BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i450–i458
doi:10.1093/bioinformatics/btm221

Modeling recurrent DNA copy number alterations in array CGH data

Sohrab P. Shah[1,*], Wan L. Lam[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]BC Cancer Research Centre, 675 W 10th Ave Vancouver, BC V5Z 1L3, Canada

ABSTRACT

Motivation: Recurrent DNA copy number alterations (CNA) measured with array comparative genomic hybridization (aCGH) reveal important molecular features of human genetics and disease. Studying aCGH profiles from a phenotypic group of individuals can determine important recurrent CNA patterns that suggest a strong correlation to the phenotype. Computational approaches to detecting recurrent CNAs from a set of aCGH experiments have typically relied on discretizing the noisy log ratios and subsequently inferring patterns. We demonstrate that this can have the effect of filtering out important signals present in the raw data. In this article we develop statistical models that *jointly* infer CNA patterns and the discrete labels by borrowing statistical strength across samples.

Results: We propose extending single sample aCGH HMMs to the multiple sample case in order to infer shared CNAs. We model recurrent CNAs as a profile encoded by a master sequence of states that generates the samples. We show how to improve on two basic models by performing joint inference of the discrete labels and providing sparsity in the output. We demonstrate on synthetic ground truth data and real data from lung cancer cell lines how these two important features of our model improve results over baseline models. We include standard quantitative metrics and a qualitative assessment on which to base our conclusions.

Availability: http://www.cs.ubc.ca/~sshah/acgh

Contact: sshah@cs.ubc.ca

1 INTRODUCTION

Genetic alterations are a hallmark of numerous human diseases such as cancer and mental retardation. Recent advances in the genome-wide identification and localization of genetic alterations by high resolution array comparative genomic hybridization (aCGH) technologies (Ishkanian *et al.*, 2004, Pinkel and Albertson, 2005) have furthered our understanding of the effect of genetic alterations on disease. Array CGH measures genetic changes as DNA copy number alterations (CNAs) of an individual's DNA against a reference at a fixed set of locations in the genome (Pinkel and Albertson, 2005). A recurrent CNA in a cohort of patients is a CNA found at the same location in multiple samples. Therefore, recurrent CNAs define a pattern that provides a molecular characterization of the cohort's phenotype, potentially identifying disrupted molecular processes, molecular targets for diagnosis, and development of novel therapeutics. A recent example is the identification of recurrent CNAs across different subtypes of lung cancer, reported in Coe *et al.* (2006). This led to the elucidation of molecular mechanisms that contribute to the distinct phenotypes in small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), which we show in this article (see Section 5).

In this study, we describe the development of statistical models for the detection of recurrent CNAs across multiple aCGH experiments from a cohort of individuals. Figure 1a–c shows an example of five aCGH NSCLC samples over small regions containing three different types of recurrent CNAs on chromosomes 8, 9 and 1. For each of the ~30 000 probes in an aCGH experiment, a log ratio of the hybridization level of the sample, relative to the reference, is produced. The log ratios have a noisy correspondence to CNAs: deletions (losses) result in negative log ratios, amplifications (gains) result in positive log ratios and no change (neutral) regions in zero log ratios.

Figure 1a shows a recurrent CNA harboring the *MYC* oncogene. One common strategy to identify such a recurrent CNA is to first pre-process individual samples to make calls of losses and gains, and then to infer recurrent CNAs using a threshold frequency of occurrence (de Leeuw *et al.*, 2004; Garnis *et al.*, 2006, Pollack *et al.*, 2002). We call this process AF for alteration frequency (see Section 3 for details). While AF may detect signals as shown in Figure 1a, pre-processing or discretizing the sequences separately may remove information by smoothing over short or low-amplification CNAs. However, by *jointly* considering all the data without pre-processing, we can borrow statistical strength (Gelman *et al.*, 2004) across the samples and identify locations where the signal is shared in the raw data. For example, in Figure 1b, we show data at the locus containing an important NSCLC gene, carbonic anhydrase IX (CA9) (Kim *et al.*, 2004; Swinson *et al.*, 2003). Log ratios of probes overlapping the gene are shown as blue stars and are indicated with arrows. This shared CNA may be hard to detect using AF because when processing individual samples, single probe CNAs are often indistinguishable from experimental noise (Veltman and de Vries, 2006). With high-dimensional arrays, many investigators require CNAs to span at least two consecutive probes (Baldwin *et al.*, 2005; Garnis *et al.*, 2006; Veltman and de Vries, 2006). However, if a single probe CNA is shared across many samples, it may correspond to an important biological feature.

Figure 1c shows a third type of signal that is a low-level or subtle shared CNA. The region includes two known lung cancer

*To whom correspondence should be addressed.

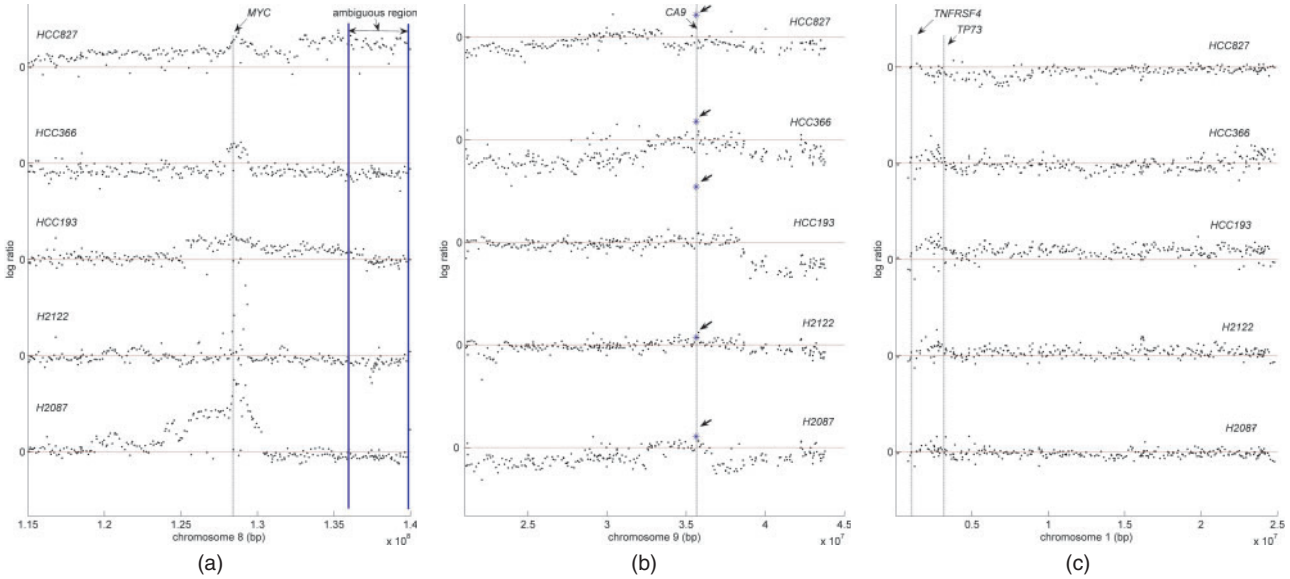

Fig. 1. aCGH profiles of three different types of recurrent CNAs for five NSCLC cell lines (labeled on the right). Horizontal red lines indicate the 0 log ratio level for each sample. Vertical black lines indicate the position of a known gene of interest in NSCLC. **(a)** a high level shared amplification of a region spanning ~3 Mb containing the *MYC* oncogene on chromosome 8 (shown with vertical line). **(b)** a single clone shared aberration at the CA9 locus on chromosome 9. **(c)** a low-level amplification on chromosome 1 including TNFRSF4 and TP73—both implicated in NSCLC. Both (b) and (c) illustrate examples of recurrent CNAs that may be undetectable if each sample is pre-processed separately.

related genes, *TNFRSF4* (Kawamata *et al.*, 1998) and *TP73*. When compared to the *MYC* region in Figure 1a, the level of amplification for *TNFRSF4* is much lower and is more difficult to distinguish from noise. However, cell lines H2122, HCC193 and HCC366 appear to share the low-level amplification. Furthermore, the *TP73* [a putative tumor suppressor involved in cell death (Alarcon-Vargas *et al.*, 2000)] loci exhibits low-level negative signals in three of the samples. These signals may be lost if each sample is pre-processed in isolation due to premature thresholding.

Most of the genome will not exhibit shared CNA patterns. Figure 1a (right end) shows a region from ~135 to 140 Mb (bounded by blue vertical lines) that is heterogeneous across the samples. One sample (HCC827) has an amplification while two are neutral (HCC193, H2087) and two are deletions (HCC366, H2122). This ambiguity in the signal across samples will be important when we develop our model in Section 3.

In this article, we present statistical models to infer recurrent CNAs from aCGH data. We extend the single sample hidden Markov model (HMM) (Fridlyand *et al.*, 2004; Shah *et al.*, 2006) to the multiple sample case. We consider three different ways to do this. The first simply modifies the observation model of the HMM so that at each location, a vector of observations is generated, one per sample. We call the state sequence of the HMM the 'master' sequence. It represents a classification of each probe location into a loss, neutral or gain state and hence it represents the canonical signal that encodes recurrent CNAs.

The second model augments this by allowing each observation in each sample to either be generated from the master sequence, or from its own private sequence. This allows for sample-specific random effects to be superimposed on the canonical signal. We demonstrate that this improves performance significantly. Finally, the third model augments the state space of the master sequence to allow undefined states, which represent locations which are highly ambiguous (such as the 135–140 MB region in Fig. 1a). This allows the master to focus on the highly conserved regions, and to ignore heterogeneous locations. We will show that the resulting output we infer is comparatively sparse, making it easier to create a short list of candidate locations for experimental follow up.

The remainder of the article is organized as follows. We discuss related work in Section 2. We describe our models in Section 3. In Section 4, we demonstrate our results on simulated data, where we know the ground truth, and in Section 5, we demonstrate results on well-studied lung cancer cell line data (Coe *et al.*, 2006). In Section 6, we conclude the article and discuss future work.

2 RELATED WORK

There has been surprisingly little work on the automatic discovery of shared patterns from aCGH data. Rouveirol *et al.* (2006) propose an algorithm that takes as input a set of discretized sequences, and which outputs a set of minimal recurrent regions. This method works by converting the sequences to a $S \times T$ binary matrix (focusing on losses and gains separately), where S is the number of samples and T is the number of probes on the array. It then tries to find short blocks that are shared by some specified fraction of the samples. The disadvantages of their approach are that it requires discretized data, and that its running time is $O(T^2)$.

(They also present an $O(T)$ version, but this tended to not work as well.) Diskin *et al*. (2006) also take a binary $S \times T$ matrix as input, and use a greedy search procedure to find regions ('stacks') that are shared across samples with statistically significant frequency. Lipson *et al*. (2006) is the only previous work we are aware of that tries to find shared patterns using the raw data (to avoid problems with premature thresholding discussed above). They present an $O(T^2)$ time algorithm for finding intervals with maximal score, although they present two other versions of the algorithm which they say in practice are $O(T^{1.5})$ and $O(T)$ complexity. An important difference between this approach and ours is that Lipson *et al*. assume the log ratios are identically and independently distributed across the chromosome, when in fact they are spatially correlated, suggesting that Markovian dynamics, encoded by HMMs, should be used. In addition, our approach is $O(T)$ and model based.

3 METHODS

We consider analyzing multiple aCGH samples from multiple chromosomes, although to simplify notation, we will concentrate on modeling a single chromosome. The data is $\mathcal{D} = (Y_{1:T}^{1:S})$, where $Y_t^s \in \mathbb{R}$ is the observed log ratio at location t in sample s, T is the number of probes and S is the number of samples. The goal is to identify patterns in the data which capture the pertinent CNAs that recur across the samples. A pattern consists (roughly speaking) of a list of locations which are highly conserved (either loss, neutral or gain). Thus, the pattern can be represented as a 'master' sequence of states $M_{1:T}$ where $M_t \in \{L,N,G\}$ (loss, neutral, gain) is a multinomial random variable and $t \in (1,2,\ldots,T)$. Since we will often be uncertain about what the pattern should be at any given location, we will summarize our uncertainty using the (marginal) posterior distributions $\phi_t = p(M_t|\mathcal{D})$, which we call a *profile*. When we have data from different groups (as in our lung cancer data), we learn a different profile for each group, $\phi_t^g = p(M_t^g|\mathcal{D}^g)$. As shown in Section 5, we analyze four different phenotypic groups of lung cancer (i.e. g=1:4). However, we will drop the g superscript for brevity.

Our task is related to learning profile HMMs for multiple sequence alignment (Durbin *et al*., 1998), but it is harder because the raw data is noisy and continuous-valued. Below, we describe four different approaches to the problem. The first is the method most widely used in current practice, and the remaining three are novel methods that we propose.

3.1 Alteration frequency (AF) model

In the simplest approach, AF, we first process each sample $Y_{1:T}^s$ into a discrete sequence $Z_{1:T}^s$, where $Z_t^s \in \{L, G, N\}$, using a standard method for discretizing single-sample aCGH data. In this article, we use the robust HMM approach described in Shah *et al*. (2006). We chose the HMM-based single-sample method for this implementation of AF to allow a more direct algorithmic comparison to the multiple sample HMMs we describe below. Note that other algorithms could be used for this step. For example, Coe *et al*. (2006) used aCGH-smooth (Jong *et al*., 2004) to preprocess the lung cancer data presented in Section 5. After preprocessing, we compute the empirical distribution over each state in each location to yield the profile $\phi_t = p(M_t|\mathcal{D})$, which can be represented as a $K \times T$ stochastic matrix, where K=3 is the number of states, T is the length of the sequence and each column sums to one. This can be further simplified to just compute the empirical probability of a recurrent CNA at each location, to yield a $1 \times T$ vector. The disadvantage of this method is that the mapping from Y^s to Z^s is done on each sample separately, so information cannot be shared across samples. Thus the method may smooth over important signals, as we will see.

3.2 Factored likelihood HMM (FL-HMM)

The second model, which we call 'factored likelihood HMM' (FL-HMM), is a standard HMM model for $M_{1:T}$ (modeling the fact that CNAs tend to occur in runs), but where we modify the likelihood function to generate multiple samples instead of a single sample. Specifically, we assume the samples are conditionally independent given M_t and use a Gaussian observation model, yielding

$$p(Y_t^{1:S}|M_t = j) = \prod_{s=1}^{S} \mathcal{N}(Y_t^s|\mu_j^s, \sigma_j^s)$$

The observation model is a product over the emission densities of the samples, hence the term 'factored likelihood'. We have one mean and variance parameter for each of the three states of the HMM. The mean and variance are sample specific, to model the fact that different samples often have quite different noise characteristics, due to quality of hybridization, different ploidy, tumor/normal admixture coefficients, etc. Note that in a given chromosome some samples may not contain any CNAs, in which case the estimates of μ_j^s and σ_j^s may be poor if j is an aberrated state (L,G). Hence we share (pool) these parameters across chromosomes for statistical strength, as described in Shah *et al*. (2006). The variable M_t has Markovian dynamics with transition matrix A_M, representing the probability of switching between the L/N/G states. The starting state distribution is denoted π_M. The model is shown as a directed graphical model in Figure 2a. Please see Bishop, (2006) for details on directed graphical models and how they relate to state transition diagrams for HMMs.

We add standard conjugate priors to all the parameters (Gelman *et al*., 2004). Specifically, for the multinomial distributions we use Dirichlet priors, $A_M \sim \mathrm{Dir}(\delta_M)$ and $\pi_M \sim \mathrm{Dir}(\delta_{\pi_M})$, where the matrix of pseudocounts δ_M encourages self-transitions, and δ_{π_M} encourages the neutral state. For the observation variance, we use $\lambda_j^s \sim Ga(\alpha_j^s, \beta_j^s)$, where $\lambda_j^s = 1/\sigma_j^s$ is the precision, and Ga is a gamma density. We set the hyper-parameters in a data driven way as follows. We set $\alpha_j^s = 1 + \sigma^s$ to encode the expected variance, where σ^s is the empirical variance of sample s, and $\beta_j^s = 1$ to reflect that this is a weak prior.

For the observation mean, we use a Gaussian prior, $\mu_j^s \sim \mathcal{N}(m_j^s, v_j^s)$. We set the hyper-parameters as follows: $m_1^s = -\sigma^s$, $m_2^s = 0$, and $m_3^s = \sigma^s$. We set the mean of state 2 to 0 based on the assumption the data has been normalized so that the neutral state usually corresponds to a log ratio of 0. We set the mean of the aberrated states to reflect the typical deviations expected in this sample. This allows the model to adapt to automatically different noise levels coming from different samples or even different platforms. We set the prior variance on the mean to $v_j^s = 10^{-3}$. This was chosen to reflect that our method for choosing m_j^s was appropriate in the majority of the data we have observed.

To maintain identifiability of the hidden states (i.e. to maintain state 1 means loss, 2 means neutral and 3 means gain), we use a truncated Gaussian on μ_j^s, to ensure $\mu_1^s < \mu_2^s < \mu_3^s$. The truncation bounds are set in a similar way to the hyper-parameters.

Let $\theta = (\mu_{1:K}^{1:S}, \sigma_{1:K}^{1:S}, A_M, \pi_M)$ be all the parameters of the model. We can estimate the parameters of this model, $p(\theta|\mathcal{D})$, using a Markov chain Monte Carlo (MCMC) algorithm called blocked Gibbs sampling (Scott, 2002). This entails alternating between sampling $M_{1:T}$ as a block using the forwards-filtering backwards-sampling (FFBS) algorithm, and sampling the parameters individually conditioned on $M_{1:T}$ and the data: see Algorithm 1 for details. Alternatively, we can compute a point estimate, $\theta^{\mathrm{MAP}} = \arg\max_\theta p(\mathcal{D}|\theta)p(\theta)$, using the EM (expectation-maximization) algorithm.

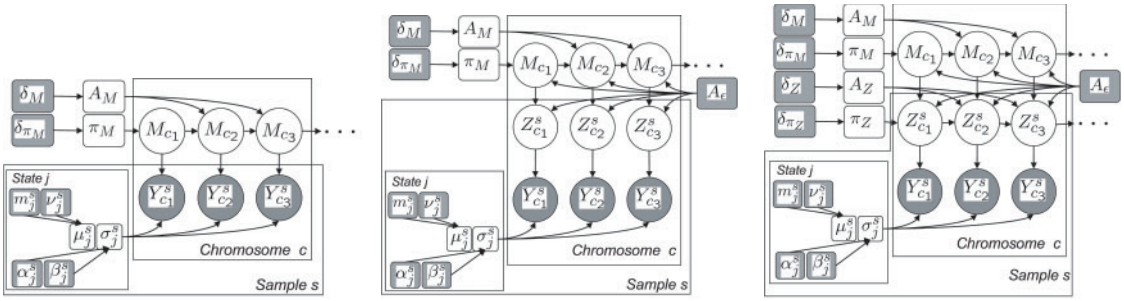

Fig. 2. The three models **(a)** FL-HMM, **(b)** BFL-HMM and **(c)** H-HMM shown as directed graphical models (Bayesian networks). Circles represent random variables and rounded squares represent parameters. We only show the models for 3 probes, but in reality, the number of random variables is proportional to the number of probes on the chromosome, T_c. Unknown quantities are unshaded and observed quantities are shaded. $Y^s_{c,t}$ represents the observed log ratio of sample s in chromosome c at location t, $M_{c,t}$ is the hidden master state and $Z^s_{c,t}$ is the hidden 'slave' state. The shaded square nodes represent fixed hyper-parameters. Arrows between nodes indicate probabilistic dependencies. Boxes around variables are called 'plates' and represent repetition of the contents inside. Thus we see that the observation parameters μ_s and σ_s are shared (tied) across chromosomes (since they are outside the c plate) but are specific to each sample (since they are inside the s plate), while the HMM parameters A_M, π_M are shared across chromosomes and samples. The differences between BFL-HMM and H-HMM are that $M_{ct} \in \{L,G,N,U\}$ in H-HMM whereas $M_{ct} \in \{L,G,N\}$ in BFL-HMM and FL-HMM. Also, H-HMM has Markovian dynamics on the Z^s_{ct} process (see the horizontal links and the new A_Z, π_Z parameters).

Algorithm 1. Blocked Gibbs sampling algorithm for H-HMM. We omit the π_M and π_Z terms for brevity. FFBS stands for forwards-filtering backwards sampling

for *iter*=1,2,... **do**
  /* Sample states (E step) */
  for t=1:T **do**
    $B(j,t) = \begin{cases} \prod_s A_\epsilon(j, Z^s_t) & \text{if} j \in \{L,G,N\} \\ \prod_s A_Z(Z^s_{t-1}, Z^s_t) & \text{if} j = U \end{cases}$
  end for
  $M_{1:T} \sim FFBS(A_M, B^{1:T})$
  for s=1:S **do**
    for t=1:T **do**
      $B(j,t) = \mathcal{N}(y^s_t | \mu^s_j, \sigma^s_j)$
      $A'_Z(i,j) = \begin{cases} A_\epsilon(M_t, j) & \text{if} M_t \in \{L,G,N\} \\ A_Z(i,j) & \text{if} M_t = U \end{cases}$
    end for
    $Z^s_{1:T} \sim FFBS(A'^{1:T}_Z, B^{1:T})$
  end for
  /* Sample parameters (M step) */
  $A_M \sim Dir(\delta_M + \sum_{c,t} I(M_{ct} = i, M_{c,t+1} = j))$
  $C_Z = \sum_{c,s,t} I(Z^s_{ct} = i, Z^s_{c,t+1} = j) I(M_t = U))$
  $A_Z \sim Dir(\delta_Z + C_Z)$
  for s=1:S **do**
    for j=1:K **do**
      $n^s_j = \sum_{c,t} I(Z^s_{ct} = j)$
      $\bar{y}^s_j = \frac{1}{n^s_j} \sum_{c,t} I(Z^s_{ct} = j) y^s_{ct}$
      $\bar{\lambda}^s_j = \frac{1}{n^s_j (\nu^s_j)^2 + (\sigma^s_j)^2}$
      $(\bar{\sigma}^s_j)^{-2} = \frac{1}{(\nu^s_j)^2} + \frac{n^s_j}{(\sigma^s_j)^2}$
      $\mu^s_j \sim \mathcal{N}(\bar{\lambda}^s_j ((\sigma^s_j)^2 m^s_j + n^s_j (\nu^s_j)^2 \bar{y}^s_j), (\bar{\sigma}^s_j)^2)$
      $\bar{\beta}^s_j = \frac{1}{2} \sum_{n=1}^{n^s_j} (I(Z^s_{ct} = j)(y^s_{cL} - \bar{y}^s_j))^2$
      $\lambda^s_j \sim Ga(\alpha^s_j + n^s_j/2, \beta^s_j + \bar{\beta}^s_j)$
    end for
  end for
end for

Parameter initialization is done by setting μ^s_j, σ^s_j using a heuristic method analogous to one iteration of K-means clustering. Using the prior means $m^s_{1:K}$ as initial values for the centroids, the data in each sample are assigned to the nearest centroid based on the Gaussian probability density function. Based on these label assignments, μ^s_k, σ^s_k are inferred using maximum likelihood estimation. We initialize A_M with 0.9 on the diagonals and 0.1 spread over the remaining entries. We initialize π_M to favor neutral states. We can then run EM/MCMC.

3.3 Buffered factored likelihood HMM (BFL-HMM)

The problem with the FL-HMM model is that M_t is summarizing the raw data $Y^{1:S}_t$. If any single sample at a given position has a large deviation from neutral, the master is likely to think that location is aberrated (because the neutral state cannot generate large aberrations). Thus large but rare deviations will be added to the profile. [This problem was also noticed by Lipson *et al.* (2006).] A simple fix to this is to add a 'buffer' to each observation, $Z^s_t \in \{L,N,G\}$, which is responsible for generating the observation Y^s_t. Now the master will summarize these discrete states rather than the raw data. A key point is that in contrast to the AF model, we estimate Z and M simultaneously. See Figure 2b.

In more detail, the BFL-HMM can be defined as follows. The 'slave' Z^s_t processes are modeled as noisy versions of the master process: $p(Z^s_t = j | M_t = k) = A_\epsilon(j,k)$, where

$$A_\epsilon = \begin{pmatrix} \epsilon & \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & \epsilon & \frac{1-\epsilon}{2} \\ \frac{1-\epsilon}{2} & \frac{1-\epsilon}{2} & \epsilon \end{pmatrix}$$

Here ϵ is the probability that the slave copies the master state. If we set ϵ=0, the slaves never copy the master, so the posterior profile will equal the prior profile, i.e. we will not have learned anything, since M_t will be disconnected from the data Y^s_t. As we increase ϵ, each slave is influenced by the master with increased strength. Thus more of the samples will get reflected in the profile. If we set ϵ=1, we are requiring that the slaves perfectly copy the master. This reduces to the FL-HMM model. In practice, we find it best to set $\epsilon \sim 0.8$. See Section 5 for further discussion on the effect of ϵ. We can estimate the parameters in this model using MCMC or EM. We simply modify the algorithm to handle the fact that the observation model is now (a product of) a mixture of

Gaussians, with mixing weights $p(Z_t^s = j|M_t = k)$. We omit details due to lack of space.

3.4 Hierarchical HMM (H-HMM)

The problem with the BFL-HMM is that the slaves *have* to copy the master with probability ϵ at every location, even if this location is highly variable. We extend the model by adding an undefined (do not-care) state U to the master. Now if M_t=U, the slaves follow their own private Markovian dynamics, modeling local runs which are not shared, as shown in Figure 2c. If $M_t \neq U$, they copy the master with probability ϵ as before:

$$p(Z_t^s|Z_{t-1}^s, M_t, A_Z, A_\epsilon) = \begin{cases} A_Z(Z_{t-1}^s, Z_t^s) & \text{if } M_t = U \\ A_\epsilon(M_t, Z_t^s) & \text{if } M_t \in \{L, N, G\} \end{cases}$$

The effect of this is that only highly conserved regions are stored in the profile; in the highly variable regions, the profile says 'undefined'. This makes the profile sparser, and easier to interpret. This is shown in Section 5. The degree of sparsity is controlled by ϵ: as we increase ϵ, the sparsity decreases, since more of the slaves influence the master.

Estimating the parameters in this model is harder, since all the Z^s chains become coupled due to explaining away [cf. factorial HMMs (Ghahramani and Jordan, 1997)]. However, conditioned on $M_{1:T}$, the $Z_{1:T}^s$ are independent and can be sampled in parallel using FFBS, so blocked Gibbs sampling is still easy. See Algorithm 1 for details. An interesting feature of this model is that there are competing processes to explain the slaves. If the slave copies the master, its conditional probability distribution (CPD) is determined by A_ϵ, otherwise the CPD is determined by A_Z. Since A_Z is potentially estimated from a large subset of the data, it tends to converge to have diagonal values near 1. In contrast, A_ϵ is fixed and therefore can be overwhelmed by the slave process. To avoid this, we use a strong prior on A_Z to discourage it from reaching near 1 on the diagonals, but still allowing it to be estimated from the data. This results in a 'fairer' competition between the A_Z process and the A_ϵ process.

We initialize the parameters as in the FL-HMM model. To initialize the states, we first sample each $Z_{1:T}^s$ using FFBS, with the master process turned off. We then initialize M_t to be the consensus majority state across $Z_t^{1:S}$. The EM algorithm for the H-HMM is similar to the MCMC algorithm except we maximize the parameters instead of sampling them. Preliminary comparisons indicate that EM tends to converge faster but gives poorer results, perhaps because it is more prone to getting stuck in local optima.

3.5 Running time

Parameter estimation in all four models takes $O(T)$ time. This makes the technique scalable for use in high density oligonucleotide arrays or SNP arrays, frequently used for DNA copy number analysis, that may contain 500 000 or more probes per experiment. In practice, the running time depends on the number of EM/MCMC iterations. For EM on the H-HMM model, we find the system converges within about 10 steps and takes ~90 min to learn a model from 20 samples with 32 000 probes each. [All experiments were performed in Matlab 7.2.0.294 (R2006a) on a Intel Xeon CPU ©2.4 GHz.] EM for the BFL-HMM and FL-HMM is much faster, since the E step can be performed exactly using the forwards–backwards algorithm, avoiding a Monte Carlo approximation.

4 QUANTITATIVE RESULTS ON SYNTHETIC DATA

Real data sets rarely have fully verified ground truth locations of recurrent CNAs. Thus, applying standard metrics to assess accuracy on real data is difficult. To overcome this, we created a synthetic data set derived from real data. We used eight mantle cell lymphoma samples originally published in de Leeuw *et al.* (2004) and used for a qualitative assessment in Rouveirol *et al.* (2006) and modified it to give us ground truth CNAs. We used the data for chromosome 20 (672 probes) which was reported to be relatively free of CNAs. We permuted the order of the data for each sample so as to remove any undetected shared signals that may be present across samples. We then inserted a recurrent CNA gain and a recurrent CNA loss at fixed positions of width w, in a fraction f of the samples. The clones within the region were shifted up/down (for gain/loss) by $\sigma_s\ \tau$ where s is one of the chosen samples, σ_s is the empirical variance of that sample and τ is the signal to noise ratio (SNR). Thus, $\sigma_s\ \tau$ preserves the sample-specific heterogeneity of the noise. In order to 'soften' the borders of the aberrations, we extended the borders by γ probes, where $\gamma \sim Gam\ (\alpha,1)$ (α proportional to w—see text below). Here, γ was sampled independently for each sample to ensure the exact borders of the aberrations were not shared. Finally, for each sample, we randomly sampled a location outside of the ground truth recurrent CNA and inserted a gain or loss (randomly chosen) of width w'. Figure 3a shows an example of the synthetic data for w=50, f=0.75, τ=0.9, w'=100 (~15% of the chromosome). The recurrent loss is at position 100-149 and the recurrent gain is at position 450-499. Comparing this figure to the real data in Figure 1, we see that the synthetic data is quite realistic and challenging.

We evaluated AF, FL-HMM, BFL-HMM and H-HMM on synthetic data for $w = (1,10,50)$, $f = (1/2,\ 3/4,\ 1)$, $\alpha = (1,5,10)$ and $\tau = (0.3,0.6,0.9,1.2)$. For the BFL-HMM and the H-HMM, we set $\epsilon = (0.8)$. Note for this large scale experiment we used (Monte Carlo) EM instead of MCMC for inference, to save time. However, preliminary results suggest that MCMC does work better, despite its increased cost.

We computed receiver operator characteristic (ROC) curves based on $p(M_t = A) = p(M_t = L) + p(M_t = G)$ where $p(M_t = A)$ is the probability that a recurrent CNA is predicted at position t. Using the ground truth labeling of the data, the false positive rate (FPR) is defined as $\frac{\text{FP}}{N}$ the number of probes incorrectly predicted as a CNA (FP) over the total number of non-CNA probes. The true positive rate is defined as $\frac{\text{TP}}{P}$, the number of correctly predicted CNA probes (TP) over the true number of CNA probes. We plotted TPR versus FPR curves and calculated area under this curve (AUC) as a measure of accuracy to test the effect over w,f,τ and ϵ across the various models.

Figure 3b shows a single summary ROC plot combining results for all values of w,f,τ and depicts the overall accuracy performance of the models. H-HMM had the highest accuracy (AUC = 0.87) followed by BFL-HMM (AUC = 0.82), AF (AUC = 0.77) and FL (0.55). Figure 3c shows the mean AUC over for every setting of w,f,τ (repeated three times). The mean and standard error AUC for the models was 0.84 ± 0.01 for H-HMM, 0.82 ± 0.01 for BFL-HMM, 0.76 ± 0.01 for AF and 0.59 ± 0.01 for FL-HMM. H-HMM and BFL-HMM were significantly more accurate than AF and FL-HMM (one way ANOVA, $p \ll 0.01$). Although H-HMM had slightly higher mean of AUC than BFL-HMM, the result was not statistically

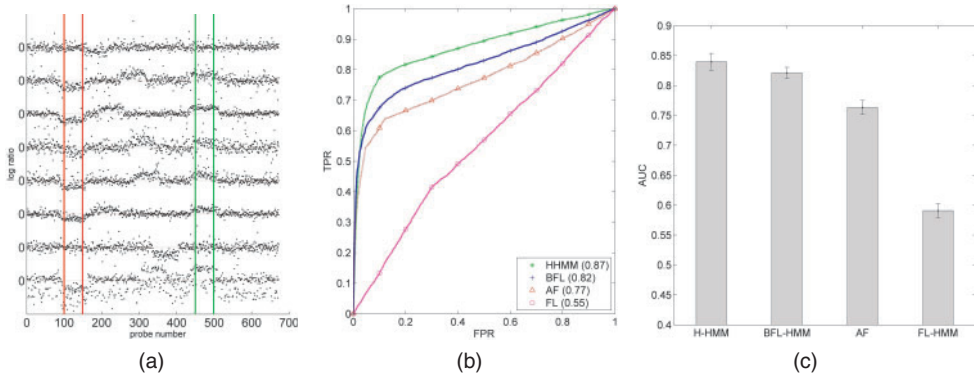

Fig. 3. **(a)** Example of the simulated data for $w=50$, $\tau=0.9$ and $f=0.75$. Green lines (on the right) bound an inserted CNA gain, and red lines (on the left) bound an inserted CNA deletion. **(b)** ROC plot for the synthetic data for H-HMM (green stars), BFL-HMM (blue crosses), FL (red triangles) and AF (purple circles). TPR and FPR were calculated using results for all data, and therefore represent a summary of how the models compare over w, τ, f and ϵ. AUC for each model is indicated in brackets in the legend. H-HMM had the best performance overall (AUC = 0.87), followed by BFL-HMM (AUC = 0.82), AF (AUC = 0.77) and FL (0.55). **(c)** Distributions of AUC for H-HMM, BFL-HMM, FL-HMM and AF over all values of w, τ, f and ϵ. H-HMM and BFL-HMM had statistically significantly better performance than AF and FL-HMM (one way ANOVA ($p \ll 0.01$). Whiskers indicate standard error bars. The mean and standard error AUC for the models was 0.84±0.01 for H-HMM, 0.82±0.01 for BFL-HMM, 0.76±0.01 for AF and 0.59±0.01 for FL-HMM.

significant. However, we show in the next section on lung cancer data that in practice, the H-HMM is considerably more useful to the investigator as it returns sparser, yet accurate predictions.

5 QUALITATIVE RESULTS ON LUNG CANCER DATA

Ultimately, we are interested in applying a model to aCGH data from clinically relevant samples. To compare the output characteristics of the various models, we ran the algorithms on aCGH samples from 39 well-studied lung cancer cell lines, originally published in (Coe *et al*., 2006; Garnis *et al*., 2006). This data is particularly relevant since phenotype-specific patterns of recurrent CNAs have been experimentally validated. The samples can be subdivided into four groups: NSCLC adenocarcinoma (NA), NSCLC squamous cell carcinoma (NS), SCLC classical (SC) and SCLC variant (SV). Eighteen samples are NA, seven are NS, nine are SC and five are SV. This data has been rigorously studied and discordant shared patterns validated using PCR and gene expression have been identified across the major and minor groups (Coe *et al*., 2006; Garnis *et al*. 2006). We fit separate profiles $\phi^{g}_{1:T}$, one per group, using each of the four models and we qualitatively assess the characteristics and biological relevance of the output, using results reported in Coe *et al*. (2006) as a guide.

The experiments on synthetic data showed H-HMM and BFL-HMM are the best models. In this section, we show how the explicit modeling of the ambiguity in the data by H-HMM displays a clear advantage over the other models. Recall that in Figure 1 we showed parts of chromosomes 8, 9, and 1 to illustrate different types of recurrent CNAs at important locations. Figures 4 and 5 show the output of H-HMM ($\epsilon=0.8$), BFL-HMM, FL-HMM and AF on the full chromosome 8 and the p-arm of chromosome 9. $p(M_t=\text{gain})$ is plotted in green and $p(M_t=\text{loss})$ is plotted in red. The clear trend is that H-HMM has sparser output and clearly predicts important regions in isolation. Note arrows at *MYC* and *CA9* for comparison to Figure 1. A similar result was seen for chromosome 1 at the *TPFRSF4* and *TP73* loci (see Fig. 6 for results of H-HMM). Notice that BFL-HMM and FL-HMM also predict CNAs at these important genes. However it is quite evident that they both overpredict, making it hard for an investigator to discern biologically relevant CNAs from spurious predictions. From Figure 4, we also see that AF has a peak at the *MYC* locus, but is unable to detect the recurrent

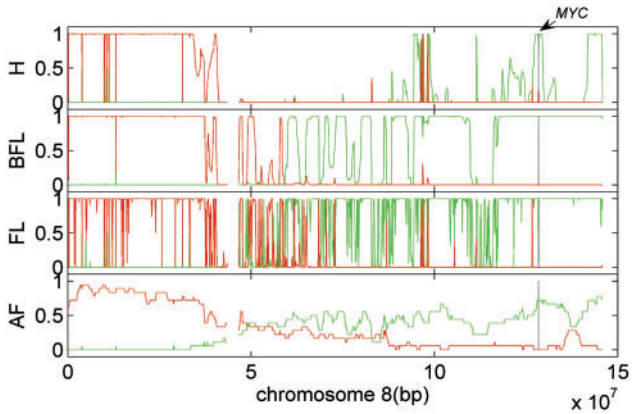

Fig. 4. Output from top to bottom of H-HMM, BFL-HMM, FL-HMM and AF for the NA group, chromosome 8. The x-axis is the chromosomal position and the y-axis is predicted probability. Red plots indicate $p(M_t=L)$ and green plots indicate $p(M_t=G)$. Note the sparse, yet accurate predictions for the H-HMM at the MYC locus (recall Fig. 1 c) and the p-arm loss prediction which recapitulates known results (Garnis *et al*., 2006). The other models either overpredict (BFL-HMM, FL-HMM) or underpredict (AF) the shared aberrations.

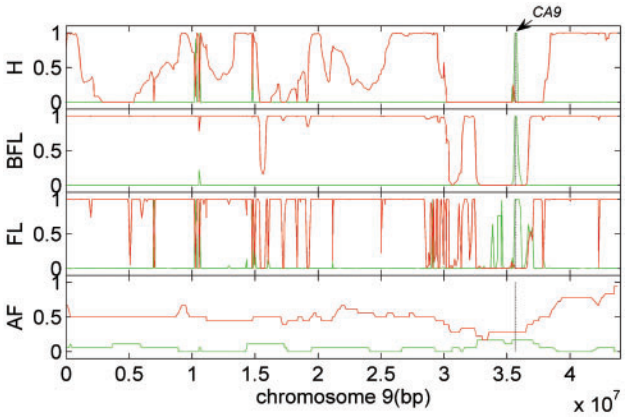

Fig. 5. Output from top to bottom of H-HMM, BFL-HMM, FL-HMM and AF for the NA group for the p-arm of chromosome 9. (see Fig. 4 for axes description). Similar to Fig. 4, notice the sparse, yet accurate predictions for the H-HMM especially at the single probe *CA9* locus (recall Fig. 1b). The AF method does not predict CA9. BFL-HMM and FL-HMM both predict CA9, however, they are over--predicting many other regions not likely to be shared CNAs.

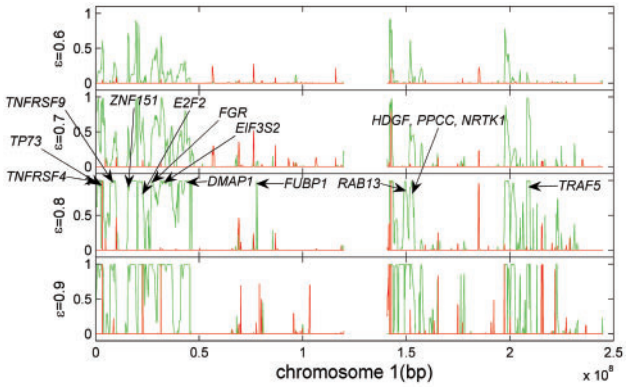

Fig. 6. Output from H-HMM on chromosome 1 for different values of ϵ for SC group. For $\epsilon \geq 0.8$, gain probability 'peaks' correspond to locations of several genes (annotated with arrow) implicated in lung or other cancers.

CNA at *CA9* (Fig. 5) with high frequency. In all three generative models, the signal is clearly predicted.

Considering Figure 4 in more detail, the p-arm (left) has a relatively high frequency of deletion and this is cleanly predicted by all models. In contrast, the centromeric half of the q-arm shows ambiguity in the AF plot. Both BFL-HMM and FL-HMM are unable to resolve the ambiguity as they are forced into a $\{L,N,G\}$ state, while the H-HMM can 'opt-out' of making a consensus prediction at these locations, choosing only to predict a CNA when the data cleanly support one (e.g. *MYC* locus). This illustrates the sparsity of the H-HMM compared to the other models.

The combination of sparsity due to modeling ambiguity and the ability to tune ϵ allows the user to effectively set the FPR of the H-HMM. An example of the value of this is shown in Figure 6, displaying the results for group SC for various values

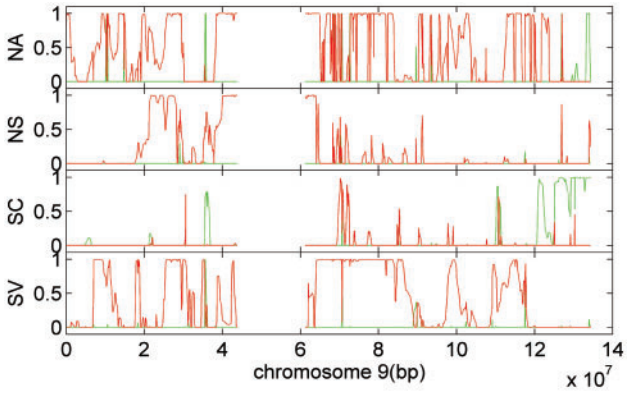

Fig. 7. H-HMM output for chromosome 9 showing discordant patterns among the lung cancer groups (NA, NS, SC, SV).

of ϵ. The sparse output for $\epsilon=0.8$ reveals isolated peaks of high probability at locations of genes (*TNFRSF4*, *TP73*, *TNFRSF9*, *ZNF151*, *E2F2*, *FGR*, *EIF3S2*, *DMAP1*, *FUBP1*, *RAB13*, *HDGF*, *PPCC*, *NTRK1*, *TRAF5*), whose expression is known to be altered in lung or other cancers. For example, *ZNF151* and *E2F2* were found to have copy number induced gene expression changes in Coe *et al.* (2006). Interestingly, the H-HMM predicts the *TP73* region as a narrow loss embedded within the gain region harboring *TNFRSF4* shown in Figure 1c. *TP73* was detected at only 22% frequency in AF and was not detected at all in BFL-HMM. Additional relatively narrow but high probability peaks correspond to the *EIF3S2* locus, which mediates the TGF-β pathway, *FUBP1* a transcriptional activator of *MYC* and the coamplification of *TNFRSF4* and *TRAF5*, which are known interactors and activators in the NF-κB pathway (Kawamata *et al.*, 1998). These results are computational predictions, yet many provide compelling evidence that they merit experimental follow up.

To investigate whether H-HMM recapitulates the results in (Coe *et al.*, 2006), we examined a subset of genes reported to be differentially disrupted in the two major groups, NSCLC and SCLC. These 22 genes are involved in key lung cancer pathways and therefore represent a highly relevant set of markers as a reference to assess our output. The H-HMM predicted shared aberrations in regions harboring 14 of the 22 genes in at least 1 of the subgroups of NSCLC and SCLC. We counted a prediction if $p(M_t=L) > 0.5$, or $p(M_t=G) > 0.5$ for losses and gains, respectively. The predicted genes included *STMN1*, *E2F2*, *SC*, *ZNF151*, *ID2*, *MAPK9*, *EGFR*, *CDK2NA*, *KNTC1*, *HMGB1*, *HSPH1*, *JJAZ1*, *NLK*, *JUNB*, *TIAM1*, *DSCAM*. Five of the regions were detected at $\epsilon \leq 0.7$, eleven at $\epsilon \leq 0.9$ and the remaining regions at $\epsilon=0.95$. This gives us a reasonable estimate for how to calibrate ϵ in order to predict relevant CNAs. The H-HMM did not predict recurrent CNAs harboring the remaining genes *PRDM2*, *SOX11*, *MAP3K4*, *ING1*, *SMAD4*, *CCDC5*, *TCF4*.

We assessed if the H-HMM could determine differences in the profiles of the phenotypic groups (NA, NS, SC, SV), as this was part of the focus of the study of Coe *et al.* (2006) and Garnis *et al.* (2006). Figure 7 shows that the H-HMM produces

very different profiles for chromosomes 9 across the different subgroups. This example chromosome was chosen as it was previously shown to have different patterns of CNA (Coe *et al*., 2006; Garnis *et al*., 2006). Although anecdotal, our qualitative results give us confidence that the H-HMM is predicting biologically relevant recurrent CNAs. Combined with the result that the H-HMM is sparser in its output, we believe the H-HMM has the right characteristics of presenting biologically meaningful results to the investigator while maintaining a low FPR.

6 DISCUSSION AND FUTURE WORK

We developed three novel methods that extend the single sample HMM for aCGH to the multiple sample case in order to infer recurrent CNAs. Our results indicate that the H-HMM, which simultaneously infers discrete labels for the samples and promotes sparsity by modeling ambiguity in the data is quantitatively and qualitatively better than simpler models and standard methods. In an informal qualitative assessment, we showed that the H-HMM produces meaningful biological output when compared to a list of experimentally validated genes. The H-HMM was able to detect previously reported discordant patterns among the lung cancer groups—a key requirement to determine phenotype specific CNA patterns.

6.1 Subgroup discovery

A natural extension to the H-HMM model is to consider the case where the samples in the data come from two or more groups. This suggests unsupervised clustering of the data, a problem recently examined by Liu *et al*. (2006), who compared distance metrics in a hierarchical clustering framework. We will investigate this problem by considering a mixture of master sequences that determine subgroups and simultaneously infer recurrent CNAs specific to each subgroup.

6.2 Copy number variations

A fraction of the recurring CNAs identified by our algorithm can be attributed to segmental copy number variations (CNV) in the human population (Redon *et al*. 2006). These variations are not the consequence of somatic alteration in tumor DNA but are naturally occurring copy number states. This suggests they should be filtered out of the output, but it is important to consider the potential contribution of CNVs to disease susceptibility, as such segmental variations overlap with genes associated with phenotypes in humans (Wong *et al*., 2007). The frequent occurrence of a given CNV in a cohort of cancer patients relative to healthy individuals would raise the possibility of an association between the copy number state and cancer susceptibility. The H-HMM will facilitate the detection of such occurrences, and the identification of susceptibility genes through CNV status will become feasible as databases of aCGH profiles of tumors expand.

6.3 Epigenomic arrays

In addition to gene dosage alterations in CNAs, epigenetic alterations such as changes in DNA methylation status of gene may result in aberrant silencing or inappropriate activation of genes in cancer. Recent development of microarray-based methods for whole genome analysis of methylation status (or methylome profiling) has enabled a new approach to cancer gene discovery (Weber *et al*., 2005). Future adaptation of the statistical strategy described in this article will expedite the detection of recurring epigenetic alterations, and facilitate the integration of genetic and epigenetic data.

ACKNOWLEDGEMENTS

This work is supported by Genome BC/Genome Canada. S.P.S. is supported by the Michael Smith Foundation for Health Research. We thank Bradley Coe (BC Cancer Research Centre) for a critical review of this manuscript.

Conflict of Interest: none declared.

REFERENCES

Alarcon-Vargas,D. *et al*. (2000) p73 transcriptional activity increases upon cooperation between its spliced forms. *Oncogene*, **19**, 831–835.

Baldwin,C. *et al*. (2005) Multiple microalterations detected at high frequency in oral cancer. *Cancer Res*., **65**, 7561–7567.

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

Coe,B.P. *et al*. (2006) Differential disruption of cell cycle pathways in small cell and non-small cell lung cancer. *Br. J. Cancer*, **94**, 1927–1935.

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.

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.

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

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

Garnis,C. *et al*. (2006) High resolution analysis of non-small cell lung cancer cell lines by whole genome tiling path array CGH. *Int. J. Cancer*., **118**, 1556–1564.

Gelman,A. *et al*. (2004) *Bayesian Data Analysis*. 2nd edition edn. Chapman and Hall.

Ghahramani,Z. *et al*. (1997) Factorial hidden Markov models. *Mach. Learn*., **29**, 245–273.

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

Jong,K. *et al*. (2004) Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics*, **20**, 3636–3637.

Kawamata,S. *et al*. (1998) Activation of OX40 signal transduction pathways leads to tumor necrosis factor receptor-associated factor (TRAF) 2- and TRAF5-mediated NF-kappaB activation. *J. Biol. Chem*., **273**, 5808–5814.

Kim,S.J. *et al*. (2004) Carbonic anhydrase IX in early-stage non-small cell lung cancer. *Clin. Cancer Res*., **10**, 7925–7933.

Lipson,D. *et al*. (2006) Efficient calculation of interval scores for DNA copy number data analysis. *J. Comput. Biol*., **13**, 215–228.

Liu,J. *et al*. (2006) Distance-based clustering of CGH data. *Bioinformatics*, **22**, 1971–1978.

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

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.

Redon,R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

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

Scott,S. (2002) Bayesian methods for hidden Markov models: recursive computing in the 21st century. *J. Am. Stat. Assoc.*

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

Swinson,D.E. *et al.* (2003) Carbonic anhydrase IX expression, a novel surrogate marker of tumor hypoxia, is associated with a poor prognosis in non-small-cell lung cancer. *J. Clin. Oncol.*, **21**, 473–482.

Veltman,J.A. *et al.* (2006) Diagnostic genome profiling: unbiased whole genome or targeted analysis? *J. Mol. Diagn.*, **8**, 534–537.

Weber,M. *et al.* (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nat. Genet.*, **37**, 853–862.

Wong,K. K. *et al.* (2007) A comprehensive analysis of common copy-number variations in the human genome. *Am. J. Hum. Genet.*, **80**, 91–104.