BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 12 2011, pages 1645–1652
doi:10.1093/bioinformatics/btr199

*Gene expression*

Advance Access publication April 21, 2011

# Exploiting prior knowledge and gene distances in the analysis of tumor expression profiles with extended Hidden Markov Models

Michael Seifert[1,*], Marc Strickert[2], Alexander Schliep[3] and Ivo Grosse[4]

[1]Department of Molecular Genetics, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, [2]Science and Technology, University of Siegen, Siegen, Germany, [3]Department of Computer Science and BioMaPS Institute for Quantitative Biology, Rutgers, The State University of New Jersey, Piscataway, USA and [4]Institute of Computer Science, Martin Luther University Halle, Halle (Saale), Germany

Associate Editor: Joaquin Dopazo

## ABSTRACT

**Motivation:** Changes in gene expression levels play a central role in tumors. Additional information about the distribution of gene expression levels and distances between adjacent genes on chromosomes should be integrated into the analysis of tumor expression profiles.

**Results:** We use a Hidden Markov Model with distance-scaled transition matrices (DSHMM) to incorporate chromosomal distances of adjacent genes on chromosomes into the identification of differentially expressed genes in breast cancer. We train the DSHMM by integrating prior knowledge about potential distributions of expression levels of differentially expressed and unchanged genes in tumor. We find that especially the combination of these data and to a lesser extent the modeling of distances between adjacent genes contribute to a substantial improvement of the identification of differentially expressed genes in comparison to other existing methods. This performance benefit is also supported by the identification of genes well known to be associated with breast cancer. That suggests applications of DSHMMs for screening of other tumor expression profiles.

**Availability:** The DSHMM is available as part of the open-source Java library Jstacs (www.jstacs.de/index.php/DSHMM).

**Contact:** seifert@ipk-gatersleben.de

**Supplementary information:** Supplementary data are available at *Bioinformatics* online. Supplementary data files are available at the Jstacs's web site.

Received on December 12, 2010; revised on April 5, 2011; accepted on April 8, 2011

## 1 INTRODUCTION

Chromosomal mutations like amplifications and deletions of DNA segments are one of the key genetic mechanisms leading to changes of gene expression levels in tumors. Different studies have shown that between 40% and 60% of genes in highly amplified regions tend to be overexpressed (Heidenblad *et al.*, 2005; Hyman *et al.*, 2002; Pollack *et al.*, 2002). Also epigenetic changes of DNA methylation or histone modifications are known to locally bias expression levels on chromosomes (Frigola *et al.*, 2006; Stransky *et al.*, 2006). Due to such mutations, gene expression levels of adjacent genes on chromosomes tend to be positively correlated. Moreover, gene expression levels of adjacent genes in close chromosomal proximity tend to be higher correlated than those of adjacent genes in greater distance. This is illustrated in Figure 1 for breast cancer gene expression data from (Pollack *et al.*, 2002).

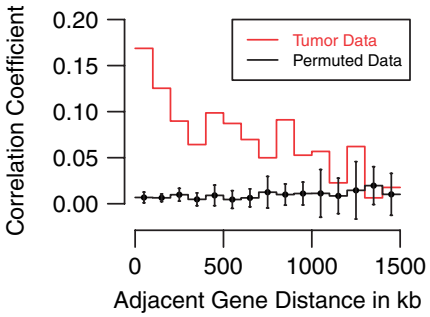


**Fig. 1.** Distance-based spatial correlations of log-ratios measured for directly adjacent genes on chromosomes in the breast cancer gene expression dataset by (Pollack *et al.*, 2002). Adjacent genes on each chromosome were grouped into distance classes between 100 and 1500 kb in steps of 100 kb. The Pearson's correlation coefficient has been computed for the log-ratios of all pairs of adjacent genes in one distance class for the tumor data (red) and for 100 random permutations of log-ratios per chromosome of the tumor data (black).

In recent years, different approaches have been proposed for the analysis of gene expression data in the context of chromosomal locations of genes. The Human Transcriptome Map (Caron *et al.*, 2001) was the first large-scale approach to study genome-wide human gene expression profiles. This mapping of gene expression data to chromosomal locations revealed a complex organization of the human genome in which highly expressed genes tend to be localized in clusters. Besides this, methods like CGMA (Crawley and Furge, 2002), or MACAT (Toedling *et al.*, 2004), and LAP (Callegaro *et al.*, 2006), both modeling chromosomal distances of genes, have been developed to improve the analysis of gene expression profiles in the context of chromosomal locations. A common characteristic of all these methods is the requirement of replicates of two defined separate samples (e.g. tumor and healthy tissue) for the identification of differentially expressed chromosomal regions based on specific test statistics coupled with permutation tests. These methods cannot be applied to studies designed to analyze individual tumor expression profiles using two-color microarrays without replicates (e.g. Pollack *et al.*, 2002). In this situation, often

*To whom correspondence should be addressed.

standard log-fold-change thresholds are applied to individual tumor expression profiles for characterizing differentially expressed genes (e.g. Hasegawa *et al.*, 2002; Pollack *et al.*, 2002).

A similar problem, closely related to the analysis of individual tumor expression profiles, is the analysis of data measured by comparative genomic hybridization experiments (e.g. Beroukhim *et al.*, 2010; Pinkel and Albertson, 2005). In the last years, different methods have been developed for the analysis of these data in the context of chromosomal locations of probes for identifying chromosomal regions with decreased or increased copy numbers in tumor compared with healthy tissue. Two studies by Lai *et al.* (2005) and Willenbrock and Fridlyand (2005) have put great efforts on the comparison of these methods, and the best-performing methods have been made available by the ADaCGH web server (Diaz-Uriarte and Rueda, 2007) enabling their application with standardized input and output. Especially, ChARM (Myers *et al.*, 2004) has been demonstrated to work on tumor expression profiles. Thus, all these different methods could be useful for the identification of differentially expressed genes in tumor.

In order to incorporate spatial correlations between gene expression levels of adjacent genes in tumor, we develop a Hidden Markov Model (HMM) for the analysis of tumor expression profiles in the context of chromosomal locations of genes. Motivated by the trend that adjacent genes in close chromosomal proximity tend to have more similar expression levels than adjacent genes in greater distance (Fig. 1), we further extend this model to an HMM with scaled transition matrices (SHMM). Adapting the use of HMMs with switched transition matrices (Knab *et al.*, 2003), the idea of scaling transition matrices by state duration was initially proposed by Seifert (2006) for improved modeling of spatial correlations. The resulting SHMM was further extended in Seifert *et al.* (2009) to distinguish between gene pair orientations in the analysis of promoter-array ChIP-chip data. Here, the SHMM is modified to integrate chromosomal distances of adjacent genes on chromosomes into the identification of differentially expressed genes in tumor. This is done by distinguishing between adjacent genes in close chromosomal proximity and adjacent genes in greater distance. To overcome the potential limitations of this fixed separation, the SHMM is further extended to an HMM with distance-scaled transition matrices (DSHMM) that directly integrates individual distances of adjacent genes into the state-transition process.

Moreover, we make use of prior knowledge about the distribution of measurements of underexpressed, unchanged and overexpressed genes during the training of the three proposed HMM-based approaches. This is realized by extending the standard Baum–Welch algorithm (e.g. Durbin *et al.*, 1998; Rabiner, 1989) to a Bayesian Baum–Welch algorithm.

We apply the HMM, the SHMM and the DSHMM to a breast cancer gene expression dataset by Pollack *et al.* (2002). Based on this, we investigate the effect of incorporating prior knowledge into the training on the identification of overexpressed genes in breast cancer. We further analyze the impact of modeling dependencies and distances between adjacent genes on chromosomes by comparing the HMM approaches against a mixture model (e.g. Bilmes, 1998) ignoring dependencies and distances. Additionally, we compare our models against existing related approaches for the analysis of comparative genomic hybridization data including two other HMM-based approaches by Fridlyand *et al.* (2004) and Marioni *et al.* (2006). The first of these two approaches integrates chromosomal locations and the second additionally models chromosomal distances. But, both approaches do not incorporate prior knowledge into the training. Moreover, we also determine hot spots of underexpression and overexpression in the breast cancer dataset and evaluate these genes based on the Breast Cancer Database (Telikicherla *et al.*, 2008), additional literature searches and database searches using Oncomine (Rhodes *et al.*, 2007).

## 2 METHODS

First, the breast cancer gene expression dataset is introduced. Then, the HMM for analyzing tumor expression profiles in the context of chromosomal locations of genes is developed. Next, this model is further extended to the SHMM that integrates chromosomal distances of adjacent genes into the analysis of tumor expression profiles by making use of two fixed scaled transition matrices. This model is further extended to the DSHMM utilizing distance-scaled transition matrices. Then, the integration of prior knowledge into the training of model parameters is considered. Finally, details of the initialization and other basic settings are given.

### 2.1 Breast cancer gene expression dataset

The breast cancer gene expression dataset by Pollack *et al.* (2002) is used to identify genes that are differentially expressed in breast cancer in comparison to a reference sample consisting of a mixture of normal tissues. This dataset contains gene expression levels for 4 breast cancer cell lines and 37 tumors across 6095 genes of the 23 human chromosomes leading to $k \in \{1,\ldots,943\}$ tumor expression profiles. That is, for each chromosome in each cell line and in each tumor, a chromosome-specific tumor expression profile $\vec{o}(k) = (o_1(k),\ldots,o_{T_k}(k))$ was measured. This profile contains the relative expression level of each gene $t \in \{1,\ldots,T_k\}$ given by the $\log_2$-ratio $o_t(k)$ of its expression level in tumor divided by its corresponding expression level in the reference sample. All log-ratios in a tumor expression profile are ordered from the p-arm to the q-arm of the chromosome based on the chromosomal locations of the corresponding genes. Summary statistics of the dataset are shown in Figure 2.

### 2.2 Standard HMM

A three-state HMM with state-specific Gaussian emission densities is used to identify differentially expressed genes in tumor expression profiles. The set of hidden states of the HMM is denoted by $S := \{-,=,+\}$. Motivated by the quantile–quantile plot of the log-ratios in Figure 2b, the three states are defined to model the following gene categories. State '=' models unchanged genes with log-ratios of about zero, underexpressed genes in tumor with log-ratios much less than zero are represented by state '−' and overexpressed genes in tumor with log-ratios much greater than zero are modeled by state '+'. The state of gene $t$ is denoted by $q_t \in S$. A state sequence $\vec{q} = (q_1,\ldots,q_{T_k})$ underlying the tumor expression profile $\vec{o}(k)$ is assumed to be generated by a homogeneous first-order Markov model. This model is parameterized by the initial state distribution $\vec{\pi} := (\pi_i)_{i\in S}$ with initial state probability $\pi_i \in (0,1)$ so that $\sum_{i\in S}\pi_i = 1$ and by the stochastic transition matrix $A := (a_{ij})_{i,j\in S}$ with state-transition probability $a_{ij} \in (0,1)$ and $\sum_{j\in S} a_{ij} = 1$ for each $j \in S$. The state sequence $\vec{q}$ belonging to the tumor expression profile $\vec{o}(k)$ is hidden. To model the log-ratio $o_t(k)$ of gene $t$ by the HMM, it is assumed that $o_t(k)$ is generated by a state-specific Gaussian emission density $b_i(o_t(k)) := 1/(\sqrt{2\pi}\sigma_i)\exp(-0.5(o_t(k)-\mu_i)^2/\sigma_i^2)$ of state $q_t = i \in S$. The corresponding emission parameters are represented by the matrix $B := (\mu_i,\sigma_i)_{i\in S}$ defining the state-specific mean $\mu_i \in \mathbb{R}$ and the state-specific standard deviation $\sigma_i \in \mathbb{R}^+$ for each state $i \in S$.

In summary, the parameters of the specified HMM are denoted by $\lambda := (\vec{\pi}, A, B)$. The three-state architecture of this model is illustrated in

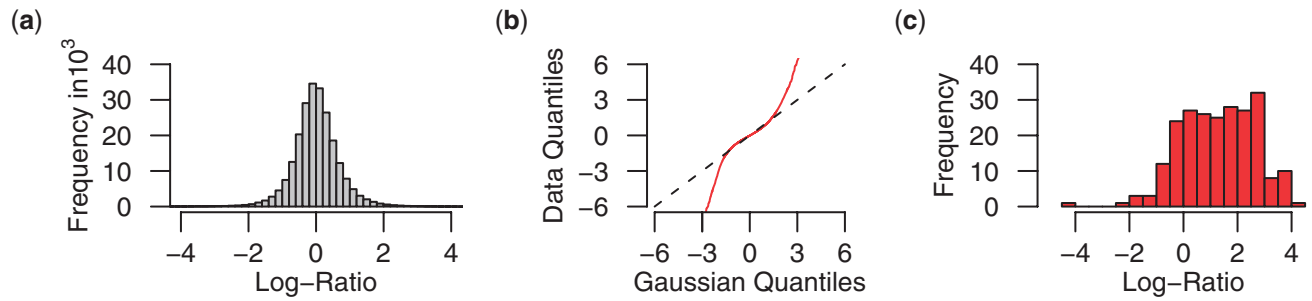


**Fig. 2.** Characteristics of the breast cancer gene expression dataset by Pollack *et al.* (2002). (**a**) Histogram of log-ratios representing gene expression levels in tumor compared with healthy tissue. Log-ratios about zero characterize genes with unchanged expression levels in tumor. Differentially expressed genes in tumor have log-ratios much less (underexpressed) or much greater (overexpressed) than zero. (**b**) Quantile–quantile plot comparing the quantiles of the log-ratios in the dataset with those obtained by a Gaussian density with mean 0.01 and standard deviation 0.7 estimated from the data. The black dashed line shows the expected quantiles if the log-ratios would follow the estimated Gaussian density. The red curve shows the deviation of the log-ratio distribution from the black dashed line indicating an enrichment of log-ratios in the far tails that represent putatively underexpressed and overexpressed genes in tumor. (**c**) Histogram of log-ratios showing the coupling between breast cancer gene expression levels of genes having at least 3-fold increased copy number in tumor compared with healthy tissue. The majority of these genes tends to be overexpressed indicated by log-ratios much greater than zero. Measurements of gene copy numbers were taken from the corresponding dataset by Pollack *et al.* (2002).

Supplementary Figure S1. The HMM realizes dependencies between log-ratios $o_t(k)$ and $o_{t+1}(k)$ of adjacent genes $t$ and $t+1$ on a chromosome via its internal first-order Markov model.

For the identification of differentially expressed genes, the HMM is used to compute the probability with which a gene $t$ in tumor expression profile $\vec{o}(k)$ is assigned to a state $i \in S$ of the HMM. This is done by computing the state-posterior probability $\gamma_t^k(i) := P[q_t = i | \vec{o}(k), \lambda]$ using the Forward–Backward algorithm (Rabiner, 1989). That enables the ranking of individual classes of genes (e.g. overexpressed genes) as well as the assignment of the most likely underlying state of each gene using the state-posterior decoding algorithm (Rabiner, 1989).

## 2.3 HMM with scaled transition matrices

The extension of the standard HMM to an HMM with scaled transition matrices (SHMM) enables the integration of chromosomal distances of adjacent genes into the analysis of tumor expression profiles. This allows to model the trend shown in Figure 1 that two adjacent genes in close chromosomal proximity tend to have more similar expression levels in tumor than two adjacent genes in greater distance. Inspired by HMM-based approaches for financial time-series (Knab *et al.*, 2003), tumor expression (Seifert, 2006) and ChIP-chip data analysis (Seifert *et al.*, 2009), the integration of this trend is realized by extending the homogeneous first-order Markov model, responsible for the state transitions of the HMM, to a specific inhomogeneous first-order Markov model that integrates the chromosomal distance of adjacent genes in a tumor expression profile into the state-transition process of the SHMM.

To include these distance information, each pair of adjacent genes $t$ and $t+1$ measured in a tumor expression profile $\vec{o}(k)$ of a chromosome is assigned to a transition class

$$c_t(k) := \begin{cases} 2, & \text{genes } t \text{ and } t+1 \text{ in distance } d_t \leq b \\ 1, & \text{otherwise} \end{cases}$$

in dependency of the chromosomal distance $d_t$ of both genes and a globally predefined distance threshold $b$. This transition class is used to select one of the two transition matrices $A_1 := (a_{ij}(1))_{i,j \in S}$ and $A_2 := (a_{ij}(2))_{i,j \in S}$ realizing the state-transition process of the inhomogeneous first-order Markov model of the SHMM. $A_1$ is assumed to be identical to a basic transition matrix $A := (a_{ij})_{i,j \in S}$ as specified for the HMM, and $A_2$ is computed based on $A$ by increasing the state duration $1/(1-a_{ii})$ of each state $i \in S$ by a predefined scaling factor $f_2 > 1$. Based on this, the parameters of transition matrix $A_c$ of transition class $c \in \{1, 2\}$ are expressed by

$$A_c := (a_{ij}(c)) := \begin{cases} \frac{a_{ii} - 1 + f_c}{f_c}, & i = j \\ \frac{a_{ij}}{f_c}, & i \neq j \end{cases} \quad (1)$$

with respect to the parameters of the basic transition matrix $A$ and a scaling factor $f_c \geq 1$, fulfilling $A_1 = A$ for $f_1 := 1$.

Thus, the SHMM transitions from state $q_t \in S$ of gene $t$ to state $q_{t+1} \in S$ of gene $t+1$ in a tumor expression profile $\vec{o}(k)$ using the transition probability $a_{q_t q_{t+1}}(c_t(k))$ of the corresponding transition matrix $A_{c_t(k)}$. This realizes that adjacent genes in close chromosomal proximity modeled by $A_2$ have a higher probability to be represented by the same state of the SHMM than gene pairs in greater distance modeled by $A_1$, because the self-transition probability $a_{ii}(2)$ of each state $i \in S$ in $A_2$ is greater than $a_{ii}(1)$ in $A_1$. Similarly, the general usage of different transition classes was first proposed by Knab *et al.* (2003) for the analysis of financial time-series data. In contrast to that, the SHMM has a reduced number of transition parameters by making use of scaled transition matrices enabling the separation of adjacent genes into two fixed distance-specific transition classes.

In spite of the extension of the state-transition process, the initial state distribution and the state-specific Gaussian emission densities defined for the HMM can still be used without modifications for the SHMM illustrated in Supplementary Figure S2. The variable parameters of the SHMM are denoted by $\lambda := (\vec{\pi}, A, B)$. This model reduces to the standard HMM by setting $f_2 = 1$. The identification of differentially expressed genes is done in analogy to the HMM under consideration of the specific transition classes assigned to pairs of adjacent genes.

## 2.4 HMM with distance-scaled transition matrices

A conceptual extension of the SHMM is achieved by replacing the fixed transition matrices by a continuous transition function of gene distance. This leads to an HMM with distance-scaled transition matrices (DSHMM) directly integrating the distance of adjacent genes into the state-transition process. Such a model avoids potential discretization effects caused by non-continuous changes of transition probabilities for small variations in gene distances.

For realizing this, the transition class $c_t(k)$ of each pair of adjacent genes $t$ and $t+1$ in a chromosome-specific tumor expression profile $\vec{o}(k)$ is redefined

to represent a standardized gene distance

$$c_t(k) := \begin{cases} d_t/b, & \text{genes } t \text{ and } t+1 \text{ in distance } d_t \leq b \\ 1, & \text{otherwise} \end{cases}$$

in dependency of the chromosomal gene distance $d_t$ and a globally predefined distance threshold $b$. This standardized gene distance defines a linear ramp function of gene distance taking values in the interval $[0,1]$. The ramp characteristic allows to limit the relationship between distance of adjacent genes and similarity of their gene expression levels to biologically relevant chromosomal distances. Based on that, a distance-specific scaling factor is defined by

$$f_{c_t(k)} := 1+(F-1)\cdot(1-c_t(k))$$

with respect to the standardized gene distance $c_t(k)$ and a globally predefined maximal scaling factor $F \geq 1$. This distance-specific scaling factor is used to rescale the state duration $1/(1-a_{ii})$ of each state $i \in S$ in a basic transition matrix $A := (a_{ij})_{i,j \in S}$ utilizing the transition matrix $A_c$ defined in Equation (1).

That is, for each transition from gene $t$ to gene $t+1$, the corresponding distance-scaled transition matrix $A_{c_t(k)}$ is used. This transition matrix defines a continuous function of gene distance for which adjacent genes in close chromosomal proximity have a higher probability to be represented by the same state of the DSHMM than adjacent genes in greater distance. The distance-scaled transition matrix $A_{c_t(k)}$ reduces to the basic transition matrix $A$ for adjacent genes in distance greater than $b$. A continuous reduction to the basic transition matrix is realized by the distance-specific scaling factor $f_{c_t(k)}$.

In comparison to the SHMM with two fixed transition matrices, the DSHMM with distance-scaled transition matrices improves the modeling of adjacent genes in close chromosomal proximity having more similar expression levels than adjacent genes in greater distance (e.g. Fig. 1). Similar concepts were proposed by Marioni *et al.* (2006) and Rueda and Diaz-Uriarte (2007) to integrate chromosomal distances of adjacent probes into the HMM-based analysis of Array-CGH data.

As for the SHMM, the same initial state distribution $\vec{\pi}$ and state-specific Gaussian emission densities with parameters $B$ are used for the DSHMM. The variable parameters of the DSHMM are denoted by $\lambda := (\vec{\pi}, A, B)$. Differentially expressed genes in tumor are identified in analogy to the standard HMM under consideration of the distance-scaled transition matrices. The DSHMM reduces to the standard HMM that ignores chromosomal distance of genes by using a maximal scaling factor $F := 1$.

## 2.5 Modeling of prior knowledge

A problem-specific characterization of the model parameters of HMM, SHMM and DSHMM is achieved by including prior knowledge about tumor expression data into the training. This is done by defining a prior distribution

$$P[\lambda|\Theta] := D_1(\vec{\pi}|\Theta_1)\cdot D_2(A|\Theta_2)\cdot D_3(B|\Theta_3) \qquad (2)$$

over the parameters of the model $\lambda := (\vec{\pi}, A, B)$ with respect to hyperparameters $\Theta := (\Theta_1, \Theta_2, \Theta_3)$. This prior represents a product of independent prior distributions for the initial state distribution $\vec{\pi}$, the transition matrix $A$ and the emission parameters $B$ of $\lambda$. For each class of model parameters, a conjugate prior distribution is used enabling the analytical parameter estimation and the integration of prior knowledge about potential parameter values during the training. The following prior distributions used for the initial state distribution, the transition matrix and the Gaussian emission densities represent the usual ones used for HMMs (e.g. Bishop, 2006; Durbin *et al.*, 1998). A transformed Dirichlet distribution is used as prior for the initial state distribution (MacKay, 1998). The prior for the transition matrix is a product of transformed Dirichlet distributions. A product of Gaussian-Inverted-Gamma distributions is used as prior for the state-specific Gaussian emission densities (Evans *et al.*, 2000). Details to these distributions are given in the Appendix A of the Supplementary Material.

Generally, the choice of hyperparameters for the prior should enable the model to distinguish between underexpressed, unchanged and overexpressed genes in tumor. Appropriate choices are problem-specific and depend on the size of the dataset. A histogram of log-ratios (e.g. Fig. 2a) helps to characterize each state of the model. Details to used hyperparameters are reported in the Appendix A of the Supplementary Material.

## 2.6 Training using prior knowledge

The training of an HMM, a SHMM or a DSHMM is typically performed with the Baum–Welch algorithm (Durbin *et al.*, 1998; Rabiner, 1989) belonging to the class of Expectation–Maximization (EM) algorithms (Dempster *et al.*, 1977). Starting from initial model parameters, the Baum–Welch algorithm locally maximizes the likelihood by a two-step procedure. This procedure adapts the parameters of the model iteratively to the tumor expression profiles without integrating prior knowledge.

To overcome this limitation, a specific extension of the Baum–Welch algorithm to a Bayesian Baum–Welch algorithm has been developed. The Bayesian Baum–Welch algorithm integrates prior knowledge specified by the prior distribution of the model. This is realized by iteratively determining new model parameters

$$\lambda(h+1) = \underset{\lambda}{\operatorname{argmax}} \left( Q(\lambda|\lambda(h)) + \log(P[\lambda|\Theta]) \right)$$

maximizing the posterior density of the model $\lambda$ with respect to the current parameters of the model $\lambda(h)$ ($h=1$ initial model). The estimation of the new parameters is done based on Baum's auxiliary function $Q(\lambda|\lambda(h))$ defined in analogy to Rabiner (1989) in combination with the logarithm of the prior distribution $P[\lambda|\Theta]$ specified in (2). This combination enables the iterative estimation of new model parameters with respect to prior knowledge about the data.

Details of the parameter estimation for the HMM are given in the Appendix B of the Supplementary Material. Due to the usage of different transition classes by the SHMM for distinguishing between adjacent genes in close chromosomal proximity and that in greater distance, the estimation of the transition parameters must be modified. Details to this are given in the Appendix C of the Supplementary Material. The estimation of the transition parameters for the DSHMM is outlined in the Appendix D of the Supplementary Material. This parameter estimation is done in analogy to that of the SHMM, but the computations are slightly more complex and require more memory because of the usage of distance-scaled transition matrices.

The process of estimating new parameters $\lambda(h+1)$ is iterated until the log-posterior density increases less than a predefined threshold in comparison to the value obtained for the previous parameters $\lambda(h)$. The training is stopped if the increase of the log-posterior density is less than $10^{-3}$ for two successive iteration steps. This iterative scheme reaches at least a local optimum in dependency of the initial parameters $\lambda(1)$ (Dempster *et al.*, 1977).

## 2.7 Initialization and basic settings

The initial HMM, SHMM and DSHMM must be able to differentiate between differentially expressed genes and genes with unchanged expression levels in tumor. A histogram of log-ratios (e.g. Fig. 2a) helps to find appropriate initial parameters.

Proportions of underexpressed and overexpressed genes in tumor are much less than that of genes with unchanged expression levels. Thus, the initial state distribution was set to $\vec{\pi} = (0.1, 0.8, 0.1)$ using $\pi_- = \pi_+ = 0.1$ and $\pi_= = 0.8$.

The initial transition matrix $A = (a_{ij})_{i,j \in S}$ was chosen to have a stationary distribution identical to $\vec{\pi}$ by setting all diagonal elements to $a_{ii} = 1 - s/\pi_i$ and all non-diagonal elements to $a_{ij} = s/(2\pi_i)$ using $s = 0.05$ to control the state durations.

State-specific Gaussian emission densities were characterized by proper means and standard deviations for representing the measured log-ratios. This was done using the means $\mu_- = -2$, $\mu_= = 0$, and $\mu_+ = 2$ and the corresponding standard deviations $\sigma_- = 0.3$, $\sigma_= = 0.5$ and $\sigma_+ = 0.3$.

Additionally, the fixed global distance threshold $b$ and the fixed scaling factor $f_2$ have to be specified for the SHMM. Both parameters are problem-specific depending on the dataset. The relation between the distances of adjacent genes on a chromosome and the correlations of their log-ratios (Fig. 1) helps to find appropriate settings. For the breast cancer gene expression data, the log-ratios of adjacent genes in distance greater than 1000 kb showed generally only weak positive correlations comparable to that obtained under permuted data. For that reason, the maximal distance threshold was set to 1000 kb. The scaling factor $f_2$ allows to adjust the probability that two directly adjacent genes are represented by the same state of the SHMM. Too high values of $f_2$ could lead to undesired predictions like pairs of overexpressed genes in which one gene has a log-ratio slightly less than zero. Under consideration of this, each global distance threshold $b \in \{10, 20, \ldots, 1000\}$ kb was assessed in combination with each scaling factor $f_2 \in \{1.1, 1.2, \ldots, 2.0\}$ for the breast cancer gene expression data.

In analogy to the SHMM, each global distance threshold $b \in \{10, 20, \ldots, 1000\}$ kb was tested in combination with each maximal scaling factor $F \in \{1.1, 1.2, \ldots, 2.0\}$ for the DSHMM.

Each initial model was trained with all tumor expression profiles of the breast cancer dataset using the developed Bayesian Baum–Welch algorithm.

## 3 RESULTS AND DISCUSSION

First, the coupling between gene expression levels and gene copy numbers is considered to provide the basics for comparing different methods. Next, the influence of prior knowledge on the training and the identification of differentially expressed genes by HMM, SHMM and DSHMM is investigated. Then, the effect of modeling dependencies and distances between adjacent genes on chromosomes is analyzed. Next, the best-performing method is compared with existing methods. Finally, hotspot-genes of under- and overexpression in breast cancer are investigated.

### 3.1 Coupling of gene expression and gene copy numbers

Gene expression profiles of individual breast tumors are known to be highly diverse (Perou *et al.*, 2000). Information about differentially expressed genes in individual tumors are widely lacking. This complicates the comparison of different methods for their ability to identify such genes. The coupling between gene expression levels and directly underlying gene copy numbers is considered to overcome this, because a large proportion of genes located in highly amplified chromosomal regions were identified as overexpressed in breast cancer compared with healthy tissue (Hyman *et al.*, 2002; Pollack *et al.*, 2002).

Motivated by this observation, information about copy numbers of genes in each of the 4 breast cancer cell lines and each of the 37 breast tumors are used to determine each gene in the corresponding tumor expression profile with an increased copy number in comparison to healthy tissue. This is done using the comparative genomic hybridization experiment by Pollack *et al.* (2002) providing a measurement of the copy number of each gene in each individual tumor expression profile. Each gene in a tumor expression profile is labeled as a potential candidate gene for overexpression if its underlying gene copy number is at least 3-fold increased in comparison to healthy tissue. This leads to 228 candidate genes of overexpression with increased copy numbers in the breast cancer gene expression dataset. The majority of these genes has increased expression levels in tumor (Fig. 2c), providing a good basis for evaluating the accuracy of different methods to identify these genes. These candidate genes of overexpression are considered in the following studies. Corresponding results for candidate genes of overexpression with a at least 2-fold or at least 4-fold increased copy number in tumor are provided in the Supplementary Material.

### 3.2 Influence of prior knowledge: comparison of Baum–Welch and Bayesian Baum–Welch training

The standard Baum–Welch algorithm trains the parameters of an HMM, a SHMM and a DSHMM without including prior knowledge about the distributions of log-ratios representing underexpressed, unchanged and overexpressed genes. The proposed Bayesian Baum–Welch algorithm integrates prior knowledge using the specified prior distribution. Exemplarily, the influence of both training algorithms on the emission parameters of the HMM is shown in Figure 3. Nearly identical emission parameters were obtained for the SHMM and the DSHMM (Supplementary Table S1). The distributions of log-ratios modeled by the states '−' and '+' clearly overlap with each other for the HMM trained by the Baum–Welch algorithm. A much better separation of these distributions is obtained for the same model trained by the Bayesian Baum–Welch algorithm. This leads to a better modeling of underexpressed and overexpressed genes that are expected to have log-ratios much different from zero (Fig. 2). Figure 4a shows that this is also applicable to the identification of candidate genes of overexpression by the HMM. As expected from the nearly identical emission parameters, this is also observed for the identification of candidate genes of overexpression by the SHMM and the DSHMM (Supplementary Fig. S3).

Consequently, only models trained by the Bayesian Baum–Welch algorithm are considered in the following studies.

### 3.3 Influence of modeling dependencies and distances between adjacent genes

Gene expression levels of adjacent genes on chromosomes in the breast cancer dataset are positively correlated and tend to be more similar for adjacent genes in close chromosomal proximity in comparison to adjacent genes in greater distances (Fig. 1).

To investigate the impact of modeling chromosomal locations and distances of genes, the abilities of HMM, SHMM and DSHMM to identify candidate genes of overexpression are compared with a mixture model. A mixture model of three Gaussian densities (e.g. Bilmes, 1998) for modeling underexpressed, unchanged and overexpressed genes is a basic model for analyzing tumor expression profiles. This model neither integrates chromosomal locations nor chromosomal distances of genes into the identification of differentially expressed genes.

To enable the comparison with the HMM-approaches, a mixture model was trained on the tumor expression profiles using the same basic initial settings and incorporating the same prior knowledge as for the HMM-approaches. The parameters of the Gaussian densities of this mixture model (Supplementary Table S2) were very similar to that obtained for HMM, SHMM and DSHMM trained using the Bayesian Baum–Welch algorithm (see Fig. 3 for HMM). This is expected, because these models are specific extensions of the mixture model.

The performance of these four models to identify candidate genes of overexpression was initially compared at a fixed false positive

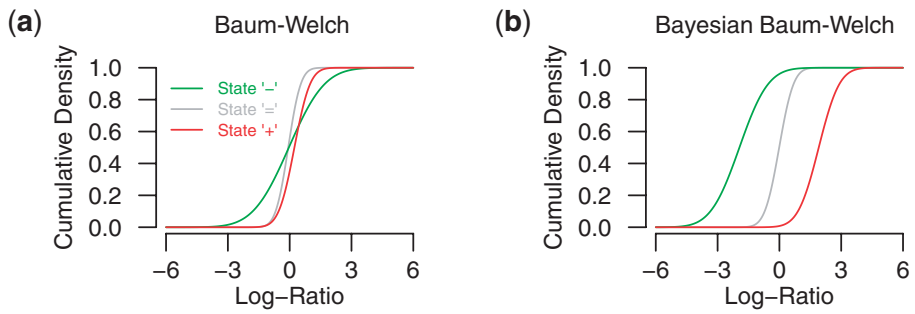


**Fig. 3.** Comparison of the Gaussian emission densities of the HMM obtained by Baum–Welch and Bayesian Baum–Welch training. For each of the three states of the HMM, the cumulative distribution function of each Gaussian emission density is shown. (**a**) The standard Baum–Welch algorithm clearly fails to characterize underexpressed genes modeled by state '−' (green) and overexpressed genes modeled by state '+' (red). (**b**) Cumulative distribution functions obtained by the proposed Bayesian Baum–Welch training. The emission densities are well separated and characterize differentially expressed genes as observed in Figure 2a and b.

rate of 3% (Supplementary Fig. S4). The HMM identifies candidate genes of overexpression clearly better than the mixture model. This is further improved by the SHMM and the DSHMM that both achieve nearly identical results with a small tendency that some DSHMMs can identify candidate genes of overexpression slightly better than corresponding SHMMs. That indicates that the DSHMM might be better suited for the analysis of breast cancer tumor expression profiles.

Based on this, one of the best DSHMMs with distance threshold $b = 350$ kb and maximal scaling factor $F = 1.9$ was selected for a more detailed comparison. The identifications of candidate genes of overexpression by the mixture model, the HMM and this DSHMM are shown in Figure 4b (see Supplementary Fig. S5 for all methods). The HMM modeling dependencies between adjacent genes on a chromosome reaches a higher accuracy for identifying these candidate genes in comparison to the mixture model ignoring these dependencies. This accuracy is further improved by the DSHMM that additionally models distances between adjacent genes on chromosomes.

In summary, the largest proportion of the improved identification of candidate genes of overexpression is obtained by the transition from the mixture model to the HMM modeling dependencies between adjacent genes on a chromosome. Additional benefit is reached due to the modeling of distances between adjacent genes by the SHMM and the DSHMM. Both characteristics are modeled best by the DSHMM.

### 3.4 Comparison of DSHMM to existing methods

The analysis of tumor expression profiles in the context of chromosomal locations of genes is closely related to the analysis of data coming from comparative genomic hybridization experiments (e.g. Beroukhim *et al.*, 2010; Pinkel and Albertson, 2005). Such data are typically analyzed by modeling dependencies between adjacent probes on a chromosome to identify probes with log-ratios much greater or much less than zero (Lai *et al.*, 2005). This is comparable to the identification of differentially expressed genes in tumor. For that reason, the DSHMM with distance threshold $b = 350$ kb and maximal scaling factor $F = 1.9$ is compared to the best-performing related methods of two comparison studies (Lai *et al.*, 2005; Willenbrock and Fridlyand, 2005) provided by the ADaCGH web server (Diaz-Uriarte and Rueda, 2007), ChARM by (Myers *et al.*, 2004) and the basic log-fold-change analysis (LFC) used by Pollack *et al.* (2002). All these methods are summarized in Supplementary Table S3 and were applied with their standard settings to the breast cancer tumor expression profiles to identify candidate genes for overexpression. We are aware that these standard settings could also be fine-tuned by expert knowledge.

The predictions and corresponding scores obtained by all methods are available as Supplementary Material. The results are shown in Figure 4c. In comparison to all these different methods, the DSHMM reaches the highest accuracy for identifying candidate genes of overexpression.

### 3.5 Hotspots of under- and overexpression in the breast cancer dataset

The publicly available Breast Cancer Database (BCD) (Telikicherla *et al.*, 2008) contains genes identified to play a role in breast cancer. In the BCD, 1361 genes overlap with the genes measured in the breast cancer gene expression dataset by Pollack *et al.* (2002). However, the individuality of breast tumors in terms of duplications, deletions and other mutations (Perou *et al.*, 2000) does not allow to use all these genes simultaneously as candidate genes for underexpression or overexpression in each individual tumor expression sample by Pollack *et al.* (2002). Yet, genes that are frequently identified as being underexpressed or overexpressed in different tumor expression profiles can be compared with the BCD. This provides additional support that a method is able to identify tumor-relevant genes, and it also provides the opportunity to further investigate frequently identified genes that are not contained in the BCD. In the previous sections, the DSHMM with distance threshold $b = 350$ kb and maximal scaling factor $F = 1.9$ has outperformed the other HMM-approaches and also related methods for the identification of differentially expressed genes. For that reason, the following analysis is only done for this model.

To identify genes with frequently altered expression levels in the breast cancer dataset, the expression status of each gene (underexpressed, unchanged, overexpressed) in each tumor expression profile was initially computed under the DSHMM using the state-posterior decoding algorithm (Rabiner, 1989). For a stringent selection, each gene that was identified at least 7

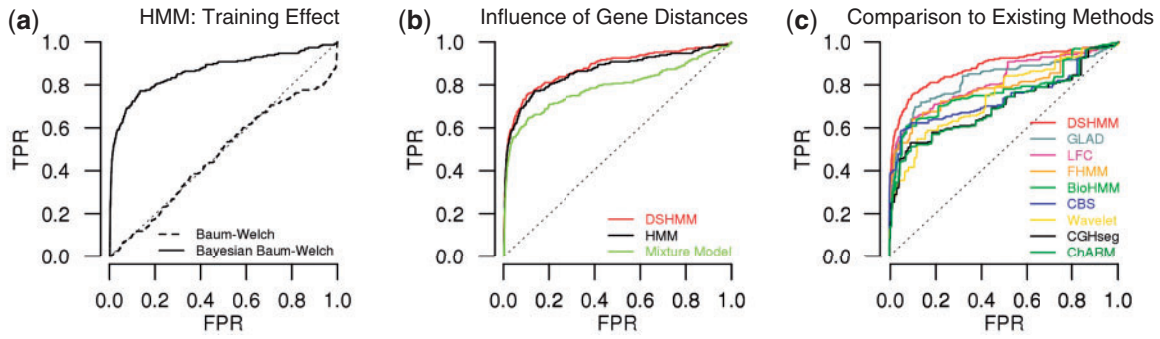


**Fig. 4.** Receiver operator characteristic (ROC) curves characterizing the ability of different methods to identify candidate genes of overexpression with at least 3-fold increased copy numbers in tumor compared with healthy tissue. (**a**) HMM: Training Effect: comparison of the HMM trained by the standard Baum–Welch algorithm ignoring prior knowledge (dashed line) to the HMM trained by the proposed Bayesian Baum–Welch algorithm incorporating prior knowledge (solid line). The HMM trained by the Baum–Welch algorithm clearly fails to identify candidate genes of overexpression, whereas the HMM trained by the Bayesian Baum–Welch algorithm reaches a much higher accuracy. Very similar trends were observed for the SHMM and the DSHMM (Supplementary Fig. S3). (**b**) Influence of Gene Distances: comparison of the mixture model neither modeling chromosomal locations nor distances of genes on chromosomes to the HMM and the DSHMM. All three models were trained using the Bayesian Baum–Welch algorithm integrating prior knowledge. The HMM analyzes tumor expression profiles in the context of chromosomal locations of genes and reaches a higher accuracy for the identification of candidate genes of overexpression than the mixture model. This is further improved by the DSHMM that additionally models distances between adjacent genes on a chromosome. The combination of integrating prior knowledge into the training by the Bayesian Baum–Welch algorithm and the modeling of gene distances by the DSHMM leads to an additional improvement of the identification of candidate genes of overexpression. The largest proportion of this improvement is obtained by the integration of prior knowledge into the training (compare Fig. 4a). (**c**) Comparison to Existing Methods: comparison of the DSHMM to existing related approaches summarized in Supplementary Table S3. The DSHMM reaches the highest accuracy for identifying candidate genes of overexpression. Generally, very similar results were obtained for the identification of candidate genes of overexpression with at least 2-fold or 4-fold increased copy numbers (Supplementary Fig. S5).

times as underexpressed or at least 7 times as overexpressed across all 41 breast cancer tumor samples has been considered for further analyses. In total, 62 genes fulfill this criterion and 43 of these genes are contained in the BCD (Supplementary Table S4). The remaining 19 genes have been further investigated by additional literature searches and independent database searches using Oncomine (Rhodes *et al.*, 2007). Based on this, 18 genes could be directly associated with breast cancer and 1 gene was found to play a role in an other type of cancer (Supplementary Table S5). This provides further support that the DSHMM is useful for the identification of differentially expressed genes in breast cancer.

## 4 CONCLUSIONS

We analyzed breast cancer expression profiles by integrating three sources of prior knowledge into extended HMMs for identifying differentially expressed genes in tumor.

First, the analysis of tumor expression profiles in the context of chromosomal locations of genes has been motivated by the observation of strong positive correlations between expression levels of adjacent genes. This has inspired the application of the HMM modeling dependencies between adjacent genes on a chromosome.

Second, the trend that expression levels of adjacent genes in close chromosomal proximity are higher correlated than that of adjacent genes in greater distance has led to the extension of the HMM to the SHMM. The SHMM additionally incorporates chromosomal distances of genes on a chromosome for identifying differentially expressed genes in tumor. This is done by two fixed transition matrices enabling the separation of adjacent genes into genes in close chromosomal proximity and genes in greater distance. To avoid potential discretization effects caused by this fixed separation, the SHMM has been further extended to the DSHMM that directly integrates individual distances of adjacent genes into the state-transition process.

Third, the modeling of expected log-fold changes for underexpressed, unchanged and overexpressed genes in tumor has been realized by including a prior distribution into the training of the three HMM-based approaches. This has led to the extension of the standard Baum–Welch training that does not integrate prior knowledge to a Bayesian Baum–Welch training that incorporates prior knowledge.

The three developed HMM-based approaches trained by the Bayesian Baum–Welch algorithm integrating prior knowledge on expected log-fold changes clearly improved the characterization of differentially expressed genes in tumor in comparison to corresponding models trained by the standard Baum–Welch algorithm ignoring these prior knowledge. This is also applicable to the identification of candidate genes of overexpression in breast cancer. Models trained by the Bayesian Baum–Welch algorithm substantially improved this identification.

The comparison of the HMM-based approaches to a mixture model neither modeling chromosomal locations nor distances of adjacent genes has revealed that the SHMM and the DSHMM that both make use of these two features achieve the highest accuracy for identifying candidate genes of overexpression. Both models performed nearly identical, but the DSHMM might represent the better suited model because of the direct integration of individual distances of adjacent genes. This comparison has also revealed that the integration of prior knowledge into the training is more important than the additional modeling of chromosomal distances of genes.

Moreover, the DSHMM has shown the best identification of candidate genes of overexpression in comparison to other existing methods. Additionally, genes frequently identified by the DSHMM as being differentially expressed in breast cancer have been

compared to two public breast cancer databases. In combination with additional literature searches, this comparison has provided further support that the DSHMM is useful for the identification of tumor-relevant genes. All these results indicate that the DSHMM could also be considered for the analysis of other tumor expression profiles.

## ACKNOWLEDGEMENTS

We thank Anton Wellstein for valuable discussions on hotspot-genes. We are grateful for the valuable comments of the reviewer triggering the development of the DSHMM. We thank Jens Keilwagen and Jan Grau for providing the basics in Jstacs.

*Funding*: Ministry of Culture Saxony-Anhalt (grant XP3624HP/0606T), DFG graduate school 1564.

*Conflict of Interest*: none declared.

## REFERENCES

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Bilmes,J.A. (1998) A gentle tutorial of the EM algorithm and its applications to parameter estimation for Gaussian mixture and Hidden Markov Models. *Technical Report ICSI-TR 97-021*.

Bishop,C.M. (2006) Pattern recognition and machine learning. In Jordan,M. *et al.* (eds) *Information Science and Statistics*. Springer.

Callegaro,A. *et al.* (2006) A locally adaptive statistical procedure (lap) to identify differentially expressed chromosomal regions. *Bioinformatics*, **22**, 2658–2666.

Caron,H. *et al.* (2001) The Human Transcriptome Map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.

Crawley,J.J. and Furge,K.A. (2002) Identification of frequent cytogenetic aberrations in hepatocellular carcinoma using gene-expression microarray data. *Genome Biol.*, **3**, research0075.1–research0075.8.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, **39**, 1–38.

Diaz-Uriarte,R. and Rueda,O.M. (2007) ADaCGH: a parallelized web-based application and R package for the analysis of aCGH data. *PLoS One*, **2**, e737.

Durbin,R. *et al.* (1998) *Biological Sequence Analysis - Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Evans,M. *et al.* (2000). *Statistical Distributions,* 3rd edn. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

Fridlyand,J. *et al.* (2004) Hidden Markov models approach to the analysis of array CGH data. *J. Multivar. Anal.*, **90**, 132–153.

Frigola,J. *et al.* (2006) Epigenetic remodeling in colorectal cancer results in coordinate gene suppression across an entire chromosome band. *Nat. Genet.*, **38**, 540–5490.

Hasegawa,S. *et al.* (2002) Genome-wide analysis of gene expression in intestinal-type gastric cancers using a complementary DNA microarray representing 23,400 genes. *Cancer Res.*, **62**, 7012–7017.

Heidenblad,M. *et al.* (2005) Microarray analyses reveal strong influence of DNA copy number alterations on the transcriptional patterns in pancreatic cancer: implications for the interpretation of genomic amplifications. *Oncogene*, **24**, 1794–1801.

Hyman,E. *et al.* (2002) Impact of DNA amplification on gene expression patterns in breast cancer. *Cancer Res.*, **62**, 6240–6245.

Knab,B. *et al.* (2003) Model-based clustering with Hidden Markov Models and its application to financial time-series data. In Schader,M. *et al.* (eds) *Between Data Science and Applied Data Analysis*. Springer, pp. 561–569.

Lai,W.R. *et al.* (2005) Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**, 3763–3770.

MacKay,D.J.C. (1998) Choice of basis for Laplace approximation. *Mach. Learn.*, **33**, 77–86.

Marioni,J.C. *et al.* (2006) BioHMM: a heterogeneous hidden Markov model for segmenting array CGH data. *Bioinformatics*, **22**, 1144–1146.

Myers,C.L. *et al.* (2004) Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, **20**, 3533–3543.

Perou,C.M. *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

Pinkel,D. and Albertson,D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**, S11–S13.

Pollack,J.R. *et al.* (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA*, **99**, 12963–12968.

Rabiner,L.R. (1989) A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Rhodes,D.R. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.

Rueda,O.M. and Diaz-Uriarte,R. (2007) Flexible and accurate detection of genomic copy-number changes from aCGH. *PLoS Comput. Biol.*, **3**, e122.

Seifert,M. (2006) Analysing microarray data using homogeneous and inhomogeneous Hidden Markov Models. Diploma Thesis, Martin Luther University Halle-Wittenberg.

Seifert,M. *et al.* (2009) Utilizing gene pair orientations for HMM-based analysis of ChIP-chip data. *Bioinformatics*, **25**, 2118–2125.

Stransky,N. *et al.* (2006) Regional copy number-independent deregulation of transcription in cancer. *Nat. Genet.*, **38**, 1386–1396.

Telikicherla,D. *et al.* (2008) A resource of molecular alterations in breast cancer. In *Proceedings of the Human Genome Meeting*. Hyderabad, India.

Toedling,J. *et al.* (2004) MACAT - microarray chromosome analysis tool. *Bioinformatics*, **21**, 2112–2113.

Willenbrock,H. and Fridlyand,J. (2005) A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 4084–4091.