BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 20 2010, pages 2571–2577
doi:10.1093/bioinformatics/btq406

Gene expression

Advance Access publication July 14, 2010

Probabilistic analysis of gene expression measurements from heterogeneous tissues

Timo Erkkilä[1,2,*], Saara Lehmusvaara[3], Pekka Ruusuvuori[1,2], Tapio Visakorpi[3], Ilya Shmulevich[1,2] and Harri Lähdesmäki[1,4,*]

[1]Department of Signal Processing, Tampere University of Technology, Finland, [2]Institute for Systems Biology, Seattle, WA, USA, [3]Institute of Medical Technology, University of Tampere and Tampere University Hospital and [4]Department of Information and Computer Science, Helsinki University of Technology, Finland

Associate Editor: Trey Ideker

ABSTRACT

Motivation: Tissue heterogeneity, arising from multiple cell types, is a major confounding factor in experiments that focus on studying cell types, e.g. their expression profiles, in isolation. Although sample heterogeneity can be addressed by manual microdissection, prior to conducting experiments, computational treatment on heterogeneous measurements have become a reliable alternative to perform this microdissection *in silico*. Favoring computation over manual purification has its advantages, such as time consumption, measuring responses of multiple cell types simultaneously, keeping samples intact of external perturbations and unaltered yield of molecular content.

Results: We formalize a probabilistic model, DSection, and show with simulations as well as with real microarray data that DSection attains increased modeling accuracy in terms of (i) estimating cell-type proportions of heterogeneous tissue samples, (ii) estimating replication variance and (iii) identifying differential expression across cell types under various experimental conditions. As our reference we use the corresponding linear regression model, which mirrors the performance of the majority of current non-probabilistic modeling approaches.

Availability and Software: All codes are written in Matlab, and are freely available upon request as well as at the project web page http://www.cs.tut.fi/~erkkila2/. Furthermore, a web-application for DSection exists at http://informatics.systemsbiology.net/DSection.

Contact: timo.p.erkkila@tut.fi; harri.lahdesmaki@tut.fi

Received on January 28, 2010; revised on June 7, 2010; accepted on July 6, 2010

1 INTRODUCTION

For being able to fully utilize capabilities of high-throughput measurement techniques that often have to deal with physically small but also heterogeneous tissue samples, attention should be paid as to how heterogeneity, the presence of multiple cell types in tissue, is addressed. In many studies the focus of interest hovers around identifying behavioral differences across cell types, and in such cases sample heterogeneity clearly has a confounding effect on downstream experiments and analysis.

Although laser-capture microdissection (LCM; Emmert-Buck *et al.*, 1996) offers a direct way to address tissue heterogeneity by allowing for isolation of morphologically distinguishable cell types, there are occasions when it is not feasible. Yield of biological content (e.g. mRNA) for conducting experiments becomes consequently lowered, which often needs to be compensated for with either more sensitive measurement devices or amplification of molecular quantities (Sooriakumaran *et al.*, 2009). However, amplification of mRNA from small albeit pure cell samples has its shortcomings, most notably nonlinearity (Otsuka *et al.*, 2007), obscuring the underlying profiles for distinct cell types.

Several authors have already studied performing computational microdissection for heterogeneous tissues, and proposed promising methods for microarray expression data. Initial attempts stem from Venet *et al.* (2001), who proposed a linear model for estimating both cell-type proportions and cell-type-specific gene expression profiles; the model assumes that, as prior information, there exist known, exclusively expressed genes for each cell-type. Subsequent studies have then demonstrated that the linearity assumption and prior information on either gene expression profiles, cell-type proportions, or both, can yield meaningful interpretations for the constituents of heterogeneous tissues (Abbas *et al.*, 2009; Gosink *et al.*, 2007; Hoffmann *et al.*, 2006; Jacobsen *et al.*, 2006; Lähdesmäki *et al.*, 2005; Quon and Morris, 2009; Stuart *et al.*, 2004).

In real experiments, conducted on the basis of heterogeneous tissue samples, having *precise* prior information is unrealistic, even though current models consistently rely on such information. We incorporate this missing functionality into the already-familiar linear regression framework through Bayesian prior densities whose shapes reflect the uncertainties associated with the prior information, such as cell-type proportions or cell-type-specific expression profiles.

For all model parameters, an efficient Markov chain Monte Carlo (MCMC) sampler is proposed. In addition to existing microdissection models, we further assume that the heterogeneous tissues have been measured under various experimental conditions, having a possible impact on cell-type-specific expression profiles. As cell-type-specific profiles are assumed to be different across both cell types and experimental conditions, assessment of statistically significant differential expression is performed with the two-sample *t*-test, though other tests for differential expression can be used.

We use simulated and real gene expression data for assessing the performance of the Bayesian model in contrast to a linear

[*]To whom correspondence should be addressed.

regression model that essentially captures properties common to the aforementioned, deterministic approaches. A series of case studies are used for demonstrating that the proposed method is capable of (i) de-noising uncertain prior information about cell-type proportions, (ii) more accurate estimation of replication variance, consequently leading to (iii) more accurate identification of differential expression across cell types and experimental conditions.

2 METHODS

2.1 Experimental design

We denote the tissue sample index with j and assume that there are J tissue samples in total. The number of cell types represented in the J samples needs to be known, and it is crucial that each of the J samples have *the same* cell types represented. We denote the cell type index by t and assume that there are T cell types in total. Lastly, we denote the number of probes (a generic term, e.g. a gene or miRNA) in an experiment by I so that the modeled data, which we denote by $\mathcal{D}$, consists of $I*J$ probe measurements,[1] y_{ij}, one for each probe i and tissue sample j.

In the simplest form this is all that is required. In addition, samples are often prepared under various experimental conditions, say, under 'No treatment', 'Treatment 1', 'Treatment 2', etc. and the analysis may be focused on finding differences in probe measurements across experimental conditions. Therefore, we incorporate the condition information into the model with variable $c(j)$ that takes on values $1,2,\dots,C$, being linked to the C different experimental conditions. For instance, if tissue samples 2 and 4 were measured under experimental condition 'No treatment', that information could be encoded by assigning $c(2)=c(4)=1$; thus, condition 'No treatment' would be associated with index 1, and so on.

2.2 Data likelihood

For tissue sample j under experimental condition $c(j)$, the data point for probe i, y_{ij}, is modeled as a sum of pure probe readings of all cell types, $\mathbf{x}_{ic(j)}=(x_{1ic(j)},x_{2ic(j)},\dots,x_{Tic(j)})$, weighted by the respective cell type proportions, $\mathbf{p}_j=(p_{1j},p_{2j},\dots,p_{Tj})$, plus an additive, normally distributed noise term, ϵ_{ij}, reflecting replication noise with variance $1/\lambda_i$:

$$y_{ij}=\sum_{t=1}^{T}p_{tj}x_{tic(j)}+\epsilon_{ij}, \tag{1}$$

so that the likelihood of data point $y_{ij}\in\mathcal{D}$ becomes $y_{ij}|\mathbf{p}_j,\mathbf{x}_i,\lambda_i\sim$ Normal$(\sum_{t=1}^{T}p_{tj}x_{tic(j)},1/\lambda_i)$. Thus, we model the replication variance, $1/\lambda_i$, as heteroscedastic across probes and homoscedastic across cell types and experimental conditions. Assuming independent and identically distributed (IID) measurements (elements in $\mathcal{D}$), a factorized form for the joint data likelihood can then be written as $f(\mathcal{D}|\boldsymbol{\theta})=\prod_{i=1}^{I}\prod_{j=1}^{J}f(y_{ij}|\mathbf{p}_j,\mathbf{x}_i,\lambda_i)$, where $\boldsymbol{\theta}$ is a collection of all model parameters, i.e. p_{tj}'s, x_{tic}'s, and λ_i's. The assumptions of additive, normally distributed noise and IID measurements is standard practice, although there is statistical evidence that at least the IID assumption may not always be valid (Efron, 2009).

2.3 Prior specifications

The model is next extended to account for parameter priors, so that the posterior distribution of all unknown model parameters required for sampling could be formulated. The prior assignments are done in a way that allows for easy sampling, and the shapes of the prior distributions are chosen to reflect the assumed variability of parameters.

We impose a normal prior $x_{tic}\sim\mathrm{N}(\mu_{tic},\nu)$ for the cell type and condition-specific probe measurement i, where the prior expression means and precision, μ_{tic} and ν, are extracted from the least-squares solution to the corresponding linear regression model assuming cell-type proportions known (see Supplementary Material for details). Normality is preferred so as to make use of the property of conjugate priors (posterior for x_{tic} will be a normal density, given that the prior and likelihood densities are also normal). Furthermore, a shared Gamma prior, Gamma(α,β), is placed on the inverses of replication variances, i.e. precisions, $\lambda_1,\dots,\lambda_i,\dots,\lambda_I$. Positive support and flexibility of Gamma$(\cdot,\cdot)$ make it useful in modeling precision parameters in a Bayesian framework (Gelman, 2006). Furthermore, the shared prior shrinks posterior estimates of λ_i's toward their common prior mean, α/β, regularizing estimates especially when dealing with small sample sizes (Smyth, 2004).

The mixing proportions for tissue sample j, $\mathbf{p}_j=(p_{1j},\dots,p_{Tj})$, are limited to a T-simplex; all elements in $\mathbf{p}_j$'s are non-negative and, vector-wise, sum up to one. A natural prior density for such vectors is the Dirichlet density, which we parameterize with w_0 and $\mathbf{p}_{0j}$ as $\mathbf{p}_j\sim$ Dirichlet$(w_0\mathbf{p}_{0j})$. The parametrization is done in a way that allows for prior knowledge on p_{tj}'s to be plugged into the model in a straightforward manner. Namely, we assume that a user has obtained prior information on the cell-type proportion in the J samples (e.g. by looking at the histology slides of the samples and making rough estimates or in an automated manner using digital microscopy images of the samples, or with flow cytometry, etc.), and these prior proportions are stored in $\mathbf{p}_{0j}$. Moreover, the belief of the correctness of prior proportions is specified by the multiplicative weight w_0. This way the user can tune the peakedness of the prior density around the prior guess, $\mathbf{p}_{0j}$; increasing w_0 increases the peakedness and vice versa. For compactness, we encapsulate the aforementioned parameters in a vector $\boldsymbol{\xi}=(\alpha,\beta,\mu_{111},\dots,\mu_{\mathrm{TIC}},w_0,\mathbf{p}_{01},\dots,\mathbf{p}_{0J})$.

2.4 Posterior sampling

Unknown parameters, i.e. $\boldsymbol{\theta}$, in our model are estimated in an MCMC fashion, which means we first must devise a sampling scheme under which samples from the posterior density of our parameters, given data and fixed parameters,$f(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\xi})\propto f(\mathcal{D}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{\xi})$, are drawn. Assuming S samples drawn from the posterior, the samples are subsequently used for summarization, i.e. approximating the expected value of the parameters with Monte Carlo integration (Gelman *et al.*, 2004), $\mathbb{E}[\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\xi}]\approx 1/S\sum_{s=1}^{S}\boldsymbol{\theta}^{(s)}$. Gibbs sampling (Gelman *et al.*, 2004) is one such sampling method, employing the idea of drawing a value from a conditional posterior for the respective parameters one at a time, while conditioning on all other model parameters, being set to previously sampled values, and data.

Next, we will construct a hybrid Gibbs and Metropolis–Hastings (M–H) sampler for all the model parameters; detailed derivations are shown in the Supplementary Material. The posterior for x_{tic} is

$$x_{tic}|\cdot\sim\mathrm{Normal}\left(\frac{P_{tic}}{Q_{tic}},\frac{1}{Q_{tic}}\right), \tag{2}$$

where the parameters of that distribution are $P_{tic}=\lambda_i\sum_{j:c(j)=c}(y_{ij}p_{tj}-p_{tj}\sum_{t'\neq t}p_{t'j}x_{t'ic})+\nu\mu_{tic}$ and $Q_{tic}=\lambda_i\sum_{j:c(j)=c}p_{tj}^2+\nu$. In a similar fashion, one finds the posterior for λ_i to be

$$\lambda_i|\cdot\sim\mathrm{Gamma}\left(\alpha+\frac{J}{2},\beta+\frac{1}{2}\sum_{j=1}^{J}e_{ij}^2\right), \tag{3}$$

where e_{ij} is the model residual $e_{ij}=y_{ij}-\sum_{t=1}^{T}p_{tj}x_{tic(j)}$. However, one cannot find such a density for the cell-type proportions since the normalizing constant for that posterior is computationally infeasible to solve. Thus, we cannot proceed with Gibbs sampling in this particular case but make use of M–H sampling (Gelman *et al.*, 2004) instead; Gibbs sampling is a special

[1] Data in linear form is preferred as modeling assumptions may otherwise become violated; see Section 4 for further discussion.

case of M–H, thus, both Gibbs and M–H sampling can be utilized in the same framework (Andrieu *et al.*, 2003).

For employing M–H sampling, one needs an un-normalized posterior of $\mathbf{p}_j$ and a transition kernel. The un-normalized posterior is

$$f(\mathbf{p}_j|\cdot) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{I}\lambda_i e_{ij}^2 + s_j\right\}, \tag{4}$$

where e_{ij} is, again, the model residual and $s_j = \sum_{t=1}^{T}(w_0 p_{0tj}-1)\ln(p_{tj})$. Dirichlet density as the transition kernel for M–H works well in our case since the sampler for the posterior of $\mathbf{p}_j$ must stay within the T-simplex, as previously explained. Now, if the previous value in the Markov chain is denoted by $\mathbf{p}_j^*$, a proposal value, denoted by $\mathbf{p}_j$, will be drawn from Dirichlet($w\mathbf{p}_j^*$), and the corresponding kernel, i.e. Dirichlet density function, is denoted by $K(\mathbf{p}_j^* \to \mathbf{p}_j)$. The role of w is analogous to that of w_0, as w is used to control the peakedness of the transition kernel around the previously sampled value, $\mathbf{p}_j^*$. The acceptance of the proposed, newly sampled value then depends on the factor

$$\rho_j(\mathbf{p}_j^* \to \mathbf{p}_j) = \frac{f(\mathbf{p}_j|\cdot)K(\mathbf{p}_j^* \to \mathbf{p}_j)}{f(\mathbf{p}_j^*|\cdot)K(\mathbf{p}_j \to \mathbf{p}_j^*)}, \tag{5}$$

and the probability of acceptance is determined by $\mathbb{P}[\text{accept}] = \min\{1, \rho_j(\mathbf{p}_j^* \to \mathbf{p}_j)\}$.

3 RESULTS

In computing the forthcoming results with DSection, we used the following values for controlling parameters of our model. Namely, we set peakedness around prior cell-type proportions to $w_0 = 10$, peakedness of transition kernel to $w = 100$, burn-in period to $B = 2000$ iterations, and chain length to $S = 500$ iterations. Along sampling, we also computed and visualized estimates of autocovariance functions of the sampled parameters, which indicated that our choice for the chain length was reasonable, i.e. covariance diminished relatively rapidly as lag was increased (data not shown) (Cowles and Carlin, 1996; Rasmussen, 2000).

3.1 Simulation

In order to demonstrate full functionality of DSection, we designed a simulation experiment containing both multiple cell types and experimental conditions; an analysis of simpler, real data will follow. Expression profiles of 700 genes of three cell types under two experimental conditions were created. The expressions, x_{tic}, were chosen so that there existed probes for which expression profiles were either identical across cell types and conditions, differed only across cell types, differed only across conditions, or both, and expressions were set to vary within the range 100...1600; thus, the theoretically maximum, achievable fold-change is $\log_2(1600/100) = 4$. Next, for each gene, a precision, λ_i, was drawn from Gamma(5, 1/0.0003) (mean precision 0.0015); justification for using the Gamma density is the same as with prior densities. In total, 14 samples, 7 per experimental condition, were created and normally distributed noise with variance $1/\lambda_i$ was added.

Performance of the models is assessed on the basis of their ability to identify differential expression across cell types and experimental conditions—that is, probe i may be differentially expressed across some cell types and experimental conditions, at most in $\binom{6}{2} = 15$ different ways, which are tested separately with the two-sample t-test (see Supplementary Material for more details).

The data are analyzed with the two models, linear regression and DSection, where the latter is utilized both with fixed cell-type proportions and by sampling from posterior of cell-type proportions. Simulation results (Fig. 1) show an increase in identification accuracy of differential expression for DSection, in contrast to our reference, the linear regression model. Thus, the analysis results indicate that our method with uncertainty in proportions incorporated actually attains an accuracy comparable with the 'best-case' scenario, i.e. cell-type proportions are known precisely and a linear regression model is used.

The methods differ mostly in estimation of replication variance, $1/\lambda_i$. Actually the discrepancy between ground-truth and estimates is sometimes so high that we visualize replication standard deviation (SD), $\sqrt{1/\lambda_i}$, instead. As the visuals suggest, only those models assuming fixed and precisely known cell-type proportions suffer from these high biases (Fig. 1c–e), whereas for DSection, which assumes noisy cell-type proportion priors, this bias is absent (Fig. 1f). Importantly, the bias is most strongly present in probes for which differential expression across cell types and experimental conditions is high; to elucidate this, we labeled each SD estimate with a color, and the intensity of that color increased along with average differential expression.

3.2 Affymetrix data

Next, we analyzed a publicly available dataset from Affymetrix oligonucleotide arrays [data downloaded from Affymetrix (2009)], consisting of over 15 000 genes whose heterogeneous expressions comprising of human brain and heart cells were summarized using robust multi-array averaging (RMA) procedure (Irizarry *et al.*, 2003). There are 33 samples in the dataset in total, each sample being designed to contain specific proportions of the distinct cell types. Table 1 contains all the samples provided within the Affymetrix dataset, but we only use those that contain cell types with ratio 25% : 75% and vice versa. Other samples—especially the ones with pure samples that we used for reference—were discarded from the analysis, for better reflecting the scarcity of repeated measurements and heterogeneity within samples, which is usually the case. Moreover, we use the procedure described in the Supplementary Material for deriving noisy estimates for cell-type proportions, in turn reflecting inaccurate prior proportion predictions.

Although no ground-truth for replication variances of Affymetrix data truly exists, we can exploit the samples for each mixture experiment to at least derive good estimates (see Supplementary Material for details). Using these derived ground-truth estimates, Figure 2 shows, again, a similar bias pattern to what is observable with simulated data (Fig. 1). Bias in SD estimation accuracy for most highly differentially expressed genes is visible for the linear regression model that assumes fixed cell-type proportions, whereas DSection, which accounts for noisy cell-type proportion priors, reduces such biases.

Moreover, no ground-truth for truly differentially and non-differentially expressed genes exist for Affymetrix data. However, as we have samples representing pure cell types, they can be derived as well (see Supplementary Material for details). As can be seen in Figure 2b, the receiver operating characteristic (ROC) curves clearly have a similar pattern to what we observed with simulated data. DSection not only outperforms the linear regression model in terms of ROC, but also the performance of DSection is comparable with the 'best-case', which we computed by plugging the true cell-type proportions into the linear regression model, as described earlier.

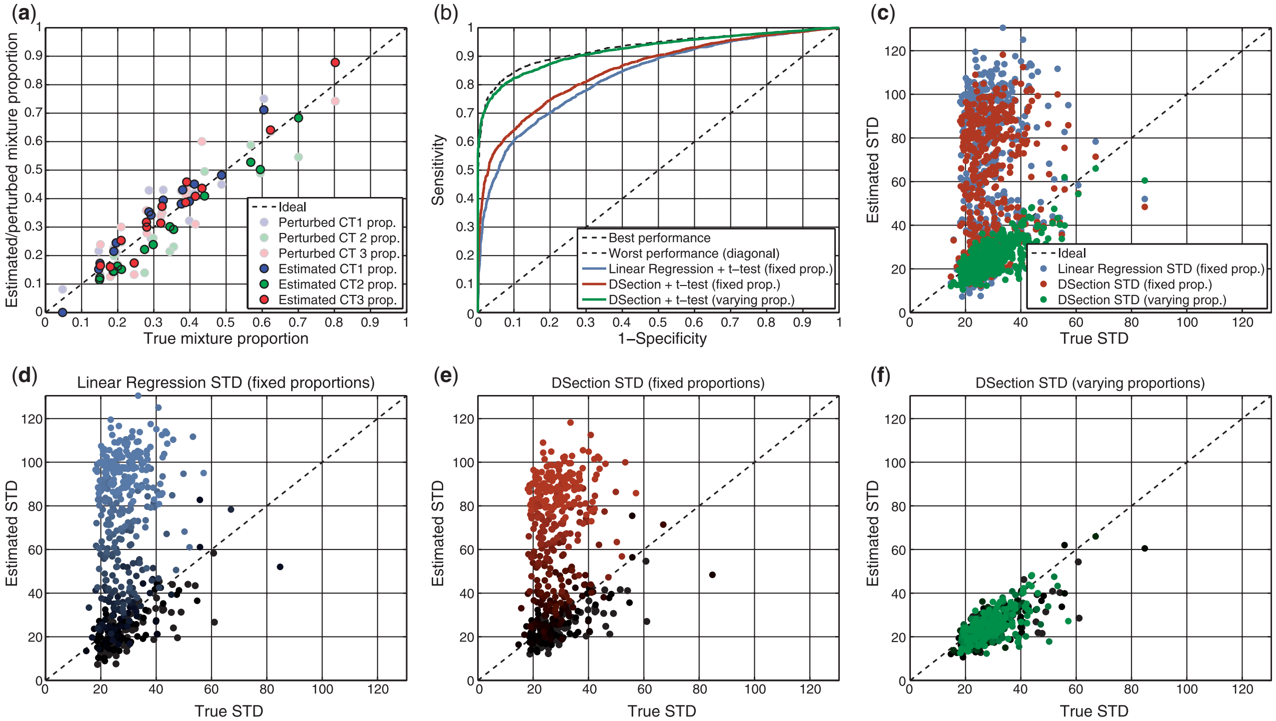

Fig. 1. Analysis results with simulated data—3 cell types, 2 experimental conditions, 700 genes and 14 samples (seven for each experimental condition). (**a**) Estimation of cell-type proportions (bright spots), given noisy priors (faint spots). (**b**) ROC curves of the compared methods (solid lines). As a reference, best performance, obtained by plugging the true cell-type proportions into the linear regression model and performing the analysis, along with the worst performance (diagonal in ROC plots) are visualized as dashed lines. (**c**) Estimation of measurement SD (given as $\sqrt{1/\lambda_i}$). Estimation of measurement SD for (**d**) The linear regression model with fixed cell-type proportions, (**e**) DSection with fixed cell-type proportions and (**f**) DSection with varying cell-type proportions, where estimates are colored depending on true, average differential expressions of probes—higher color intensity means higher average differential expression. Clearly, SD estimation accuracy for highly differentially expressed genes is poor when uncertainty in cell-type proportions are not properly accounted for [(**d**) and (**e**) versus (**f**)].

Table 1. Known cell-type proportions for each sample in Affymetrix data

Sample (j)	1−3	4−6	7−9	10−12	13−21	22−24	25−27	28−30	31−33
Brain (p_{1j})	0.00	0.05	0.10	0.25	0.50	0.75	0.90	0.95	1.00
Heart (p_{2j})	1.00	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.00

For each mixing experiment (one column of the table), a triplet of measurements have been conducted except for samples 13–21, which all have 50%/50% mixing ratio. Samples 10–12 and 22–24 were used for estimating cell-type-specific gene expression profiles, and the expression estimates were then compared with the pure cell-type-specific gene expressions (samples 1–3 and 31–33). Furthermore, we included samples 7–9 and 25–27 when testing how increasing the number of heterogeneous samples for analysis with DSection affects the model performance.

3.2.1 Increasing sample size Additionally, we assessed the effect an increase in sample size has on both cell-type proportion estimation and expression profiling. In addition to the six samples (25%/75% and vice versa) we already used in the previous case study, we augment that data by the ones which contain cell types with ratio 10%/90% and vice versa—that is, 6 more samples making 12 samples in total.

The assessment of improvement was made in the following manner. The six samples of 25%/75% etc. purity were augmented by (i) a subset of $0,1,\dots,6$ samples of 10%/90%, etc. purity, (ii) noise was added to the ground-truth cell-type proportions of the selected samples with the previously used method, (iii) linear regression model and DSection was fitted to the data and (iv) this was repeated 10 times.

For each iteration, mean absolute differences (MAD) between the estimates and ground-truth cell-type proportions and expression profiles were computed, followed by computing a sample mean over the 10 iterations. MAD was preferred as it essentially captures both bias and variance into single quantity. As we increased the number of samples from 6 to 12, MAD was consistently lower for DSection than that for the noisy estimates of cell-type proportions (those used directly with the linear regression model) (Fig. 3). A decreasing trend for MAD is observable while more samples were added, however, that is due to our way of adding noise to cell-type proportions. Namely, the closer the true cell-type proportions are to $1/T$, i.e. as heterogeneous sample as possible, the more noise is added. And since the augmented samples were less heterogeneous in contrast to 25%/75% ones, increasing sample size in turn decreased the

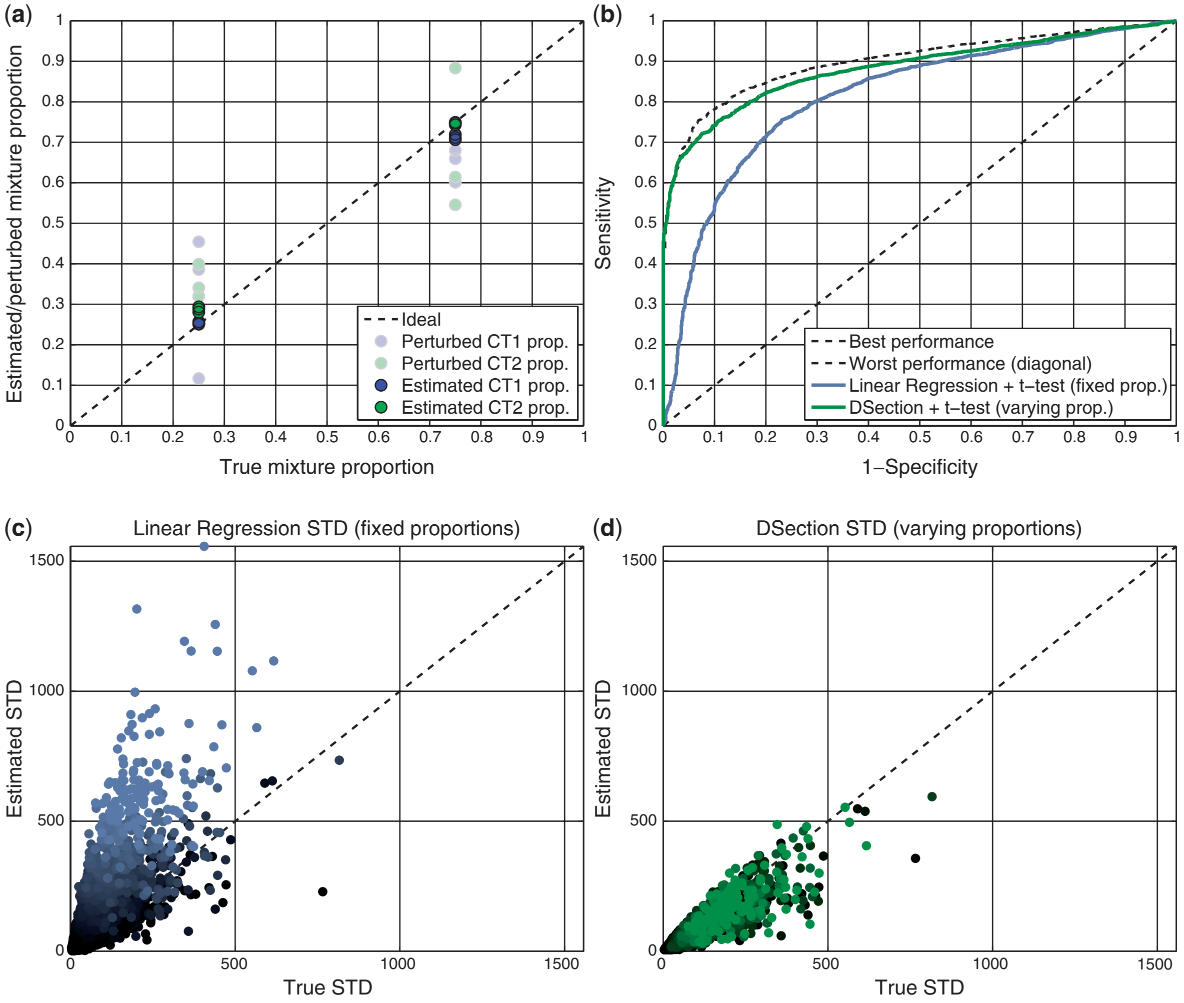

Fig. 2. Analysis results with Affymetrix data—2 cell types, 1 experimental condition, ~15 000 genes and 6 samples (25%/75% and vice versa). (**a**) Estimation of cell-type proportions (bright spots), given noisy priors (faint spots). (**b**) ROC curves of the compared methods. Estimation of measurement STD for (**c**) The linear regression model with fixed cell-type proportions and (**d**) DSection with varying cell-type proportions, where estimates are colored depending on true, average differential expressions of probes. Again, as with simulated data, STD estimation accuracy for highly differentially expressed genes is poor when uncertainty in cell type proportions are not properly accounted for [(**c**) versus (**d**)].

average MAD of noisy cell-type proportions, in turn decreasing the MAD of DSection estimates. We did not observe any significant difference of MAD for expression profiling between the two models (data not shown), indicating that DSection relies heavily upon the priors derived using the deterministic linear regression counterpart.

4 DISCUSSION

Previous studies, including this, have almost exclusively been considering microarray gene expression data. However, due to recent revolutionizing improvements in sequencing techniques, gene expression measurements by sequencing, or RNA-seq (Wang *et al.*, 2009; Wilhelm and Landry, 2009), has become a serious competitor to standard probe-based microarray alternatives, not only due to increased genome coverage offered by RNA-seq, but also due to increased measurement reproducibility (Marioni *et al.*, 2008). Although data preprocessing and normalization steps between microarray and RNA-seq data are different, there are no fundamental factors that would directly make current modeling approaches obsolete. In fact, since a strong linear relationship between RNA concentrations and sequence reads has been reported (Mortazavi *et al.*, 2008), in contrast to not-so-linear microarrays (Quackenbush, 2002), one would expect the modeling transition from array-based analysis to RNA-seq to be rather effortless for any model, including ours.

We propose a framework under which measurements, arising from heterogeneous tissues, can be analyzed without having to rely upon manual—and possibly time consuming—sample preprocessing steps such as LCM. Instead, DSection assumes that measurements contain profiles of all cell types of interest with varying proportions in the tissue samples. Furthermore, as without constraints this task would contain no unique solution for expression profiles and cell-type proportions, uncertain information is assumed to be available on the cell-type proportions. In realistic situations where information

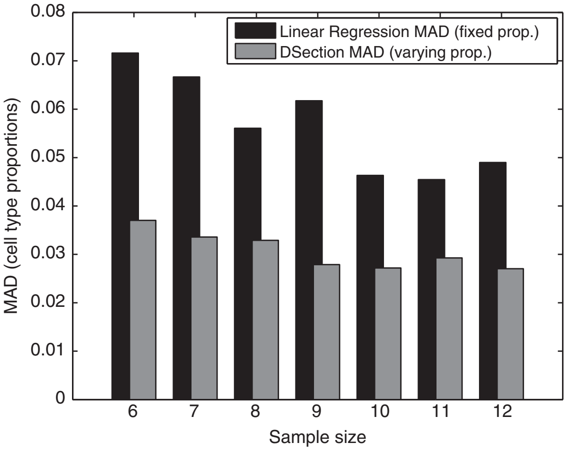

Fig. 3. MAD for cell-type proportion estimates (referenced against the ground-truth). MAD for the linear regression model basically stands for the baseline, i.e. cell-type proportion estimation was not supported by the model, and anything below that (black bars) is considered as improvement. In terms of MAD, DSection (gray bars) is able to recover true cell-type proportions under noisy estimates.

about cell-type proportions is extracted on the basis of, say, microscopy or flow cytometry, it is evident that such estimates are prone to inaccuracy. We showed that, under the Bayesian framework, not only the passing of uncertain information to our model is straightforward due to the notion of prior information, but also that our model is capable of 'de-noising' that uncertain information, thus resulting in more accurate overall modeling performance in contrast to traditional models without this functionality implemented.

The extraction of information about cell-type proportions was not addressed in this article, although it is a crucial part required to make the model work as intended. In real experiments, i.e. those including real tissue samples with unknown cell-type proportions, as opposed to data we used, such precise information as cell-type proportions does not exist. However, as our results suggest, prior information about the proportions of different cell types can be exploited in modeling even though the estimates of proportions would include uncertainty. Thus, including image-based prior estimation could provide a valuable addition into the current analysis framework, but in order to be useful the image analysis needs to be done in an automated manner. Numerous tissue image analysis methods have been presented in the literature, such as those in Kleiner *et al.* (2009); Newberg and Murphy (2008) and Strömberg *et al.* (2007), and incorporating similar methods as a part of the analysis pipeline is one of our main objectives.

Imposing $w_0 = 10$ results in a lightly concentrated density surface around the prior cell-type proportions, $\mathbf{p}_{0j}$, which along with the results suggest that having strong prior information, at least on cell-type proportions, is not required. However, constraining model parameters albeit vaguely is required as the model would otherwise become unidentifiable. If proportions for some cell types are missing, due to morphological indistinguishability, for instance, one could consider pooling those cell types together and model them as one; this approximation would be accurate only in cases where pooled cell types share similar expression profiles. On the other hand, if the precise value for T is debatable but now cell-type proportions for different values of T existed, cross-validation, reversible-jump MCMC (Green, 1995), etc., for determining most suitable T could be utilized.

Although the assumed linearity may not strictly hold for some or even most of the genes being considered, it is still expected that such a linear model can, to some extent, capture nearly linear responses with sufficient accuracy (Hoffmann *et al.*, 2006). In fact, during parameter estimation, we used Affymetrix data with and without log-transform (results shown here are for non-log data) with comparable accuracy in terms of ROC, suggesting that the linearity assumption indeed is fairly robust. Furthermore, Gaussian processes (Rasmussen and Williams, 2006) are currently under investigation as part of incorporating nonlinear responses into the model.

ACKNOWLEDGEMENTS

Academy of Finland (application number 134290; application number 213462, Finnish Programme for Centres of Excellence in Research 2006–2011); Finnish Graduate School in Computational Sciences; National Institutes of Health (application number NIHP50GMO76547).

Conflict of Interest: none declared.

REFERENCES

Abbas,A.R. *et al*. (2009) Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One*, **4**, e6098.

Affymetrix (2009) Available at http://www.affymetrix.com/support/technical/sample_data/gene_1_0_array_data.affx. (last accessed date June 22, 2009).

Andrieu,C. *et al*. (2003) An introduction to mcmc for machine learning. *Mach. Learn.*, **50**, 5–43.

Cowles,M.K. and Carlin,B.P. (1996) Markov chain monte carlo convergence diagnostics: a comparative review. *J. Am. Stat. Assoc.*, **91**, 883–904.

Efron,B. (2009) Are a set of microarrays independent of each other? *Ann. Appl. Stat.*, **3**, 922–942.

Emmert-Buck,M.R. *et al*. (1996) Laser capture microdissection. *Science*, **274**, 998–1001.

Gelman,A. *et al*. (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman,A. (2006) Prior distributions for variance parameters in hierarchical models. *Bayesian Anal.*, **1**, 1–19.

Gosink,M.M. *et al*. (2007) Electronically subtracting expression patterns from a mixed cell population. *Bioinformatics*, **23**, 3328–3334.

Green,P. (1995) Reversible jump Markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.

Hoffmann,M. *et al*. (2006) Robust computational reconstitution - a new method for the comparative analysis of gene expression in tissues and isolated cell fractions. *BMC Bioinformatics*, **7**, 369.

Irizarry,R.A. *et al*. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Jacobsen,M. *et al*. (2006) Deconfounding microarray analysis - independent measurements of cell type proportions used in a regression model to resolve tissue heterogeneity bias. *Methods Inf. Med.*, **45**, 557–563.

Kleiner,H.E. *et al*. (2009) Tissue microarray analysis of eif4e and its downstream effector proteins in human breast cancer. *J. Exp. Clin. Cancer Res.*, **28**, 5.

Lähdesmäki,H. *et al*. (2005) In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, **6**, 54.

Marioni,J.C. *et al*. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods*, **5**, 621–628.

Newberg,J. and Murphy,R.F. (2008) A framework for the automated analysis of subcellular patterns in human protein atlas images. *J. Proteome Res.*, **7**, 2300–2308.

Otsuka,Y. *et al*. (2007) Correlating purity by microdissection with gene expression in gastric cancer tissue. *Scand. J. Clin. Lab. Invest.*, **67**, 367–379.

Quackenbush,J. (2002) Microarray data normalization and transformation. *Nat. Genet.*, **32** (Suppl.), 496–501.

Quon,G. and Morris,Q. (2009) Isolate: a computational strategy for identifying the primary origin of cancers using high-throughput sequencing. *Bioinformatics*, **25**, 2882–2889.

Rasmussen,C.E. (2000) The infinite gaussian mixture model. *Adv. Neural Inf. Process. Syst.*, **12**, 554–560.

Rasmussen,C.E. and Williams,C.K.I. (2006) *Gaussian Processes for Machine Learning*. The MIT Press.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Sooriakumaran,P. *et al*. (2009) A novel method of obtaining prostate tissue for gene expression profiling. *Int. J. Surg. Pathol.*, **17**, 238–243.

Strömberg,S. *et al*. (2007) A high-throughput strategy for protein profiling in cell microarrays using automated image analysis. *Proteomics*, **7**, 2142–2150.

Stuart,R.O. *et al*. (2004) In silico dissection of cell-type-associated patterns of gene expression in prostate cancer. *Proc. Natl Acad. Sci. USA*, **101**, 615–620.

Venet,D. *et al*. (2001) Separation of samples into their constituents using gene expression data. *Bioinformatics*, **17** (Suppl. 1), S279–S287.

Wang,Z. *et al*. (2009) RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Wilhelm,B.T. and Landry,J.-R. (2009) RNA-seq-quantitative measurement of expression through massively parallel rna-sequencing. *Methods*, **48**, 249–257.