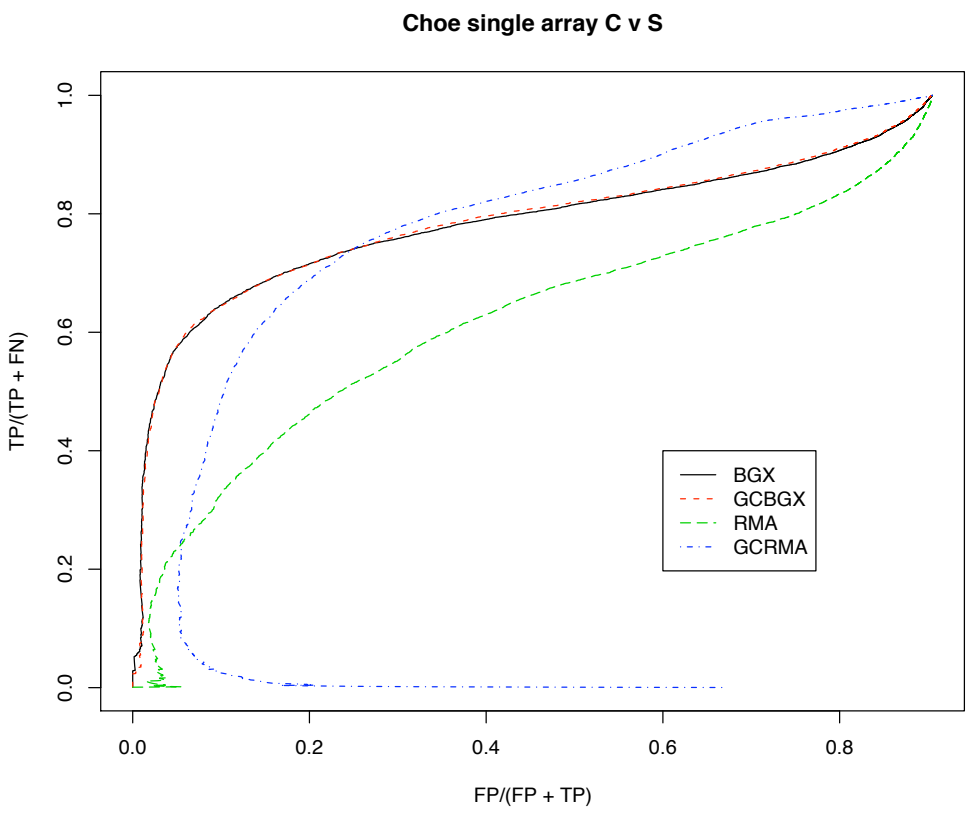
- We categorise each probe by its GC content into "probe affinity categories"
- Non-specific hybridisation is calculated separately for each category



Wu et al 2004

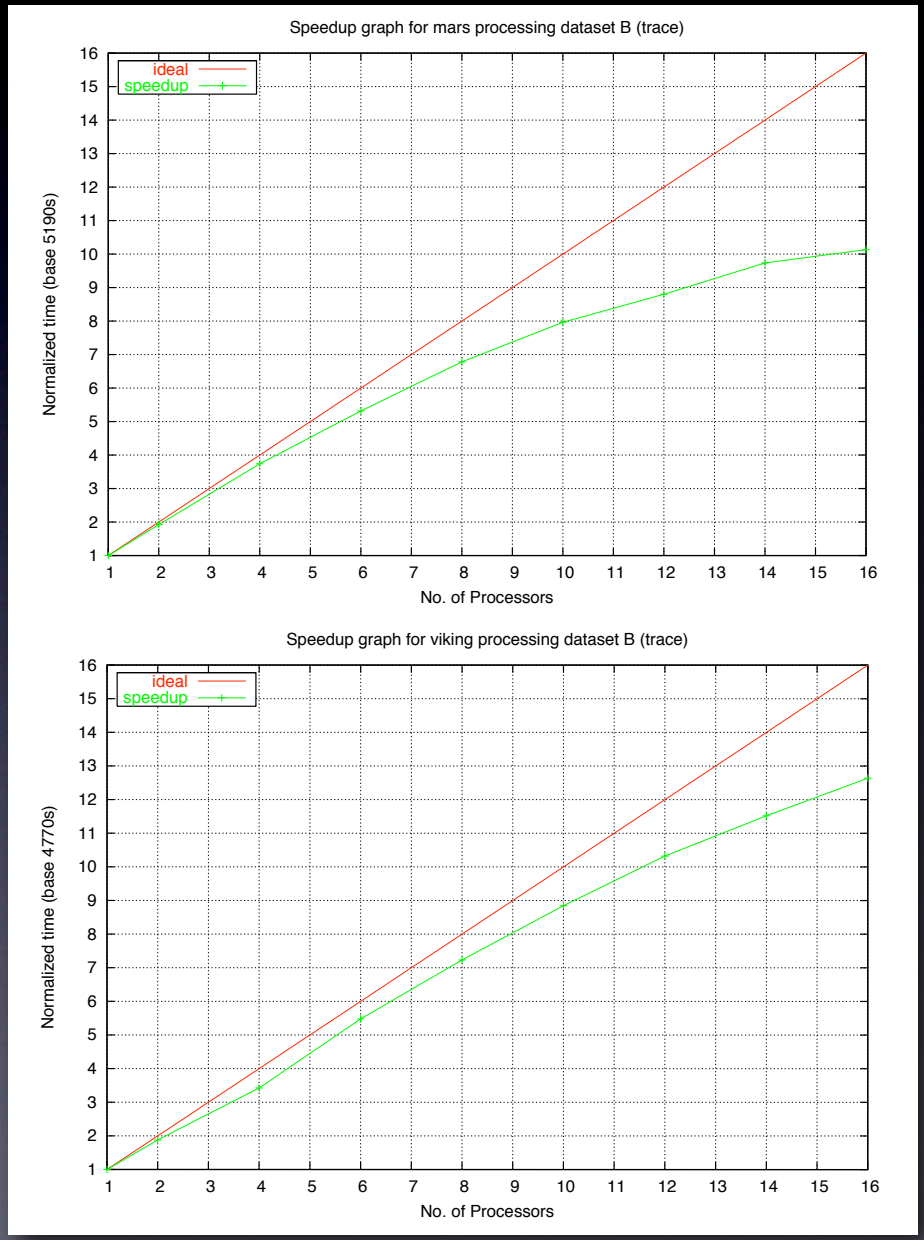
BGX performance





Speed

- BGX is computationally intensive
- A parallel version was developed
- It can be run on a computer cluster and offers significant speed-ups
- Only available for non-GC version



Exon arrays

- New Affymetrix arrays work at the exon level
- Exon arrays you to distinguish between different isoforms of a gene
- Probes on different exons can be summarised into an expression value of all transcripts from the same gene

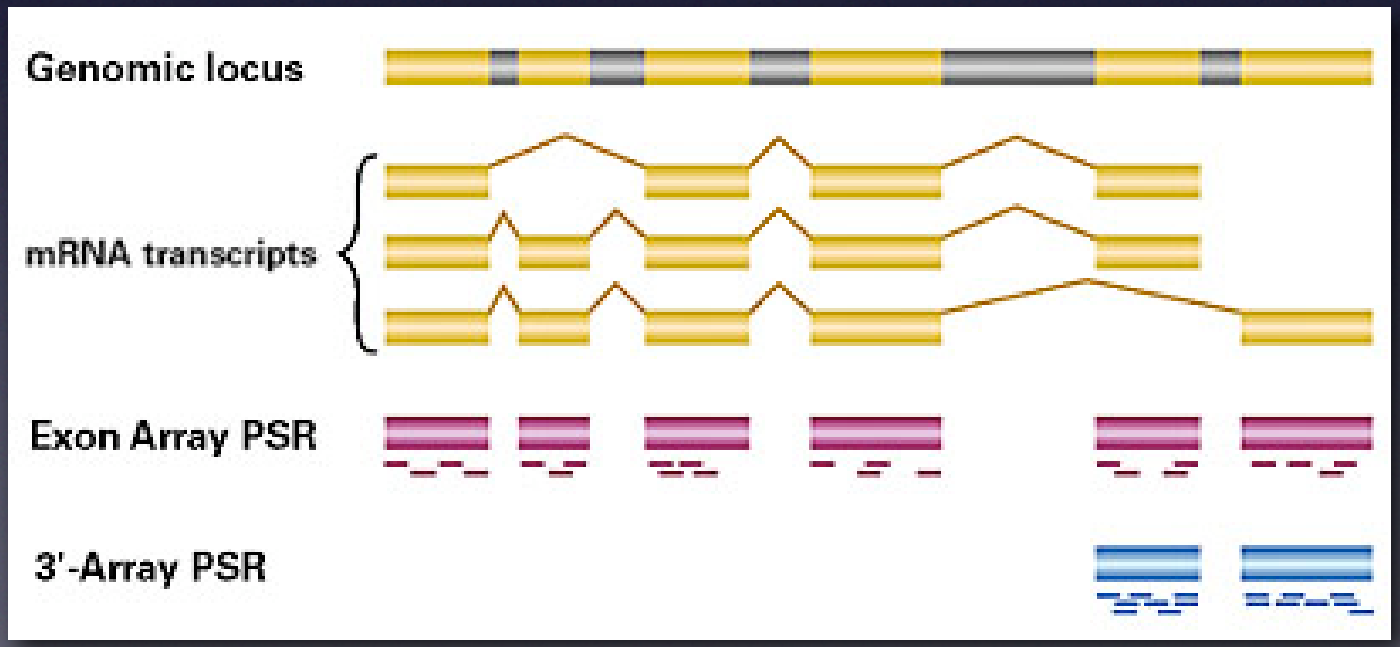


Image courtesy of Affymetrix

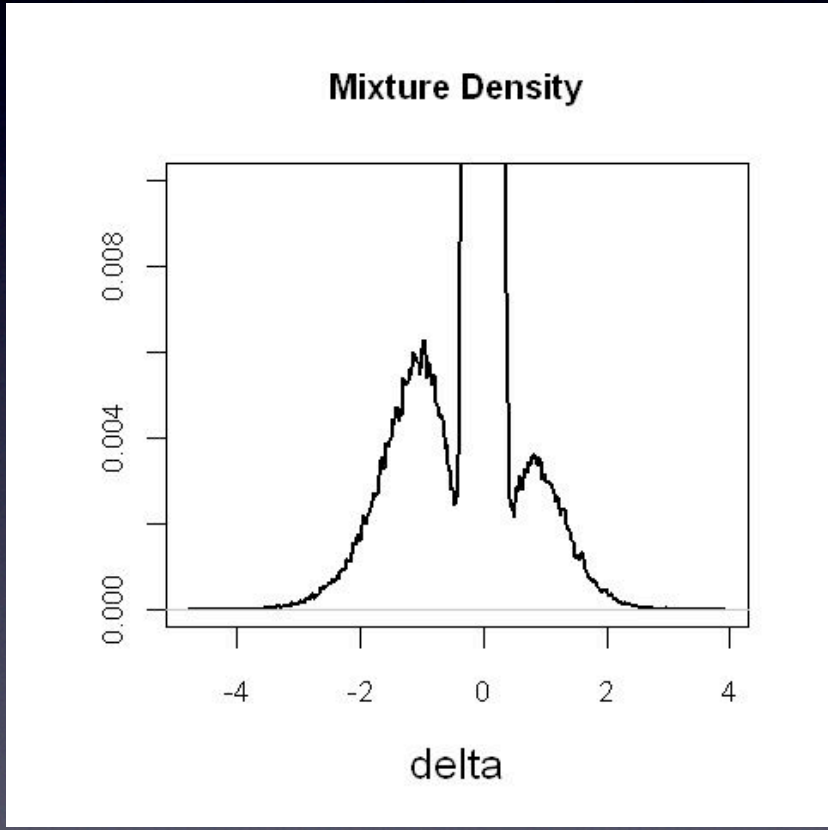
BGMix: mixture model for differential expression

Integrated modelling of
Affymetrix GeneChip data

Lewin et al 2006, Biometrics
Lewin et al 2007, under review

Modelling Assumptions

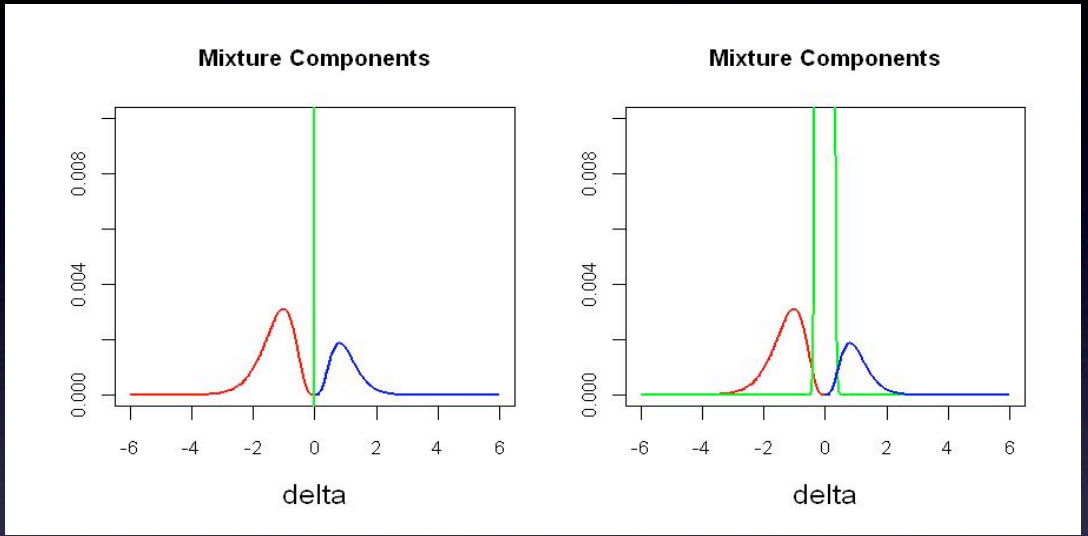
- Groups genes into 3 classes:
  - non-DE
  - over-expressed
  - under-expressed
- Gene-dependent errors
- Estimation and classification is simultaneous



Modelling Assumptions

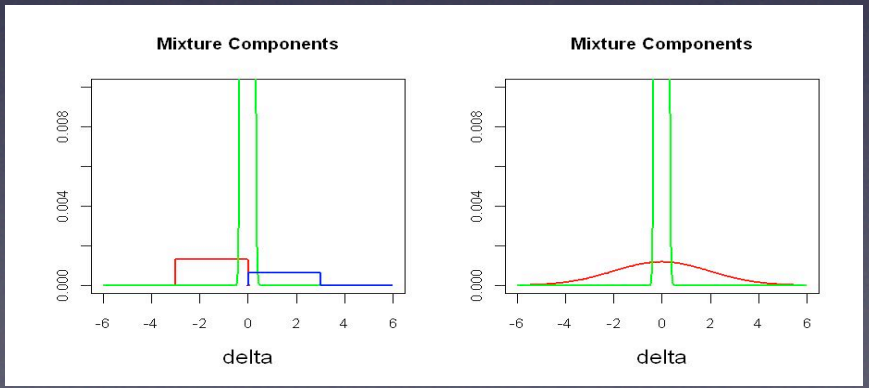
Choice of Null Distribution

- True log fold changes = 0
- 'Nugget' null: true log fold changes = small but not necessarily zero



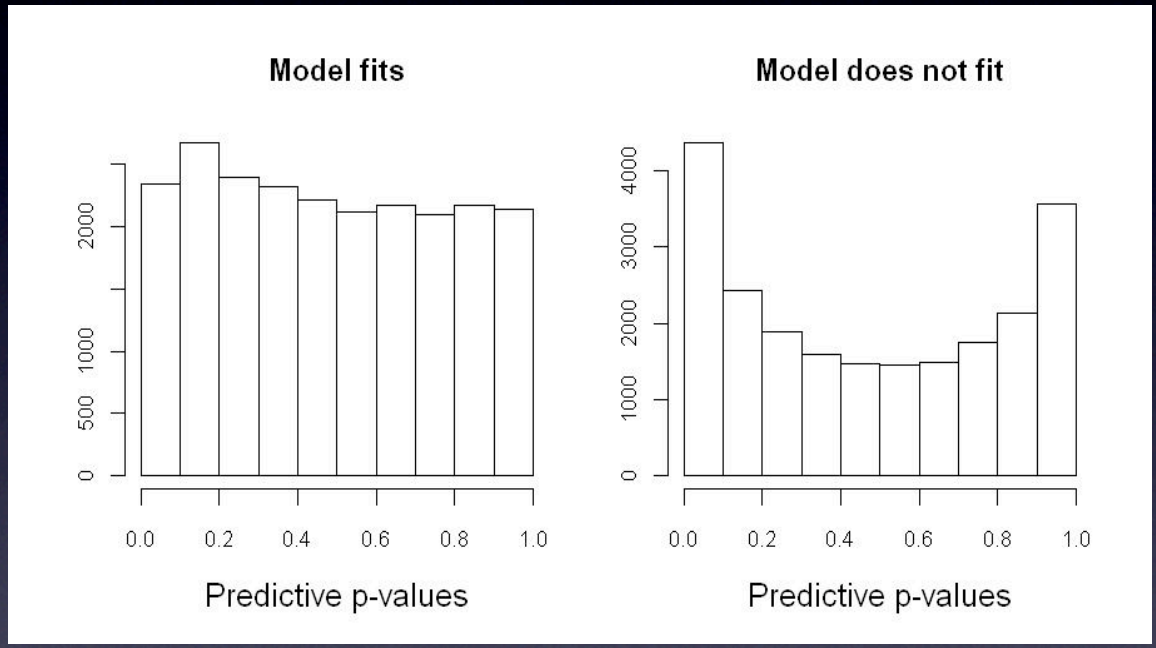
Choice of DE genes distributions

- Gammas
- Uniforms
- Normal



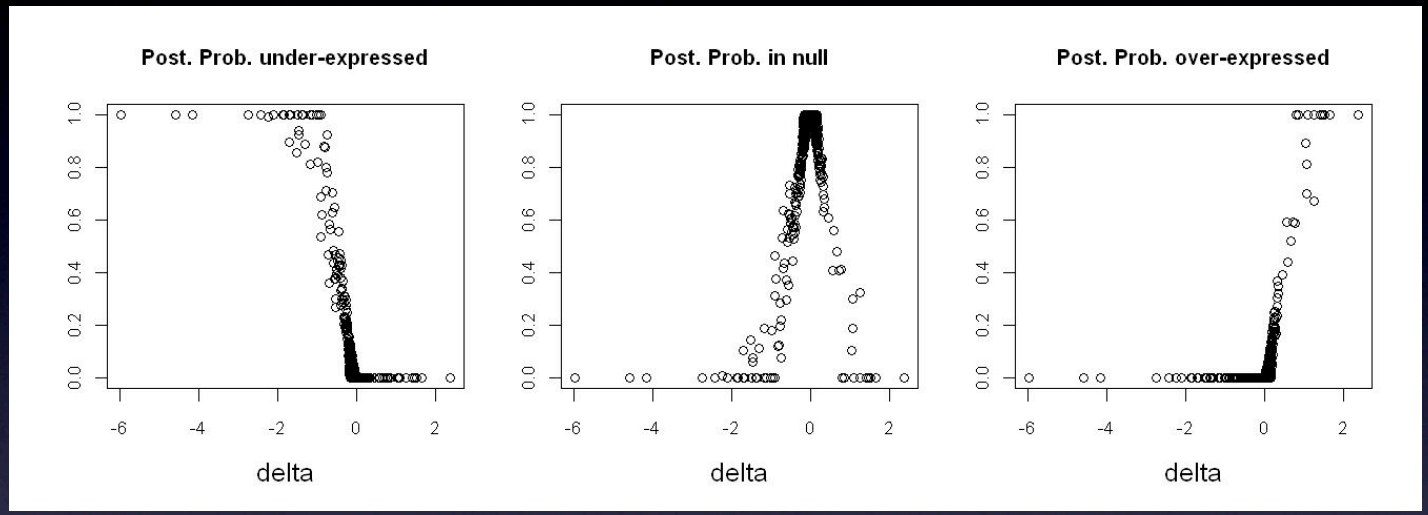
Model checks

- Predict new data from the model
- Compare predicted with observed data
- Predictive p-values close to Uniform if model is ok



CURRENT WORK

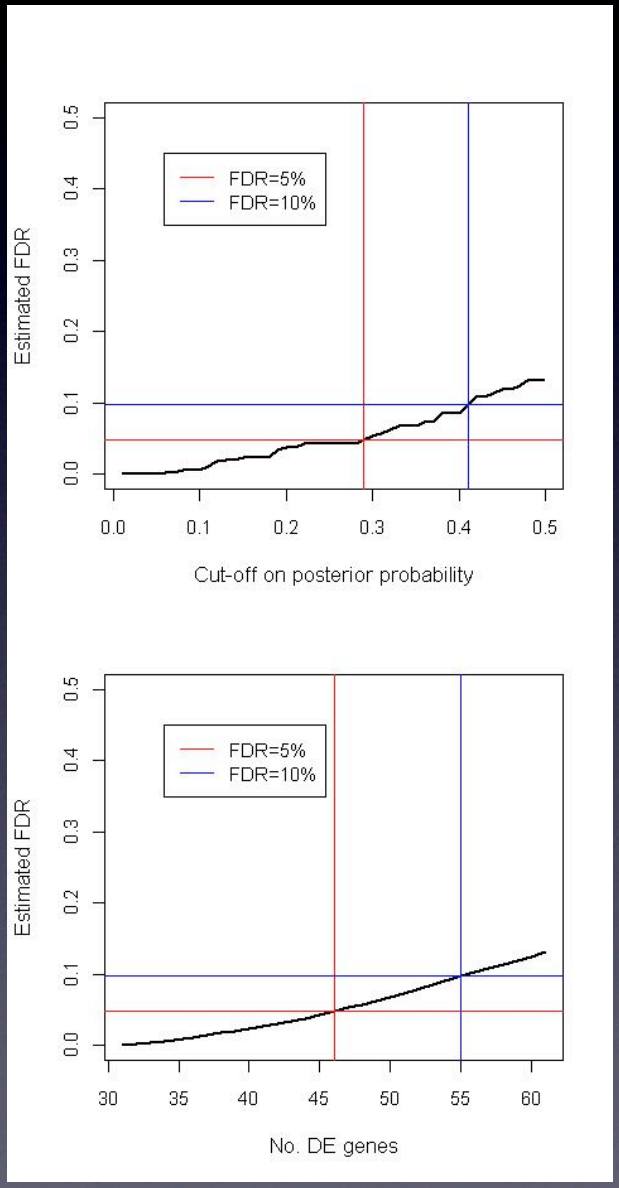
Outputs from model



- Point estimates (and s.d.) of log fold changes (stabilised and smoothed)
- Posterior probability for gene to be in each group
- Estimate of proportion of differentially expressed genes based on grouping (parameter of model)

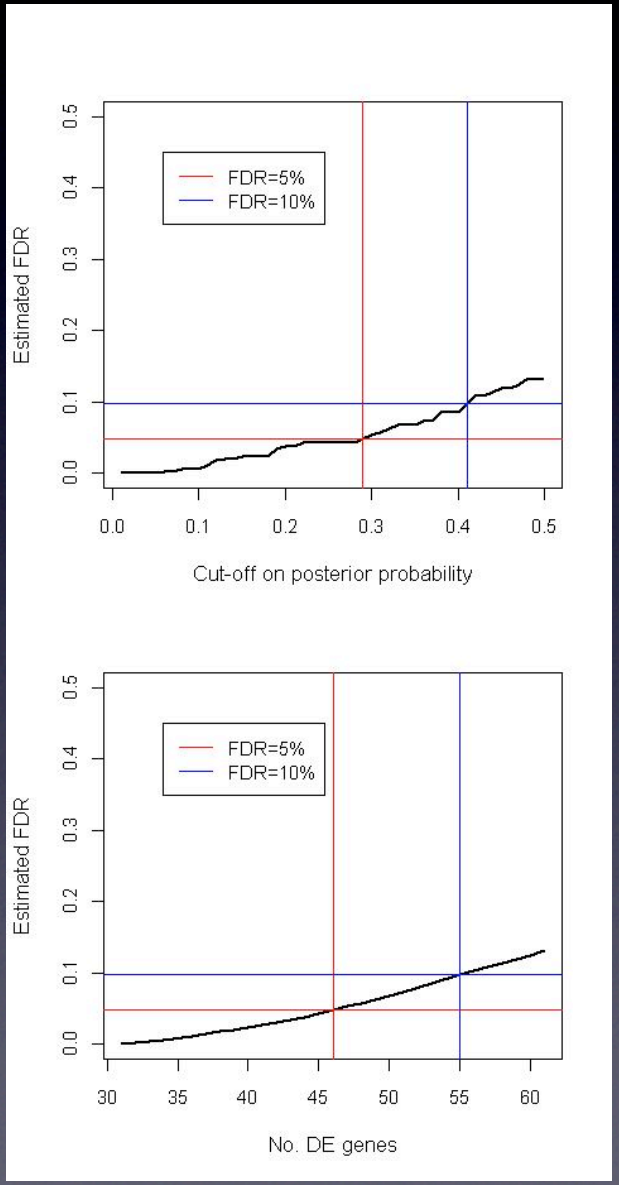
Obtaining gene lists

- Threshold on posterior probabilities
- Posterior probability of classification in the null < threshold → gene is DE
- Estimate of False Discovery Rate for any gene list (estimate = average of posterior probabilities)
- Very simple estimate!



Obtaining gene lists

- Choice of decision rule:
  - Bayes Rule (threshold=0.5)
  - Fix False Discovery Rate
  - More complex rules for mixture of 3 components



Thanks to

Sylvia Richardson

Tim Aitman

Anne-Mette Hein

Natalia Bochkina

Marta Blangiardo

Peter Green

Demo