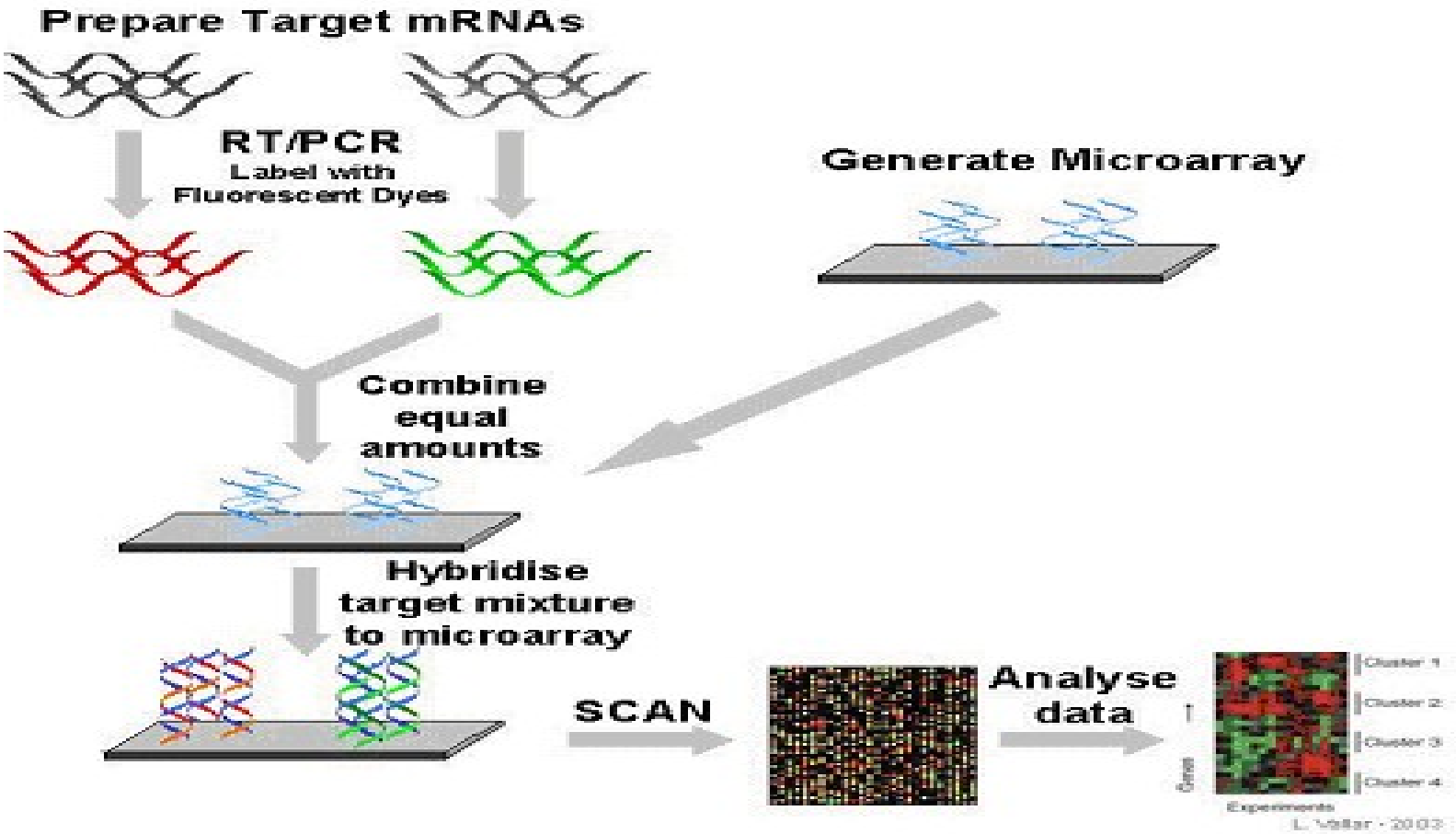
A Bayesian Hierarchical Model for Correlation in Microarray Studies

Bernard Omolo
University of South Carolina-Upstate
email: bomolo@uscupstate.edu
Strathmore International Math Research Conference, Nairobi, Kenya

July 23 - 27, 2012

Outline

- Microarrays are miniaturised biological devices consisting of molecules (e.g. DNA or protein), called “probes”, that are orderly arranged at a microscopic scale onto a solid support such as a nylon membrane or a glass slide.
- The array elements (probes) bind specifically to labeled molecules, called ”targets”, into complex molecular mixtures, thereby generating signals that reveal the identity and the concentration of the interacting labeled cells.
- Microarray analysis has a broad range of applications that involve different types of probes and/or targets (cDNA or oligos).


Prepare Target mRNAs
RT/PCR
Label with
Fluorescent Dyes
Generate Microarray
Combine
equal
amounts
Hybridise
target mixture
to microarray
SCAN
Analyse
data
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Genes
Experiments

- Question: are the (two) studies reproducible?
- Assessment of reproducibility may help in deciding whether to use an original expression data or one updated with additional samples, for differential gene expression analysis.
- One approach is to model parameters measuring association between two independent datasets, e.g gene-specific correlations.
- A major challenge is that the numbers of probes for each gene are small and these numbers are different across the datasets.

- Correlations computed using mean expression values for each common gene and cell-line between the two datasets.
- Replicates for each gene not used, thereby ignoring the effect of multiple probes per gene.
- Within cell-line variability not accounted for, which may lead to poor estimates of the correlations.
- We propose a multi-level model that will better assess gene-specific correlations between the two datasets.

- Primary data available at the UNC Microarray Database (https://genome.unc.edu).
- Dataset A (16 melanomas) accessible in GEO (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE7469 and contains 15,749 probes.
- Dataset B (35 melanomas) contains 19,734 probes.
- The merged dataset has 11,271 genes on 16 melanomas.

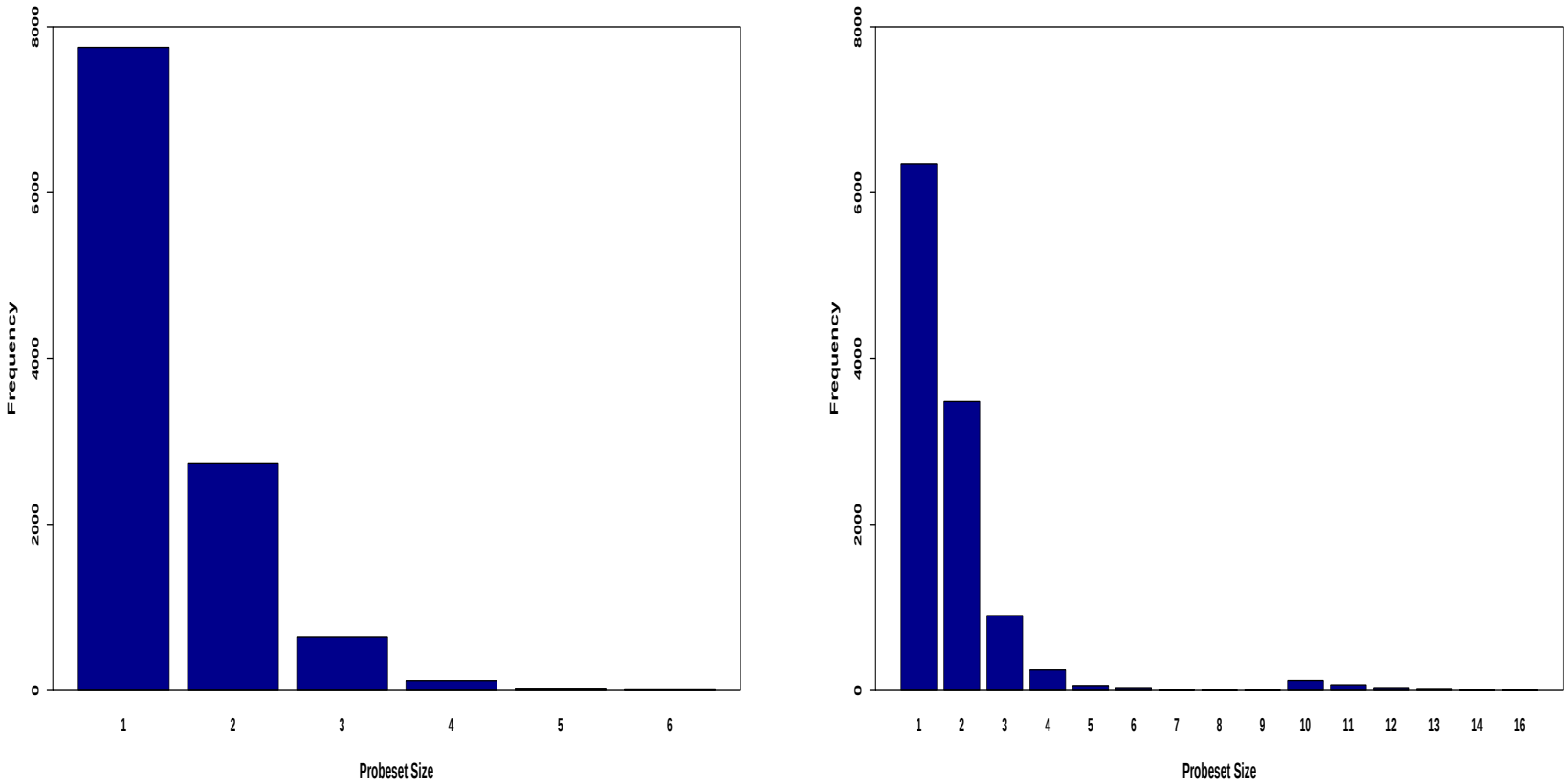


Figure: 1

Probeset sizes for datasets A (left) and B (right)

- Even though most genes were represented by just one probe in both datasets, the frequency of replicated genes were quite high (over 3500 for dataset A and over 4000 for dataset B).
- There were more probes among genes in dataset B than dataset A.
- These results provide further motivation for considering replicates in modeling gene-specific correlations.

- $\mathbf{x} = (x_{gij},\ j = 1, \ldots, n_{x_g}, i = 1, \ldots, S, g = 1, \ldots, G)$
- $\mathbf{y} = (y_{gik},\ k = 1, \ldots, n_{y_g},\ i = 1, \ldots, S,\ g = 1, \ldots, G)$
- $\overline{x}_{gi} = \frac{1}{n_{x_g}} \sum_{j=1}^{n_{x_g}} x_{gij}$
- $\overline{y}_{gi} = \frac{1}{n_{y_g}} \sum_{k=1}^{n_{y_g}} y_{gik}$
- $D = \{x_{gij},\ y_{gik}\}$, full expression data
- $D_{mean} = \{\bar{x}_{gi}, \bar{y}_{gi}\}$, collapsed data

- $\mathbf{x}_{gi}$'s are independent for all g and i and $\mathbf{y}_{gi}$'s are independent for all g and i.
- For a given g and i, x_{gij} and $x_{gij'}$ may be dependent for $j \neq j'$ and y_{gik} and $y_{gik'}$ may also be dependent for $k \neq k'$.
- $\mathbf{x}_{gi}$ and $\mathbf{y}_{gi}$ are dependent within the same gene g and independent across different genes for all g.

- $(\overline{x}_{gi}, \overline{y}_{gi})' \sim N_2(\tilde{\mathbf{m}}_g, \tilde{H}_g)$
- $\tilde{\mathbf{m}}_g = (\tilde{m_{\overline{x}_g}}, \tilde{m_{\overline{y}_g}})'$
- $\tilde{H}_g = \begin{pmatrix} \tilde{h}^2_{\overline{x}_g} & \tilde{\rho}_g \tilde{h}_{\overline{x}_g} \tilde{h}_{\overline{y}_g} \\ \tilde{\rho}_g \tilde{h}_{\overline{x}_g} \tilde{h}_{\overline{y}_g} & \tilde{h}^2_{\overline{y}_g} \end{pmatrix}$
- Pearson correlation as an estimate of $\tilde{\rho}_g$

- $(\overline{x}_{gi}, \overline{y}_{gi})' \sim N_2(\tilde{\mathbf{m}}_g, \tilde{H}_g)$.
- $\tilde{\mathbf{m}}_g = (\tilde{m_{\overline{x}_g}}, \tilde{m_{\overline{y}_g}})'$
- $\tilde{m_{\overline{x}_g}} \sim \mathcal{N}(m_{x_0}, \tau^2_{x_0})$
- $\tilde{m_{\overline{y}_g}} \sim \mathcal{N}(m_{y_0}, \tau^2_{y_0})$
- $\tilde{H}_g \sim \mathrm{IW}(\nu_0, H_0)$

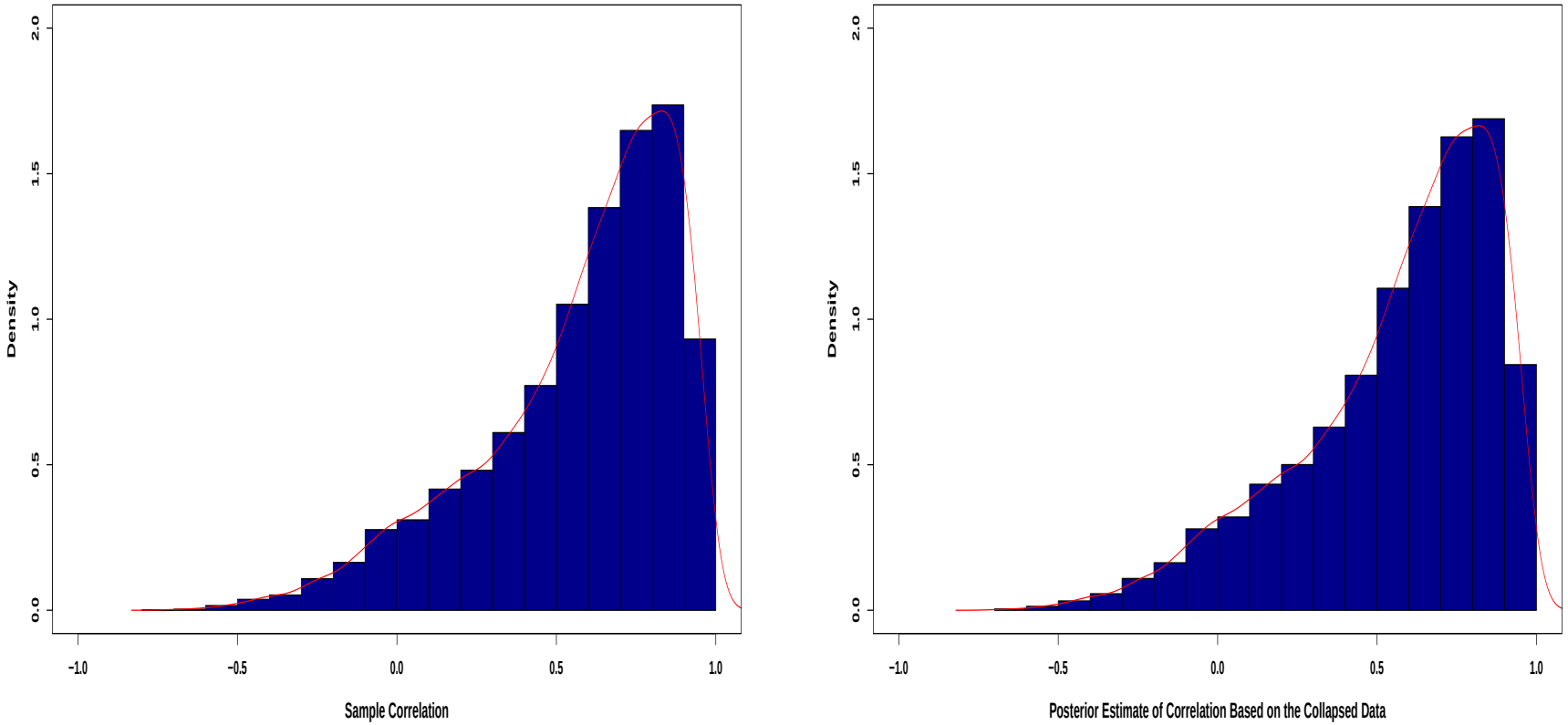


Figure: 3

Density Histograms of Correlations (Frequentist vs Bayesian)

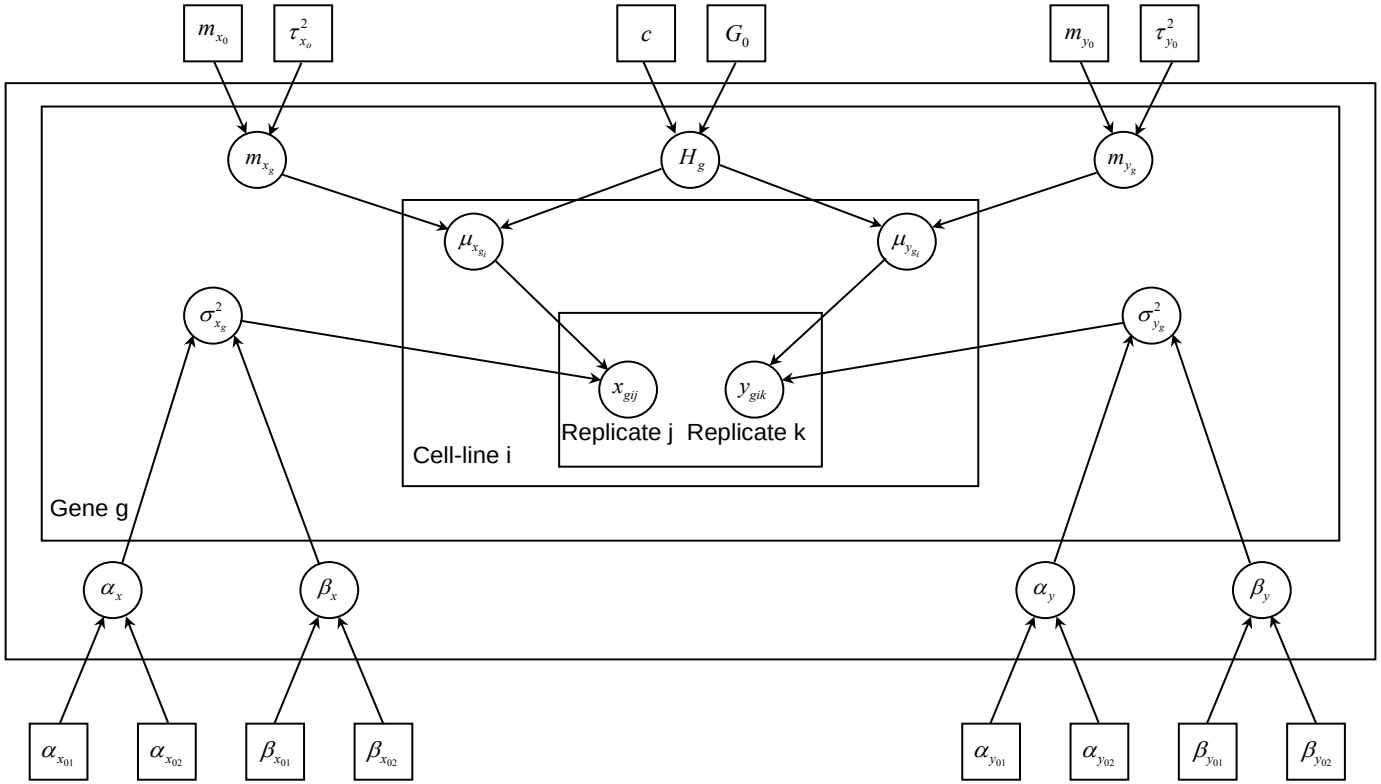


Figure: 4

DAG of the BHM (circle: stochastic; square: specified hyperparameter)

Level 1

- $x_{gij} = \mu_{x_{gi}} + \epsilon_{xgij},\ \epsilon_{xgij} \sim \mathcal{N}(0, \sigma^2_{x_g})$
- $y_{gik} = \mu_{y_{gi}} + \epsilon_{ygik},\ \epsilon_{ygik} \sim \mathcal{N}(0, \sigma^2_{y_g})$
- $\mu_{x_{gi}}$ and $\mu_{y_{gi}}$ denote the "true" expression levels of x_{gij} and y_{gik}, assumed to be dependent and random.
- $\mu_{x_{gi}}$ and $\mu_{y_{gi}}$ capture the dependence of x_{gij}'s and y_{gik}'s within gene g and cell-line i, respectively.
- $\sigma^2_{x_g}$ and $\sigma^2_{y_g}$ quantify the within-cell-line variabilities of the expression values.

Level 2

- $\boldsymbol{\mu}_{gi} = (\mu_{x_{gi}}, \mu_{y_{gi}})' \sim \mathcal{N}_2(\mathbf{m}_g, H_g)$
- $\mathbf{m}_g = (m_{x_g}, m_{y_g})'$
- $H_g = \begin{pmatrix} h^2_{x_g} & \rho_g h_{x_g} h_{y_g} \\ \rho_g h_{x_g} h_{y_g} & h^2_{y_g} \end{pmatrix}$
- $\sigma^2_{x_g} \sim IG(\alpha_x, \beta_x)$
- $\sigma^2_{y_g} \sim IG(\alpha_y, \beta_y)$
- (α_x, β_x), (α_y, β_y) are unknown.

Level 3

- $H_g \sim \text{IW}(\nu_0, H_0)$
- $m_{x_g} \sim \mathcal{N}(m_{x_0}, \tau^2_{x_0})$; $m_{y_g} \sim \mathcal{N}(m_{y_0}, \tau^2_{y_0})$
- $\alpha_x \sim \Gamma(\alpha_{x_{01}}, \alpha_{x_{02}})$; $\beta_x \sim \Gamma(\beta_{x_{01}}, \beta_{x_{02}})$
- $\alpha_y \sim \Gamma(\alpha_{y_{01}}, \alpha_{y_{02}})$; $\beta_y \sim \Gamma(\beta_{y_{01}}, \beta_{y_{02}})$
- m_{x_0}, $\tau^2_{x_0}$, m_{y_0}, $\tau^2_{y_0}$, $\alpha_{x_{01}}$, $\alpha_{x_{02}}$, $\alpha_{y_{01}}$, $\alpha_{y_{02}}$, $\beta_{x_{01}}$, $\beta_{x_{02}}$, $\beta_{y_{01}}$, $\beta_{y_{02}}$: pre-specified hyperparameters.

Posterior

Parameters: $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{H}, \mathbf{m}, \alpha_x, \alpha_y, \beta_x, \beta_y)$,
Hyperparams:
$\gamma_0 = (H_0, m_{x_0}, \tau^2_{x_0}, m_{y_0}, \tau^2_{y_0}, \alpha_{x_{01}}, \alpha_{x_{02}}, \alpha_{y_{01}}, \alpha_{y_{02}}, \beta_{x_{01}}, \beta_{x_{02}}, \beta_{y_{01}}, \beta_{y_{02}})$

$$\begin{aligned}\text{Posterior: } \pi(\boldsymbol{\theta}|\mathbf{x}, \mathbf{y}, \gamma_0) \propto\ & L(\boldsymbol{\mu}, \boldsymbol{\sigma}^2|\mathbf{x}, \mathbf{y})\pi(\boldsymbol{\mu}|\mathbf{m}, \mathbf{H})\pi(\mathbf{m}|m_{x_0}, \tau^2_{x_0}, m_{y_0}, \tau^2_{y_0})\pi(\mathbf{H}|\nu_0, H_0) \\ & \times \pi(\boldsymbol{\sigma}^2|\alpha_x, \beta_x, \alpha_y, \beta_y)\pi(\alpha_x|\alpha_{x_{01}}, \alpha_{x_{02}})\pi(\alpha_y|\alpha_{y_{01}}, \alpha_{y_{02}}) \\ & \times \pi(\beta_x|\beta_{x_{01}}, \beta_{x_{02}})\pi(\beta_y|\beta_{y_{01}}, \beta_{y_{02}})\end{aligned}$$

Part I

- $\bar{x}_g = \frac{1}{S}\sum_{i=1}^{S}\bar{x}_{gi},\ \ \bar{y}_g = \frac{1}{S}\sum_{i=1}^{S}\bar{y}_{gi}\ \ (g = 1, \ldots, G)$
- $s^2_{x_g} = \frac{1}{Sn_{x_g}-1}\sum_{i=1}^{S}\sum_{j=1}^{n_{x_g}}(x_{gij} - \bar{x}_g)^2$
- $s^2_{y_g} = \frac{1}{Sn_{y_g}-1}\sum_{i=1}^{S}\sum_{k=1}^{n_{y_g}}(y_{gik} - \bar{y}_g)^2$
- $\bar{s}^2_x = \frac{1}{G}\sum_{g=1}^{G}s^2_{x_g},\ \ s^2_{s^2_x} = \frac{1}{G-1}\sum_{g=1}^{G}(s^2_{x_g} - \bar{s}^2_x)^2$
- $\bar{s}^2_y = \frac{1}{G}\sum_{g=1}^{G}s^2_{y_g},\ \ s^2_{s^2_y} = \frac{1}{G-1}\sum_{g=1}^{G}(s^2_{y_g} - \bar{s}^2_y)^2$

Part II - EB (Chen et al., 2008)

- $\alpha_{x01} = k_{x01}\left[2 + (\bar{s}_x^2)^2/s_{s_x^2}^2\right],\ \alpha_{x02} = k_{x01}$
- $\beta_{x01} = k_{x02}\left[1 + (\bar{s}_x^2)^2/s_{s_x^2}^2\right]\bar{s}_x^2,\ \beta_{x02} = k_{x02}$
- $\alpha_{y01} = k_{y01}\left[2 + (\bar{s}_y^2)^2/s_{s_y^2}^2\right],\ \alpha_{y02} = k_{y01}$
- $\beta_{y01} = k_{y02}\left[1 + (\bar{s}_y^2)^2/s_{s_y^2}^2\right]\bar{s}_y^2,\ \beta_{y02} = k_{y02}$
- $(k_{x01}, k_{x02}, k_{y01}, k_{y02})$ pre-specified s.t. $\alpha_{x01} \geq 1$ and $\alpha_{y01} \geq 1$.

Part III

- $\bar{\bar{x}} = \frac{1}{G}\sum_{g=1}^{G}\bar{x}_g,\ \ \bar{\bar{y}} = \frac{1}{G}\sum_{g=1}^{G}\bar{y}_g$
- $s^2_{\bar{x}} = \frac{1}{G-1}\sum_{g=1}^{G}(\bar{x}_g - \bar{\bar{x}})^2$
- $s^2_{\bar{y}} = \frac{1}{G-1}\sum_{g=1}^{G}(\bar{y}_g - \bar{\bar{y}})^2$
- $m_{x_0} = \bar{\bar{x}},\ \tau^2_{x_0} = w_{x_0}s^2_{\bar{x}}$
- $m_{y_0} = \bar{\bar{y}},$ and $\tau^2_{y_0} = w_{y_0}s^2_{\bar{y}}$
- $H_0 = \frac{h_0}{G(S-1)}\sum_{g=1}^{G}\sum_{i=1}^{S}(\bar{x}_{gi} - \bar{x}_g, \bar{y}_{gi} - \bar{y}_g)'(\bar{x}_{gi} - \bar{x}_g, \bar{y}_{gi} - \bar{y}_g)$
- $w_{x_0} > 0,\ w_{y_0} > 0, h_0 > 0$ pre-specified.

Step 0. Set the initial values of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{m}, \mathbf{H}, \alpha_x, \alpha_y, \beta_x, \beta_y)$.

Step 1. Update $\mathbf{H}$ from the conditional posterior distribution $\pi(\mathbf{H}|\boldsymbol{\theta}_{(-\mathbf{H})}, \gamma_0)$.

Step 2. Update $\boldsymbol{\theta}_{(-\mathbf{H})}$ from the conditional posterior distribution $\pi(\boldsymbol{\theta}_{(-\mathbf{H})}|\mathbf{H}, \mathbf{x}, \mathbf{y}, \gamma_0)$.

Step 3. Go back to Step 1.

In Step 1 above,

$$H_g|\boldsymbol{\theta}_{(-\mathbf{H})}, \gamma_0 \sim \pi(H_g|\nu_0 + S, H_{0g}),$$

$$\pi(H_g|\nu_0 + S, H_{0g}) \propto |H_g|^{-\frac{\nu_0+S+2+1}{2}} \exp\left\{ -\frac{1}{2}\text{tr}(H_{0g}H_g^{-1})\right\},$$

$$H_{0g} = H_0 + \sum_{i=1}^{S}(\boldsymbol{\mu}_{gi} - \mathbf{m}_g)(\boldsymbol{\mu}_{gi} - \mathbf{m}_g)'.$$

Step 2 requires sampling from each conditional below in turn:

1. $[\boldsymbol{\mu}, \mathbf{m}|\boldsymbol{\sigma}^2, \mathbf{H}, \mathbf{x}, \mathbf{y}, \gamma_0]$;
2. $[\boldsymbol{\sigma}^2|\boldsymbol{\mu}, \alpha_x, \alpha_y, \beta_x, \beta_y, \mathbf{x}, \mathbf{y}]$(Inverse Gamma);
3. $[\mathbf{H}|\boldsymbol{\mu}, \mathbf{m}]$ (Inverse Wishart);
4. $[\alpha_x|\beta_x, \boldsymbol{\sigma}^2, \gamma_0]$(log-concave);
5. $[\beta_x|\alpha_x, \boldsymbol{\sigma}^2, \gamma_0]$(Gamma);
6. $[\alpha_y|\beta_y, \boldsymbol{\sigma}^2, \gamma_0]$(log-concave);
7. $[\beta_y|\alpha_y, \boldsymbol{\sigma}^2, \gamma_0]$(Gamma).

- For (1) above, use the collapsed Gibbs method (Liu (1994); Chen et al. (2000))
- (1.a) $[\mu|\mathbf{m}, \sigma^2, \mathbf{H}, \mathbf{x}, \mathbf{y}]$ (Normal)
- (1.b) $[\mathbf{m}|\sigma^2, \mathbf{H}, \mathbf{x}, \mathbf{y}, \gamma_0]$ (Normal),
- Identity:

$$[\mu, \mathbf{m}|\sigma^2, \mathbf{H}, \mathbf{x}, \mathbf{y}, \gamma_0] = [\mu|\mathbf{m}, \sigma^2, \mathbf{H}, \mathbf{x}, \mathbf{y}][\mathbf{m}|\sigma^2, \mathbf{H}, \mathbf{x}, \mathbf{y}, \gamma_0].$$

- Fit the Bayesian naïve model to the collapsed cDNA microarray data D_{mean} and the Bayesian hierarchical model to the full cDNA microarray data D.
- $(k_{x_{01}}, k_{x_{02}}, k_{y_{01}}, k_{y_{02}}) = (0.5, 0.1, 0.5, 0.1)$, $w_{x_0} = w_{y_0} = 1000$, $h_0 = 0.0001$.
- Resulting hyper-parameters: $\alpha_{x_{01}} = 1.215$, $\alpha_{x_{02}} = 0.5$, $\alpha_{y_{01}} = 1.203$, $\alpha_{y_{02}} = 0.5$, $\beta_{x_{01}} = 0.143$, $\beta_{x_{02}} = 0.1$, $\beta_{y_{01}} = 0.141$, $\beta_{y_{02}} = 0.1$, $m_{x_0} = -0.219$, $\tau^2_{x_0} = 7303.5$, $m_{y_0} = -0.031$, $\tau^2_{y_0} = 1264.1$, $h_0 = 0.0001$,
$$H_0 = \begin{pmatrix} 7.04 \times 10^{-5} & 1.86 \times 10^{-5} \\ 1.86 \times 10^{-5} & 0.97 \times 10^{-5} \end{pmatrix}.$$

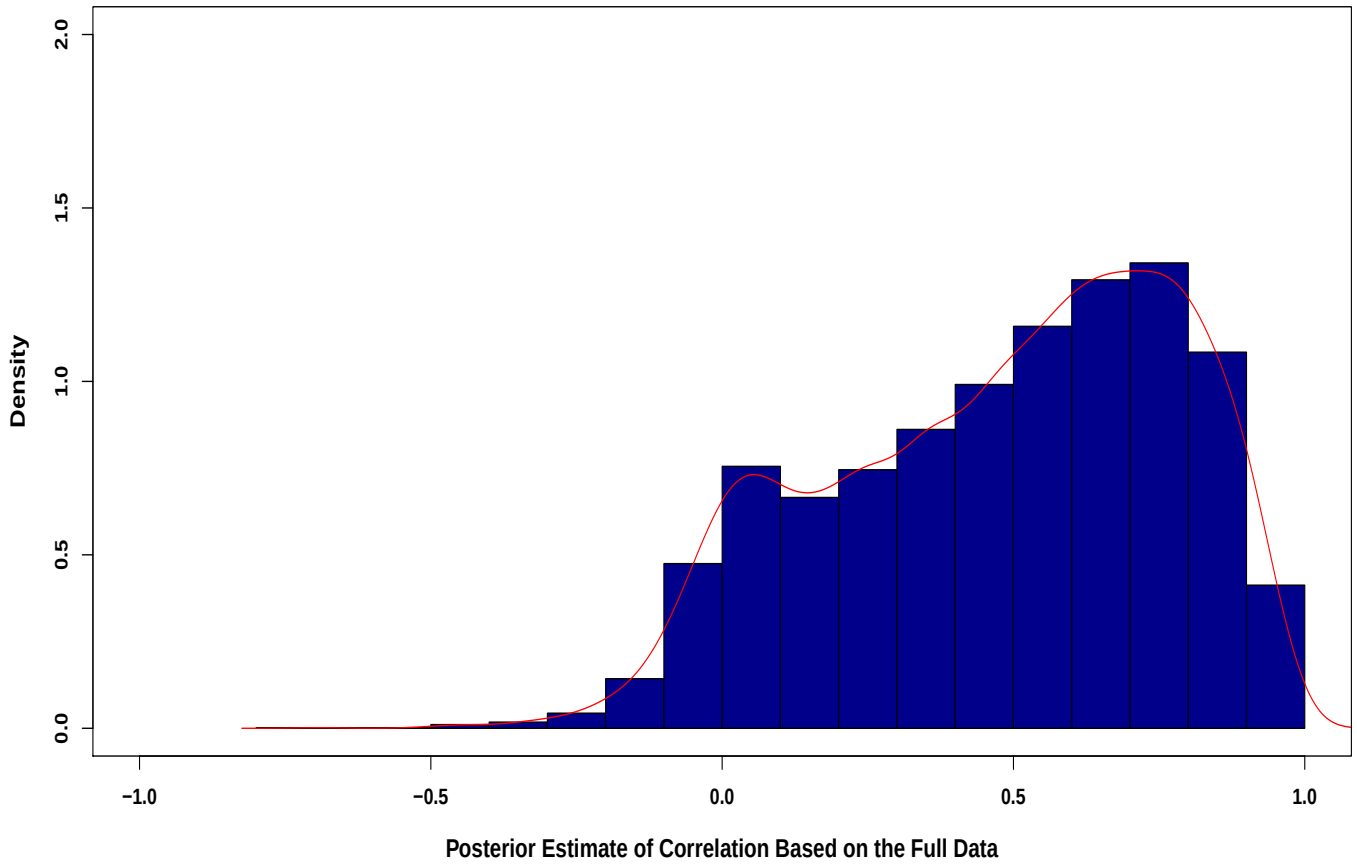


Figure: 5

Posterior Estimates of Correlations based on the Full Data

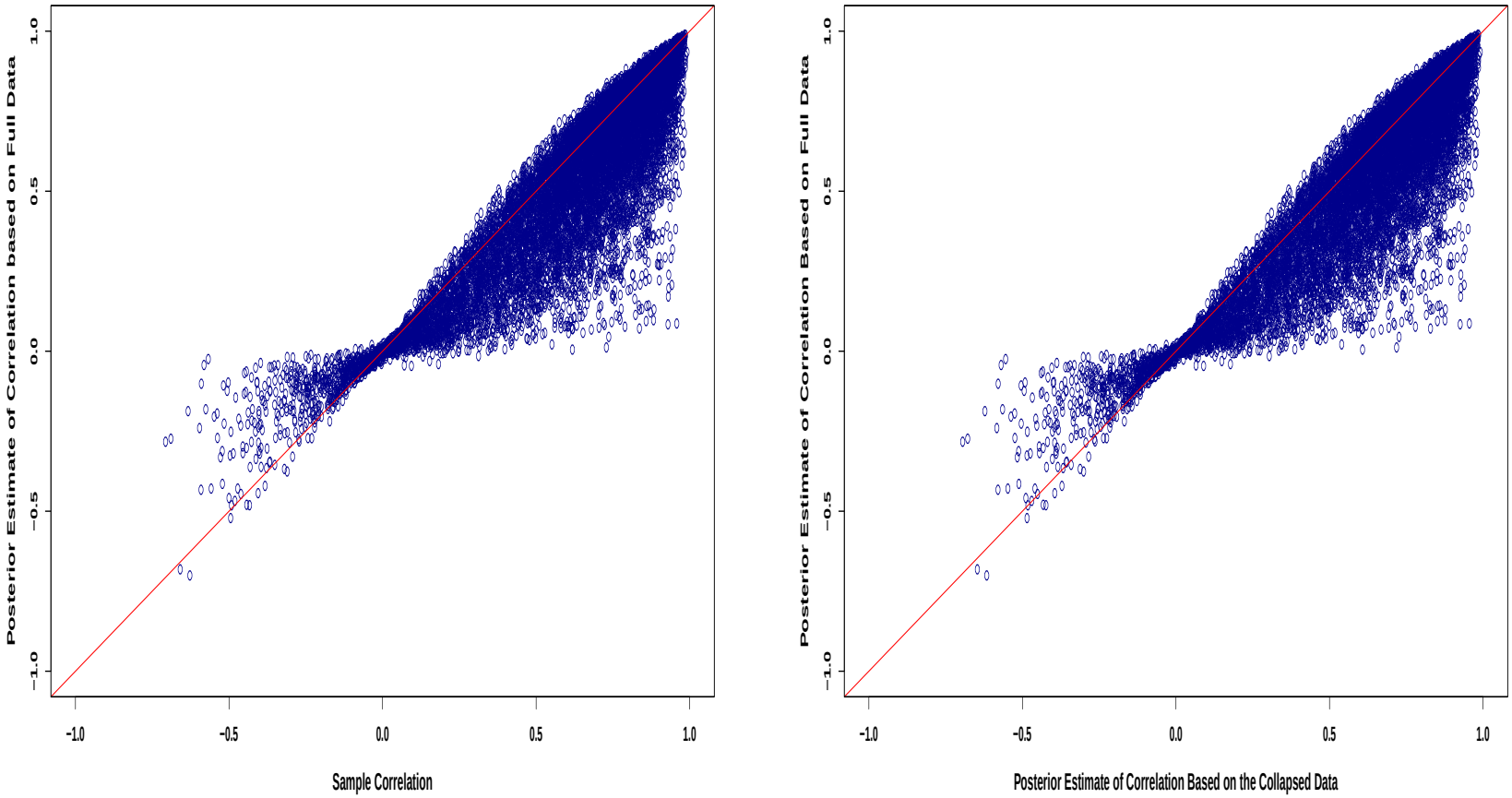


Figure: 6

Post. Corr. Full data vs Sample Correlations (left) and Post. Corr. Collapsed data (right)

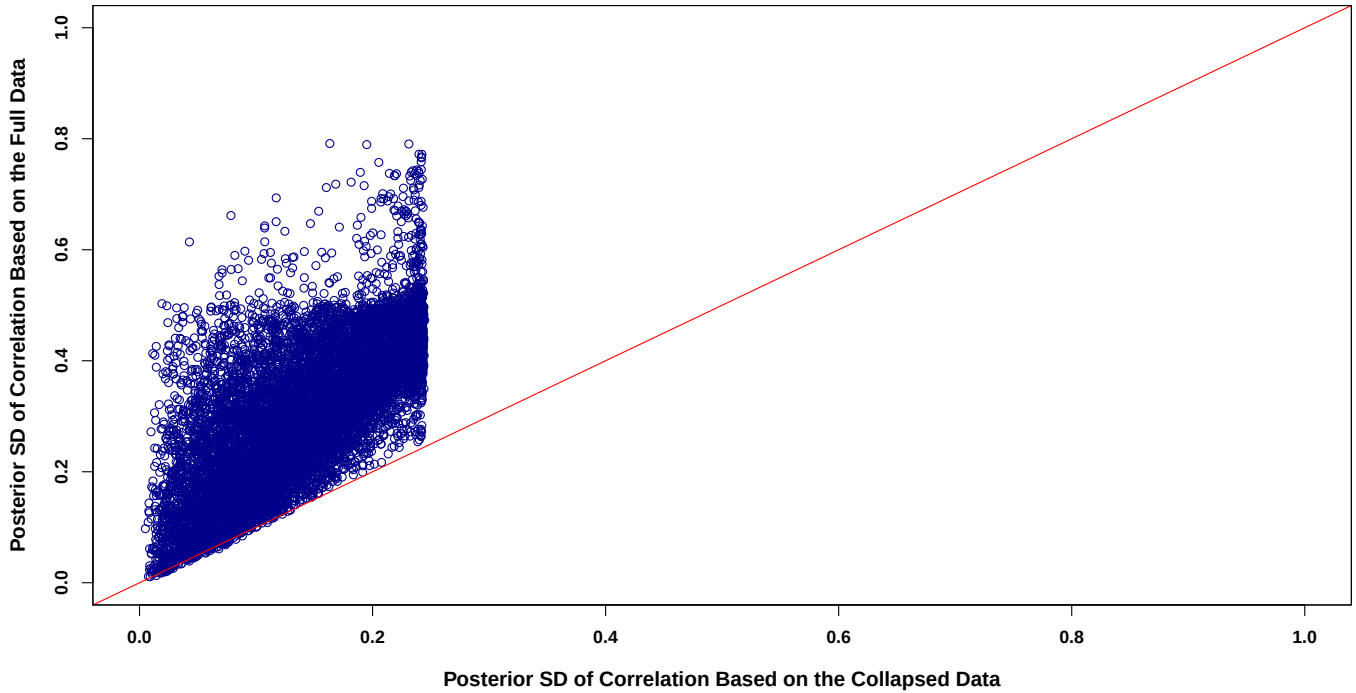


Figure: 7

Posterior Standard Deviations (SDs) for the Full Data vs the Collapsed Data

- The correlation between the sample correlations and the posterior estimates of correlations based on the full data was 0.8980.
- The correlation between the posterior estimates of correlations based on the collapsed data and the posterior estimates of correlations based on the full data was 0.8989.
- The sample correlations or the posterior estimates of correlations based on the collapsed data were larger than those posterior estimates of correlations based on the full data.
- The posterior SDs of gene-specific correlations based on the full data were much larger those based on the collapsed data.
- This result is expected as the mean cDNA expression values had a much smaller variability than the original expression values.

- We have proposed a three-level Bayesian hierarchical model for the gene-specific correlation coefficient between two independent datasets that utilizes replicated expression values for each gene.
- A comparison with a naïve approach indicates that the Bayesian hierarchical model is more appropriate and thus more preferable for differential gene expression analysis.
- The Bayesian hierarchical model allows borrowing strength across genes.
- The analysis of the cDNA microarray data empirically shows that the use of the mean cDNA expression values led to over-estimation of correlations and under-estimation of the variability of the estimates of gene-specific correlations.

- A simulation study is to be conducted.
- As there were more cell-lines in the new dataset, a natural extension of this research is to develop a Bayesian procedure to analyze the new dataset by using the old dataset to elicit an informative prior.
- Another extension will be to develop a mixture model for the correlation coefficients.
- We have focused on the inference of correlation coefficients. It is of practical interest to develop a Bayesian procedure to compare the mean expression levels between two datasets.

References

- Chen et al. (2008), *JSPI*, 138: 387-404
- Chen et al. (2000), New York: Springer
- Gilks and Wild (1992), *App.Stat*, 41: 337-348
- Ibrahim et al. (2002), *JASA*, 97: 88-99
- Ishwaran and James (2001), *JASA*, 96: 161-173
- Kaufmann et al. (2008), *J. Invest. Dermatol.*, 128: 175-187
- Lindley et al. (1972),*JRSSb*, 34: 1-41
- Liu (1994), *JASA*, 89: 958-966
- Scharpf et al. (2009), *JASA*, 104: 1295-1310

- Collaborators:
  - Dr. Joe Ibrahim (Biostatistics, UNC Chapel Hill)
  - Dr. Ming-Hui Chen (Statistics, University of Connecticut)
  - Dr. Haitao Chu (Biostatistics, University of Minnesota)
- Host:
  - CARMS (Strathmore University)

Thanks for your attention!