BIOINFORMATICS ORIGINAL PAPER Vol. 30 no. 10 2014, pages 1417–1423
doi:10.1093/bioinformatics/btu044

Gene expression

Advance Access publication January 24, 2014

Modeling tissue contamination to improve molecular identification of the primary tumor site of metastases

Martin Vincent[1,*,†], Katharina Perell[2,†], Finn Cilius Nielsen[3], Gedske Daugaard[2] and Niels Richard Hansen[1]

[1]Department of Mathematical Sciences, University of Copenhagen, [2]Department of Oncology, and [3]Center for Genomic Medicine, Copenhagen University Hospital, Rigshospitalet, 2100 Copenhagen Ø, Denmark

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: Contamination of a cancer tissue by the surrounding benign (non-cancerous) tissue is a concern for molecular cancer diagnostics. This is because an observed molecular signature will be distorted by the surrounding benign tissue, possibly leading to an incorrect diagnosis. One example is molecular identification of the primary tumor site of metastases because biopsies of metastases typically contain a significant amount of benign tissue.

Results: A model of tissue contamination is presented. This contamination model works independently of the training of a molecular predictor, and it can be combined with any predictor model. The usability of the model is illustrated on primary tumor site identification of liver biopsies, specifically, on a human dataset consisting of microRNA expression measurements of primary tumor samples, benign liver samples and liver metastases. For a predictor trained on primary tumor and benign liver samples, the contamination model decreased the test error on biopsies from liver metastases from 77 to 45%. A further reduction to 34% was obtained by including biopsies in the training data.

Availability and implementation: http://www.math.ku.dk/~richard/msgl/.

Contact: vincent@math.ku.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 22, 2013; revised on October 21, 2013; accepted on January 21, 2014

1 INTRODUCTION

Several studies have considered molecular predictors for primary tumor site identification (see Lu *et al.*, 2005; Ramaswamy *et al.* 2001; Rosenfeld *et al.*, 2008). All these studies report misclassification percentages of ~10% in predicting the primary tumor site from primary tumor samples, which is consistent with our findings (see Section 3). However, the performance of molecular predictors on metastatic samples is less clear. Most studies assess the performance of their predictor using a combination of primary tumor and metastatic samples, with an unbalanced metastatic sample set. Metastases are often more difficult to classify by conventional diagnostics, compared with primary tumors. Yet, samples of metastatic cancers are generally underrepresented in the validation. To correctly validate the performance of a predictor, the validation samples must be representative of the samples that the predictor is intended to be used on. For the majority of patients with metastatic cancer, identification of the primary tumor site relies on small formalin-fixed paraffin-embedded (FFPE) needle biopsies (core biopsies) from metastatic lesions. Hence, we argue that a molecular predictor designed to assist in the diagnostic work-up of patients with metastatic cancer must be compatible with, and validated on, FFPE core biopsies from metastatic lesions.

To develop such a predictor for primary tumor site identification of metastases, it would seem preferable to train it exclusively on core biopsies constituting metastatic tissue. However, it may be difficult to obtain a sufficient amount of metastatic biopsies of known primary tumor origin. Furthermore, a larger technical variation because of the smaller amount of processed material and the varying tumor content make it difficult to rely on core biopsies only. Therefore, previous studies, as well as ours, have relied on primary tumor samples as a predominant part of the training set.

These considerations lead to the central problem that we address in this article: *when core biopsy samples are scarce or completely absent, how can we best adapt primary tumor samples to build a molecular predictor able to identify the primary tumor site of metastatic tumor samples*?

Such an adaptation is generally referred to as a *domain adaptation*. The problem being that the distribution of the cases to be predicted (the target domain) differs from the distribution of the cases used for training (the source domain) (see e.g. Daumé *et al.*, 2006; Mansour *et al.*, 2009). To overcome this problem, we explicitly model how the target domain is related to the source domain, and use this model to train a predictor. A central difference between the source and target domains in our setting is caused by the fact that a core biopsy is often contaminated by tissue surrounding the tumor cells, that is, benign tissue from the biopsy site unrelated to the tumor (see Fig. 1). Another difference can arise if the molecular signature of the metastatic tumor deviates from the signature of the primary tumor; see, e.g. Ramaswamy *et al.* (2002) or Albini *et al.* (2008) for a discussion of the biology of metastases.

Recent findings by Elloumi *et al.* (2011) suggest that tissue contamination may result in a decline of performance for molecular predictors. The results presented in the present article

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

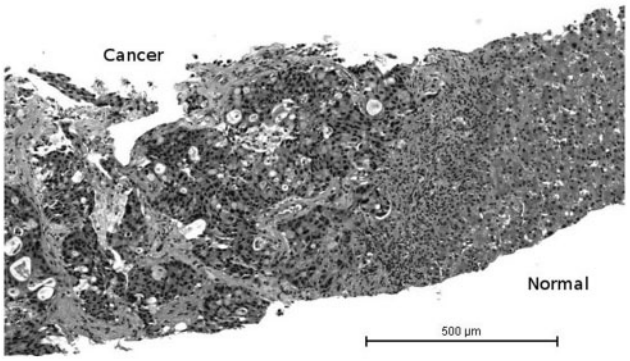


Fig. 1. Micrograph of a liver core biopsy with cancer. Cancer as well as normal (benign) tissue can clearly be seen

demonstrate that tissue contamination hampers primary tumor site identification of non-microdissected metastatic liver core biopsies. We developed our contamination model to correct this. However, our suggested methodology is not specific to primary tumor site identification, the biomarkers used—in our case microRNA (miRNA) expression—or the technological platform. It is a general computational model, which is broadly applicable whenever sample contamination constitutes a problem.

We found that a non-adapted primary tumor site predictor, trained solely on miRNA expression measurements from primary tumors and benign liver samples, misclassified 60% of the liver core biopsies in our sample set. Notably, the performance of the predictor was even worse for the subset of liver core biopsies constituting metastases. By applying our contamination model and suggested domain adaptation procedure, this performance was improved. For a predictor based on 50 miRNAs, the contamination model decreased the misclassification percentage for metastatic liver core biopsies from 77 to 45%. A further improvement was obtained by combining the contamination model with a few biological metastatic liver core biopsies. This combined approach reduced the misclassification percentage for metastatic liver core biopsies down to 34%.

2 METHODS

2.1 Domain adaptation

To present our contamination model in an appropriate context, we briefly review the domain adaptation terminology. The goal is to predict samples drawn from a distribution $\mathcal{T}$ called the *target distribution* or *target domain*. However, none or only a small number of samples drawn from $\mathcal{T}$ are available for training. The philosophy of domain adaptation is to use a related distribution $\mathcal{S}$, called the *source distribution* or *source domain*, to construct a target domain predictor. Typically, either the source distribution or a predictor trained on samples from the source distribution are adapted to the target domain, perhaps, using (the few) available target samples.

In our setting, the source distribution is the distribution of molecular signatures for primary tumor and benign liver samples (resections). The target distribution is the distribution of molecular signatures from liver core biopsies. The relation between the source distribution and the target distribution is made explicit in Section 2.2 below. We suggest to apply the following general domain adaptation strategy to our problem.

(1) Specify a domain adaptation *model*, that is, specify a map

$$F : \mathcal{F}_{\text{source}} \to \mathcal{F}_{\text{target}}$$

where $\mathcal{F}_{\text{source}}$ and $\mathcal{F}_{\text{target}}$ denote the sets of relevant source and target distributions, respectively.

(2) With $\hat{\mathcal{S}}$ the empirical distribution of source samples, use the plug-in estimate $F(\hat{\mathcal{S}})$ as an estimate of the target distribution.

(3) Generate an approximate target distribution $\mathcal{T}_{\text{sim}}$ by simulation of artificial target samples based on $F(\hat{\mathcal{S}})$.

(4) Use $\mathcal{T}_{\text{sim}}$ for training of a predictor.

The domain adaptation model F provides a special kind of relation between source and target; for any given source, $\mathcal{S}$, there is only one related target, the *model target* $F(\mathcal{S})$. In practice, the model F will contain unknown components that have to be specified or estimated from data as well. The following sections describe how the steps of the general procedure above are carried out in the context of tissue contamination. Supplementary Figure S1 shows a flow chart of the entire implementation.

2.2 Contamination models

A simple model of the molecular signature from a liver-contaminated core biopsy is

$$\alpha \times \text{primary tumor signature} + (1-\alpha) \times \text{normal liver signature}$$

where $\alpha \in [0, 1]$ is the relative amount of tumor content. This is a plausible model on the molecular level, but the contamination is not necessarily additive on the measured scale. Therefore, we need to transform the measurements using a suitable *scale function* $f : \mathbb{R}^p \to \mathbb{R}^p$, which is a function

$$f(x_1, \ldots, x_p) = (g(x_1), \ldots, g(x_p))$$

for a continuous strictly monotone function $g : \mathbb{R} \to \mathbb{R}$.

Letting $(X, Y) \in \mathbb{R}^p \times \{1, \ldots, K\}$ denote a pair of random variables, with X representing the molecular signature and Y the site label, the distribution of (X, Y) is the source distribution. With $Z \in \mathbb{R}^p$ a random variable representing the molecular signature of the contaminating tissue, let

$$U(y) \stackrel{\text{def}}{=} f\big(\alpha f^{-1}(X) + (1-\alpha) f^{-1}(Z)\big) \,|\, Y = y$$

for f a scale function and $\alpha \in [0, 1]$ another random variable representing the relative amount of tumor content. The variables X, Z and α are assumed conditionally independent, given Y, as illustrated in Figure 2a. With $Y_{\mathcal{T}}$ the marginal distribution of class labels in the target distribution, the distribution of $(U(Y_{\mathcal{T}}), Y_{\mathcal{T}})$ constitutes our model target. The model is specified by choosing a scale function and conditional distributions of Z and α given Y.

We will consider domain adaptation models for two particular scale functions. The *linear scale* function is given by simply taking f to be the identity, thus assuming that the contamination is additive on the measured scale. The data used in this article consist of miRNA (a small non-coding RNA molecule) expression measurements, obtained by quantitative polymerase chain reaction (qPCR); see, e.g., Vaerman *et al.* (2004) or VanGuilder *et al.* (2008) for an introduction to qPCR. For qPCR, a logarithmic scale function that models the relation between the actual miRNA concentration and the measured quantity is appropriate. The *log scale* function is given by

$$g(\delta) \stackrel{\text{def}}{=} -R \log \delta$$

for $\delta > 0$ proportional to the molecular count and $R > 0$ a constant. The log scale function can be derived based on theoretical considerations for qPCR reactions. The constant $e^{\frac{1}{R}} - 1$ is the amplification efficiency, and we use a standard value of 0.8 throughout the article.

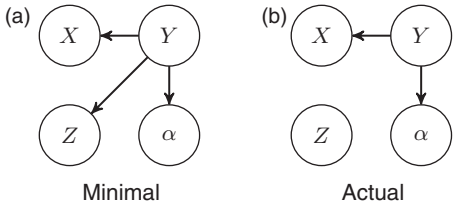


Fig. 2. The random variables α, X, Y and Z represent the tumor proportion, the molecular signature, the site label and the molecular signature of the contaminating tissue, respectively. A minimal assumption (**a**) of our domain adaptation model is that X, Z and α are conditionally independent, given Y. The actual assumption made (**b**) is that Z is marginally independent of the other variables, and that X and α are conditionally independent, given Y

2.3 Simulation of artificial core biopsies

The approximate target distribution $\mathcal{T}_{\text{sim}}$ is generated as a weighted empirical distribution of a total of M simulated samples. First, the class labels $y_1, \ldots, y_M$ are sampled or chosen. Based on the source dataset (primary tumors, normal and cirrhotic liver), we sample $x_1, \ldots, x_M$ independently, given $y_1, \ldots, y_M$, such that $x_i|y_i$ is sampled from the conditional empirical distribution of the source data given y_i. That is, x_i is drawn with replacement from the source data with class label y_i. In addition, we draw M samples with replacement $z_1, \ldots, z_M$, from the contamination data and M samples $\alpha_1, \ldots, \alpha_M$ independently given $y_1, \ldots, y_M$ such that $\alpha_i|y_i$ has the desired distribution. For a given scale function f, we compute

$$u_i = f\big(\alpha_i f^{-1}(x_i) + (1-\alpha_i) f^{-1}(z_i)\big)$$

and the empirical distribution of the samples $(u_1, y_1), \ldots, (u_M, y_M)$ with weights $\omega_1, \ldots, \omega_M$ form the approximate target distribution $\mathcal{T}_{\text{sim}}$. The weights can be chosen to achieve a desired distribution of class labels in $\mathcal{T}_{\text{sim}}$, for example, to match the distribution of class labels for the primary tumor dataset. The simulation assumes the conditional independence structure illustrated in Figure 2b. If the less restrictive conditional independence structure illustrated in Figure 2a is assumed, we need to adjust the simulation to draw z_i conditionally on y_i, but this requires knowledge of a class-dependent contamination distribution. Note that the biological liver core biopsies do not need to completely fulfill the assumptions on which the simulations are based. For our purpose, the usefulness of the contamination model and the resulting artificial target data is judged by its ability to improve the performance of the predictor.

2.4 Class prediction

For class prediction, we use multinomial logistic regression, which is the multiclass extension of logistic regression. The predictor is trained using a group-lasso penalized likelihood approach as described in detail in Vincent *et al.* (2014).

We briefly review the multinomial group-lasso regression method. Consider a prediction problem with K classes, N samples and p features. Assume, given a dataset $(x_1, y_1), \ldots, (x_N, y_N)$ where, for all $i = 1, \ldots, N$, $x_i \in \mathbb{R}^p$ is the observed feature vector, and $y_i \in \{1, \ldots, K\}$ is the categorical class label. With $h : \{1, \ldots, K\} \times \mathbb{R}^K \to \mathbb{R}$ defined as

$$h(l, \eta) \stackrel{\text{def}}{=} \frac{\exp(\eta_l)}{\sum_{k=1}^{K} \exp(\eta_k)}$$

the (symmetric) multinomial model is given by

$$P(Y = l|x) = h(l, \beta^{(0)} + \beta x)$$

Table 1. Number of samples included in the study. For the metastatic liver core biopsies from the non– liver-related cancers the numbers a and b shown as (a/b) are the numbers of low and high tumor content samples, respectively

Class description	Resections (primaries)	Liver core biopsies
Breast cancer	17	7 (5/2)
Colorectal cancer	20	12 (8/4)
Gastric/cardia cancer	18	12 (8/4)
Pancreatic cancer	20	10 (5/5)
Squamous cell cancers (of different origins)	16	12 (6/6)
Hepatocellular carcinoma	17	3
Cholangiocarcinoma	20	4
Subtotal	**128**	**60**
Cirrhotic liver	17	8
Normal liver	20	7
Total	**165**	**75**

Here, the parameters are organized as the K-dimensional vector $\beta^{(0)}$ of intercept parameters and the $K \times p$ matrix

$$\beta \stackrel{\text{def}}{=} \big(\beta^{(1)} \cdots \beta^{(p)}\big)$$

with $\beta^{(j)} \in \mathbb{R}^K$ the parameters associated with the jth feature. The group-lasso maximum likelihood estimator of $(\beta^{(0)}, \beta)$ is the minimizer of the group-lasso penalized negative log-likelihood,

$$-\sum_{i=1}^{N} \omega_i \log h(y_i, \beta^{(0)} + \beta x_i) + \lambda \sum_{J=1}^{p} \big\|\beta^{(J)}\big\|_2 \qquad (1)$$

Where $\omega_1, \ldots, \omega_N$ are sample weights. Here, $\|\cdot\|_2$ is the 2–norm on $\mathbb{R}^K$. The penalization results in feature selection, meaning that for some features the corresponding parameter vector is estimated to be 0. The regularization parameter $\lambda > 0$ is a tuning parameter, and the larger the λ, the fewer features are selected.

For comparison, we also apply the procedure used in Ferracin *et al.* (2011) for class prediction. They used analysis of variance (ANOVA) for selection of miRNAs, and then prediction analysis of microarrays (PAM) for the actual prediction; see Ferracin *et al.* (2011) for the details.

2.5 Biological samples

A total of 240 samples were included in the study. These samples consist of 128 resected primary tumors of different origin (representing seven primary tumor classes), 60 liver core biopsies from metastatic or primary liver tumors of known origin (representing the same predefined seven primary tumor classes), 37 liver resections and 15 liver core biopsies from normal and cirrhotic liver (see Table 1). The five classes breast, colorectal, gastric/cardia, pancreatic and squamous cell cancers are referred to as the *non–liver-related cancers*, and the corresponding 53 liver core biopsies are referred to as the *metastatic liver core biopsies*. Samples were obtained from The University Hospital of Copenhagen, Denmark. The sample set is a subset of the training samples used in Perell *et al.* (unpublished data) and available from the Gene Expression Omnibus with accession number GSE51429. Samples were archived FFPE tissues (dated 2000–2012). All primary tumor resections were cut into one section of 10 µm and microdissected before being processed to remove the surrounding benign tissue. Core biopsies were cut into two sections of 5 µm according to the standard pathological procedure—no microdissection was performed. Malignant core biopsy samples were required to have a minimum tumor content of 10%. Tumor content

was defined as tumor and stromal cells. All samples were from independent patients; thus, no patient overlap between primary tumor samples and liver core biopsy samples was accepted. All samples were reviewed by an independent pathologist to confirm the reference diagnosis and to estimate the proportion of tumor (see Table 2). Based on these tumor percentage estimates, the core biopsy samples were divided into two groups representing high (>50%) and low (<50%) tumor content. For each sample, the expression levels of 377 miRNAs were measured using quantitative real-time PCR (TaqMan low-density array cards, human MicroRNA array A, Applied Biosystems) according to the manufacturer's instructions.

Before a predictor was trained, the artificial core biopsy data were simulated and data were preprocessed. Simulation and preprocessing were executed in the following order.

(1) Controls and miRNAs not expressed in the primary tumor samples were removed.
(2) Linear and log scale artificial core biopsies were simulated (see Section 2.3).
(3) All samples, such as primary tumor samples, core biopsy samples and the artificial core biopsy samples, were normalized and then standardized as described below.

The data were first *normalized* by centering and scaling the individual samples to mean 0 and variance 1. The purpose of normalization is to remove technical (non-biological) variation. It was demonstrated in Mestdagh *et al.* (2009) that for qPCR, mean centering outperforms alternative normalization strategies. Supplementary Figure S2 shows the effect of the normalization on the sample distribution. To ensure that differences in scale do not influence the variable selection, data were *standardized* by centering and scaling the expression measurements of each miRNA across the samples.

The centers and scales were estimated using the primary tumor samples and applied for standardization of the primary tumor samples as well as the core biopsy and artificial core biopsy samples. That is, the standardized sample $\tilde{x} \in \mathbb{R}^p$ of a sample $x \in \mathbb{R}^p$ is given by

$$\tilde{x}_i = \frac{x_i - \hat{\mu}_i}{\hat{\sigma}_i} \quad \text{for } i = 1, \ldots, p$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ denote the empirical mean and standard deviation, respectively, for the ith miRNA in the primary tumor dataset. Note that the order in which normalization and standardization are applied matters.

3 RESULTS

In this section, we present results obtained by testing the suggested domain adaptation method on the miRNA expression dataset described in Section 2.5. The domain adaptation method was used with a linear and a log scale contamination model, as described in Section 2.2. Predictors were constructed using multinomial group lasso, outlined in Section 2.4, and by using the ANOVA + PAM method from Ferracin *et al.* (2011).

Table 2. Distribution of tumor percentages (visual estimates) in core biopsies

Tumor content (%):	0–20	20–40	40–60	60–80	80–100
Number of samples:	4	21	8	9	11

Results obtained with and without using the domain adaptation method were compared.

To assess the generalization ability of the predictor method, a multinomial group-lasso predictor was trained and cross-validated solely on the 165 resections. This predictor achieved an overall 10-fold cross-validation error of 11% on the resected primary tumors, normal liver and cirrhotic liver samples. However, on liver core biopsies—our target domain—this predictor had an overall test error of ~60%.

Note that results are reported in terms of misclassification percentages of the predictors, which we also refer to as the *errors* of the predictors. The errors are either estimated on test data by cross-validation or by subsampling as described in detail in the following sections. Furthermore, all errors reported henceforth—except Section 3.5—originate from validation on our target domain, i.e., samples from one of the following categories: liver core biopsies of metastatic tumors, liver core biopsies of primary liver cancers and core biopsies of normal and cirrhotic liver.

3.1 Predictors based on primary tumor samples

The results presented in this section were obtained using primary tumor and benign liver resections for training, i.e. no core biopsy samples were used for training. Errors were estimated solely on liver core biopsy samples. We compared multinomial group-lasso predictors obtained by training on one of the following three datasets:

(a) Primary tumor and benign liver samples.
(b) Artificial core biopsies obtained using the linear scale contamination model.
(c) Artificial core biopsies obtained using the log scale contamination model.

For the simulation of artificial core biopsies, as described in Section 2.3, we used the 20 normal liver resections as contamination data. The distribution of α was taken as a beta distribution with shape parameters (2, 2) for the seven cancer classes, and degenerate at one for the two benign liver classes. That is, for the benign liver classes, no contamination was added. The two artificial core biopsy datasets were generated using either the linear or the log scale contamination model described in Section 2.2. The simulation was carried out with 750 samples from each class, and the weights were chosen as

$$\omega_i = \frac{N_{\text{prim}, y_i}}{750}$$

where, N_{prim, y_i} is the number of primary samples of class y_i. For the two benign liver classes, the simulation amounts to sampling with replacement from their empirical distribution. In practice, the simulation step for these two classes was therefore skipped, and the 37 benign liver resections were just included, all with weight 1.

Figure 3 shows the test error plotted against the number of miRNAs included in the predictor. A larger number of miRNAs correspond to a lower value of the tuning parameter λ in (1). The predictors trained directly on the primary tumor, normal liver and cirrhotic liver samples performed poorly when applied to the

liver core biopsy samples. The best predictors achieved an overall error of ~60%. Figure 4 further shows that the error was even larger on the metastatic liver core biopsies. From Figure 3, we see that the overall error dropped to ~50% for the predictors trained on the artificial core biopsies using the linear scale contamination model. Using the log scale contamination model, the overall error could be further reduced to ~35%. Figure 4 shows that the overall improvements embraced notable differences between the sample subgroups. In particular, the log scale contamination model performed best on core biopsies with low tumor content, whereas it did not improve the error rate as much for the high tumor content samples. In contrast, the linear scale contamination model resulted in marginally better predictors for the high tumor content samples.

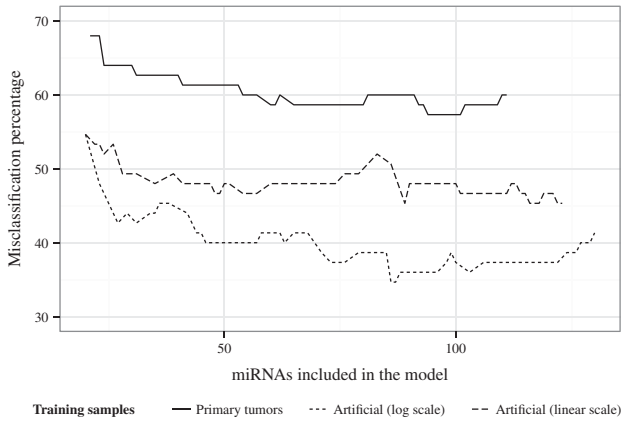


Fig. 3. Test error for prediction of liver core biopsies. The error is shown as a function of miRNAs included in the predictor. The predictors were trained using primary tumor samples or artificial core biopsy samples derived from primary tumor samples using either the log or the linear scale contamination model

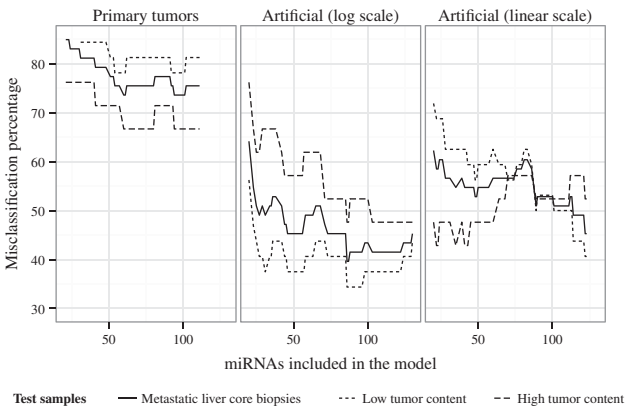


Fig. 4. Test error for primary tumor site prediction of metastatic liver core biopsies. Predictors were trained on the resections (primary tumor, normal and cirrhotic liver samples) or artificial core biopsies. The artificial core biopsies were obtained using either the log or the linear scale contamination model. The error is shown as a function of miRNAs included in the predictor model. The error of low tumor content and high tumor content samples is shown in addition to the overall error of metastatic liver core biopsies

3.2 Comparison of distributions

The main purpose of introducing the contamination model was to improve the performance of the predictor. In addition, we validated the model itself by comparing the distributions of the simulated artificial core biopsies, the biological core biopsies and the primary tumor resections. Because the contamination model does not alter the normal and cirrhotic liver signature, these classes were kept out of the comparison. Hence, we compared the distribution of the seven cancer classes. Figure 5 shows a bivariate projection of the distributions. It shows that the distribution of the primary tumor samples clearly differed from the distribution of the biological core biopsies. The figure also shows that the distribution of the artificial core biopsies matched the distribution of the biological core biopsies better. The figure was constructed as follows: for each of the nine classes, we computed the class mean of the primary tumor samples and projected the samples onto the two first principal components of these nine class means.

3.3 Predictor based on metastatic tumor samples

For comparison, we trained a predictor solely on the 75 available liver core biopsies. The leave-one-out cross-validation error was ~55% for the best predictor, using ~100 miRNAs. It is likely that the error could be further reduced if additional core biopsies were available for training.

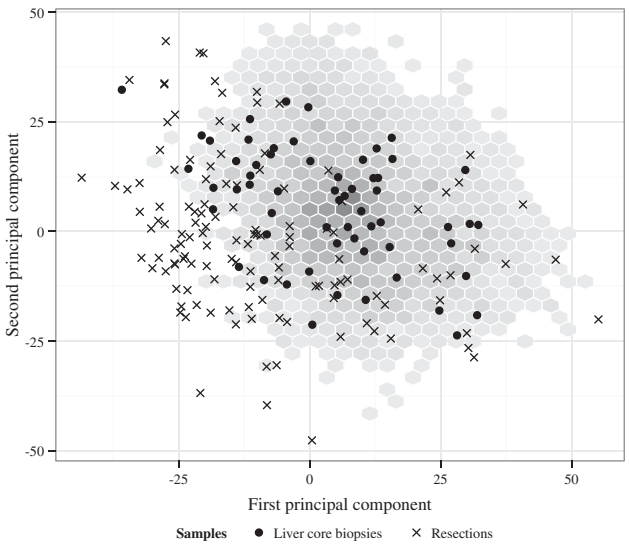


Fig. 5. A bivariate projection of the liver core biopsies, primary tumor resections and artificial core biopsies (the latter shown using hexagonal binning). Only samples from the seven cancer classes are shown. The projection is onto the first two principal components of the class means for the primary samples

3.4 Predictors based on primary and metastatic tumor samples

To further improve the predictors, we investigated the effect of training on either primary tumor samples or artificial core biopsies in combination with available biological core biopsies. Thus, for training we included $n = 1, 2, 3$ or four core biopsies from each of the non–liver-related classes. To assess the performance of this procedure, we used the following subsampling scheme—repeated 100 times.

(1) Randomly split the core biopsy samples into a training and a test dataset, such that each contains approximately half of the samples per class.

(2) Randomly sample n core biopsies from the training dataset for each of the five non–liver-related classes.

(3) Train the predictor on the combined training dataset (artificial core biopsies or resections + biological core biopsies selected for training).

(4) Estimate the error using the core biopsy test data.

Figure 6 shows that the combination of primary tumors and core biopsies resulted in predictors with a smallest achievable error ~40% for the metastatic liver core biopsies. Combining the artificial core biopsies with the biological core biopsies, this error was reduced to just above 30% using the log or the linear scale contamination model. Table 3 shows that the 50 miRNA predictor based on the artificial core biopsies and four biological core biopsies per class and using the log scale, achieved an error of 34%. Figure 6 further shows that the optimal predictors were achieved with 40–60 miRNAs using the log scale and 60–80 miRNAs using the linear scale. For the log scale contamination model, a minor improvement on low tumor content samples was seen when including biological core biopsies. However, a considerable improvement was seen for high tumor content samples. In contrast, for the linear scale contamination model, a considerable improvement was seen for low, as well as high tumor content samples, when biological core biopsies were included.

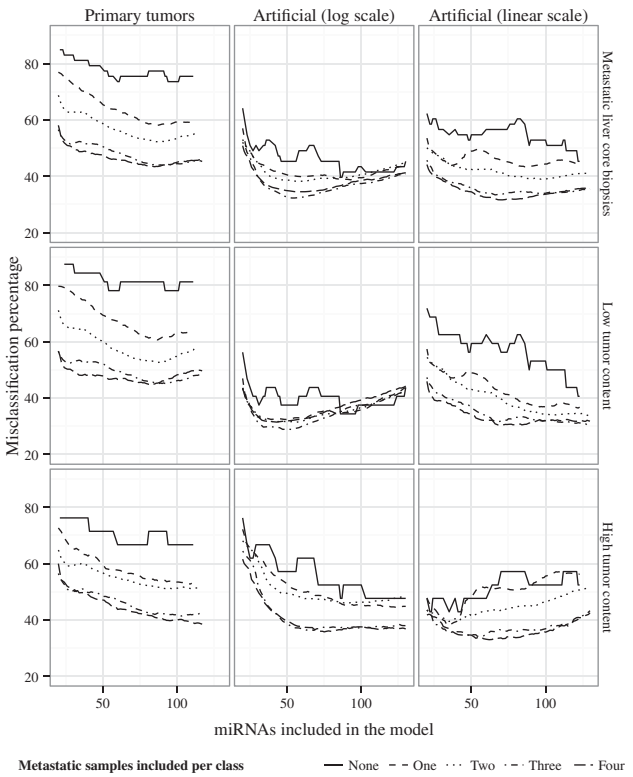


Fig. 6. The error of primary tumor site prediction of metastatic liver core biopsies. The error is shown as a function of miRNAs included in the predictor model. The plots show the error for predictors trained on resections (primary tumor samples) or artificial core biopsies in combination with biological core biopsies. The line type represents the number of biological core biopsies included per class during the estimation procedure. Biological core biopsies were not included in the hepatocellular carcinoma, cholangiocarcinoma, normal and cirrhotic liver classes

3.5 Comparing predictors

We compared multinomial group-lasso predictors with predictors obtained using the ANOVA + PAM method applied in Ferracin *et al.* (2011). For comparison, the test errors on the 53 metastatic liver core biopsies were used, and predictors using 50 or 100 miRNAs were considered. Table 3 shows the results.

We trained ANOVA+PAM predictors on our 165 resections of primary tumors, normal and cirrhotic liver. The predictor worked as expected on the primary tumor samples with cross-validation errors on the non–liver-related classes being 23% (50 miRNA predictor) and 16% (100 miRNA predictor). However, as seen in Table 3 the corresponding errors were 81 and 77% when the predictor was applied to metastatic liver core biopsies.

Large errors were seen for all of the predictors trained solely on primary tumor, normal and cirrhotic liver samples, as shown in Table 3. These errors decreased, if we either included core biopsies in the training, or if we trained on the artificial samples from our log scale contamination model, instead of the primary

Table 3. Misclassification percentages for predicting the primary site of metastatic liver core biopsies using 50 or 100 miRNAs

Principal training data	Number of core biopsies	ANOVA + PAM		Multinomial group lasso	
		Number of miRNAs			
		50	100	50	100
Primaries	0 (0)	81%[a]	77%[a]	77%	74%
	2 (10)	74%	71%	59%	54%
	4 (20)	64%	64%	48%	45%
Artificial	0 (0)	60%	57%	45%	43%
	2 (10)	–[b]	–[b]	39%	41%
	4 (20)	–[b]	–[b]	34%	39%

Note: The table summarizes the errors for two predictor models using various training data setups. The number of core biopsies included for training is denoted by a (b), where a is the number per class and b is the total number of included core biopsies. Core biopsies were only included for training in non–liver-related classes.
[a]This predictor is constructed similarly to the predictor presented in Ferracin *et al.* (2011). [b]Individual sample weighting is required—ANOVA + PAM does not directly support this.

resections. The improvements of the multinomial group-lasso predictors were larger than those obtained by the ANOVA + PAM predictors.

4 DISCUSSION

Contamination of samples can affect the performance of molecular predictors. This has been reported in other studies (e.g. Elloumi *et al*., 2011). In the present article, we have seen that normal liver contamination hampers primary tumor site identification of liver biopsies. This was seen regardless, whether multinomial group lasso or ANOVA + PAM was used.

Several molecular predictors for identification of the primary tumor site have been published, which include Ferracin *et al*. (2011), Meiri *et al*. (2012), Rosenfeld *et al*. (2008), Lu *et al*. (2005) and Ramaswamy *et al*. (2001). Commonly, these predictors rely on microdissection to minimize tissue contamination, and they generally require samples with high tumor content. In Meiri *et al*. (2012), a tumor content >60% was required, whereas 33 of the 53 metastatic liver core biopsies in our dataset had a tumor content <60%. Notably, microdissection may not always be applied to core biopsies or may cause delay in the diagnostic work-up.

Previously reported primary tumor site predictors commonly result in misclassification percentages of ~25–30%. The predictor reported in Meiri *et al*. (2012) was validated on the combination of 334 primary and 146 metastatic tumor samples, with the samples being a mixture of resections and biopsies. A correct single prediction was reported for 74% of the samples, corresponding to a misclassification percentage of 26%. The ANOVA + PAM predictor reported in Ferracin *et al*. (2011) obtained a misclassification percentage of 27% when tested on 45 microdissected metastases. Furthermore, Ferracin *et al*. (2011) reported a misclassification percentage of 31% when this predictor was applied to the data from Rosenfeld *et al*. (2008). In the present article, we show, how an ANOVA + PAM predictor exclusively trained on primary tumor, normal and cirrhotic liver samples, results in a misclassification percentage of 77% when validated on the 53 metastatic liver core biopsies. Hence, the ANOVA + PAM predictor is, like the multinomial group-lasso predictor, not able to generalize well from the primary tumor samples to non-microdissected metastatic liver core biopsies with a heterogenous tumor content distribution. We did not find that any of the previous studies systematically addressed the problem of identifying primary tumor site of non-microdissected core biopsies. Addressing this problem, we found that novel methodology was needed to explicitly incorporate tissue contamination.

We have developed a computational approach that deals with tissue contamination from surrounding tissue, and we have shown that the method improves the performance of the molecular predictor. Notably, training of such predictors does not require metastatic samples from the specific biopsy site, e.g. liver. Comparable improvements could be obtained by including liver core biopsies in the training data for each of the non–liver-related classes. Moreover, including liver core biopsies *in combination with the contamination model* was shown to reduce the error even further. With a sufficiently large number of metastatic liver core biopsies, the improvement because of the contamination model is likely to be small. However, an advantage of the contamination model is that it reduces the number of metastatic samples from the specific biopsy site needed for training. Combining the contamination model with 20 core biopsies, the misclassification percentage of the non-microdissected metastatic liver core biopsies was reduced to 34%.

Although we only considered liver core biopsies in the present article, it is natural to assume that our contamination model will work for other biopsy sites as well. With a limited number of benign tissue resections, representing the biopsy site of interest, our contamination model could easily be used to develop a predictor adapted to core biopsies of other metastatic sites. In addition, even though we only address tissue contamination, our contamination model has the potential to be used to model background contamination in other types of samples.

ACKNOWLEDGEMENTS

The authors thank Prof Ben Vainer, DMSc, Bodil Laub Pedersen, DMSc, and Birgitte Federspiel, DMSc, from the Department of Pathology, University Hospital of Copenhagen, for their substantial work reviewing and scoring the samples included in the study.

Funding: M.V. was supported by the Ministry of Science, Technology and Innovation (09-049337). M.V. and N.R.H. were supported by the University of Copenhagen Program of Excellence: Statistical Methods for Complex and High Dimensional Models.

Conflict of interest: None declared.

REFERENCES

Albini,A. *et al.* (2008) Metastasis signatures: genes regulating tumor-microenvironment interactions predict metastatic behavior. *Cancer Metastasis Rev.*, **27**, 75–83.

Daumé,H. III and Marcu,D. (2006) Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.*, **26**, 101–126.

Elloumi,F. *et al.* (2011) Systematic bias in genomic classification due to contaminating non-neoplastic tissue in breast tumor samples. *BMC Med. Genomics*, **99**, 54.

Ferracin,M. *et al.* (2011) MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin. *J. Pathol.*, **255**, 43–53.

Lu,J. *et al.* (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.

Mansour,Y. *et al.* (2009) Domain adaptation: learning bounds and algorithms. *In Proceedings of The 22nd Annual Conference on Learning Theory (COLT 2009), Montréal, Canada, June 2009*, Omnipress.

Meiri,E. *et al.* (2012) A second-generation microRNA-based assay for diagnosing tumor tissue origin. *Oncologist*, **17**, 801–812.

Mestdagh,P. *et al.* (2009) A novel and universal method for microRNA RT-qPCR data normalization. *Genome Biol.*, **10**, R64.

Ramaswamy,S. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15149–15154.

Ramaswamy,S. *et al.* (2002) A molecular signature of metastasis in primary solid tumors. *Nature Genet.*, **33**, 49–54.

Rosenfeld,N. *et al.* (2008) MicroRNAs accurately identify cancer tissue origin. *Nat. Biotechnol.*, **26**, 462–469.

Vaerman,J.L. *et al.* (2004) Evaluation of real-time PCR data. *J. Biol. Regul. Homeost. Agents*, **18**, 212–214.

VanGuilder,HD. *et al.* (2008) Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, **44**, 619–626.

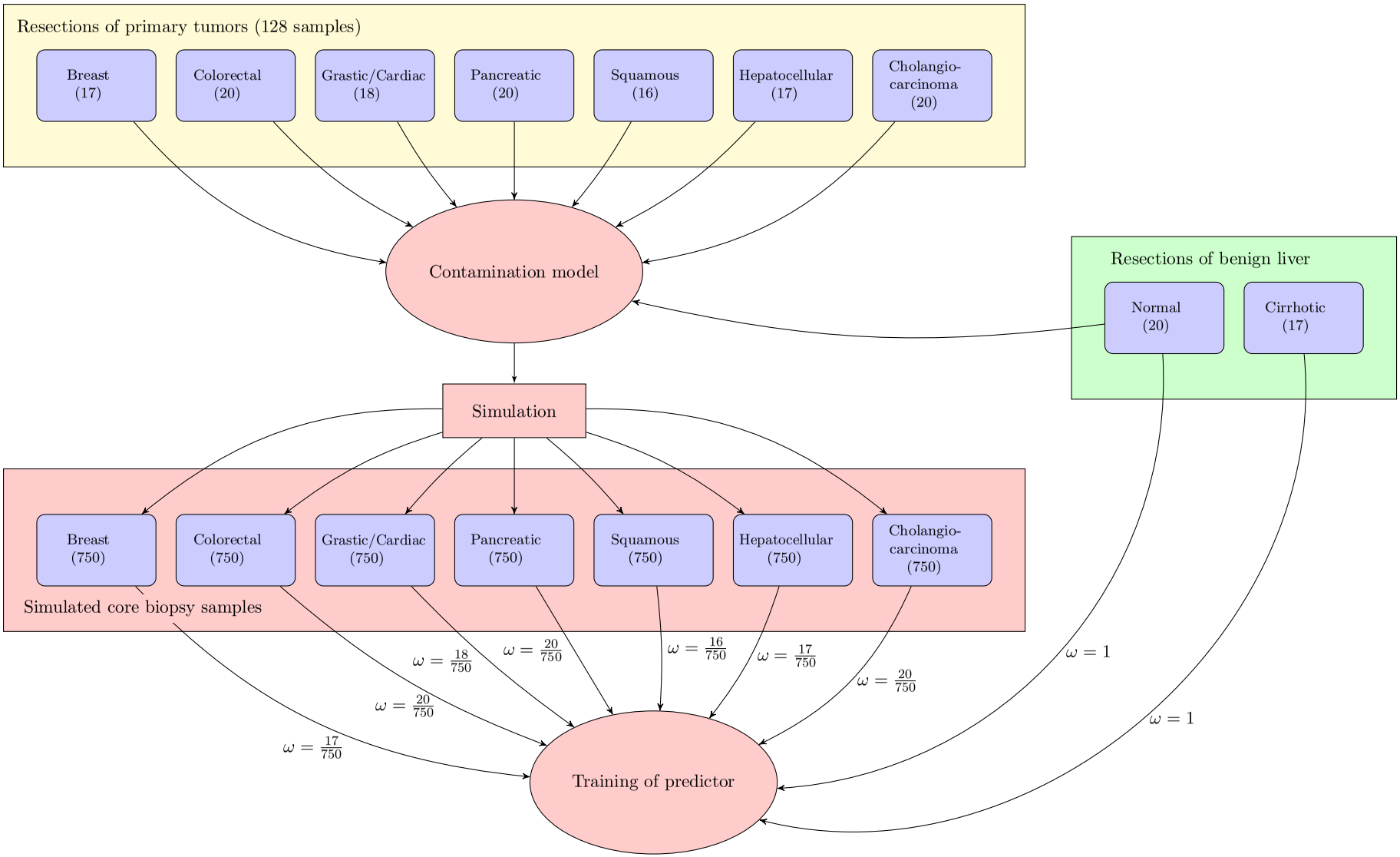
Vincent,M. and Hansen,N. (2014) Sparse group lasso and high dimensional multinomial classification. *Comp. Stat. Data Anal.*, **71**, 771–786.

Supplementary Figures:

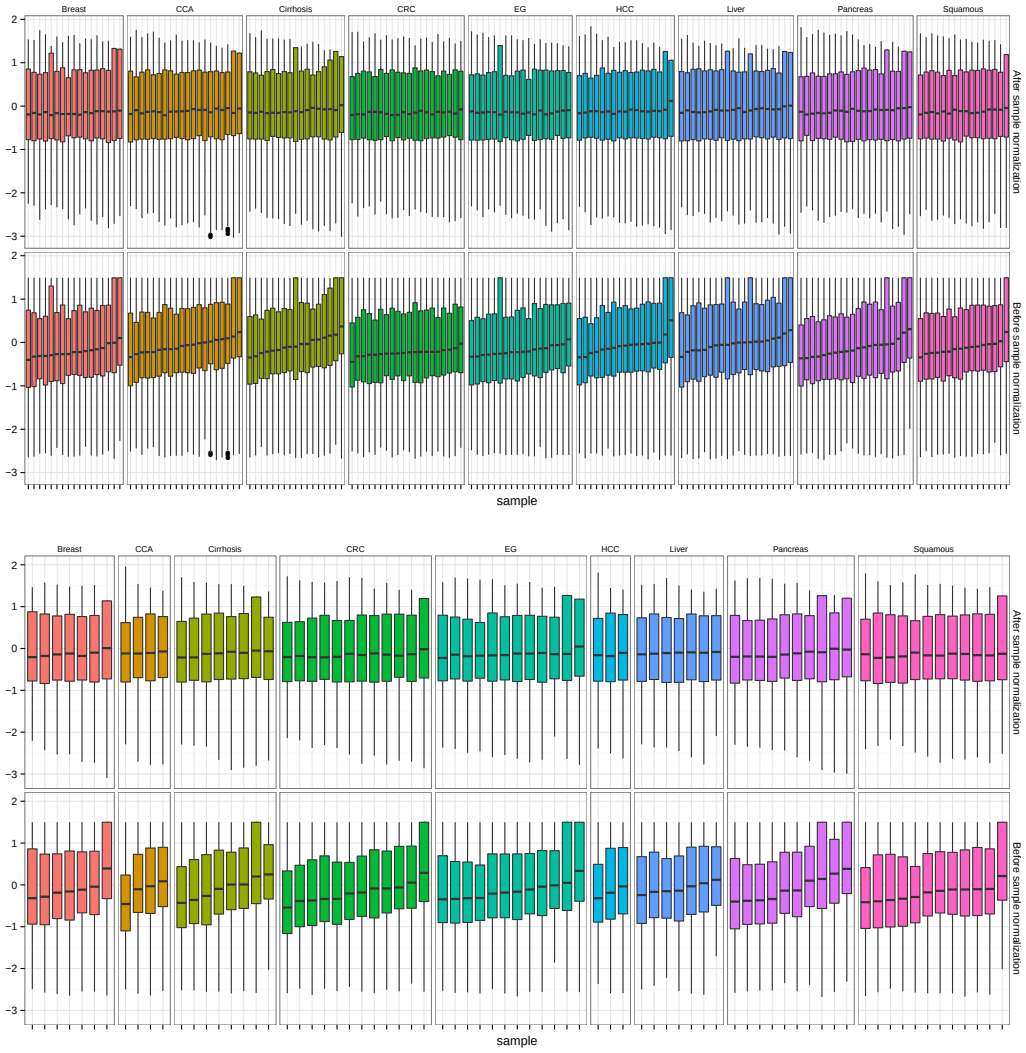
Modeling tissue contamination to improve molecular identification of the primary tumor site of metastases

Martin Vincent, Katharina Perell, Finn Cilius Nielsen,
Gedske Daugaard and Niels Richard Hansen

October 21, 2013



Supplementary Figure 1: Flow chart representation of the implementation of the domain adaptation procedure used for the results reported in Section 3.1. The contamination model is constructed on the basis of samples from the 7 cancer classes and the normal liver class. The artificial core biopsy samples are simulated from this model. In the implementation we simulate 750 samples from each of the 7 cancer classes. The predictor is trained on the combination of the simulated samples with individual sample weights ω and the 37 biological samples from the 2 benign liver classes.



Supplementary Figure 2: Boxplots showing the effect of the normalization on the distribution within samples. The two top panels show data from the resections. The two bottom panels show data from the core biopsies. The samples have been ordered within each of the 9 classes according to the median of the unnormalized measurements within each sample.