Genome-wide estimation of transcript concentrations from spotted cDNA microarray data.

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

Supplementary Material

1 Model building

The binomial model for the molecular selection process is simple and natural. At each experimental step we start out with a certain number of molecules, and a selection of those which travel along to the next step takes place. In each experimental step, a molecule has a certain probability to 'survive', and that probability depends on sample t, gene g and array a and is modulated with several covariates.

Below we discuss which covariates act in the different experimental steps, how conditional independence among molecules can be assumed, and hence argue that in each step, a binomial model is suitable.

Looking to the experiment from cDNA synthesis to hybridisation and washing as a whole, the sequence of binomials will nest up. The final binomial formula connects the number of molecules in sample t that in the end is available for imaging on spot s on array a to the initial number of transcripts from the corresponding gene in sample t, via the over all survival probability $p_s^{t,a}$, appearing in Eq. (1) below.

Preparation of the mRNA solution.

The known quantity of material for sample t on array a is denoted $q^{t,a}$, for example the weight of mRNA after amplification. For each gene g let K_g^t denote the unknown number of transcripts per weight unit in sample t.

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 (order of a few hundreds) to allow for such an approximation. The expected number of actually bound CyX-dUTP's does however depend on dye, since there is a chemical dye effect.

The $q^{t,a} \cdot K_g^t$ molecules reverse transcribe and are labelled independently of each other with a certain probability depending on gene and sample specific covariates (like purity of the sample). The resulting number of labelled cDNA molecules (or target molecules) for sample t, gene g and array a, then follows the binomial distribution with parameters $q^{t,a} \cdot K_g^t$ and a certain gene and sample dependent success probability.

Purification.

The two solutions are mixed. Excessive CyX-dUTP molecules are washed away. During this process also some of the target molecules will be lost. For sample t, gene g and array a, given the number of molecules from the step above, the number of molecules independently remaining in the solution after purification, is then again binomial with a certain success probability (success now means that a molecule is not washed away, but remains in the solution). We suggest that this probability might depend 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 probability of remaining in the solution specifically caused by target length will instead be absorbed in a gene specific covariate (β_g).

Hybridisation.

The variability of probe material and microarray production modulates the probability of successful hybridisation. Both array and pen information are included as covariates in the model, in addition to probe quantity and quality 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 target molecules do not cluster nor repulse. Let n_s^a be the number of pixels in spot s on array a. A proportion $c \cdot n_s^a$ of the gene molecules in the purified solution candidates to reach the correct spot s for hybridisation. Each of these molecules has a certain probability to hybridise and it is reasonable to claim independence because of 'probe in excess'. Hence again the number of molecules hybridised in spot s follows a binomial distribution. The success (hybridisation) probability 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, but also that some hybridised molecules might disappear. The number of remaining molecules is binomial and the success probability depends on probe length, reflecting the binding strength, and on microarray effects. Let the remaining number of molecules of sample t hybridised in spot s on array a, participating in the following imaging process, be denoted $H_s^{t,a}$. Because of nesting of binomials, we get

$$H_s^{t,a} \sim \text{Binomial}(c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t,\, p_s^{t,a}),$$

where g is the gene spotted in spot s on array a and

$$\begin{aligned} p_s^{t,a} &= \min[1, \exp\{\beta_0 + \beta_e + \beta_a + \beta_p + \beta_g + \beta_{\text{RID}} + \beta_{\text{PID}} \\ &\quad +\beta_l \cdot [\text{probe length}] + \beta_q \cdot [\text{probe quantity}] + \beta_m \cdot [\text{purity}_t]\}]. \end{aligned} \tag{1}$$

The notation for the various covariates is described in the paper.

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

On top of the binomial selection model for molecules described above (Layer (i) in the paper), we model the scanning process and measurement and residual errors, described in detail in the paper (Layers (ii) and (iii)). The hierarchical model then reads

$$\begin{aligned} H_s^{t,a} &\sim \text{Binomial}(c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t, p_s^{t,a}), \\ p_s^{t,a} &= \min[1, \exp\{\beta_0 + \beta_e + \beta_g + \beta_m \cdot [\text{purity}_t] + \bar{\beta} X_s^a\}] \\ \mu_s^{t,a} &= 2^{f_{\text{dye}}\,\text{PMT}^{t,a}} H_s^{t,a} \alpha_{\text{dye}}, \\ L_{j,s}^{t,a} &= \frac{\mu_s^{t,a}}{n_s^a} + \varepsilon_{j,s}^{t,a}, \quad \varepsilon_{j,s}^{t,a} \sim \text{Normal}(0, (\sigma_s^{t,a})^2) \end{aligned}$$

where

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

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

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

For obtaining easier convergence of the MCMC (see Supplemental Methods 3) we reparameterize in such a way that all parameters appearing in the expectations can actually be estimated on the basis of the expectation only; all the other parameters appear instead only in the variances. To explain this reparametrisation, it is easier to use the link function $\exp(x)$ instead of $\min(1, \exp(x))$. Note that identifiability under the relaxed link function implies identifiability under the $\min(1, \exp(x))$ link function, because the parameter space is smaller due to the constraint. 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 $E[L_{j,s}^{t,a}] = C^{t,a} \cdot \tilde{K}_g^t \alpha'_{\text{dye}} \exp(\bar{\beta} X_s^a)$, where $C^{t,a}$ is a product of known constants, 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 constraints (see below) and with a design which has the following characteristics: 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 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 performed, since otherwise estimated uncertainties of the concentrations will be shrunk. 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 (single data points). 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 concentrations K_g^t are estimated using the estimates of $\tilde{K}_g^t$, β_0, β_e, β_m, α and β_g.

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-connected experiments and the gene related parameters β_g. For each set of connected experiments e, $\sum \beta_a = 0$, where the sum runs over all arrays of the 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 $\beta_{\mathrm{PID}} \sim \mathrm{Normal}(0, (\sigma_{\mathrm{PID}})^2)$ for all probe sequences. Similarly, we have $\beta_{\mathrm{RID}} \sim \mathrm{Normal}(0, (\sigma_{\mathrm{RID}})^2)$ for all replications of probe. Otherwise, all hyper-parameters are equipped with flat improper 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-connected designs are allowed.

For completeness, we mention that the covariates probe length, SYBR green, and purity were normalised to mean 0 and standard deviation 1.

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 $\beta K^t_{\mathrm{gene}(s)}$ in the log-probability. Then, we expect β to be negative: the larger $K^t_{\mathrm{gene}(s)}$, 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 (MCMC) 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 and the prior for the parameter is $\mathrm{Normal}(0, \sigma_p^2)$, draw from

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

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.

Initial parameter estimates for α'_{Cy3} and the β's (except for β_0, β_m, the β_e's and the β_g's) are found from the data using linear regression. Initial values for 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 substituted by 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).

4 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), 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. Total RNA was isolated by use of Trizol reagent (Life Technologies) from samples from 12 cervical cancers. Labeled cDNA was synthesized from 20 μg total RNA using Superscript II transcriptase (Life technologies) and FairPlay Micriarray Labeling Kit (Stratagene) in the presence of either Cy3-dUTP or Cy5-dUTP (Amersham Pharmacia). Each sample was co-hybridized with a reference sample pooled from 10 cancer cell lines (Stratagene) in a dye-swap design, yielding totally 24 data sets in the analysis.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 in the paper. 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).

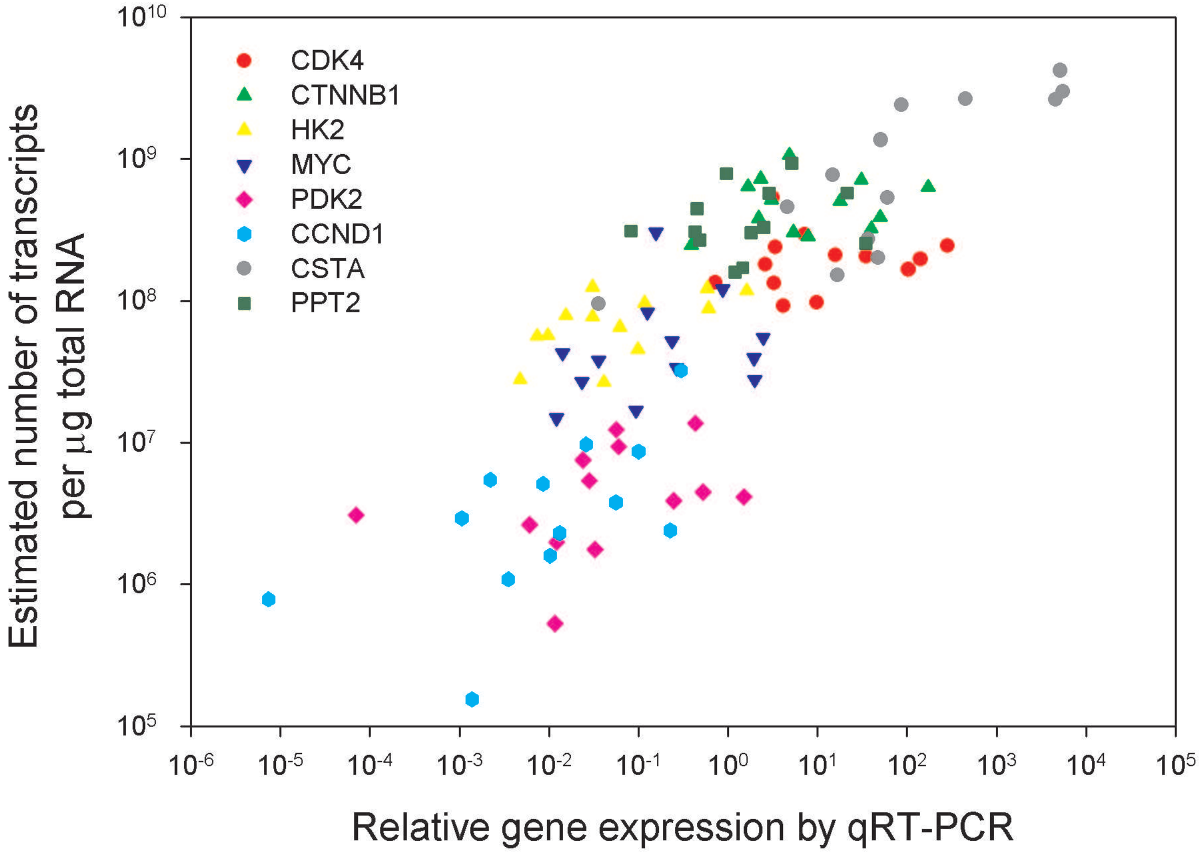
Scanning and Image Analysis. The slides were imaged at a resolution of 10 μm using an Agilent G2565BA scanner (Agilent Technologies), using a laser power and PMT voltage of 100%. Saturated spot intensities were corrected using the algorithm described previously in Lyng et al (2004) (reference (19) in the paper). The GenePix 4.1 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.

5 Quantitative real-time PCR

The pre-designed, gene-specific TaqMan gene expression assays (Applied Biosystems) Hs00364847 m1 (CDK4), Hs00170025 m1 (CTNNB1), Hs00606086 m1 (HK2), Hs00153408 m1 (MYC), Hs00193257 m1 (CSTA), Hs00607118 m1 (PPT2), Hs00277039 m1 (CCND1), Hs00176865 m1 (PDK2) and 4326322E (TBP) were used. All assays except 4326322E had a FAM reporter dye at the 5 end of the probe and a non-fluorescent quencher at the 3 end, whereas a VIC reporter dye was used in the TBP-assay. Conditions for amplification were one cycle of 95 degrees Celsius, 10 min, followed by 40 cycles of 95 degrees Celsius, 15 sec and 60 degrees Celsius, 1 min.

6 Figure 1

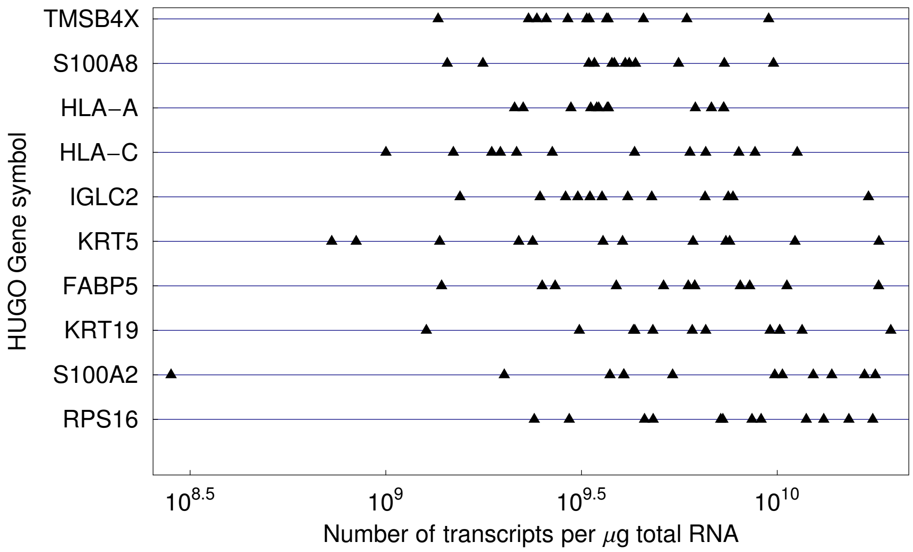
This figure should be seen together with Figure 3 in the paper. Figure below is based on background corrected data.



Supplementary Material Figure 1: qRT-PCR validation of the methodology to estimate absolute numbers of transcripts. Transcript concentration (number of transcripts per μg total RNA) of 10157 genes and ESTs in twelve cervical tumors and a pool of ten cancer cell lines was determined with our technique, and the data of eight genes covering the whole concentration range were plotted against the corresponding data achieved with qRT-PCR. There was a strong correlation between our estimates and the qRT-PCR data ($r = 0.79, p < 0.0001$, Pearson Product Moment Correlation). There was also a significant correlation for individual genes in some cases, despite a limited concentration range ($r = 0.60, p = 0.03$, CCND1; $r = 0.83, p = 0.0005$, CSTA; $r = 0.58, p = 0.04$, HK2).

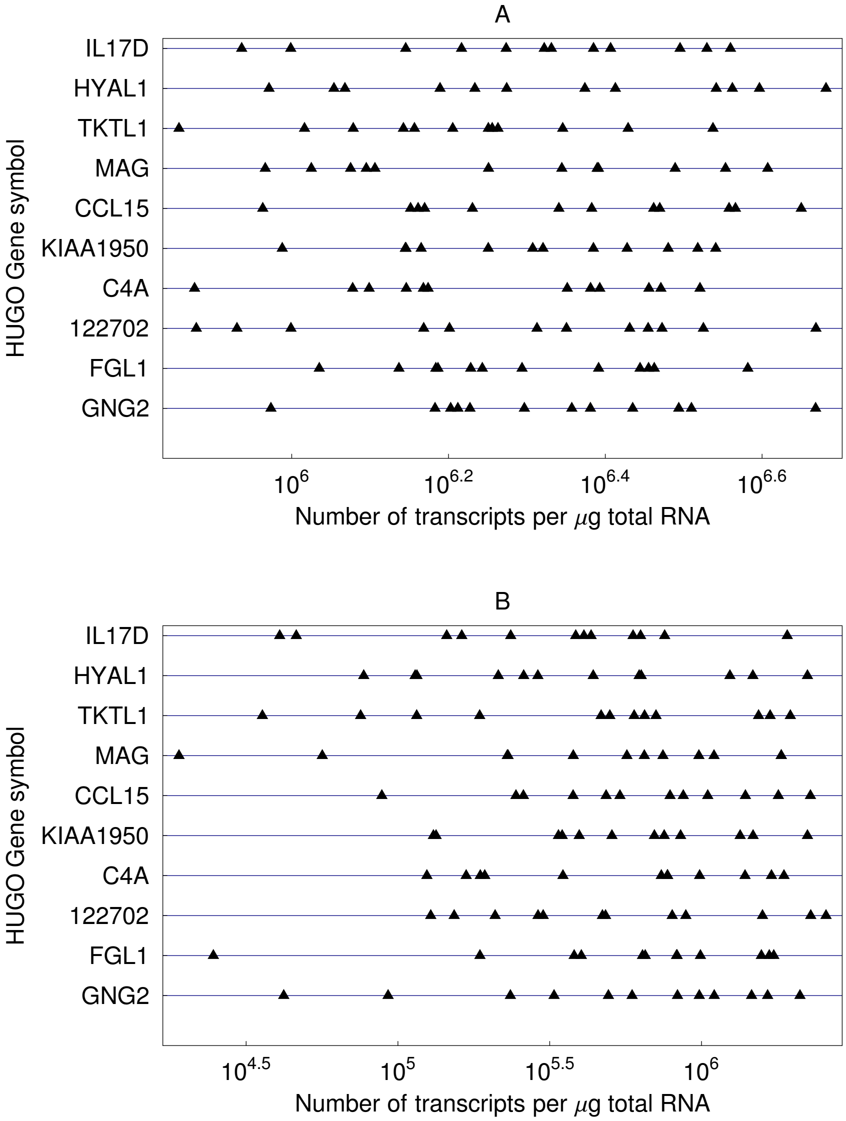
7 Figure 2

Supplementary Material Figure 2: This figure should be seen together with Figure 5 in the paper. It shows transcript concentration (number of transcripts per μg total RNA) for the same ten highly concentrated genes in cervix cancer. The figure below is based on background corrected data. As expected, background correction hardly influenced the estimates for high concentrations.



8 Figure 3

Supplementary Material Figure 3: Transcript concentration (number of transcripts per μg total RNA) for ten genes in cervix cancer with estimated mean concentration without (A) and with (B) background correction. Note that the x-axes correspond to different scales. All genes were selected so to have lowest concentration but valid data for all 12 tumors. Each point represents the estimated value of a single tumor. Large differences in transcript concentration among the tumors were found, especially for background corrected data (B), which seems to increase variability a lot. The within gene range (max-min) varies from 2.5-fold (FGL1) to 10-fold (122702) for uncorrected concentrations and 20-fold (C4A) to 100-fold (MAG) for background corrected concentrations.


A
IL17D
HYAL1
TKTL1
MAG
CCL15
KIAA1950
C4A
FGL1
GNG2
HUGO Gene symbol
10^{6}
$10^{6.2}$
$10^{6.4}$
$10^{6.6}$
Number of transcripts per μg total RNA
B
IL17D
HYAL1
TKTL1
MAG
CCL15
KIAA1950
C4A
FGL1
GNG2
HUGO Gene symbol
$10^{4.5}$
10^{5}
$10^{5.5}$
10^{6}
Number of transcripts per μg total RNA