BIOINFORMATICS

Vol. 25 ISMB 2009, pages i30–i38
doi:10.1093/bioinformatics/btp205

Model-based clustering of array CGH data

Sohrab P. Shah[1,2]*, K-John Cheung Jr[2], Nathalie A. Johnson[2], Guillaume Alain[1], Randy D. Gascoyne[2], Douglas E. Horsman[2], Raymond T. Ng[1] and Kevin P. Murphy[1]

[1]Department of Computer Science, University of British Columbia, 201-2366 Main Mall Vancouver, BC V6T 1Z4 Canada and [2]British Columbia Cancer Agency, 600 W 10th Ave Vancouver, BC V5Z 4E6 Canada

ABSTRACT

Motivation: Analysis of array comparative genomic hybridization (aCGH) data for recurrent DNA copy number alterations from a cohort of patients can yield distinct sets of molecular signatures or profiles. This can be due to the presence of heterogeneous cancer subtypes within a supposedly homogeneous population.

Results: We propose a novel statistical method for automatically detecting such subtypes or clusters. Our approach is model based: each cluster is defined in terms of a sparse profile, which contains the locations of unusually frequent alterations. The profile is represented as a hidden Markov model. Samples are assigned to clusters based on their similarity to the cluster's profile. We simultaneously infer the cluster assignments and the cluster profiles using an expectation maximization-like algorithm. We show, using a realistic simulation study, that our method is significantly more accurate than standard clustering techniques. We then apply our method to two clinical datasets. In particular, we examine previously reported aCGH data from a cohort of 106 follicular lymphoma patients, and discover clusters that are known to correspond to clinically relevant subgroups. In addition, we examine a cohort of 92 diffuse large B-cell lymphoma patients, and discover previously unreported clusters of biological interest which have inspired followup clinical research on an independent cohort.

Availability: Software and synthetic datasets are available at http://www.cs.ubc.ca/~sshah/acgh as part of the CNA-HMMer package.

Contact: sshah@bccrc.ca

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

1 INTRODUCTION

Copy number alterations (CNA) are structural variations expressed in the form of DNA copy number differences at a particular region in the genome. The search for 'driver' CNAs in genetic material derived from cancerous tissues is a major goal in diagnostic and cytogenetic cancer research (Aguirre *et al.*, 2004; Chin and Gray, 2008; Michels *et al.*, 2007; Tonon *et al.*, 2005). Putative driver CNAs are genomic amplifications or deletions ranging in size from a few kilobases to whole chromosome arms that are recurrent in a larger than expected proportion of patients. Their detection provides candidate genetic markers that may play a role in tumorigenesis and/or have clinicopathologic significance. In contrast, 'passenger' CNAs arise during the evolution of the tumor and may be present due to genomic instability or other mechanisms. In the context of defining the driver CNAs, passenger CNAs represent (often ubiquitous) 'biological' noise that might obscure driver signals. Using high-resolution array comparative genomic hybridization (aCGH) (Pinkel and Albertson, 2005), data consisting of tens to hundreds of thousands of probes, putative driver CNAs can be detected by identifying the subset of probes they span using a number of algorithmic and statistical tools (Diskin *et al.*, 2006; Klijn *et al.*, 2008; Rouveirol *et al.*, 2006; Shah *et al.*, 2007). These analyses lead to a molecular profile of recurrent CNAs that help define the molecular characteristics of the disease.

A challenging phenomenon is that, frequently, patient cohorts exhibit heterogeneity in their molecular profiles. This has been demonstrated in breast (Perou *et al.*, 2000), ovarian (Khalique *et al.*, 2007) and prostate cancers, as well as lymphomas (Cheung *et al.*, 2008; Höglund *et al.*, 2004), suggesting that the patients should be stratified into molecular subtypes, where the patients within a group share a common group-specific driver CNA profiles. This concept has been successfully applied many times over using gene expression data (Perou *et al.*, 2000; Wright *et al.*, 2003), however it has been relatively under-studied in aCGH data.

Considering a cohort of patients as a composite of a fixed set of molecular subtypes has distinct advantages when determining recurrent CNAs. By grouping or clustering the patients, recurrent CNAs that might otherwise go undetected can be revealed. This approach has the potential of determining CNAs that co-occur within a subtype and CNAs that are mutually exclusive between subtypes. Moreover, groups of patients can be assessed for distinct clinical outcomes. Molecular subtypes often correlate with clinical outcomes and in fact can, once identified, be considered as distinct disease entities (Sorlie, 2004) with different prognoses and/or response to therapy.

Recent discovery of clinically relevant molecular subtypes by aCGH (Chin *et al.*, 2007; Idbaih *et al.*, 2008) suggest that the inventory of CNA-derived molecular subtypes in cancer is not complete. Large-scale projects such as the Cancer Genome Atlas Project (Collins and Barker, 2007) and the International Cancer Genome Consortium (ICGC: http://www.icgc.org) are now generating genomic array datasets from tumors from hundreds of patients for specific cancer types, thereby providing excellent potential for the discovery of new CNA-derived subtypes. In order to take full advantage of these data, robust and accurate computational algorithms for discovering molecular subgroups must be developed to keep pace with the data generation.

In this article, we propose an approach to this problem based on a mixture of HMMs (hidden Markov models); we call our approach HMM-Mix. This extends our previous work (Shah *et al.*, 2006, 2007) by defining multiple HMMs, one per cluster and automatically assigning samples to clusters while simultaneously inferring the profile of each cluster. Although the profiles are defined in terms of

*To whom correspondence should be addressed.

'called' data (i.e. each location is classified as a loss, a gain or neutral/no change), the model works directly with the raw aCGH data, and can recall ambiguous data in the context of the cluster to which it is assigned. This increases the statistical power of our method to detect shared, but subtle, CNAs that may be lost by methods that require discretization of the data as a preprocessing step, as shown in our previous work (Shah *et al.*, 2007) and by Klijn *et al.* (2008).

In a simulation study, with realistic data, we show how our method is more accurate than other clustering methods, including hierarchical clustering (van Wieringen and van de Wiel, 2008) and K-medoids (KM) (an approach not previously applied to data of this kind). More importantly, we show how HMM-Mix reveals clinically relevant subgroups in data derived from a cohort of 106 follicular lymphoma (FL) patients, originally reported in Cheung *et al.* (2008), and reveals previously unreported patterns of alteration in a cohort of 92 diffuse large B-cell lymphoma (DLBCL) patients (Johnson *et al.*, 2008).

2 METHODS

We first describe our probabilistic model, and then how we perform inference in this model. We also describe three other approaches against which we compare our method: a simple K-medoids method (WKM) a weighted K-medoids method, and a previously described hierarchical clustering algorithm designed for aCGH (van Wieringen and van de Wiel, 2008).

2.1 The HMM-Mix model

We represent the aCGH logratios as $Y_t^p \in \mathbb{R}$ for each probe $t \in (1,\ldots,T)$ in the array and for each patient $p \in (1,\ldots,P)$. Each probe maps to unique genomic coordinates and is positionally ordered along the chromosomes. $Y_{1:T}^{1:P}$ represents the full data matrix. For each datapoint, we assume there is a discrete mapping from $Y_{1:T}^{1:P} \rightarrow Z_{1:T}^{1:P}$ where $Z_t^p \in k$ and k is a discrete copy number state $\in \{L,N,G\}$ representing loss, neutral and gain.

The HMM-Mix model is a probabilistic generative model of $Y_{1:T}^{1:P}$. We illustrate our conditional independence assumptions using a graphical model in Figure 1, and we define all the conditional distributions in Figure 2. See also Table 1 for a summary of the notation.

The model generates the data as follows. First we sample a group or cluster label for each patient, denoted $G^p \in \{1,\ldots,G\}$, from a Multinomial with parameter π^g. Here, G is the number of clusters (see below for how we choose this), and π^g is the vector of mixing weights. Next, each group G generates a profile which is represented as a sequence of states, $M_t^g \in \{L,B,G\}$, $t=1:T$, representing loss, background or gain at probe t in the array. Probes which are labeled loss are expected to contain mostly losses; probes which are labeled gain are expected to contain mostly gains; probes which are labeled background are expected to contain whatever the background distribution of loss, gains and neutrals is. Thus, the non-background probes are the interesting ones.[1] Since CNAs occur in runs (span contiguous sets of probes), we model correlation between consecutive locations using a first-order Markov chain on the M_t^g variables. The transition matrix, A^g is a 3×3 matrix whereby $A^g(i,j)$ represents $p(M_t^g = j | M_{t-1}^g = i)$. We expect this matrix to have large elements on the diagonal encouraging self-transitions [which we model with a Dirichlet prior with parameters δ_A (see Fig. 1 and Table 1)], and thus runs of repeated states. Of course the quantities of A^g are unknown at run time and are estimated by fitting the model to the data (see Section 2.2). Therefore, the off-diagonal elements of the matrix, including for example the transitions $\{B\rightarrow L, B\rightarrow G, L\rightarrow B, \ldots\}$, are fully represented and estimated accordingly.

[1]Indeed, one of the primary goals of inference is to find the probes for which $p(M_t^g \neq B|\mathcal{D})$ is high; these probes represent a sparse profile defining the signature for group g. Thus, our model is somewhat similar to approaches that perform simultaneous feature selection and clustering (Law *et al.*, 2004; Raftery and Dean, 2006).

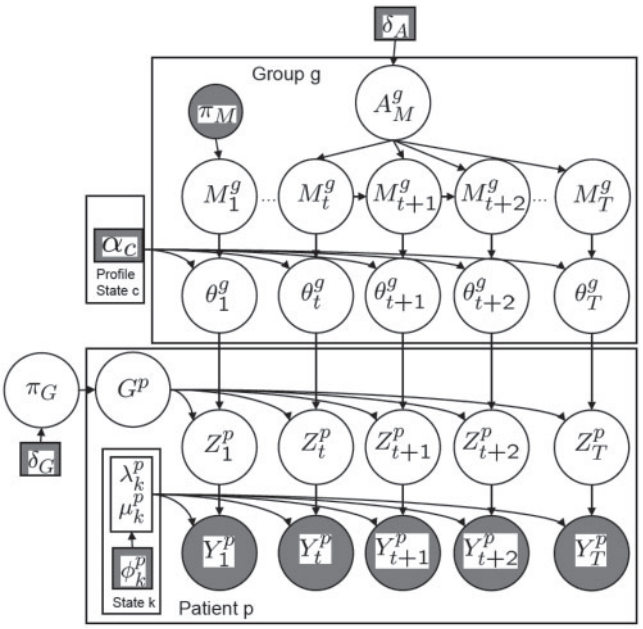

Fig. 1. Proposed HMM-Mix model for clustering aCGH data, represented as a directed graphical model (Gilks *et al.*, 1996). Shaded nodes are observed/fixed, unshaded nodes are hidden (unknown). The two boxes represent repetition over patients and groups. $Y_t^p \in \mathbb{R}$ is the observed aCGH logratio at probe t in patient p. $Z_t^p \in \{L,N,G\}$ is the discrete state, representing whether probe t is a loss, neutral or gain. Given $Z_t^p = k$, Y_t^p is assumed to be sampled from a class conditional Student-t distribution with parameters μ_k^p, λ_k^p and ν_k. $G^p \in \{1,\ldots,G\}$ is the group that patient p belongs to, which is sampled from a Multinomial with parameter π_G. θ_t^g is the Multinomial parameter over Z_t^p, which is sampled from a Dirichlet with parameter $\alpha_{M_t^g}$, where $M_t^g \in \{1,\ldots,C\}$ represents the state of the sparse profile for probe t in group g. A^g is the transition matrix for the profile model. Conditional probability distributions are shown in Figure 2. Description of variables is given in Table 1.

$$
\begin{aligned}
p(M_t^g | M_{t-1}^g = i, A) &= Mult(M_t^g | A(i,:), 1) \\
p(M_1^g | \pi_M) &= Mult(M_1^g | \pi_M, 1) \\
p(\theta_t^g | M_t^g = c, \alpha_{1:3}) &= Dir(\theta_t^g | \alpha_c) \\
p(Z_t^p | G^p = g, \theta_t^{1:G}) &= Mult(Z_t^p | \theta_t^g, 1) \\
p(G^p | \pi_G) &= Mult(G^p | \pi_G, 1) \\
p(Y_t^p | Z_t^p = k, \mu_k^p, \lambda_k^p) &= St(Y_t^p | \mu_k^p, \lambda_k^p, \nu_k) \\
p(\mu_k^p | \lambda_k^p, \phi) &= N(\mu_k^p | m_k, \frac{\eta_k}{\lambda_k^p}) \\
p(\lambda_k^p | \phi) &= Gam(\lambda_k^p | S_k, \gamma_k) \\
p(A(i,.) | \delta_A) &= Dir(A(i,.) | \delta_A) \\
p(\pi_G | \delta_\pi) &= Dir(\pi_G | \delta_G)
\end{aligned}
$$

Fig. 2. List of conditional probability distributions of HMM-Mix.

Table 1. Summary of variables

Symbol	Meaning
S_k	Set of probability vectors of length k
T	Number of probes (measurements)
G	Number of groups (clusters)
P	Number of patients (samples)
$t \in \{1,\ldots,T\}$	Probe location
$g \in \{1,\ldots,G\}$	Group index
$p \in \{1,\ldots,P\}$	Patient index
$c \in \{L,B,G\}$	State index
$k \in \{L,N,G\}$	Call index
$M_t^g \in \{L,B,G\}$	State of profile
$\theta_t^g \in S_3$	Distribution over calls
$Z_t^p \in \{L,N,G\}$	Called aberation
$Y_t^p \in \mathbb{R}$	Raw data (log ratios)
$G^p \in \{1,\ldots,G\}$	Group assignment
$\pi^G \in S_G$	Prior over groups
$A^g \in S_{3\times 3}$	Transition matrix
$\mu_k^p \in \mathbb{R}$	Mean of observations
$\lambda_k^p \in \mathbb{R}^+$	Precision of observations
$\nu_k \in \mathbb{R}^+$	DOF for observations (fixed)
π_M	Multinomial parameters (fixed)
$\alpha_c, \delta_G, \delta_A$	Dirichlet hyper-parameters (fixed)
$\phi = (m_k, \eta_k, S_k, \gamma_k)_{k=1}^3$	Normal-Gamma hyper-parameters (fixed)

DOF = degrees of freedom.

Once we have generated a discrete profile for each group, we convert it into a distribution over calls. Specifically, state M_t^g of the Markov chain 'emits' a probability vector θ_t^g, representing a probability distribution over the 'letters' $\{L,N,G\}$, representing 'called' aCGH states. In other words, θ_t^g represents the relative frequencies of calls we would expect at location t in group g. If $M_t^g = L$, then θ_t^g is sampled from a Dirichlet with parameters $\alpha_L = (a_L, 1, 1)$, which is biased toward the letter L (by setting $a_L \gg 1$). Similarly, if $M_t^g = G$, then θ_t^g is sampled from a Dirichlet with parameters $\alpha_G = (1, 1, a_G)$, which is biased toward the letter G (by setting $a_G \gg 1$). If $M_t^g = B$, then θ_t^g is set equal to θ_0^g, representing the overall background, which is shared across locations; θ_0^g is itself sampled from a Dirichlet with parameters $\alpha_B = (1, a_B, 1)$, which is biased toward the letter N (by setting $a_N \gg 1$). Once we have generated the continuous profile for each group, θ_t^g, we are able to generate data for each patient. We sample a call $Z_t^p \in \{L,N,G\}$ from a Multinomial with parameter θ_t^g. Here, it would be appropriate to model $Z_{1:T}^p$ as a Markov chain to capture the spatial correlation in the data at the level of each patient. However, as shown in our previous work (Shah *et al.*, 2007), this makes inference expensive since all the Z chains become coupled. Instead, we *initialize* each $Z_{1:T}^p$ using Markov chains (see below) to capture the patient level spatial correlation and find that this is sufficient for our task of capturing the group-specific *recurrent* CNAs which are explicitly modeled as a Markov chain $M_{1:T}^g$.

Finally, we convert the discrete call into a continuous observation, $Y_t^p \in \mathbb{R}$, by sampling from a Student-t distribution; this is more robust to outliers than a Gaussian. Specifically, if $Z_t^p = k$, we use mean μ_k^p, precision λ_k^p and fixed degrees of freedom $\nu = 3$. (We fix the degrees of freedom to simplify the inference procedure; we have found that our results are reasonably robust to the value of ν.) Note that the parameters of the observation density are patient specific, but are shared across locations. The observation parameters μ_k^p and λ_k^p are sampled from a standard conjugate prior. Details on how we set the hyper-parameters are outlined in Shah *et al.* (2007).

2.2 Inference

Although the model was described in terms of M_t^g generating θ_t^g, which in turn generates the Z_t^p calls, it turns out to simplify inference if we analytically integrate out θ_t^g. This is valid since θ_t^g is just a nuisance parameter, i.e. it is not a variable we are interested in estimating. (Several other variables are also nuisance parameters, but eliminating them would make inference harder, not easier.) The modified conditional distribution is

$$p(Z_t^p | M_t^{1:G}, \alpha_{1:3}, G^p = g) = \int p(Z_t^p | \theta_t^g) p(\theta_t^g | \alpha_c) d\theta_t^g = \frac{1}{\sum_k \alpha_c^k} \prod_{k=1}^{K} \frac{\Gamma(I(Z_t^p = k) + \alpha_c^k)}{\Gamma(\alpha_c^k)} \quad (1)$$

where $c = M_t^g$ is the state of the Markov chain, and $\Gamma()$ is the Gamma function (see Brown *et al.*, 1993, for details) and $I(Z_t^p = k)$ is an indicator function stating that the copy number call for patient p at probe t is k. Henceforth, we assume θ_t^g has been removed from the model in this way.

Our primary objective is to infer a clustering, $p(G^p|\mathcal{D})$, and a profile for each cluster, $p(M_{1:T}^g|\mathcal{D})$. One approach would be to use Markov chain Monte Carlo (MCMC) to draw samples from the full posterior, but this is too slow for our application, which has about $P \sim 100$ patients, and about $T \sim 27\,000$ probes (over all the chromosomes) per patient.

An alternative would be to use the expectation maximization (EM) algorithm (Dempster *et al.*, 1977). A natural approach would be to treat all the unknown discrete variables (i.e. M_t^g, Z_t^g and G^p) as 'hidden variables', and treat the rest (i.e. $A^g, \pi^g, \mu_k^p, \lambda_k^p$) as 'parameters'. Unfortunately, this makes the E step computationally intractable, since all the HMMs $M_{1:T}^g$ become coupled in the posterior. However, conditional on a known clustering (i.e. setting of G^p), the HMMs become independent. Hence we can estimate the posterior profile for group g using the data that belongs to group g using the forwards–backwards algorithm. (This requires marginalizing out Z_t^g as well, in order to derive the observation model $p(Y_t^p|M_t^g)$, but this is straightforward.) Note that this requires that we treat G^p as a 'parameter' in the sense that we estimate it in the M step rather than the E step. This requires that we perform a hard clustering of the patients, rather than a soft clustering.

It turns out that even EM is too slow for our application, because of the need to marginalize out Z_t^g, and because of EM's relatively slow convergence. We therefore decided to use the iterative conditional modes (ICM) algorithm (Besag, 1986). This is a simple coordinate ascent algorithm, in which we set each variable to its most probable value given its neighbors in the graph. This can be thought of as a deterministic version of Gibbs sampling. Alternatively, it can be thought of as a version of Viterbi EM, in which we compute the most probable value of $M_{1:T}^g$ using the Viterbi algorithm instead of computing posterior marginals using forwards–backwards. More details on the algorithm can be found below. Its complexity is $O(TGP)$ per iteration, where T is the number of probes, G is the number of groups and P is the number of patients. In practice, it takes about 320 s to fit the model to our DLBCL data (92 patients, 5 groups and 30 000 probes) on a MacBook Pro with 2.6 GHz Intel Core Duo 2 using a Matlab implementation.

We now give a full description of the algorithm.

2.2.1 HMM-mix algorithm—main loop The basic procedure iterates over each node, and either samples from, or maximizes, each full conditional distribution (details in Section 2.2.3).

(1) Estimate profile: $p(M_{1:T}^g|A, \pi_M, Z_{1:T}^{1:P}, G^{1:P})$
(2) Assign to cluster: $p(G^p|\pi_G, Z_{1:T}^p, M_{1:T}^g)$
(3) Call data: $p(Z_t^p|Y_t^p, G^p, M_t^g, \mu_{1:3}^p, \lambda_{1:3}^p)$
(4) Fit observation model: $p(\mu_k^p, \lambda_k^p|Y_{1:T}^p, Z_{1:T}^p, \phi)$
(5) Fit transition model: $p(A^g|M_{1:T}^g, \delta_A)$
(6) Fit group prior: $p(\pi_G|G^{1:P}, \delta_G)$

2.2.2 HMM-mix algorithm—initialization

1. Set the hyper-parameters ϕ in a data-driven way, as explained in Shah *et al.* (2007).

2. Estimate $Z^p_{1:T}$ for each patient separately using Shah *et al.* (2006).
3. Estimate G^p using WKM (see Section 2.4) on $Z^{1:P}_{1:T}$.
4. Estimate M^g_t as follows. Given $Z^{1:P}_{1:T}$ for the patients in group g, compute the entropy of each column. If the entropy is low and most calls are losses, set $M^g_t = L$; if the entropy is low and most calls are gains, set $M^g_t = G$; otherwise set $M^g_t = B$.

2.2.3 HMM-mix algorithm details We now explain each step in more detail.

(1) The most expensive step is the first one, which takes $O(TGP)$ time using the Viterbi algorithm. To compute this, we need the observation likelihoods for each location, which are given by

$$B^g(t,c) = \prod_{p=1}^{P} I(G^p = g) p(Z^p_t | M^g_t = c)$$

where $p(Z^p_t | M^g_t = c)$ is the likelihood obtained by integrating out θ^g_t using Dirichlet hyper-parameter α_c (Equation 1). We then compute

$$M^g_{1:T} = \text{Viterbi}(B^g(:,:), A^g, \pi_m)$$

(2) Posterior over cluster assignments:

$$p(G^p = g|\cdot) \propto \pi^g_G \prod_{t=1}^{T} p(Z^p_t | M^g_t, \alpha_{1:3})$$

(3) Posterior over calls

$$p(Z^p_t = k|\cdot) \propto p(Z^p_t = k | M^{1:G}_t, \alpha_{1:3}, G^p = g) p(Y^p_t | \mu^p_k, \lambda^p_k)$$

(4) Update observation model parameters (as specified in Archambeau, 2005), but for the 1D case for each patient p. Use a Normal Gamma prior for $p(\mu^p_k, \lambda^p_k)$ (Fig. 2), with hyper-parameters $(m_k, \eta_k, S_k, \gamma_k)$. Compute the following quantities:

$$\bar{u}^p_t(k) = \frac{1+\nu_k}{(Y^p_t - \mu^p_k)^2 \lambda^p_k + \nu_k)}$$

$$\rho^p_t(k) = p(Z^p_t = k|\cdot)$$

where $\rho^p_t(k)$ is computed in step 3. The *maximum a posteriori* update equations then become:

$$\mu^p_k = \frac{\sum_{t=1}^{T} \rho^p_t(k) \bar{u}^p_t(k) Y^p_t + \eta_k m_k}{\sum_{t=1}^{T} \rho^p_t(k) \bar{u}^p_t(k) + \eta_k}$$

$$\lambda^p_k = \left\{ \frac{\sum_{t=1}^{T} \rho^p_t(k) \bar{u}^p_t(k) (Y^p_t - \mu^p_k)^2}{\sum_{t=1}^{T} \rho^p_t(k) + \gamma_k - 1} + \frac{\eta_k (\mu^p_k - m_k)^2 + S_k}{\sum_{t=1}^{T} \rho^p_t(k) + \gamma_k - 1} \right\}^{-1}$$

(5) Posterior over transition matrix. Define the sufficient statistics as

$$N_{ij} = \sum_{t=2}^{T} I(M^g_{t-1} = i, M^g_t = j)$$

Then

$$p(A^g|\cdot) = \prod_{i=1}^{3} Dir(A^g(i,:) | N_{i,:} + \delta_M)$$

(6) Posterior over group prior. Define the sufficient statistics as

$$N_g = \sum_{p=1}^{P} I(G^p = g)$$

Then

$$p(\pi_G|\cdot) = Dir(\pi_G | N_1 + \delta_{G,1}, \ldots, N_G + \delta_{G,G})$$

2.3 K-medoids

To compare HMM-Mix to a simpler method, we decided to use the KM algorithm applied to precalled data, i.e. the input is Z^p_t rather than Y^p_t. [We used our own HMM method (Shah *et al.*, 2006) to discretize each sample separately, but other methods could be used.] As such, KM (as well as WKM and WECCA, both described below) are two-step or sequential methods where in the first step, the raw data are called as discrete copy number states and in the second step, the patients are clustered based on the called data. KM is just like K-means, except each cluster is represented using one of the original samples (a discrete sequence of calls), rather than as an arithmetic average of the samples, which does not make sense for categorical data. KM requires a distance metric between a sample and a cluster center (prototype). We used Hamming distance: $d(i,j) = \sum_{t=1}^{T} I(Z^i_t = Z^j_t)$. Since KM is prone to getting stuck in local minima, we used 100 restarts, and returned the clustering with the lowest overall distortion. To choose K (the number of clusters), we used the Silhouette coefficient (van der Laan *et al.*, 2003) (see Section 2.6).

2.4 Weighted KM

The KM algorithm described above treats all probes (features) equivalently when computing the distance function. However, we assume that only a small subset of features are important in determining the distance between two patients. We therefore also tried a weighted distance function, $d(i,j) = \sum_{t=1}^{T} w_t I(Z^i_t = Z^j_t)$. We call the resulting method WKM.

The weights are chosen in the following heuristic way. We first compute the empirical distribution over calls at each location, f_t. We then compute the entropy of this distribution, $E_t = -\sum_{k=L,N,G} f_t(k) \log f_t(k)$. Finally, we assign high weights to locations which are highly entropic: $w_t = \sigma(E_t/\alpha)$, where $\sigma(\eta) = \frac{1}{1+e^{-\eta}}$ is the sigmoid function, and α is a constant that controls the steepness of the sigmoid. (We found $\alpha = 0.25$ gave good results.) The use of the sigmoid function ensures $0 \leq w_t \leq 1$.

The reason that we assign high weights to the entropic locations is as follows: locations which are useful for distinguishing the groups must differ across patients, and hence are likely to have a multimodal distribution, whereas locations which are not discriminative are likely to have all possible values (be closer to uniform), and therefore have lower entropy.

In our experimental results below, we show that WKM is much better than KM, although not as good as our model-based approach. However, because of its simplicity and speed, we use it as a way to initialize our model-based approach.

2.5 Hierarchical clustering

In recent work, van Wieringen and van de Wiel (2008) introduce a system called 'Weighted clustering of called array CGH data' (WECCA). This represents the first clustering approach to be tailored specifically to the aCGH data and is a specialized implementation of hierarchical agglomerative clustering. The authors define a weighted form of similarity, similar in spirit to the weighted Hamming distance described above, although the weights are expected to be provided by the user, rather than automatically calculated.

2.6 Choosing the number of groups

The KM and our HMM-Mix model both require that the user specify the number of clusters G. (Hierarchical clustering does not need this information, although one must specify some other mechanism for choosing where to cut the dendogram.) Since KM is not a probabilistic model, one can only use heuristics methods for picking G. We use the Silhouette coefficient (Tan *et al.*, 2005), which computes a measure of quality that considers both cohesion (how similar the points in a cluster are) and separation (how different the

clusters are). In particular, we compute $S(G)$ for a range of values of G, and pick the G with maximum score.

3 DATA

3.1 Simulated data

To test and compare performance of the various algorithms where the true clustering was known, we generated data and embedded group-specific patterns of recurrent CNAs. To avoid circularity that can arise from generating data from the model directly, we created datasets based on real aCGH data derived from mantle cell lymphoma cell lines reported in de Leeuw *et al.* (2004) and used similarly in Shah *et al.* (2007). We first extracted the data from chromosome 21 (chosen because it was reported to have relatively few alterations), resulting in a dataset of 8 samples each with 672 probes. For each simulated dataset, we performed 100 random draws (simulating patients) from the eight cell lines. For each of the 100 patients, we shuffled the 672 probes and randomly assigned the patient to one of G groups. For each group, we preset coordinates of one recurrent gain and one recurrent loss. These group-specific coordinates defined the profile for the group. The alterations were embedded into each patient's data at their group-specific coordinates, plus a random offset number of probes (sampled from a Gamma distribution with a mean of 10 probes). This offset was meant to simulate the fact that recurrent CNAs often have different patient-specific start and end coordinates, but have segments that intersect across patients. Losses were generated by shifting 1 SD down from the neutral state, and gains were shifts of 1 SD up. Finally, for each patient, we randomly embedded alterations of length L at locations different than the group-specific alterations in order to simulate patient-specific 'passenger' alterations expected to be unrelated to the group profile. We created 10 replications with $G=3,5,10$ and $L=50,75$ yielding 60 datasets. These data and the ground truth cluster assignments are included in the Supplemental Material.

With these ground truth datasets in hand, we evaluated clustering accuracy using the Jaccard coefficient as described by Tan *et al.* (2005). (This is a number between 0 and 1, where 1 is the best possible score, corresponding to perfect correspondence to the true clustering.)

3.2 Clinical data

We use two clinical datasets: FL (Fig. 4) and DLBCL (Fig. 5).

The FL data were derived from 106 samples taken at time of diagnosis from patients with FL. These data were previously reported in Cheung *et al.* (2008) and were expected to fall into at least four genetic subtypes (Höglund *et al.*, 2004). A characteristic of FL is that in a subset of patients, the tumor undergoes a transformation to a more aggressive subtype that consistently correlates with inferior survival outcome. Developing a prognostic CNA profile predictive of transformation is therefore of great clinical interest.

The DLBCL data (Johnson *et al.*, 2008), contains aCGH data for 92 patients with *de novo* DLBCL, all treated uniformly with multi-agent chemotherapy (CHOP) and anti-CD20 monoclonal antibody rituximab.

All clinical data were produced using the SMRT array platform (Ishkanian *et al.*, 2004) and contain approximately 27 000 probes per sample.

4 RESULTS

4.1 Simulated data

Figure 3 shows the distribution of the Jaccard coefficient resulting from using WECCA, KM, WKM and HMM-Mix on the 10 replicates for each setting of G, the number of groups and L, the length of the distracting patient-specific passenger alterations. Table 2 contains the mean and standard error for each of the six datasets for the four methods. HMM-Mix showed the highest accuracy for all six settings. When $G=3, L=50$ (Fig. 3A), HMM-Mix and WKM were more accurate than WECCA at recovering the ground truth classes, and statistically more accurate than KM (one-way ANOVA, $P<0.01$). For $G=3$, $L=75$ (Fig. 3D) and $G=5$, $L=50$ (Fig. 3B), HMM-Mix was more accurate than WKM and statistically more accurate than both KM and WECCA ($P<0.01$). For $G=5, L=75$ (Fig. 3E), $G=10, L=50$ (Fig. 3C) and $G=10, L=75$ (Fig. 3F), HMM-Mix was statistically more accurate than all other methods ($P<0.01$). However, for $G=10, L=75$ all methods performed poorly, since this problem is much harder than the others: there are only 10 samples per group, and each sample is 'corrupted' with a fairly long ($L=75$) random CNAs. We repeated these experiments using $P=500$ patients, and all methods improved in their accuracy, although the overall relative rankings are the same (data not shown).

HMM-Mix was generally more robust to the size L of the randomly placed passenger alterations than the other methods, suggesting that the model is able to maintain its ability to detect group-specific alterations in the presence of additional structured noise.

We also tested the robustness of HMM-Mix to initialization. In particular, we initialized with both KM and WKM, and found that the final results were nearly identical, despite the fact that WKM was significantly more accurate than KM.

This suggests that in these settings, HMM-Mix is able to overcome a poor initialization, most likely due to its ability to re-estimate the calls and adapt the feature selection during inference. We suspect that these characteristics allow it to escape from local optima more readily than WKM, which cannot re-estimate the calls and requires the feature selection to be fixed ahead of time. Thus, these results suggest that the joint inference of group assignments and copy number calls used by HMM-Mix is more robust than the sequential methods of WECCA, KM and WKM, all of which perform a two-step method of first calling the data, then clustering.

4.2 FL data

We applied HMM-Mix to the FL cohort of 106 patients (Cheung *et al.*, 2008). We initialized the model using WKM with 100 multiple restarts and we determined the number of groups to be 6 using the maximum Silhouette coefficient over $G=(2,\ldots,8)$. Figure 4 shows the WKM initializations, and the final results of HMM-Mix. In particular, Figure 4A shows the initial $Z_{1:T}^{1:P}$ matrix where rows are patients and columns are probes. The rows are ordered according to their WKM cluster assignments. The green, red and black probes are predicted losses, gains and neutrals, respectively. Figure 4B shows the converged estimates of HMM-Mix where the rows have been ordered according to the HMM-Mix cluster assignments, and the data displayed are the re-estimated calls in the presence of the profiles. Figure 4B (top) shows the profiles of each group and it

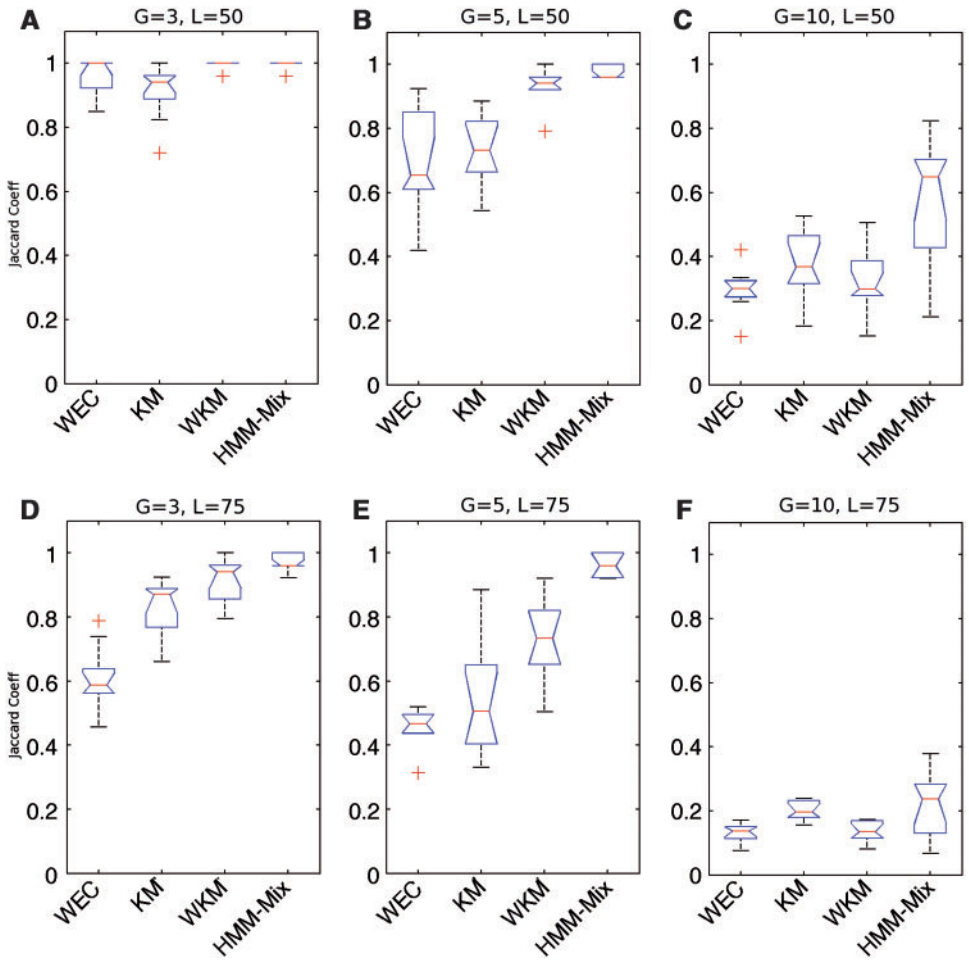

Fig. 3. Distribution of accuracy of WECCA, KM, WKM and HMM-Mix for synthetic data generated with six different parameter settings. HMM-Mix was the most accurate for all six settings (see Table 2 for details). Each dataset was composed of $P=100$ patients with 672 probes each. From left to right there were $G=3,5,10$ embedded groups in the data. The top row had randomly placed CNAs of $L=50$ and the bottom row with $L=75$. Distributions of Jaccard coefficient over 10 replicates of the G,L settings are shown as notched box plots where non-overlapping notches indicate statistical difference of the medians (red horizontal lines) with 95% CI.

Table 2. Results for simulation study showing means and standard errors of the Jaccard coefficient

Dataset	WECCA	KM	WKM	HMM-Mix	ANOVA P-value
G = 3 L = 50	0.959±0.018	0.916±0.027	0.996±0.004	0.996±0.004	4.2×10^{-3}
G = 5 L = 50	0.692±0.048	0.734±0.034	0.932±0.018	0.976±0.007	5.7×10^{-8}
G = 10 L = 50	0.296±0.022	0.375±0.033	0.317±0.031	0.580±0.065	6.7×10^{-5}
G = 3 L = 75	0.611±0.030	0.828±0.029	0.923±0.022	0.965±0.009	4.3×10^{-12}
G = 5 L = 75	0.460±0.019	0.548±0.057	0.730±0.043	0.964±0.011	6.2×10^{-11}
G = 10 L = 75	0.131±0.010	0.202±0.010	0.138±0.010	0.223±0.032	1.3×10^{-3}

is clear that the re-estimated calls are heavily influenced by their corresponding profiles.

The resulting groups can be summarized as follows: (1) +7 (meaning gain of chromosome 7) (7 patients); (2): a 'null' group with no recurrent alterations (67 patients); (3): a group with +18 (19 patients); (4): a group with +1q and a small loss at 1p36 (7 patients); (5): a singleton outlier (1 patient); and (6): +6p/6q- (5 patients). Notably, +1p, +6p/6q−, +7, and +18 have previously been established as cytogenetic pathways to the initiation and development of FL using principal component analysis applied to data generated by a difference laboratory technique called G-banded karyotyping (Höglund *et al.*, 2004).

The clusters produced by HMM-Mix set mirror those reported in Cheung *et al.* (2008). In that paper, the WKM method was used

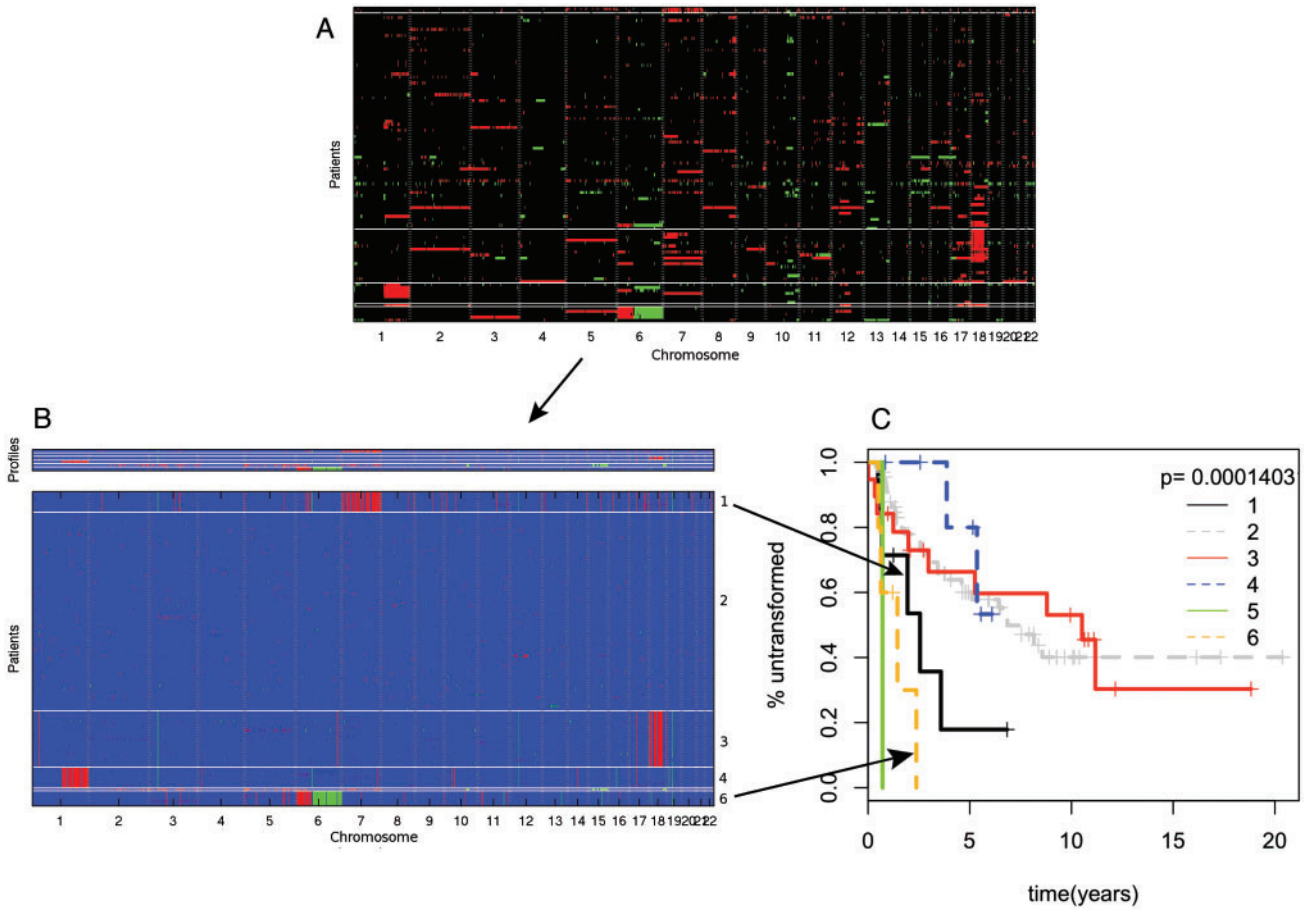

Fig. 4. Clustering of FL data showing the initial calls and WKM clusters (**A**), the converged estimates of the calls (**B**), clusters and profiles by HMM-Mix and the associated time to transformation Kaplan-Meier plots for each group (**C**). (A) The calls and clusters depicted as a heat map for WKM with $G = 6$. The rows of the data indicate the patients and the columns indicate the probes. Red indicates gain, green loss and black neutral. The rows are ordered according to their assigned groups as predicted by WKM. (B) The posterior probability of the calls (where red represents $p(Z_t^p = red)$, blue represents $p(Z_t^p = neutral)$) and green represents $p(Z_t^p = loss)$, the clusters and the profiles (top) for the $G = 6$ groups. In comparison to (A) the clusters are readily apparent from the data, they appear to be tighter and the re-estimated calls are clearly influenced by the profiles, resulting in far less noisy, and far more interpretable output. Importantly, 4 of the 6 groups (labeled on right) recapitulate the previously reported subtypes for FL. Group numbers that correspond to the time to transformation curves (C) are annotated on the right of (B). Groups 1 and 6 both had statistically significantly shorter time to transformation. (C) Time to transformation Kaplan-Meier curves for each group of patients as predicted by HMM-Mix for the FL cohort. Groups 1 and 6 (black and yellow) had significantly reduced time to transformation by log-rank test with 5 degrees of freedom. [The green curve corresponds to the singleton group shown in (B)]. These correspond, respectively, to the groups characterized by +7 and 6p-/6q+ and suggests that these recurrent CNAs confer inferior prognoses to the patients in these groups.

to perform the clustering, but the method used significant human expertise both in determining the initial called data, $Z_{1:T}^{1:P}$, and in defining the weighting terms, $w_{1:T}$. In addition, the number of groups (5) was chosen using supporting evidence from the literature. In contrast, HMM-Mix is fully automated, with no user-settable parameters, yet it managed to recover essentially the same results of this previous method.

As further validation of the biological significance of the clusters found by our method, we computed Kaplan-Meier curves for each group of the time to transformation (TTT) (defined as the time from diagnosis to clinical or pathological endpoint: transformation to the more aggressive subtype). We show the results in Figure 4C. We see that groups 1 and 6 (black and yellow curves) display a significantly shortened TTT to the others (log-rank test, $P < 0.01$), indicating the profiles characterized by +7 and +6p/6q- are potential unfavorable prognostic indicators for FL. Note that by WKM, group 1 (shown as the top group in Fig. 4) which results in the HMM-Mix group characterized by +7 only contains two patients which is inconsistent with both Cheung *et al.* (2008) and Höglund *et al.* (2004) and might therefore be considered less plausible than the HMM-Mix results. The resulting clusters for the 106 patients predicted by WKM and HMM-Mix are included in the Supplemental Material.

4.3 DLBCL data

Figure 5 shows the results of applying WKM and HMM-Mix to the 92 patients in the DLBCL cohort. We see that HMM-Mix is achieving the desired effect of focusing on putative driver or highly recurrent within-group alterations, while ignoring non-recurrent passenger alterations, thus clearly separating signal from noise.

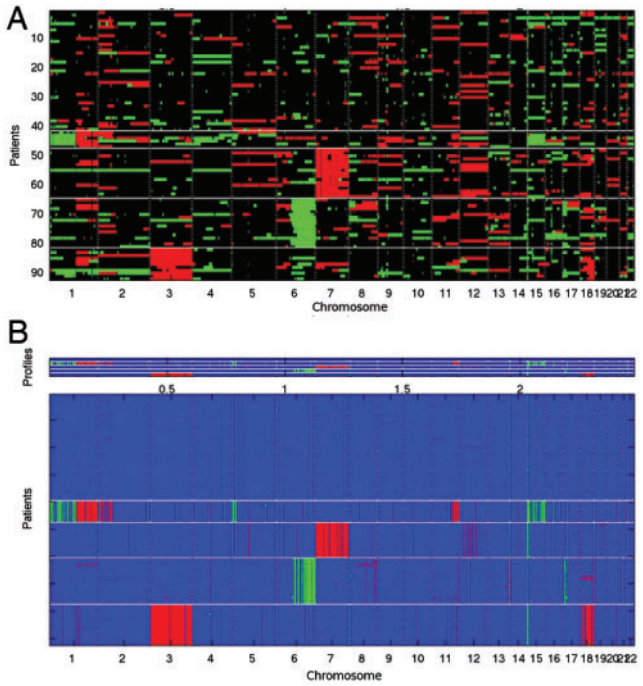

Fig. 5. Clustering of 92 DLBCL profiles into five groups. Comparison between WKM initialization (**A**) and HMM-Mix (**B**) clearly shows HMM-Mix ability to reduce noise and report only highly conserved within-group patterns. The bottom cluster for HMM-Mix (B) shows a potentially novel subtype with gain of chromosome +3/+18. The colors for both (A) and (B) are as described in Figure 4.

The data fell into five distinct groups characterized by a 'null' group with no discernible pattern, and four groups characterized by 1p-/+1q/+2p/+11q/15-, +7, 6q− and +3/+18. The last group is a previously unreported pattern of alteration in DLBCL. Previous work had identified that both changes show increased frequency in the so-called activated B cell (ABC) subtype of DLBCL (Bea *et al.*, 2005), but had not recognized that these two alterations travel together and may indeed define a unique molecular subgroup.

5 DISCUSSION AND FUTURE WORK

The HMM-Mix model presented in this article is effectively able to discover subgroups and their defining profiles given a set of aCGH data derived from a patient cohort. We showed the model's capability of finding clinically relevant subtypes in an FL cohort and a previously undescribed subtype in the DLBCL cohort. We demonstrated how the joint inference procedure of inferring copy number calls, cluster assignments and profiles, coupled with adaptive feature selection, makes HMM-Mix significantly more accurate than partitioning and hierarchical clustering methods. Future work will entail experimental validation and further exploration of the +7 and 6p-/6q+ subgroups detected in the FL cohort for prognostic significance for TTT, and determining clinical relevance of the DLBCL subgroups we reported.

Extension of HMM-Mix to high density SNP arrays (e.g. Affymetrix 6.0) will be of interest, as patterns of both genotype and copy number can be elucidated. HMM-based models for SNP arrays introduced in Colella *et al.* (2007) and Scharpf *et al.* (2008) will be investigated for extension to the clustering setting using the HMM-Mix framework introduced here. Compared to the BAC arrays used in this study, genotyping array probes are much less uniformly distributed across the chromosome. Thus, location-specific transition matrices with distance-based priors as suggested by Colella *et al.* (2007) will be a necessary feature of this work. (Note that most likely owing to the fact that the platform used to generate the data in this study has relatively uniformly distributed probes, we found that non-stationary transition matrices made no difference to our results.) In addition, we will be applying the model to a large cohort of breast tumors for which we have generated Affymetrix SNP 6.0 data with the goal of uncovering novel molecular subtypes. Note that the CNAs in lymphoma entities we studied as part of this article can be dominated by chromosome arm or whole chromosome events. Application to breast cancer will allow us to assess how well the model generalizes to cancers that have much more complex genomes.

Finally, we are investigating the use of variational methods Bishop (2006) for inference that will at once obviate the need to hard assign each patient to a group and preserve the computational efficiency of the inference algorithm. We expect this extension to provide full posterior distributions over the quantities of interest thus better modeling the uncertainty of these estimates. In addition, we are investigating approaches to model selection to avoid having to choose the number of groups at run time.

Funding: Michael Smith Foundation for Health Research (to S.P.S.). KPM wishes to acknowledge support from an NSERC Discovery Grant and from CIFAR. N.A.J. is a research fellow of the Terry Fox Foundation through an award from the National Cancer Institute of Canada (019005) and the Michael Smith Foundation for Health Research (ST-PDF-01793). This project was funded by NCIC grants #016003 and #019001 and Genome Canada.

Conflict of Interest: none declared.

REFERENCES

Aguirre,A.J. *et al*. (2004) High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc. Natl Acad. Sci. USA*, **101**, 9067–9072.

Archambeau,C. (2005) *Probabilistic models in noisy environments – and their application to a visual prosthesis for the blind*. PhD Thesis, Universit catholique de Louvain.

Bea,S. *et al*. (2005) Diffuse large b-cell lymphoma subgroups have distinct genetic profiles that influence tumor biology and improve gene-expression-based survival prediction. *Blood*, **106**, 3183–3190.

Besag,J. (1986) On the statistical analysis of dirty pictures. *J. R. Stat. Soc. Ser. B*, **48**, 259–302.

Bishop,C.M. (2006) *Pattern Recognition and Machine Learning*. Springer, New York, NY.

Brown,M.P. *et al*. (1993) Using Dirichlet mixture priors to derive Hidden Markov models for protein families. *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology*, AAAI, pp. 47–55.

Cheung,K.J. *et al*. (2008) Genome-wide profiling of follicular lymphoma by array comparative genomic hybridization reveals prognostically significant DNA copy number imbalances. *Blood*, **113**, 137–148.

Chin,L. and Gray,J. (2008) Translating insights from the cancer genome into clinical practice. *Nature*, **242**, 553–563.

Chin,S.F. *et al*. (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, **8**, R215.

Colella,S. *et al*. (2007) QuantiSNP: an objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.*, **35**, 2013–2025

Collins,F.S. and Barker,A.D. (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci. Am.*, **296**, 50–57.

de Leeuw,R.J. *et al*. (2004) Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. *Hum. Mol. Genet.*, **13**, 1827–1837.

Dempster,A.P. *et al*. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soci. Ser. B*, **34**, 1–38.

Diskin,S.J. *et al*. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.*, **16**, 1149–1158.

Gilks,W. *et al*. (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Höglund,M. *et al*. (2004) Identification of cytogenetic subgroups and karyotypic pathways of clonal evolution in follicular lymphomas. *Genes Chromosomes Cancer*, **39**, 195–204.

Idbaih,A. *et al*. (2008) BAC array CGH distinguishes mutually exclusive alterations that define clinicogenetic subtypes of gliomas. *Int. J. Cancer*, **122**, 1778–1786.

Ishkanian,A. *et al*. (2004) A tiling resolution DNA microarray with complete coverage of the human genome. *Nat. Genet.*, **36**, 299–303.

Johnson,N.A. *et al*. (2008) Deletion in chromosome 17p12 and gains in chromosome 9q33.3 by array comparative hybridization are associated with R-CHOP treatment failure in patients with diffuse large B cell lymphoma. *Blood*, **111**, a477.

Khalique,L. *et al*. (2007) Genetic intra-tumour heterogeneity in epithelial ovarian cancer and its implications for molecular diagnosis of tumours. *J. Pathol.*, **211**, 286–295.

Klijn,C. *et al*. (2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.*, **36**, e13.

Law,M.H.C. *et al*. (2004) Simultaneous Feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 1154–1166.

Michels,E. *et al*. (2007) ArrayCGH-based classification of neuroblastoma into genomic subgroups. *Genes Chromosomes Cancer*, **46**, 1098–1108.

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

Pinkel,D. and Albertson,D. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, **37**(Suppl), 11–17.

Raftery,A.E. and Dean,N. (2006) Variable selection for model-based clustering. *J. Am. Stat. Assoc.*, **101**, 168–178.

Rouveirol,C. *et al*. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849–856.

Scharpf,R.B. *et al*. (2008) Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. *Ann. Appl. Stat.*, **2**, 687–713.

Shah,S.P. *et al*. (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, **22**, 431–439.

Shah,S.P. *et al*. (2007) Modeling recurrent DNA copy number alterations in array CGH data. *Bioinformatics*, **23**, 450–458.

Sorlie,T. (2004) Molecular portraits of breast cancer: tumour subtypes as distinct disease entities. *Eur. J. Cancer*, **40**, 2667–2675.

Tan,P.-N. *et al*. (2005) *Introduction to Data Mining*. First edn. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Tonon,G. *et al*. (2005) High-resolution genomic profiles of human lung cancer. *Proc. Natl Acad. Sci. USA*, **102**, 9625–9630.

van der Laan,M.J. *et al*. (2003) A new partitioning around medoids algorithm. *J. Stat. Comput. Simul.*, **73**, 575–584.

van Wieringen,W.N. and van de Wiel,M.A. (2008) Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **9** 484–500.

Wright,G. *et al*. (2003) A gene expression-based method to diagnose clinically distinct subgroups of diffuse large b cell lymphoma. *Proc. Natl Acad. Sci. USA*, **100**, 9991–9996.