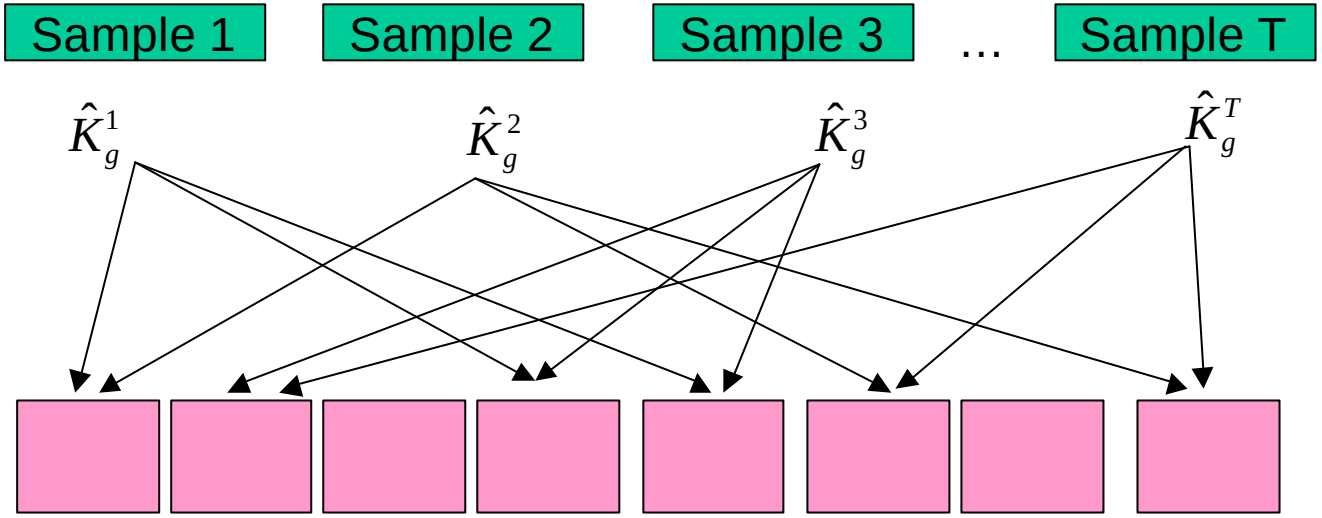
Model-based estimation of transcript concentrations from spotted microarray data

Arnoldo Frigessi, University of Oslo
Mark van de Wiel, Technische Universiteit Eindhoven
Marit Holden, Norwegian Computing Center
Ingrid K. Glad, University of Oslo
Heidi Lyng, The Norwegian Radium Hospital

Norsk Regnesentral
Norwegian Computing Center

Estimating absolute concentrations of mRNA

- Absolute concentrations of mRNA are universal and can be included in further analysis with similar estimates obtained with different techniques in other labs
- A first step towards building an annotated data base of transcript levels of cells



- Estimates $\hat{K}_g^1, \hat{K}_g^2, \hat{K}_g^3, ..., \hat{K}_g^T$ that can be used in other data analyses, together with other preparations, … etc.

Properties/advantages of our method

- Propagating uncertainty
  - Current practice
    - Divide the experiment into separate steps
      - Microarray production
      - Transcription – labelling – hybridisation
      - Image analysis
      - Estimation of intensities
      - Normalisation
      - Imputation
      - Testing, clustering,….
    - Do inference inside each task and *plug-in* results into the next step
  - We do a coherent statistical analysis and propagate uncertainties
- No normalization and imputation needed
  - Model-based normalisation
  - Handling of unbalanced data sets
- No transitivity needed (A-B B-A A-C C-D)
  - But at least one dye-swap or loop is required to estimate the dye effect
- Replications
  - Some genes must be spotted at least twice on some arrays for identifiability

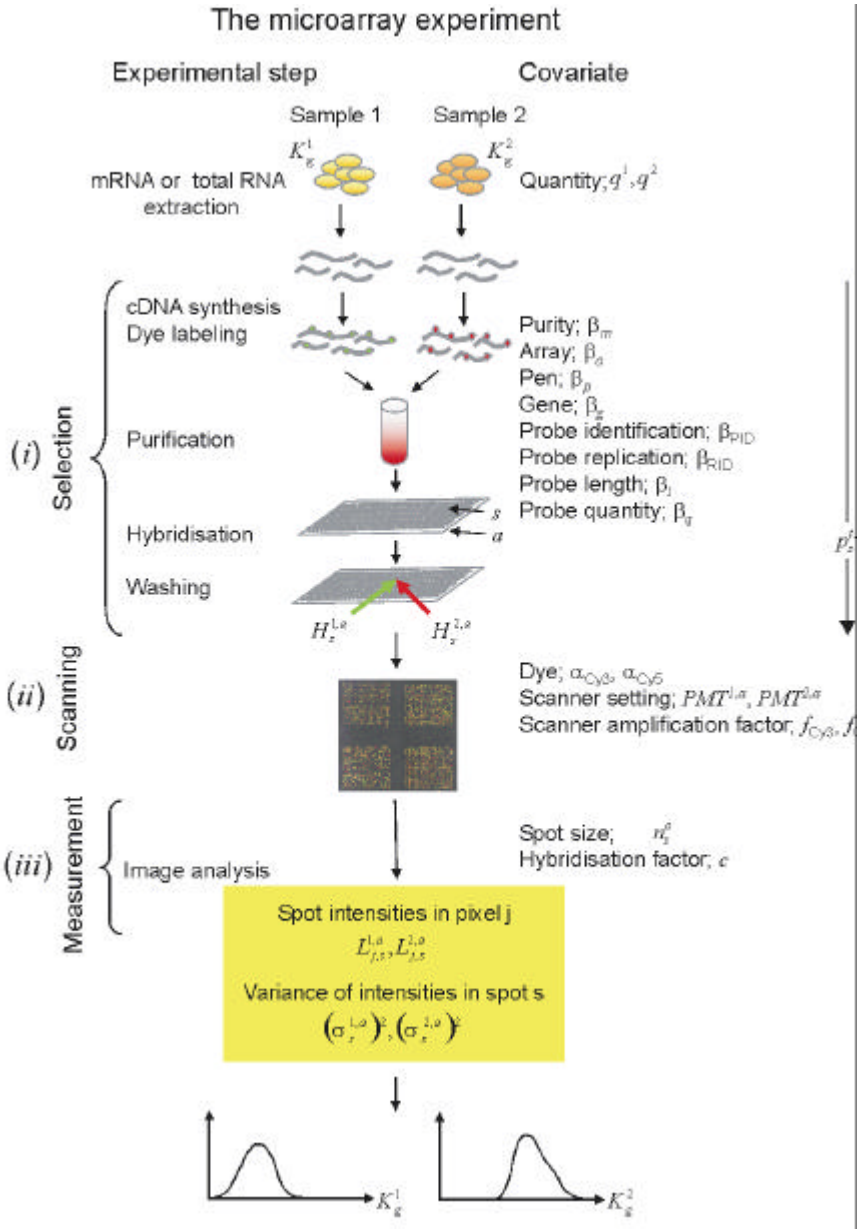
Hierarchical Bayesian approach

- We use available *covariates* describing the various steps of the experiment, from target preparation to laser scanning of the ima ges
- We try to keep the model as close as possible to the biology, ph ysics and bio-chemistry of the experiment

MCMC based inference

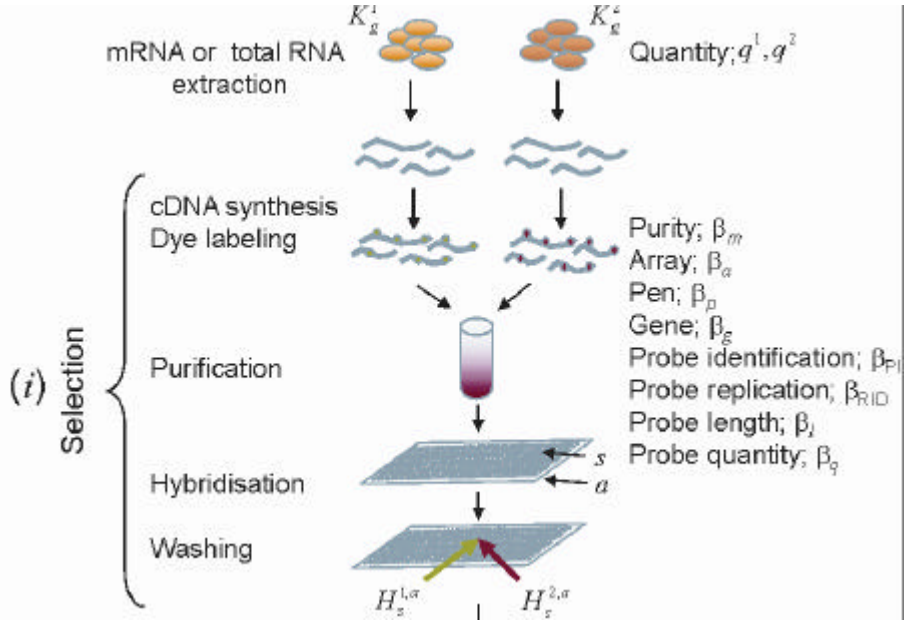
- MCMC converges slowly, as usual in complex models
- We sample full posteriors, and develop new ways of selecting interesting genes, based on absolute transcript levels

The microarray experiment



- We follow the mRNA molecules through the whole experiment.
- At each step, some molecules survive, according to a Binomial process with a success probability depending on appropriate covariates
- At the end, some molecules are scanned, and produce our data, i.e. the raw measured intensities

The selection process



K_g^t - the unknown number of transcripts of gene g per weight unit in sample t

$q^{t,a}$ - the known quantity of material for sample t on array a

n_s^a - the number of pixels in spot s, array a

$H_s^{t,a}$ - molecules of sample t hybridising on spot s, array a

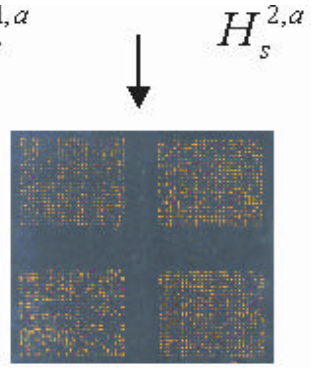
$p_s^{t,a}$ - probability to survive all selections until final washing

- cn_s^a of the $q^{t,a}K_g^t$ molecules candidate to reach the correct spot for hybridisation
- Each of the $cn_s^a q^{t,a} K_g^t$ molecules has a probability $p_s^{t,a}$ to hybridise and survive washing. This happens independently of other molecules.

$$H_s^{t,a} \sim Binomial(cn_s^a q^{t,a} K_g^t, p_s^{t,a}) \qquad \text{(gene g is spotted on spot s)}$$

$$p_s^{t,c} = \min[1, \exp(ß_0 + ß_e + ß_a + ß_p + ß_g + ß_{RID} + ß_{PID} + ß_l \cdot [\text{probe length}] + ß_q \cdot [\text{probe quality}] + ß_m \cdot [\text{purity}_t])]$$

Scanning



Dye; α_{Cy3}, α_{Cy5}
Scanner setting; $PMT^{1,a}$, $PMT^{2,a}$
Scanner amplification factor; f_{Cy3}, f_{Cy5}

Assume, based on laser physics

Setting of the laser scanner

Dye effect

Expected scanned intensity on spot s, array a, sample t

$$m_s^{t,a} = 2^{f_{dye} PMT^{t,a}} H_s^{t,a} a_{dye}$$

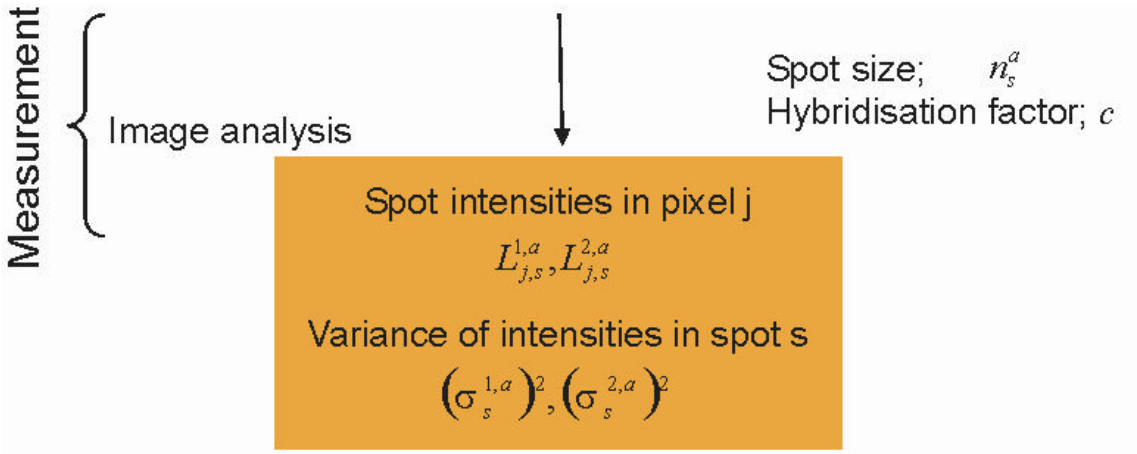
Technology dependent scanner amplification factor

A measure of the increase in intensity per unit of increase in PMT voltage during laser scanning

Determined in off-line experiment from two scans with different PMT values

Molecules of sample t hybridising on spot s, array a

Image analysis



c - used to scale the estimated values to the true number of transcripts

Technology dependent

Determined in off-line experiment

Assume for the pixel-wise intensity measurement

$$L_{j,s}^{t,a} = \frac{m_s^{t,a}}{n_s^a} + e_{j,s}^{t,a} \qquad e_{j,s}^{t,a} \sim Normal(0, (s_s^{t,a})^2)$$

$m_s^{t,a}$ - expected scanned intensity on spot s, array a, sample t

Implementation

- Hyperpriors: flat, not informative
- Identifiability constraints
- MCMC
- Produce posterior point estimates and credibility intervals of absolute concentrations and other interesting quantities
- Reparametrisation necessary
  - Approximate Binomial with Poisson
  - Find the parameters that are identifiable
  - Reparametrise K_g^t to include all other remaining parameters ($\tilde{K}_g^t$)
  - Approximate Binomial with Normal
  - Parameters that were *not* Poisson identifiable do not occur in the expectation, but only in the variance

$$H_s^{t,a} \sim Binomial(cn_s^a q^{t,a} K_g^t, p_s^{t,a})$$

$$\begin{aligned} p_s^{t,c} = \min[1, \exp(&ß_0 + ß_e + ß_a \\ &+ ß_p + ß_g + ß_{RID} + ß_{PID} \\ &+ ß_l \cdot [\text{probe length}] \\ &+ ß_q \cdot [\text{probe quality}] \\ &+ ß_m \cdot [\text{purity}_t])] \end{aligned}$$

$$m_s^{t,a} = 2^{f_{dye} PMT^{t,a}} H_s^{t,a} a_{dye}$$

$$L_{j,s}^{t,a} = \frac{m_s^{t,a}}{n_s^a} + e_{j,s}^{t,a}$$

$$\tilde{K}_g^t = K_g^t \cdot \frac{a_{Cy3}}{a_{Cy5}} \cdot \exp(ß_0 + ß_e + ß_g + ß_m \cdot [\text{purity}_t])$$

Sequential Bayesian procedure

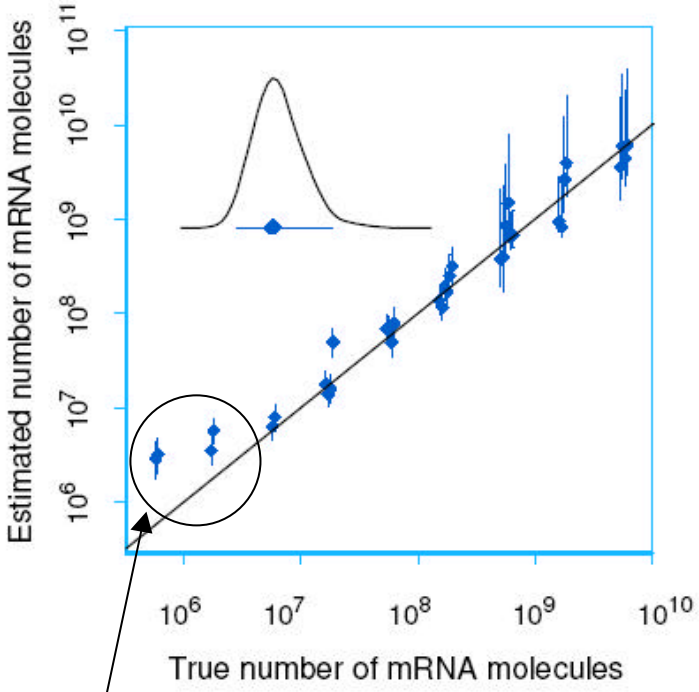
- Some parameters appear in the variance only. When there is just one piece of data (e.g. genes spotted only once, samples hybridised only once), such data must be excluded when variance related parameters are estimated, otherwise estimated variances are shrinked.
- Step 1
  - Drop single data points; estimate all parameters from the remaining data
- Step 2
  - Use only the single data points and the posterior distributions of all parameters as priors to estimate the remaining concentrations
- In practice all done within one MCMC run

Why does this work? A few simple issues

- Estimate both parameters in a binomial
- Do not use single observations to estimate its variance
- Use conditional independence in hierarchical models to model complex dependencies in a flexible way
- Start MCMC runs with *central* initial values

Results – Validation

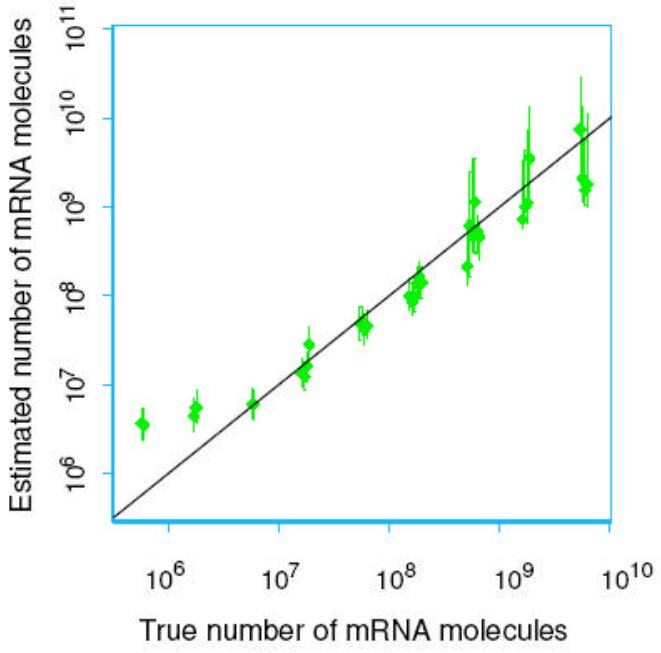
- Validate estimated concentrations in a dye-swap experiment with control samples at known concentrations
- Arrays: 17 genes spotted each 6 times on 2 arrays
- Target: 2 control samples (spikes) each with 17 different mRNA sequences at specific concentrations
- Hybridisation factor = 0.001

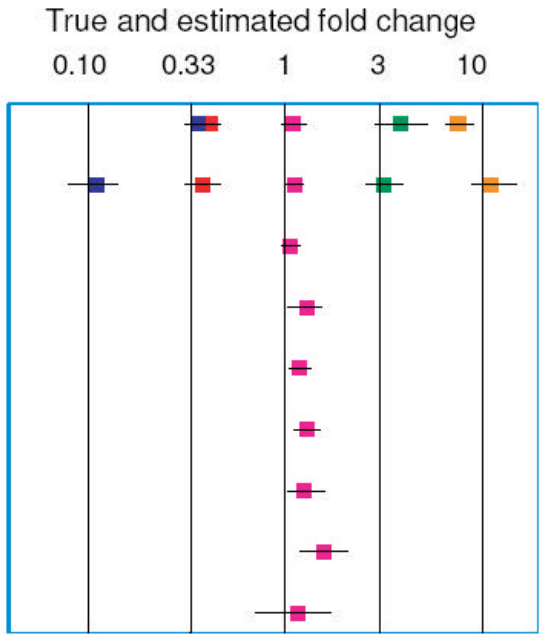


Low concentrations are overestimated

Results – Validation

- A second identical, but independent experiment, with the same hyb. factor
- Systematic underestimation of log10-concentrations by 0.1
  - OK, since scale of concentrations is 6 to 10
- True hybridisation factor = 0.0008





- Estimated fold changes for various genes are reliable

Results – Validation

- Data set with cervix tumour biopsies
- 100 genes; 158 spots per array; 27 duplicated with different probe sequences; 31 duplicated with identical probe sequence; 5 pens. Unbalanced design.
- 4 samples: A, B1 and B2 tumour biopsies; Ref reference
- B1 and B2 are obtained by dividing one biopsy in two
- Loop design
- Posterior estimates of concentrations

Micro-array	Tissue dye Cy5	Tissue dye Cy3
1	Ref	B1
2	B1	B2
3	B2	A
4	A	Ref

	Sample B1	Sample B2	Sample A
Ref	0.258	0.280	0.396
A	0.912	0.928	
B2	0.993		

Gene Number and name		Reference Mode and 95% credibility interval ($*10^6$)		Biopsy B1 Mode and 95% credibility interval ($*10^6$)		Biopsy B2 Mode and 95% credibility interval ($*10^6$)		Biopsy A Mode and 95% credibility interval ($*10^6$)	
1	*ABR*	0.308	(0.18, 0.718)	0.406	(0.227, 0.862)	0.385	(0.225, 0.858)	0.543	(0.293, 1.122)
2	*ARPC2*	0.132	(0.063, 0.258)	0.151	(0.082, 0.338)	0.2	(0.112, 0.417)	0.181	(0.098, 0.357)
3	*B4GALT1*	0.318	(0.145, 0.634)	0.213	(0.118, 0.492)	0.243	(0.118, 0.489)	0.263	(0.125, 0.498)
4	*BCL2A1*	0.059	(0.031, 0.174)	0.077	(0.03, 0.174)	0.092	(0.053, 0.264)	0.127	(0.057, 0.257)
5	*CAPZB*	0.095	(0.044, 0.236)	0.093	(0.046, 0.266)	0.087	(0.044, 0.231)	0.085	(0.039, 0.187)
6	*CASP3*	0.426	(0.176, 1.019)	0.925	(0.393, 2.056)	0.615	(0.282, 1.621)	0.544	(0.265, 1.449)
7	*CASP7*	1.209	(0.695, 3.054)	1.314	(0.739, 2.884)	1.295	(0.767, 3.33)	1.585	(0.96, 3.691)

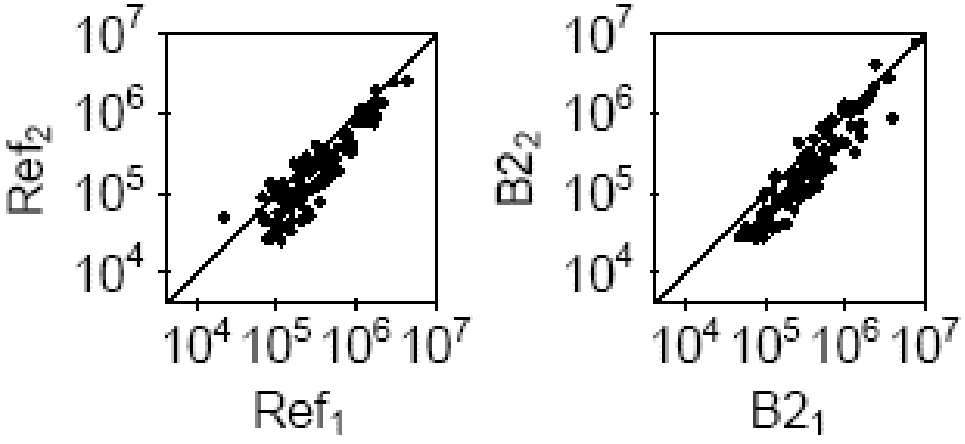
Results – Validation

- Split the data in two separate data sets

Micro-array	Tissue dye Cy5	Tissue dye Cy3	
1	Ref	B1	Data set 1
2	B1	B2	
3	B2	A	Data set 2
4	A	Ref	

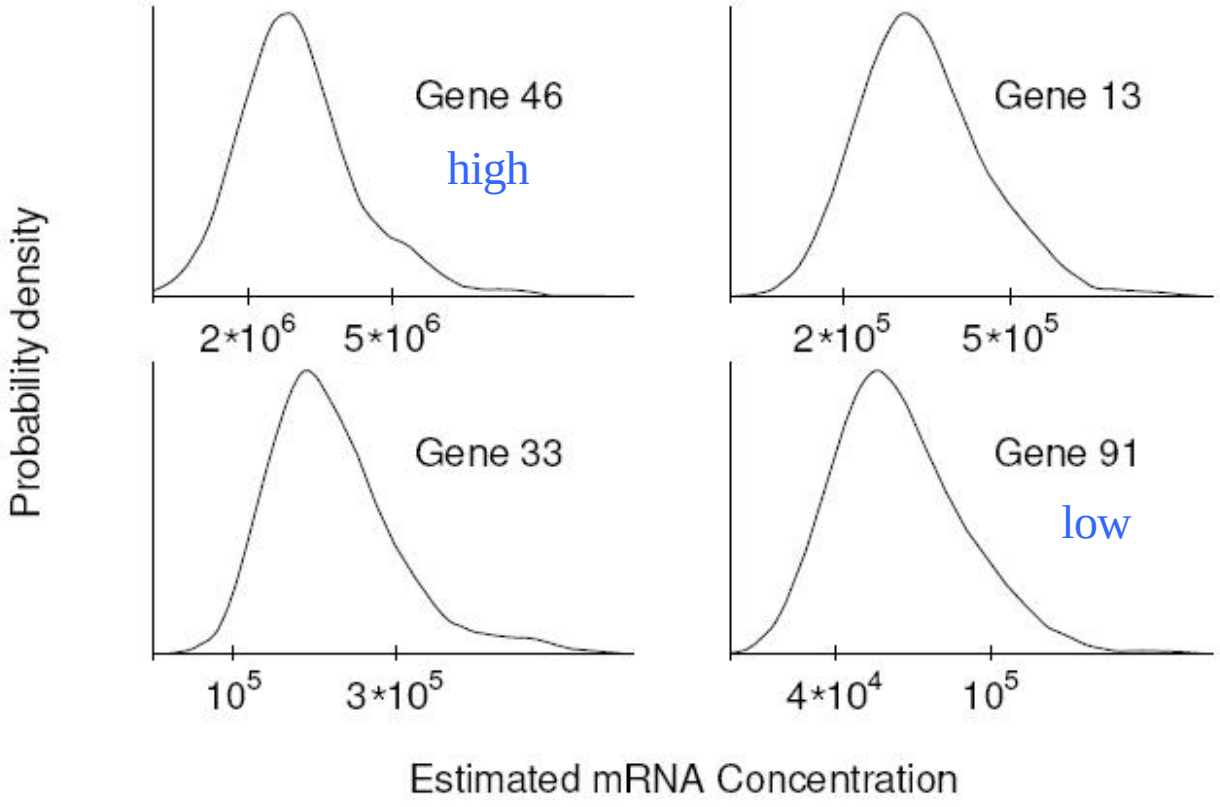
- Analyse separately and compare estimates of same concentrations

- Estimated concentrations can be compared and combined also when originating from different experiments, with no transitivity



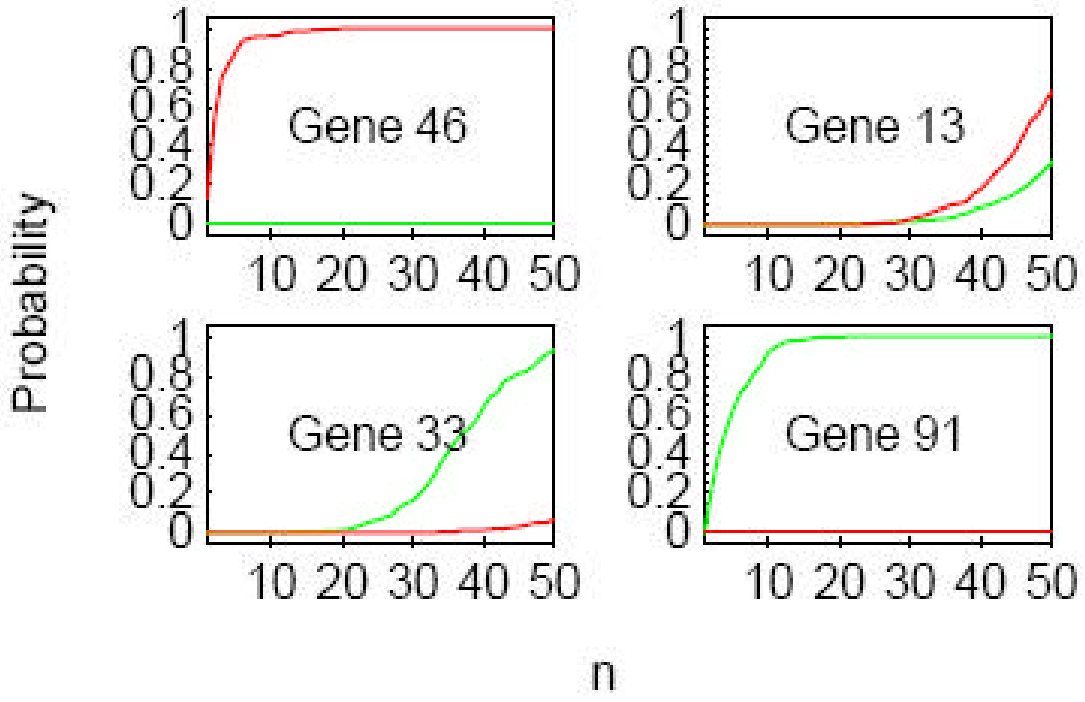
Posterior estimates of the mean mRNA concentration

- What is the estimated mean concentration of a gene in cervix tumour biopsies?
- Assume 3 measurements (A, B1, B2)



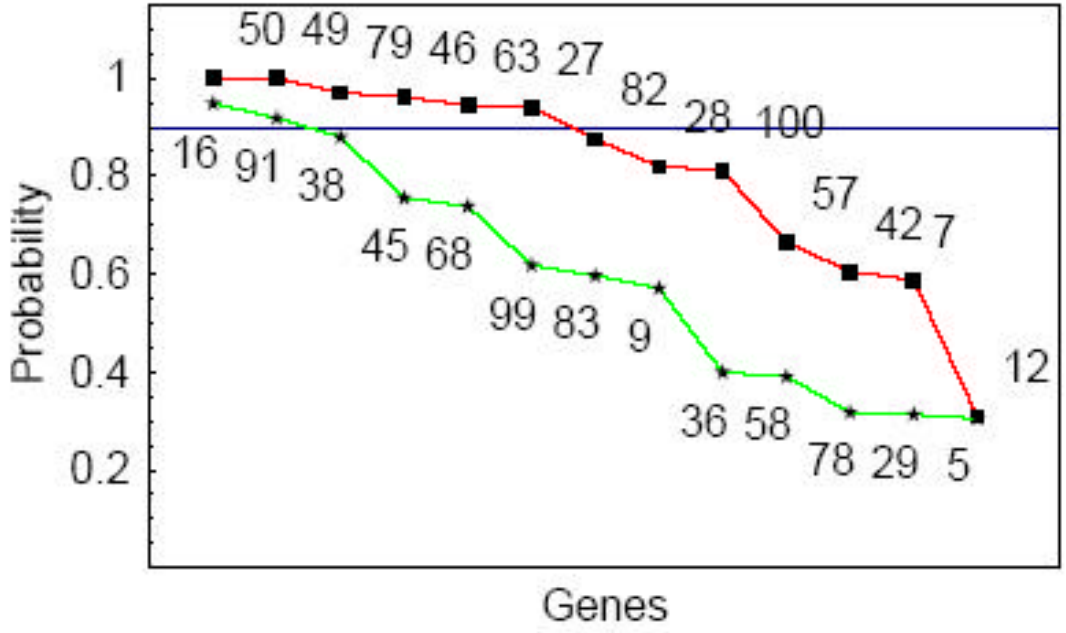
Posterior probability that a gene is top concentrated

- Estimated posterior probability that gene g is among the n genes with highest (lowest) concentrations in cervix cancer



Posterior probability that a gene is top concentrated

- Rank genes according to the estimated posterior probability that they are among the 10% with highest (lowest) concentrations in cervix cancer



Conclusion

Four main ideas:

- We use covariates explicitly
- We treat unequal number of replicates per gene
- We use the binomial process, which better describes the experimental dynamics and allows estimation of gene and dye effects
- We build a bottom-to-top coherent stochastic model, avoiding plug-in's and propagating fully uncertainty

---

Technical report

- Arnoldo Frigessi, Mark A. van de Wiel, Marit Holden, Ingrid K. Glad and Heidi Lyng. *Model-based estimation of transcript concentrations from spotted microarray data.* NR Report 999, ISBN 82-539-0507-6, May 2004.
- http://www.nr.no/files/samba/smbi/Transcount/report999.pdf

TransCount

- A prototype and not-at-all-user-friendly version of the MCMC sampler
- http://www.nr.no/pages/samba/area_emr_smbi_transcount