Model-based estimation of transcript concentrations from spotted microarray data.

Arnoldo Frigessi, Mark A. van de Wiel, Marit Holden, Ingrid K. Glad and Heidi Lyng

Supplemental material

Supplemental Methods

1 Model building

The experimental steps of the microarray experiment (Fig. 1 in the paper), seen from a statistical modelling point of view, are now discussed in turn.

cDNA synthesis and dye labelling.

Dye labelled cDNAs are achieved by incorporation of Cy3-dUTP and Cy5-dUTP during or after cDNA synthesis. The amount of dye and nucleotides are assumed to be in excess, so that all mRNA molecules can in principle be reverse transcribed and labelled. We assume that the expected number of actually bound Cy3- or Cy5-dUTP's is the same for all transcripts of all genes, since the number of binding sites, though different, is always large enough to allow for such a geometric approximation. The expected number of actually bound CyX-dUTP's does however depend on dye, i.e. there is a chemical dye effect. This effect will be important in the imaging step described below.

We assume the $q^{t,a} \cdot K_g^t$ molecules to be reverse transcribed and labelled independently of each other with probability $m_g^{t,a}$. Then, $M_g^{t,a}$, the resulting number of labelled cDNA molecules (or target molecules) for sample t, gene g and array a, follows the binomial distribution with parameters $q^{t,a} \cdot K_g^t$ and success probability $m_g^{t,a}$. The probability $m_g^{t,a}$ can depend on gene and sample specific covariates (like purity of the sample).

Purification.

The two samples are mixed. Excessive CyX-dUTP molecules are washed away. During this process also some of the target molecules will be lost. Let $V_g^{t,a}$ be the number of molecules, each independently remaining with probability $v_g^{t,a}$ in the solution after purification, for sample t, gene g and array a. Then $V_g^{t,a}$ is binomial with parameters $M_g^{t,a}$ and $v_g^{t,a}$. We expect that $v_g^{t,a}$ depends on the target sequence length of gene g, since target length possibly influences purification as longer molecules are less likely to be mistakenly washed away. After purification, the solution will still contain some remaining free CyX-dUTP's that will be washed away after hybridisation. Target length has not been included directly in the current model because target length information was not available. Differences in the $v_g^{t,a}$'s specifically caused by

target length will instead be absorbed in the gene specific covariate (β_g).

Hybridisation.

The variability of probe material and microarray production modulates the probability of successful hybridisation. For a certain spot the microarray, the pen and the probe used influence this probability. Consequently, both array and pen are included as covariates in the model, in addition to probe quantity and quality dependent covariates. Because each of the pens is used on a specific subgrid of the microarray, the pen effect may be confounded with spatial effects.

Quantity of the probe material may vary. A test slide of each printing batch is stained with SYBR green, a fluorophore with specific affinity for ssDNA, (1). The fluorescence intensity is used as an estimate of probe quantity of each spot of the arrays and is included as a covariate in the model. We do not distinguish here between spot center and periphery, assuming for simplicity that each part of a spot is equally covered by probe. Quality of the probe material may also vary. We distinguish two probe quality related covariates; the probe identification (PID) and the replication identification (RID). PID and RID distinguish genes replicated with equal or different probe sequence. PID accounts specifically for the effect of different probes, and RID for replications of equal probe.

We assume that the target is homogeneous, i.e. the spatial distribution of each target molecule is uniform over the slide and target molecules do not cluster nor repulse. Let gene(s) be the gene spotted in spot s. A proportion $c \cdot n_s^a$ of $V^{t,a}_{\text{gene}(s)}$ reaches the correct spot to candidate for hybridisation. Let $Q_s^{t,a}$ be the number of gene(s) molecules of sample t succeeding in hybridising to spot s, in array a. Each target molecule has a probability $q_s^{t,a}$ to independently hybridise. Then $Q_s^{t,a}$ is binomial with parameters $c \cdot n_s^a \cdot V^{t,a}_{\text{gene}(s)}$ and $q_s^{t,a}$. The success probability $q_s^{t,a}$ depends on probe properties and technical experimental conditions as well as on target properties. The first two classes include probe quantity, probe length, PID, RID, pen and array. Target length influences the diffusion coefficient of target molecules and could have been included here also, if available. Hybridisation is assumed to be dye independent (2) and the hybridisation probability is assumed to be constant in time. The model does not include cross-hybridisation.

Washing.

We assume that all non-hybridised material, including unbound CyX-dUTPs, is removed during microarray washing. Again we assume that the number of remaining molecules $H_s^{t,a}$ is binomial with parameters $Q_s^{t,a}$ and success probability $h_s^{t,a}$, which may depend on probe length, reflecting the binding strength, and on microarray effects. $H_s^{t,a}$ is the number of gene(s) molecules of sample t hybridised in spot s in array a, participating in the imaging process.

Scanning and image analysis.

The image achieved during scanning is gridded and segmented into spots. Each measured pixel intensity, $L_{j,s}^{t,a}$, for sample t, array a and pixel j of spot s depends on $H_s^{t,a}$, on the PMT voltage used during scanning and on a known scanner dependent amplification factor. In addition the measured intensity depends on whether sample t was labelled with Cy3- and Cy5-dUTP's. This dye dependency has both chemical and optical reasons. As described previously, the expected number of actually bound CyX-dUTP's might be different for Cy3-

or Cy5-dUTP's, i.e. there is a chemical dye effect. The optical dye effect is present because of different optical characteristics of the two dyes.

The model.

Nesting all bionomial variables results in the bionomial model presented in the paper. The success probabilities depend on covariates from all the steps mentioned above. All effects are estimated together with the unknown K_g^t's and $H_s^{t,a}$'s.

2 Reparametrisation of the model, identifiability, constraints and hyper-priors

Recall the four levels of our model:

$$\begin{aligned}
H_s^{t,a} &\sim \text{Binomial}(c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t, p_s^{t,a}),\\
p_s^{t,a} &= \max[1, \exp\{\beta_0 + \beta_e + \beta_a + \beta_p + \beta_g + \beta_{\text{RID}} + \beta_{\text{PID}}\\
&\quad + \beta_l \cdot [\text{probe length}] + \beta_q \cdot [\text{probe quantity}] + \beta_m \cdot [\text{purity}_t]\}],\\
\mu_s^{t,a} &= 2^{f_{\text{dye}}\,\text{PMT}^{t,a}} H_s^{t,a} \alpha_{\text{dye}},\\
L_{j,s}^{t,a} &= \frac{\mu_s^{t,a}}{n_s^a} + \varepsilon_{j,s}^{t,a}, \quad \varepsilon_{j,s}^{t,a} \sim \text{Normal}(0, (\sigma_s^{t,a})^2).
\end{aligned}$$

Introduce

$$\bar{\beta} X_s^a = \beta_a + \beta_p + \beta_{\text{RID}} + \beta_{\text{PID}} + \beta_l \cdot [\text{probe length}] + \beta_q \cdot [\text{probe quantity}],$$

so that

$$p_s^{t,a} = \max[1, \exp\{\beta_0 + \beta_e + \beta_g + \beta_m \cdot [\text{purity}_t] + \bar{\beta} X_s^a\}].$$

In a classical likelihood context all parameters must be identifiable, while in the Bayesian setting flat posterior densities correspond to model misspecification or lack of information in the data on parameters. MCMC convergence is then particularly slow. We require classical identifiability of the parameters and discuss now how the parameters can be estimated within our framework. We will show identifiability under the relaxed inverse link function $\exp(x)$ instead of $\max(1, \exp(x))$, which is used in practice, and assuming all parameters are fixed. This is allowed, because the censored inverse link function that we use and non-flat priors on the fixed effects (which is the Bayesian way of introducing a random effect) only restrict the size of the parameter space.

For computational purposes, it is useful to approximate binomials with normal densities:

$$H_s^{t,a} \sim \text{Normal}\big(c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t \cdot p_s^{t,a}, c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t \cdot p_s^{t,a} \cdot (1 - p_s^{t,a})\big).$$

Next, we reparameterise in such a way that parameters not estimable based on the mean alone do not occur in the mean, but only in the variance. We have

$$E[L_{j,s}^{t,a}] = E[\mu_s^{t,a}]/n_s^a = C^{t,a} \cdot K_g^t \cdot \alpha_{\text{dye}} \exp(\beta_0 + \beta_e + \beta_g + \beta_m \cdot [\text{purity}_t]) \exp(\bar{\beta} X_s^a),$$

where $C^{t,a}$ is a product of known constants. Then, let

$$\alpha_{\text{dye}} = \alpha'_{\text{dye}} \alpha$$

where $\alpha'_{\text{Cy5}} = 1$, and α'_{Cy3} and α are the new parameters to be estimated, replacing α_{Cy3} and α_{Cy5}. In addition $\tilde{H}$'s and $\tilde{K}$'s replace the H's and K's, where the $\tilde{H}$'s and $\tilde{K}$'s are defined as follows

$$\tilde{H}_s^{t,a} = H_s^{t,a} \cdot \alpha$$

$$\tilde{K}_g^t = K_g^t \cdot \alpha \, \exp(\beta_0 + \beta_e + \beta_g + \beta_m \cdot [\text{purity}_t]).$$

Then, we observe that

$$\tilde{H}_s^{t,a} \sim \text{Normal}\Big(c \cdot n_s^a \cdot q^{t,a} \cdot \tilde{K}_g^t \cdot \exp(\bar{\beta} X_s^a), \\ c \cdot n_s^a \cdot q^{t,a} \cdot \tilde{K}_g^t \exp(\bar{\beta} X_s^a)\big(1 - \exp(\beta_0 + \beta_e + \beta_g + \beta_m \cdot [\text{purity}_t] + \bar{\beta} X_s^a)\big) \cdot \alpha \Big).$$

Since $E[L_{j,s}^{t,a}] = C^{t,a} \cdot \tilde{K}_g^t \alpha'_{\text{dye}} \exp(\bar{\beta} X_s^a)$, all parameters except β_0, β_e, β_m, the β_g's and α are estimable based on the mean pixel-wise values with the described reparametrisation, when the regression of this mean on the covariates X_s^a is identifiable. This can be guaranteed by some constratints (see below) and with a design which has the following characteritics: some genes must be spotted at least in duplicate, with different pens for some of these replicates, and the whole data set must include at least one loop, i.e. a self-self array or a dye swap or a longer chain, to identify the parameters α'_{dye}, β_a and β_p.

The parameters β_0, β_e, β_m, α and the β_g are estimable from the variances and none of these occur in the expressions for the mean. Some care is required to handle the special situation of samples hybridised only once on one array. This happens for example in reference designed studies. Since there is just one piece of data relative to such samples for non-repeated genes, these data must be excluded when inference on variance related parameters is perfomed, since otherwise estimated uncertainties of the concentrations will be shrinked. We operate as follows: First we exclude all such single data points and estimate all parameters on the rest of the data. In a reference design, this corresponds to using all data of the reference and all data from the samples for repeated genes. We then use the posterior distribution of all parameters as prior in the second phase, where we consider only the rest of the data, those corresponding to samples and genes measured only ones. We thus obtain the correct estimates for all concentrations, equipped with the coherently propagated uncertainty. In practice, all is performed within MCMC: sampled values from the posterior distribution of all parameters given the repeatedly observed samples are used in the model for the uniquely observed data. This second phase is not necessary in loop designs or when dye swaps are included. Finally, transcript concentration K_g^t is estimated using the estimates of $\tilde{K}_g^t$, β_0, β_e, β_m, α and β_g by inverting the formula above.

We need to constrain the categorical parameters for identifiability. In order to assure identifiability of the pen parameters, we use the constraint $\sum \beta_p = 0$, where the summation runs over the P different values β_p may attain, when P different pens are used. A similar constraint is used for the parameter β_e describing the effect of different non-transitive experiments and the gene related parameters β_g (to ensure identifiability together with β_0). For each set of experiments e, $\sum \beta_a = 0$, where the sum runs over all arrays of the transitive set e. Moreover, we restrict the mean effect of all probes per gene to be zero which is achieved by applying the constraints $\sum \beta_{PID} = 0$, for all genes, where summation runs over all probes in the probe set of the particular gene. Similarly, we constrain $\sum \beta_{RID} = 0$, for all probes, where summation runs

over all replicates for the particular probe. In addition to these constraints we consider experiment (β_e), array (β_a), pen (β_p), gene-dependent selection (β_g), probe identification (β_{PID}) and replication identification (β_{RID}) as random effects. Then, we have $\beta_e \sim \text{Normal}(0, (\sigma_e)^2)$, $\beta_a \sim \text{Normal}(0, (\sigma_a)^2)$, $\beta_p \sim \text{Normal}(0, (\sigma_p)^2)$ and $\beta_g \sim \text{Normal}(0, (\sigma_g)^2)$. Since the number of probe products per gene is usually small, we do not use separate random effects for each gene, but instead we have $\beta_{\text{PID}} \sim \text{Normal}(0, (\sigma_{\text{PID}})^2)$ for all probe sequences. Similarly, we have $\beta_{\text{RID}} \sim \text{Normal}(0, (\sigma_{\text{RID}})^2)$ for all replications of probe. Otherwise, all hyper-parameters are equipped with flat improper non-informative priors.

The identifiability of all parameters, including the transcript concentrations K_g^t, assures that two experiments that both satisfy the identifiability conditions above can be combined, *even* when they do not share a sample and hence non-transitive designs are allowed.

On the link function We have discussed reparametrisation and identifiability under the inverse link $\exp(x)$. Within an MCMC context this can easily be adapted to a censored inverse link by not accepting proposals for which $\exp(\beta_0+\beta_e+\beta_g+\beta_m\cdot[\text{purity}_t]+\bar{\beta}X_s^a) > 1$. One might be inclined to use the proportional, non-censored inverse link instead of the censored inverse link. However, use of the censored inverse link conserves the complicated non-proportional effects of factors. We know that these exist in the original formulation with an intercept, because mapping a linear combination of factors to a number between 0 and 1 implies non-proportionality.

Overdispersion The covariates in the model should explain the selection probability as good as possible. However, some explanatory factors might be missing from the model. Moreover, individual molecules may not have completely identical selection probabilities due to differences on a molecular level. We can allow for overdispersion by adding variability to the selection probability using a spot, array and sample varying random model error $\epsilon_s^{t,a}$ with distribution $\text{Normal}(0, \sigma^2)$. We note, however, that estimation of σ^2 together with the other variance related parameters may lead to slow convergence of the MCMC. In the human cervical tumour study we included overdispersion in order to confirm that identifiability is maintained also in this case. In the validation example no overdispersion was included.

Competition We have not included competition among molecules in our model for hybridisation. This is possible to do in terms of density dependence, for example by adding the term $\beta K^t_{\text{gene}(s)}$ in the log-probability. Then, we expect β to be negative: the larger $K^t_{\text{gene}(s)}$, the more competition and hence the smaller the probability to hybridise.

Other Bayesesian microarray studies For more examples of Bayesian inference for statistical models of gene expression data we refer to Baldi and Hatfield (3) and references therein. None of these deal with absolute concentrations.

3 Initial values and proposal functions: Details on MCMC

The marginal posteriors of interest are not available in closed form and so we use Markov Chain Monte Carlo to sample from the posterior model. Specifically, we implement a single-

update random-walk Metropolis-Hastings sampler. Convergence is difficult to monitor (4) and we used very long chains, started after burn-in with different random seeds, and observed convergence to the same posterior parameter densities. A block-updating strategy might improve convergence. For all the model parameters we use a uniform proposal. More precisely, let v be the current value of the parameter p for which a new value will be proposed, and let $c_{p,0}$ and $c_{p,1}$ be two constants. If the parameter is not restricted to be positive, draw from

$$U[v - c_{p,1}\sigma_p, v + c_{p,1}\sigma_p]$$

if the prior for the parameter is Normal$(0, \sigma_p^2)$, otherwise draw from

$$U[v - (c_{p,1}|v| + c_{p,0}), v + (c_{p,1}|v| + c_{p,0})].$$

If the parameter is restricted to be positive, draw the logarithm of the parameter from

$$U[log(v) - (c_{p,1}log(v) + c_{p,0}), log(v) + (c_{p,1}log(v) + c_{p,0})].$$

The two constants for each parameter p, $c_{p,0}$ and $c_{p,1}$, were tuned such that reasonable acceptance rates were obtained, between 0.2 and 0.5. After reparametrisation the parameters to be estimated are α'_{Cy3}, α, the β's, the overdispersion ε's, σ, variances of the random effects, the $\tilde{H}$'s and the $\tilde{K}$'s. Initial parameter estimates for α'_{Cy3}, the β's (except for β_0, β_m, the β_e's and the β_g's), the overdispersion ε's and σ are found from the data using linear regression. The variances of the random effects, the $\tilde{H}$'s and the $\tilde{K}$'s are then computed from these estimates. In the computations of all these initial estimates we use formulas where all random variables are set equal to their expectations. The parameters $\beta_0, \beta_m, \beta_e$'s and the β_g's are initialised such that for each gene g, the geometric mean of the selection probabilities $p_s^{t,a}$ becomes 0.5. Finally, α is set equal to the geometric mean of

$$(\tilde{H}_s^{t,a} - c \cdot n_s^a \cdot q^{t,a} \cdot \tilde{K}_g^t \cdot \exp(\bar{\beta} X_s^a))^2 / (c \cdot n_s^a \cdot q^{t,a} \cdot \tilde{K}_g^t \cdot \exp(\bar{\beta} X_s^a) \cdot (1 - p_s^{t,a})).$$

Details on the MCMC, such as the number of iterations, are available here:

http://www.nr.no/pages/samba/area_emr_smbi_transcount.

References

1. Battaglia, C., Salani, G., Bernardi, L. R., and De Bellis G. Analysis of DNA microarrays by non-destructive fluorescent staining using SYBR green II. *Biotechniques* 29, 78 - 81 (2000).

2. Wang, Y., Wang, X, Guo, S.-W. and Ghosh, S. Conditions to ensure competitive hybridization in two-color microarray: a theoretical and experimental analysis. *Biotechniques* 32, 1342 - 1346 (2002).

3. Baldi, P. and Hatfield, G.W. *DNA microarrays and gene expression - From experiments to data analysis and modeling.* Cambridge University Press (2002).

4. Frigessi, A., Martinelli, F. and Stander, J. Computational complexity of Markov Chain Monte Carlo methods for finite Markov Random Fields. *Biometrika* 84, 1 - 18 (1997).

Supplemental Materials

Array Slides. cDNA microarray slides were produced at the cDNA Microarray Facility at The Norwegian Radium Hospital (http://www.mikromatrise.no). The probes were human cDNA clones, derived from the I.M.A.G.E. Consortium (http://image.llnl.gov) or locally prepared, amplified by PCR and printed to Corning CMT GAPS slides (Corning) by using a Microgrid II printing robot (BioRobotics) with 32 pens. Each array contained 18432 spots printed in 32 subarrays. Some probes were printed in duplicate with different pens, and some probes with different cDNA sequence representing the same genes. Probe length ranged from 525 to over 2000 base pairs; in this latter case, 2000 was used as covariate value in our models. Furthermore, for validation of our method, seventeen DNA control samples (Lucidea Universal ScoreCard, Amersham Biosciences) were printed in equal amount on six of the subarrays.

Sample Preparation and Hybridisation. Validation was performed adding two control samples, each containing 17 different mRNA sequences, pre-mixed at specific concentrations (Lucidea Universal ScoreCard). 0.5 μl of each sample was used, corresponding to a number of transcripts in the range of $5.8 \times 10^5 - 5.8 \times 10^9$. The concentration ratios achieved when hybridising the two samples together were 1:1, 1:3, 3:1, 1:10, and 10:1 at high and low level concentrations. The control samples were prepared as described by the manufacturer and subjected to cDNA synthesis and dye labelling as described below. The labelled samples were hybridized together in a dye-swap design. Hybridisation was performed overnight at 65^oC by use of Genetac hybridisation station (Perkin Elmer).

Furthermore, two tumour biopsies (A, B) and a reference sample (Ref) of total RNA (Stratagene) were used. The reference sample was pooled from ten human cell lines. The biopsies, $(5\ \text{mm})^3$ in size, were from two different locations in a human cervical tumour. Biopsy B was divided into two pieces (B1, B2) before isolation of total RNA. Total RNA was isolated from the biopsies using Trizol reagent (Invitrogen) and the recommended protocol. Fifty to sixty μg of total RNA were used to produce labelled cDNA by anchored oligo(dT)-primed reverse transcription, using SuperScript II reverse transcriptase (Invitrogen) and either Cy3-dUTP or Cy5-dUTP (Amersham Pharmacia). The labeled samples were hybridized in a loop design overnight in water bath at 65^oC (Table 1). Purity was optimal and equal for all samples in our experiments and was therefore not used in our models.

Scanning and Image Analysis. The slides were imaged at a resolution of 10 μm using an Agilent G2565BA scanner (Agilent Technologies) for slides with control samples and a ScanArray4000 scanner (GSI Lumonics) for slides with biopsies and reference. A laser power of 100% was used. The PMT voltage was adjusted for the red and green channel individually to ensure that the intensity of the weakest spots and background segments were within the linear range of the scanner. Saturated spot intensities were corrected using the algorithm described previously in Lyng et al (2004)(reference (15) in the paper). The GenePix 3.0 image analysis software (Axon Instruments) was used for spot segmentation and intensity calculation. Bad spots and regions with high unspecific binding of dye were manually flagged and excluded from the analysis.

Supplemental Table 2: Parameter estimates

We considered 100 genes in 158 spots of each array. Totally, there are 127 β_{PID}'s, since 27 genes were duplicated with different probe sequence. Because of the contraints put on the β_{PID}'s 73 of these are set to zero. The other 54 are divided into pairs which are constrained such that the sum of the β_{PID}'s for each pair is zero. In the table the estimate for one β_{PID} from each of the 27 pairs is given. Similarly, there are 158 β_{RID}'s, since 58 genes were duplicated either with different (27) or with equal (31) probe sequence. Because of the contraints put on the β_{RID}'s 96 of these are set to zero. The other 62 are divided into pairs which are constrained such that the sum of the β_{RID}'s for each pair is zero. In the table the estimate for one β_{RID} from each of the 31 pairs is given. In the paper the probe length effect was discussed. We see that it is in the same scale as the array effect and hence contributes similarly to the selection probability. The probe lengths have been scaled to zero mean and standard deviation one, making the β's comparable. β_m has not been included in the model because purity was optimal and equal for all samples. There is no experiment effect, $\beta_e = 0$, since this experiment was transitive.

Parameter	Mode	95% Credibility Interval	Parameter	Mode	95% Credibility Interval	Parameter	Mode	95% Credibility Interval
β_0	-2.343	(2.614, -2.074)	β_g, 47	0.002	(-0.301, 0.341)	β_{PID}, 6	-0.149	(-0.48, 0.143)
α	1.589	(1.195, 2.102)	48	0.004	(-0.345, 0.322)	7	-1.604	(-1.938, -1.226)
α'_{Cy3}	0.365	(0.355, 0.383)	49	0.01	(-0.352, 0.336)	8	0.199	(-0.262, 0.629)
β_l	-0.17	(-0.335, 0.036)	50	-0.01	(-0.32, 0.287)	9	-0.266	(-0.614, 0.169)
β_q	0.254	(0.106, 0.431)	51	-0.009	(-0.167, 0.617)	10	0.172	(-0.07, 0.481)
β_a,1	-0.535	(-0.585, -0.481)	52	-0.024	(-0.353, 0.266)	11	0.141	(-0.214, 0.517)
2	-0.041	(-0.086, 0.005)	53	-0.017	(-0.333, 0.359)	12	0.222	(-0.081, 0.596)
3	0.178	(0.133, 0.227)	54	0.022	(-0.321, 0.326)	13	0.216	(-0.084, 0.623)
4	0.396	(0.347, 0.44)	55	0.014	(-0.346, 0.339)	14	0.206	(-0.083, 0.525)
β_p,1	0.011	(-0.066, 0.342)	56	-0.007	(-0.316, 0.321)	15	0.061	(-0.335, 0.414)
2	0.001	(-0.124, 0.309)	57	0.022	(-0.351, 0.316)	16	1.155	(0.789, 1.633)
3	0.001	(-0.253, 0.111)	58	0.003	(-0.281, 0.315)	17	-0.086	(-0.363, 0.27)
4	-0.018	(-0.354, 0.037)	59	0.011	(-0.302, 0.283)	18	-1.366	(-1.607, -1.096)
5	-0.003	(-0.152, 0.128)	60	-0.01	(-0.316, 0.354)	19	0.45	(0.133, 0.763)
β_g,1	-0.012	(-0.374, 0.251)	61	0	(-0.288, 0.337)	20	0.45	(0.177, 0.676)
2	0.003	(-0.295, 0.337)	62	0.022	(-0.305, 0.327)	21	0.159	(-0.253, 0.479)
3	0.013	(-0.363, 0.288)	63	0.015	(-0.293, 0.299)	22	-0.559	(-0.885, -0.195)
4	-0.017	(-0.362, 0.316)	64	0.011	(-0.288, 0.327)	23	-0.475	(-0.817, -0.128)
5	-0.018	(-0.33, 0.385)	65	0.015	(-0.325, 0.35)	24	-0.395	(-0.734, -0.076)
6	-0.03	(-0.35, 0.34)	66	-0.005	(-0.37, 0.36)	25	1.267	(0.958, 1.618)
7	0.011	(-0.35, 0.333)	67	0	(-0.39, 0.284)	26	0.041	(-0.256, 0.351)
8	0.016	(-0.382, 0.327)	68	-0.007	(-0.344, 0.296)	27	-0.233	(-0.544, 0.069)
9	0.008	(-0.363, 0.333)	69	0.008	(-0.326, 0.36)	β_{RID}, 1	0.09	(-0.189, 0.348)
10	0.017	(-0.357, 0.327)	70	-0.011	(-0.335, 0.374)	2	0.379	(0.087, 0.676)
11	-0.002	(-0.342, 0.333)	71	0.018	(-0.338, 0.401)	3	-0.127	(-0.364, 0.172)
12	-0.003	(-0.352, 0.302)	72	-0.01	(-0.552, 0.083)	4	0.136	(-0.171, 0.408)
13	-0.009	(-0.36, 0.329)	73	0.016	(-0.367, 0.233)	5	0.199	(-0.054, 0.472)
14	0.003	(-0.315, 0.384)	74	-0.006	(-0.366, 0.3)	6	-0.522	(-0.82, -0.152)
15	-0.016	(-0.309, 0.294)	75	0.008	(-0.295, 0.238)	7	0.569	(0.134, 0.881)
16	-0.014	(-0.319, 0.368)	76	0.028	(-0.391, 0.34)	8	-1.19	(-1.515, -0.75)
17	-0.003	(-0.441, 0.286)	77	0.022	(-0.31, 0.348)	9	0.109	(-0.144, 0.387)
18	0.01	(-0.325, 0.374)	78	0.002	(-0.355, 0.275)	10	-0.356	(-0.72, 0.039)
19	-0.007	(-0.319, 0.354)	79	-0.005	(-0.391, 0.274)	11	-1.022	(-1.342, -0.715)
20	-0.003	(-0.361, 0.252)	80	0.008	(-0.322, 0.34)	12	-0.62	(-0.964, -0.333)
21	-0.024	(-0.344, 0.316)	81	-0.001	(-0.356, 0.385)	13	-0.093	(-0.447, 0.241)
22	-0.011	(-0.36, 0.313)	82	0.003	(-0.376, 0.289)	14	-1.334	(-1.643, -0.978)
23	-0.009	(-0.295, 0.376)	83	-0.013	(-0.33, 0.334)	15	-0.007	(-0.329, 0.368)
24	0.02	(-0.326, 0.331)	84	0.021	(-0.337, 0.354)	16	-0.057	(-0.305, 0.303)
25	-0.004	(-0.134, 0.397)	85	-0.016	(-0.337, 0.319)	17	1.277	(1.036, 1.6)
26	0.014	(-0.308, 0.354)	86	-0.004	(-0.283, 0.353)	18	0.432	(0.13, 0.83)
27	0.014	(-0.315, 0.339)	87	0.026	(-0.358, 0.29)	19	0.462	(0.191, 0.846)
28	-0.019	(-0.354, 0.329)	88	0.015	(-0.356, 0.327)	20	-0.515	(-0.78, -0.087)
29	-0.017	(-0.336, 0.341)	89	0.015	(-0.353, 0.304)	21	0.493	(0.242, 0.765)
30	-0.007	(-0.317, 0.296)	90	0.013	(-0.317, 0.324)	22	0.177	(-0.06, 0.48)
31	0.011	(-0.308, 0.325)	91	-0.014	(-0.375, 0.302)	23	-0.086	(-0.429, 0.181)
32	0.007	(-0.303, 0.371)	92	0.004	(-0.304, 0.233)	24	0.192	(-0.051, 0.526)
33	0.003	(-0.301, 0.356)	93	0.022	(-0.319, 0.358)	25	0.037	(-0.307, 0.319)
34	-0.004	(-0.136, 0.482)	94	0.001	(-0.27, 0.366)	26	0.161	(-0.135, 0.461)
35	-0.015	(-0.352, 0.308)	95	-0.01	(-0.351, 0.369)	27	-0.459	(-0.737, -0.123)
36	-0.019	(-0.311, 0.311)	96	-0.023	(-0.396, 0.317)	28	0.286	(0.06, 0.592)
37	0.017	(-0.339, 0.313)	97	0.003	(-0.342, 0.382)	29	-0.154	(-0.563, 0.214)
38	0.017	(-0.314, 0.314)	98	0.023	(-0.327, 0.321)	30	0.352	(0.024, 0.622)
39	0	(-0.278, 0.253)	99	-0.007	(-0.359, 0.296)	31	0.179	(-0.123, 0.501)
40	-0.019	(-0.324, 0.34)	100	0.02	(-0.262, 0.348)	σ_{PID}	0.599	(0.53, 0.702)
41	-0.018	(-0.3, 0.291)	β_{PID}, 1	-0.057	(-0.321, 0.333)	σ_{RID}	0.556	(0.485, 0.61)
42	-0.003	(-0.318, 0.373)	2	0.06	(-0.201, 0.323)	σ_p	0.041	(0.011, 0.346)
43	-0.002	(-0.32, 0.325)	3	0.057	(-0.186, 0.275)	σ_a	0.34	(0.2, 1.031)
44	-0.002	(-0.292, 0.363)	4	0.011	(-0.307, 0.418)	σ_g	0.034	(0.011, 0.365)
45	0.009	(-0.354, 0.311)	5	0.561	(0.301, 0.855)	σ	0.294	(0.269, 0.321)
46	-0.005	(-0.354, 0.318)						

Supplemental Table 3: Correlations between estimated concentrations in the second experiment

In the second experiment, four arrays were hybridised in a loop design with three samples (A, B1, B2) from a human cervical tumour and a reference sample (Ref) (Table 1).

Estimated concentrations for individual genes were reliable, as pairwise scatterplots (Supplemental Figure 8), and correlations in the following table show.

Supplemental Table 3: Correlation between estimated concentrations

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

These relationships were consistent with A, B1 and B2 originating from the same tumour and B1 and B2 originating from the same location within the tumour.

Supplemental Table 4: Estimates of absolute transcript concentrations

For the human cervical tumour study we considered 100 genes and four samples; reference, biopsy B1, biopsy B2 and biopsy A. The estimated number of transcripts for each gene in each sample is given together with its uncertainty. The estimates are posterior marginal modes and the uncertainties are described by 95% credibility intervals, lying between the 2.5% and the 97.5% quantiles.

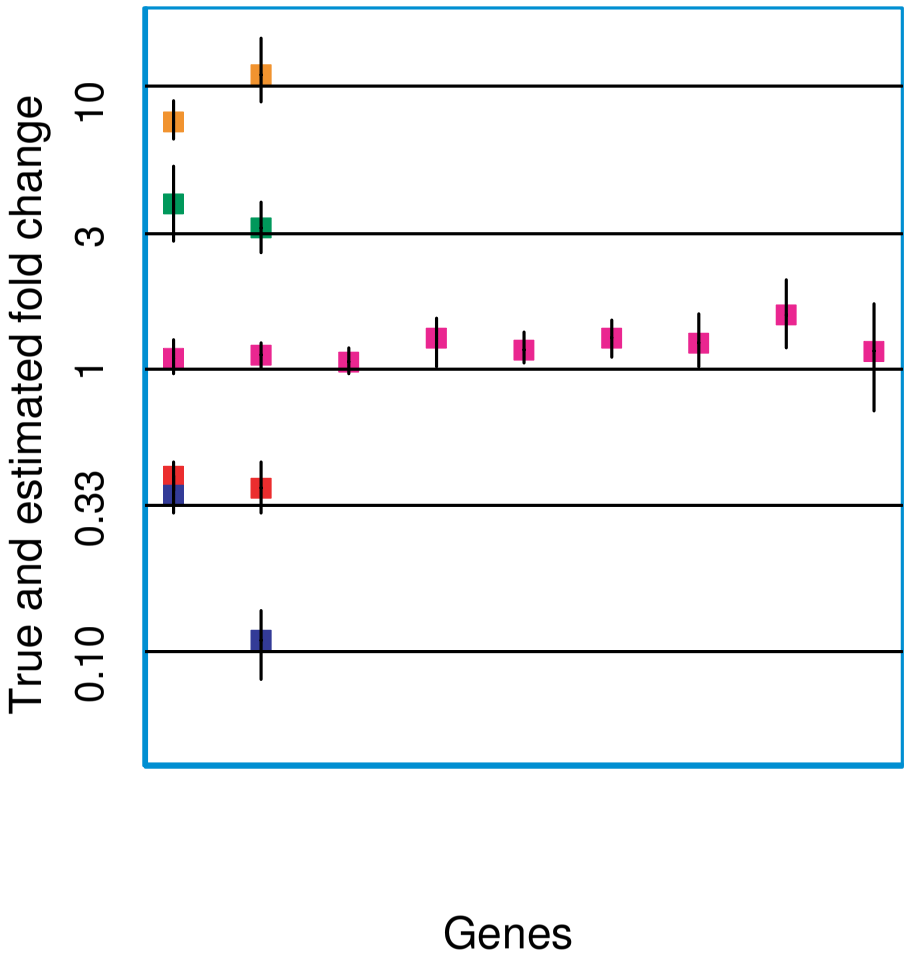
Gene Number and name		Reference Mode and 95% credibility interval ($*10^6$)		Biopsy B1 Mode and 95% credibility interval ($*10^6$)		Biopsy B2 Mode and 95% credibility interval ($*10^6$)		Biopsy A Mode and 95% credibility interval ($*10^6$)	
1	*ABR*	0.308	(0.18, 0.718)	0.406	(0.227, 0.862)	0.385	(0.225, 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)
8	*CCT6A*	1.01	(0.511, 1.887)	0.741	(0.399, 1.497)	1.06	(0.578, 2.229)	1.248	(0.644, 2.227)
9	*CD34*	0.055	(0.024, 0.157)	0.055	(0.026, 0.18)	0.064	(0.03, 0.169)	0.092	(0.042, 0.239)
10	*CD37*	0.098	(0.051, 0.33)	0.16	(0.079, 0.468)	0.177	(0.09, 0.528)	0.184	(0.092, 0.54)
11	*CD44*	3.531	(2.087, 8.298)	0.513	(0.306, 1.255)	0.711	(0.332, 1.428)	1.716	(0.82, 3.614)
12	*CD53*	0.401	(0.25, 0.977)	0.926	(0.489, 1.837)	1.151	(0.718, 2.729)	1.801	(0.999, 3.757)
13	*CDC6*	0.528	(0.284, 1.14)	0.289	(0.14, 0.616)	0.279	(0.16, 0.599)	0.329	(0.182, 0.712)
14	*CDH13*	0.561	(0.326, 1.23)	0.112	(0.064, 0.279)	0.1	(0.063, 0.251)	0.238	(0.122, 0.501)
15	*CDK8*	0.403	(0.247, 0.905)	0.498	(0.274, 0.944)	0.557	(0.295, 1.003)	0.465	(0.288, 1.008)
16	*CDK9*	0.06	(0.026, 0.144)	0.043	(0.018, 0.126)	0.047	(0.024, 0.134)	0.045	(0.023, 0.131)
17	*CDKN1B*	0.269	(0.125, 0.676)	0.453	(0.22, 1.058)	0.337	(0.189, 0.907)	0.412	(0.208, 0.957)
18	*CDKN2D*	0.243	(0.132, 0.603)	0.262	(0.135, 0.574)	0.289	(0.146, 0.601)	0.297	(0.173, 0.778)
19	*CHL1*	0.112	(0.067, 0.366)	0.159	(0.085, 0.541)	0.208	(0.112, 0.634)	0.299	(0.171, 0.779)
20	*CKS2*	0.984	(0.511, 1.838)	0.41	(0.227, 0.819)	0.583	(0.357, 1.301)	0.99	(0.565, 2.116)
21	*CLCN3*	0.324	(0.154, 0.736)	0.323	(0.159, 0.797)	0.297	(0.133, 0.579)	0.235	(0.106, 0.466)
22	*CLDN3*	0.106	(0.046, 0.278)	0.096	(0.048, 0.311)	0.113	(0.049, 0.287)	0.062	(0.032, 0.192)
23	*CLIC1*	0.081	(0.043, 0.186)	0.178	(0.092, 0.407)	0.177	(0.103, 0.431)	0.158	(0.076, 0.309)
24	*COL15A1*	0.176	(0.098, 0.384)	0.139	(0.07, 0.285)	0.131	(0.076, 0.306)	0.252	(0.148, 0.592)
25	*COL1A2*	0.26	(0.159, 0.537)	0.933	(0.56, 1.839)	0.73	(0.491, 1.479)	1.123	(0.71, 1.928)
26	*COX10*	0.172	(0.079, 0.424)	0.124	(0.053, 0.306)	0.142	(0.073, 0.355)	0.135	(0.072, 0.347)
27	*COX7C*	1.553	(0.864, 3.878)	1.823	(0.976, 4.194)	2.098	(1.075, 4.548)	1.85	(1.037, 4.788)
28	*CREG*	1.151	(0.615, 2.17)	1.733	(0.98, 3.674)	1.793	(0.894, 3.555)	1.389	(0.828, 3.171)
29	*CSK*	0.087	(0.046, 0.186)	0.114	(0.059, 0.249)	0.091	(0.05, 0.204)	0.079	(0.041, 0.166)
30	*CYP2A7*	0.216	(0.124, 0.473)	0.34	(0.186, 0.694)	0.319	(0.172, 0.626)	0.246	(0.111, 0.412)
31	*DDX16*	0.239	(0.142, 0.561)	0.373	(0.196, 0.754)	0.352	(0.216, 0.785)	0.425	(0.198, 0.737)
32	*DUSP5*	0.133	(0.065, 0.274)	0.156	(0.086, 0.359)	0.155	(0.094, 0.364)	0.096	(0.052, 0.226)
33	*EPHA1*	0.157	(0.076, 0.34)	0.187	(0.097, 0.405)	0.219	(0.105, 0.44)	0.206	(0.117, 0.497)
34	*ETS2*	0.186	(0.114, 0.407)	0.782	(0.459, 1.508)	0.674	(0.345, 1.131)	0.622	(0.345, 1.117)
35	*FLJ00023*	0.305	(0.185, 0.67)	0.333	(0.188, 0.661)	0.384	(0.229, 0.788)	0.555	(0.287, 1.002)
36	*FLJ10701*	0.063	(0.029, 0.161)	0.072	(0.033, 0.186)	0.082	(0.043, 0.217)	0.087	(0.043, 0.206)
37	*FLJ10871*	0.342	(0.16, 0.856)	0.594	(0.288, 1.535)	0.463	(0.255, 1.192)	0.288	(0.159, 0.788)
38	*FN1*	0.137	(0.064, 0.284)	0.048	(0.026, 0.129)	0.06	(0.027, 0.137)	0.077	(0.041, 0.172)
39	*FNTA*	0.227	(0.145, 0.467)	0.244	(0.191, 0.596)	0.413	(0.23, 0.758)	0.387	(0.273, 0.728)
40	*FY*	0.099	(0.049, 0.21)	0.12	(0.074, 0.296)	0.121	(0.061, 0.256)	0.09	(0.053, 0.211)
41	*GADD34*	0.203	(0.115, 0.514)	0.127	(0.066, 0.351)	0.21	(0.106, 0.496)	0.223	(0.123, 0.546)
42	*GAPD*	2.253	(1.296, 5.399)	1.31	(0.62, 2.75)	1.599	(0.81, 3.246)	1.89	(1.052, 4.546)
43	*GPM6B*	0.129	(0.08, 0.28)	0.092	(0.05, 0.218)	0.088	(0.049, 0.203)	0.091	(0.047, 0.177)
44	*GSTA2*	0.112	(0.057, 0.32)	0.173	(0.085, 0.446)	0.116	(0.054, 0.281)	0.109	(0.046, 0.235)
45	*GSTA3*	0.084	(0.043, 0.216)	0.066	(0.034, 0.211)	0.065	(0.034, 0.2)	0.049	(0.021, 0.114)
46	*GSTP1*	1.036	(0.496, 2.789)	2.727	(1.234, 6.845)	2.915	(1.292, 6.885)	2.103	(1.098, 5.482)
47	*GSTTLp28*	0.56	(0.318, 1.245)	0.665	(0.348, 1.495)	0.604	(0.403, 1.557)	0.763	(0.421, 1.621)
48	*HDGF*	1.12	(0.564, 2.156)	1.405	(0.702, 2.633)	1.242	(0.659, 2.4)	0.888	(0.521, 2.117)
49	*HLA-C*	0.528	(0.267, 1.506)	9.238	(3.8, 20.511)	8.572	(4.233, 20.923)	6.456	(2.641, 14.813)
50	*HLA-DPB1*	0.101	(0.053, 0.255)	3.127	(1.635, 6.909)	3.754	(1.909, 7.065)	4.908	(3.17, 12.066)

Continued on next page

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 ($*10^6$)	
51	*HXB*	0.667	(0.396, 1.45)	0.38	(0.229, 0.92)	0.398	(0.212, 0.798)	0.888	(0.594, 2.227)
52	*IGF1*	0.142	(0.065, 0.343)	0.379	(0.156, 0.787)	0.32	(0.141, 0.654)	0.297	(0.139, 0.657)
53	*IGF1R*	0.291	(0.138, 0.678)	0.142	(0.064, 0.345)	0.145	(0.08, 0.419)	0.151	(0.073, 0.338)
54	*IGHG3*	0.138	(0.076, 0.336)	0.187	(0.107, 0.504)	0.204	(0.104, 0.436)	0.184	(0.101, 0.433)
55	*IL10RA*	0.122	(0.059, 0.315)	0.233	(0.121, 0.542)	0.24	(0.131, 0.608)	0.223	(0.105, 0.509)
56	*IL13RA1*	0.189	(0.117, 0.582)	0.305	(0.159, 0.844)	0.309	(0.148, 0.833)	0.217	(0.114, 0.621)
57	*IL1R2*	0.508	(0.22, 1.082)	2.374	(1.05, 4.696)	1.607	(0.71, 3.429)	1.212	(0.63, 2.811)
58	*IL1RN*	0.093	(0.045, 0.258)	0.09	(0.045, 0.29)	0.08	(0.043, 0.263)	0.075	(0.034, 0.189)
59	*IL6*	0.113	(0.054, 0.279)	0.331	(0.178, 0.698)	0.263	(0.147, 0.6)	0.249	(0.136, 0.541)
60	*IL8*	0.088	(0.038, 0.232)	0.194	(0.097, 0.525)	0.219	(0.106, 0.517)	0.413	(0.203, 0.955)
61	*IRF1*	0.149	(0.088, 0.365)	0.2	(0.116, 0.491)	0.16	(0.105, 0.426)	0.213	(0.106, 0.39)
62	*JUN*	0.146	(0.066, 0.349)	0.232	(0.12, 0.6)	0.283	(0.142, 0.686)	0.451	(0.232, 1.106)
63	*JUNB*	0.538	(0.295, 1.256)	2.297	(1.181, 4.826)	1.823	(1.149, 4.443)	1.432	(0.848, 3.214)
64	*KIAA1705*	0.211	(0.112, 0.461)	0.1	(0.055, 0.244)	0.108	(0.061, 0.277)	0.147	(0.077, 0.308)
65	*KLF2*	0.117	(0.065, 0.262)	0.138	(0.076, 0.317)	0.105	(0.061, 0.249)	0.105	(0.06, 0.241)
66	*LAMB1*	0.414	(0.225, 0.812)	0.251	(0.149, 0.574)	0.324	(0.187, 0.665)	0.402	(0.22, 0.747)
67	*LMNA*	0.465	(0.283, 0.987)	0.388	(0.22, 0.823)	0.461	(0.248, 0.896)	0.427	(0.29, 1.015)
68	*LOX*	0.094	(0.048, 0.233)	0.045	(0.024, 0.125)	0.057	(0.03, 0.145)	0.125	(0.069, 0.311)
69	*MAPK14*	0.379	(0.193, 0.904)	0.425	(0.204, 1.015)	0.372	(0.214, 1.028)	0.768	(0.399, 1.946)
70	*MCAM*	2.323	(1.392, 5.289)	0.51	(0.304, 1.214)	0.598	(0.332, 1.245)	0.76	(0.448, 1.74)
71	*MEG3*	0.195	(0.108, 0.443)	0.195	(0.102, 0.387)	0.188	(0.114, 0.392)	0.176	(0.111, 0.42)
72	*MID1*	0.109	(0.064, 0.298)	0.398	(0.211, 0.872)	0.376	(0.276, 0.985)	0.507	(0.293, 1.109)
73	*MY014*	0.224	(0.116, 0.456)	0.166	(0.071, 0.29)	0.135	(0.076, 0.296)	0.169	(0.085, 0.322)
74	*NGFB*	0.206	(0.104, 0.464)	0.195	(0.093, 0.423)	0.134	(0.066, 0.292)	0.087	(0.044, 0.219)
75	*OAZ1*	0.987	(0.704, 2.485)	1.026	(0.652, 2.35)	1.049	(0.721, 2.517)	1.188	(0.731, 2.354)
76	*ODC1*	3.129	(1.763, 6.841)	0.566	(0.346, 1.249)	0.552	(0.365, 1.388)	0.803	(0.537, 2.061)
77	*OSTF1*	0.436	(0.259, 0.949)	0.603	(0.33, 1.228)	0.37	(0.241, 0.881)	0.382	(0.215, 0.76)
78	*PAPPA*	0.112	(0.067, 0.263)	0.071	(0.039, 0.18)	0.061	(0.04, 0.172)	0.119	(0.07, 0.256)
79	*PC4*	1.628	(0.964, 3.631)	1.704	(1.07, 4.083)	2.195	(1.219, 4.627)	2.183	(1.301, 5.453)
80	*PFKM*	0.302	(0.177, 0.649)	0.163	(0.083, 0.305)	0.142	(0.09, 0.327)	0.182	(0.098, 0.329)
81	*PIM1*	0.205	(0.087, 0.494)	0.19	(0.086, 0.504)	0.189	(0.107, 0.599)	0.318	(0.129, 0.62)
82	*PLAU*	0.232	(0.116, 0.59)	2.35	(1.04, 4.531)	1.878	(1.01, 4.523)	1.458	(0.74, 3.349)
83	*PLGL*	0.07	(0.031, 0.151)	0.087	(0.047, 0.209)	0.084	(0.035, 0.175)	0.063	(0.028, 0.126)
84	*PPP2R1B*	0.374	(0.218, 0.752)	0.183	(0.098, 0.392)	0.2	(0.101, 0.373)	0.208	(0.129, 0.477)
85	*PSMC4*	1.397	(0.873, 2.823)	1.263	(0.742, 2.677)	1.238	(0.643, 2.277)	1.249	(0.791, 2.628)
86	*PTE1*	0.302	(0.178, 0.606)	0.184	(0.106, 0.448)	0.18	(0.119, 0.447)	0.167	(0.097, 0.331)
87	*RAB6A*	1.014	(0.514, 1.836)	0.59	(0.298, 1.067)	0.48	(0.269, 0.915)	1.159	(0.666, 2.261)
88	*RAF1*	0.274	(0.153, 0.653)	0.242	(0.124, 0.518)	0.268	(0.145, 0.62)	0.548	(0.295, 1.137)
89	*RI58*	0.103	(0.047, 0.248)	0.261	(0.119, 0.506)	0.251	(0.128, 0.576)	0.247	(0.122, 0.581)
90	*S100A7*	0.047	(0.022, 0.155)	0.679	(0.361, 1.511)	1.035	(0.628, 2.846)	0.255	(0.135, 0.635)
91	*SERPINA1*	0.128	(0.058, 0.323)	0.052	(0.023, 0.163)	0.046	(0.02, 0.13)	0.045	(0.028, 0.149)
92	*SFRS9*	0.561	(0.361, 1.224)	0.361	(0.228, 0.814)	0.441	(0.249, 0.929)	0.622	(0.315, 1.112)
93	*SLC2A3*	0.332	(0.2, 0.812)	0.094	(0.05, 0.23)	0.085	(0.049, 0.202)	0.088	(0.044, 0.2)
94	*TM4SF3*	0.407	(0.271, 0.924)	0.465	(0.318, 1.053)	0.5	(0.302, 1.018)	0.579	(0.336, 1.127)
95	*TP53BP1*	0.418	(0.252, 0.915)	0.269	(0.127, 0.515)	0.241	(0.156, 0.594)	0.348	(0.192, 0.656)
96	*TRIP7*	0.531	(0.318, 1.122)	0.804	(0.44, 1.633)	1.038	(0.576, 2.014)	0.934	(0.522, 1.998)
97	*TSC2*	0.214	(0.106, 0.443)	0.163	(0.085, 0.37)	0.109	(0.075, 0.285)	0.129	(0.068, 0.274)
98	*UFD1L*	0.153	(0.081, 0.37)	0.165	(0.081, 0.381)	0.171	(0.088, 0.383)	0.133	(0.08, 0.341)
99	*VDR*	0.123	(0.064, 0.26)	0.081	(0.041, 0.177)	0.052	(0.028, 0.127)	0.078	(0.042, 0.171)
100	*VEGF*	1.172	(0.699, 3.381)	1.794	(0.983, 5.296)	1.91	(1.025, 5.029)	1.703	(1.008, 4.766)

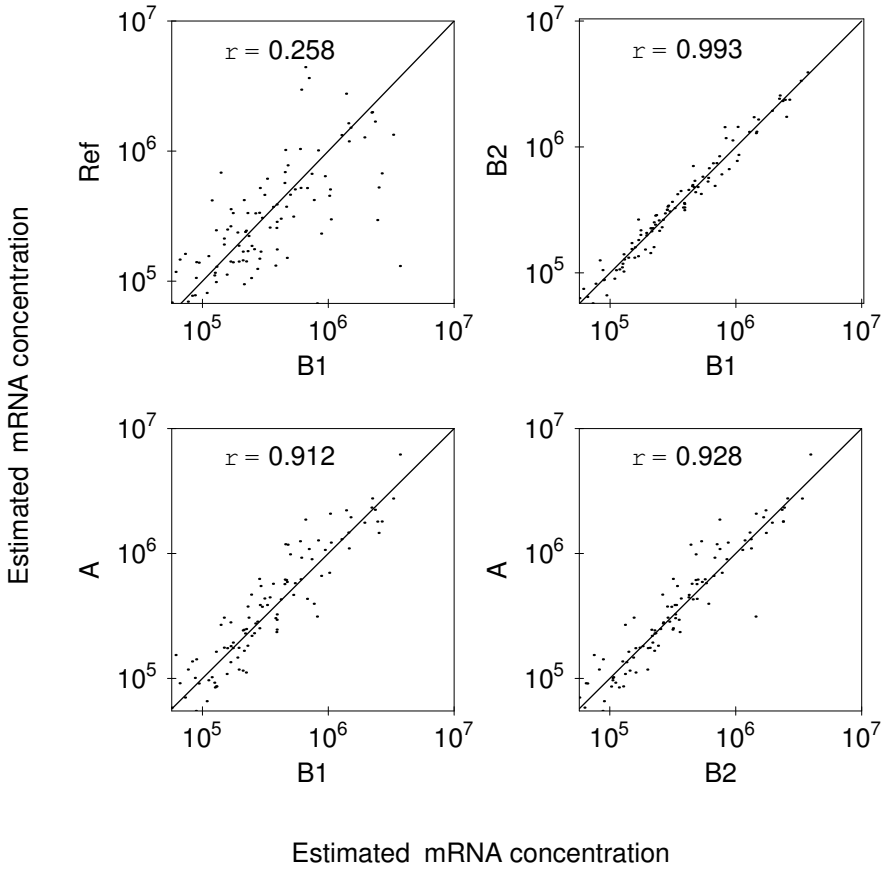
Supplemental Figure 7: validation of the methods: Estimated ratios between numbers of molecules per gene in two samples with known concentrations

Estimated ratios between numbers of molecules per gene in the two samples were obtained. The data in Figure 7 are based on the dye-swap experiment presented in Figure 2a in the paper. Estimated ratios of the numbers of mRNA molecules of the two samples are plotted for each of the 17 genes, together with their 95% credibility intervals. Horizontal lines represent true ratios, coloured squares are estimated ratios with a colour for each true ratio. Nine genes have true ratio 1. True ratios 10, 3, 1, 0.33 and 0.10 are coloured yellow, green, pink, red and blue respectively. The ratios are sufficiently well estimated; for example, when the true fold is 1, estimates range between 1 and 1.5. One 10-fold is estimated however as a 3-fold. A two- or three-fold cut-off analysis would correctly deliver 4 overexpressed and 4 underexpressed genes. Estimated ratios of number of molecules were similar to the ratios of normalised measured intensities; in both cases folds were sufficiently well estimated to guide a search for differential expressions.


True and estimated fold change
10
3
1
0.33
0.10
Genes

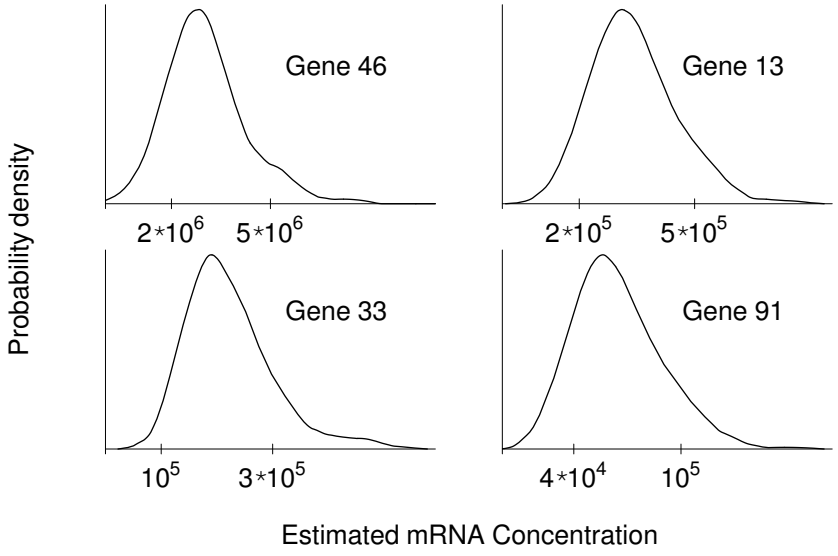
Supplemental Figure 8: Comparison of absolute transcript levels in different samples from a human cervical tumour

Comparison of absolute transcript levels in different samples from a human cervical tumour. The data are based on the experiments listed in Table 1 in the paper. Estimated mRNA concentrations (number of mRNA molecules per μg of total RNA; posterior modes) are plotted for each gene and sample. Correlation coefficients and diagonal lines are shown. The remaining two correlation coefficients are 0.280 for reference versus sample B2 and 0.396 for reference versus sample A. The estimated mRNA concentrations for the tumour samples (A, B1, B2) showed a much stronger correlation to each other than to the reference sample (Ref). Moreover, the two samples derived from the same biopsy (B1, B2) were more correlated to each other than to the sample derived from a different location in the tumour (A). Notice that the range obtained by multiplying estimated concentrations in the biopsies by the quantity of total RNA leads to estimated numbers of transcripts, which are within a range comparable to that of the validation experiment in Figure 2a and 2b in the paper.



Supplemental Figure 9: Estimated probability densities of mRNA concentration in samples from a human cervical tumour

Estimated probability densities of mRNA concentration (number of mRNA molecules per μg of total RNA) in a cervical tumour for four typical genes; gene 13, 33, 46, and 91 (see Supplemental Table 2 for the gene symbols of the 100 genes included in the analysis). The data are based on the experiments listed in Table 1, and represent the distributions of the mean concentrations of the three tumour samples (A, B1, B2). The width of the credibility intervals describe the biological variability of the three samples in addition to the uncertainty of the estimates. Gene 46 had a relatively high mRNA concentration, gene 91 low, while the other two were intermediate. Different axis scales are used for the different genes.



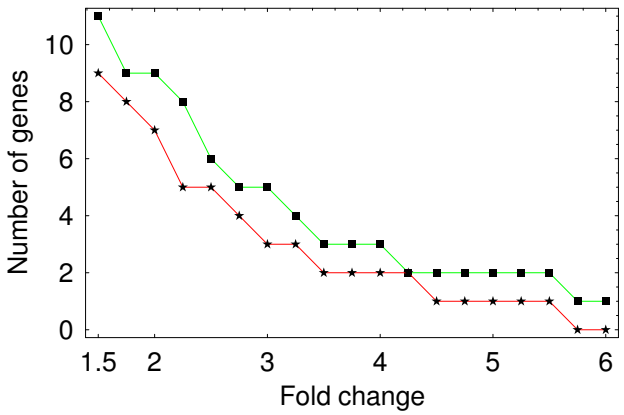


Figure 1:

Supplemental Figure 10: the number of selected genes as function of the fold k

We plotted the number of selected genes as a function of the fold k, for a given value of the probability that all but m genes were at least k-fold expressed. We used $m = 0$ and two values of the probability, 0.95 (red) and 0.80 (green), for the two curves. This plot helps to find the required balance between level of differential expression (here the fold k) and the number of selected genes, for a given level of posterior probability.