BIOINFORMATICS

Vol. 22 no. 14 2006, pages e431–e439
doi:10.1093/bioinformatics/btl238

Integrating copy number polymorphisms into array CGH analysis using a robust HMM

Sohrab P. Shah[1,*], Xiang Xuan[1], Ron J. DeLeeuw[2], Mehrnoush Khojasteh[2], Wan L. Lam[2], Raymond 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 Research Centre, 675 West 10th Avenue Vancouver, BC V5Z 1L3 Canada

ABSTRACT

Motivation: Array comparative genomic hybridization (aCGH) is a pervasive technique used to identify chromosomal aberrations in human diseases, including cancer. Aberrations are defined as regions of increased or decreased DNA copy number, relative to a normal sample. Accurately identifying the locations of these aberrations has many important medical applications. Unfortunately, the observed copy number changes are often corrupted by various sources of noise, making the boundaries hard to detect. One popular current technique uses hidden Markov models (HMMs) to divide the signal into regions of constant copy number called segments; a subsequent classification phase labels each segment as a gain, a loss or neutral. Unfortunately, standard HMMs are sensitive to outliers, causing over-segmentation, where segments erroneously span very short regions.

Results: We propose a simple modification that makes the HMM robust to such outliers. More importantly, this modification allows us to exploit prior knowledge about the likely location of "outliers", which are often due to copy number polymorphisms (CNPs). By "explaining away" these outliers with prior knowledge about the locations of CNPs, we can focus attention on the more clinically relevant aberrated regions. We show significant improvements over the current state of the art technique (DNAcopy with MergeLevels) on previously published data from mantle cell lymphoma cell lines, and on published benchmark synthetic data augmented with outliers.

Availability: Source code written in Matlab is available from http://www.cs.ubc.ca/∼sshah/acgh.

Contact: sshah@cs.ubc.ca

1 INTRODUCTION

Array comparative genomic hybridization (aCGH) is a high-throughput genetic technique to measure DNA copy number changes in a disease sample compared to a normal sample [20]. Chromosomal aberrations that exhibit DNA copy number changes are indicative of numerous diseases including cancer and mental retardation. Identifying such aberrations can help to locate diagnostically important regions in the genome, that harbour important genes. For example, oncogenes or tumour suppressor genes contained in aberrated regions could in turn exhibit differential expression due to the copy number changes in the DNA. Application of aCGH is widespread in molecular analysis of cancer and holds great promise as a technique to identify clinically relevant diagnostic biomarkers.

The aCGH technique is based on spotting clones that span a discrete region in the human genome on an array. The size and number of clones vary depending on the technological platform and the desired resolution: see Pinkel and Albertson [20] for a review. In this paper, we use aCGH data from eight mantle cell lymphoma (MCL) cell lines (see deLeeuw *et al*. [4] for details) generated using sub-megabase resolution tiling arrays (SMRT) [13]. We use the midpoint of the clone along the chromosome to denote its location. The output of all aCGH platforms is represented as a $\log_2$ ratio of the reference and tumour fluorescence intensities for every clone in the array. The $\log_2$ ratios are expected to be proportional to copy numbers. In a neutral state, one would expect to see $\log_2(2/2) = 0$; with one copy lost, one would expect to see $\log_2(1/2) = -1$; with one gain $\log_2(3/2) = 1.58$, etc. The goal of analysis techniques is to detect contiguous regions, that are expected to share the same mean $\log_2$ ratio. We call these regions segments. The identification of segments is called ''segmentation''. Once segments are identified, each segment is labeled as a loss, neutral or gain (sometimes it is useful to distinguish gains of 1 copy from gains of more than 1). This latter task is called ''classification''.

In reality, segmentation and classification of the data are much more difficult than the above description suggests. Figure 3 (A) shows a typical plot of aCGH data for chromosome 1 from MCL cell line HBL-2 (see Section 3.1 for more details on MCL). The yellow squares represent clones that are found in a region of loss identified manually by an expert [4]. Similarly, blue circles represent clones in a region of gain. The figure demonstrates that although copy number changes in DNA is a theoretically discrete process, the intensity ratios for aCGH do not produce a clean piecewise constant signal. Also note that aberrated regions tend to span contiguous sets of clones along a chromosome. This suggests that any analysis technique should exploit such spatial correlation.

In Figure 3 (A), we also depict 'outlying' clones (detected by eye) with light blue stars. Treating such points as inliers can significantly affect the remaining points, by causing over-segmentation, resulting

[*]To whom correspondence should be addressed.

in segments that span only a single clone, for example. There are several possible causes of such outliers. The first is some kind of measurement noise, or mislabeling (sometimes the locations of clones is mis-recorded). Second, there is the possibility that the single clone outliers correspond to known locations of copy number polymorphisms (CNPs). Finally, they could truly represent aberrated regions. In our experience, this is rare.

The full impact of CNPs on aCGH analysis is not yet known, however indications from two recent large scale studies by Sebat *et al.* [22] and Iafrate *et al.* [12] measuring background frequencies of copy number variations in the normal human population have revealed hundreds of loci in the genome that are polymorphic in copy number. Buckley *et al.* [2] suggest that the results produced by these two studies represent the ''tip of the CNP iceberg''. For example Sebat *et al.* report a CNP at a gene involved in food intake, suggesting a differential propensity for obesity. They also report CNPs at loci related to neurological development and at loci implicated in leukemia and breast cancer drug resistance [22]. These latter examples indicate that for cancer studies, the 'baseline' copy number should be considered when assessing aberrations. We anticipate that the impact of CNPs will be greater on high-resolution arrays and/or full genome coverage arrays, as they are intended to reveal all aberrations in a sample and will detect a larger number of CNPs. Note that the MCL data is both high-resolution and full coverage and therefore is likely to contain CNPs.

1.1 Our contributions

In this paper, we introduce a joint classification and segmentation method that is designed to handle outliers and integrate CNP knowledge into the analysis. Our method extends the standard HMM framework, outlined in Scott [21] and proposed for aCGH in Guha *et al.* [9]. The basic idea is to replace the Gaussian observation model with a mixture of Gaussians; one mixture component represents the $\log_2$ ratio we would expect from the given state (loss, neutral or gain); the other mixture component represents the $\log_2$ ratio we would expect from an outlier. This simple change makes the model much more robust.

More significantly, we can incorporate knowledge about CNPs into the mixing weights of the mixture model. That is, we can set the prior probability of using the outlier component at location i to the known frequency of CNPs at location i, if i overlaps with a known CNP location; otherwise we set it to the general background outlier probability (which is estimated from data). We explain our model in more detail in Section 2.1.

Several authors (e.g., [9,21]) propose estimating the parameters of the HMM using MCMC (Markov chain Monte Carlo) techniques, as opposed to the more common EM (expectation maximization) algorithm. The advantage of MCMC is that it provides full posterior estimates over the parameters, rather than just point estimates, thus properly modeling uncertainty (see e.g., [8] for an introduction to MCMC and Bayesian data modeling). MCMC also partly mitigates problems with local minima that EM is well known to suffer from. It also turns out to be simpler to exploit informative prior constraints in a sampling framework than in an optimization framework. We explain how to perform efficient MCMC in Section 2.4.

We first evaluate performance of our model on real data representing aCGH profiles from eight MCL cell lines published in deLeeuw *et al.* [4] in a study aimed at identifying important signature regions in non-Hodgkin's lymphoma. This data set contains ground truth annotation of regions of gain and loss, some of which are recurrent across cell lines. In addition some of these aberrations have been validated in the laboratory. Using this rich data set, we were able to assess performance quantitatively using standard performance metrics. We compare our method to DNAcopy+MergeLevels (using default parameters), which has been shown in two previous comparative studies [24,16] to be a leading current method. Henceforth we will refer to this method as MergeLevels. Having established that our method is better than current techniques, we then validate our findings on an additional synthetic data set, which we believe to be 'harder' than the real data. The advantages of using synthetic data are two-fold. First, the ground truth locations of the aberrations are known. Second, we can control the difficulty of the problem. We used data published in Willenbrock and Fridlyand [24]. This data is considerably harder (but more realistic) than other synthetic datasets used in earlier papers. We make the Willenbrock and Fridlyand data even harder by adding outliers, to check the robustness of our method and to validate results obtained using MCL data. Our results are in Section 3, which we discuss in Section 4.

1.2 Related work

A recent survey paper by Lai *et al.* [16] describes and evaluates eleven algorithms for aCGH data analysis. We can loosely group these methods into three main approaches: smoothing, segmentation, and combined segmentation and classification. Smoothing approaches such as Quantreg, developed by Eilers and Menezes [5], and the wavelet approach of Hsu *et al.* [10], attempt to fit a curve to the data, while handling abrupt changes. Smoothing methods generally filter the data using a fixed size window, and therefore will be unable to detect outliers or CNPs that span a single clone. In addition, they are primarily designed as a visual aid to interpret the data and do not accomplish the main objective of automatically identifying aberrated clones.

As mentioned previously, segmentation methods identify contiguous sets of clones (segments) that share the same mean $\log_2$ ratio. The output of the segmentation methods usually consists of the boundaries and means of the segments. The clones within a segment are assumed to share the same copy number. We refer to the boundaries of segments as breakpoints. Examples of segmentation algorithms include DNACopy [18], which is based on a recursive circular binary segmentation algorithm; CGHSeg [19] which uses a penalised likelihood model to determine breakpoints; aCGH-Smooth [14], which uses a genetic algorithm to find breakpoints; and the GLAD method of Hupe *et al.* [11], which includes a median absolute deviation model to explicitly treat outliers as separate from its surrounding segment. In Lai's comparison, CGHSeg and DNA-Copy are consistently the best. Willenbrock and Fridlyand [24] compared performance of DNACopy and GLAD and report better performance with DNACopy. We therefore use DNAcopy as our baseline model. Note that Lai *et al.* [16] determined that as the noise level in the data increases, all segmentation methods—including DNAcopy, show less than satisfactory results.

A general limitation of segmentation is that the output needs to be further analysed in order to infer which segments are aberrated regions, i.e., to ''call'' the gains and losses. Methods such as GLADMerge [11] and MergeLevels [24] perform this post-processing task by merging together segments with ''similar'' mean levels, and then classifying them. However, as noted by

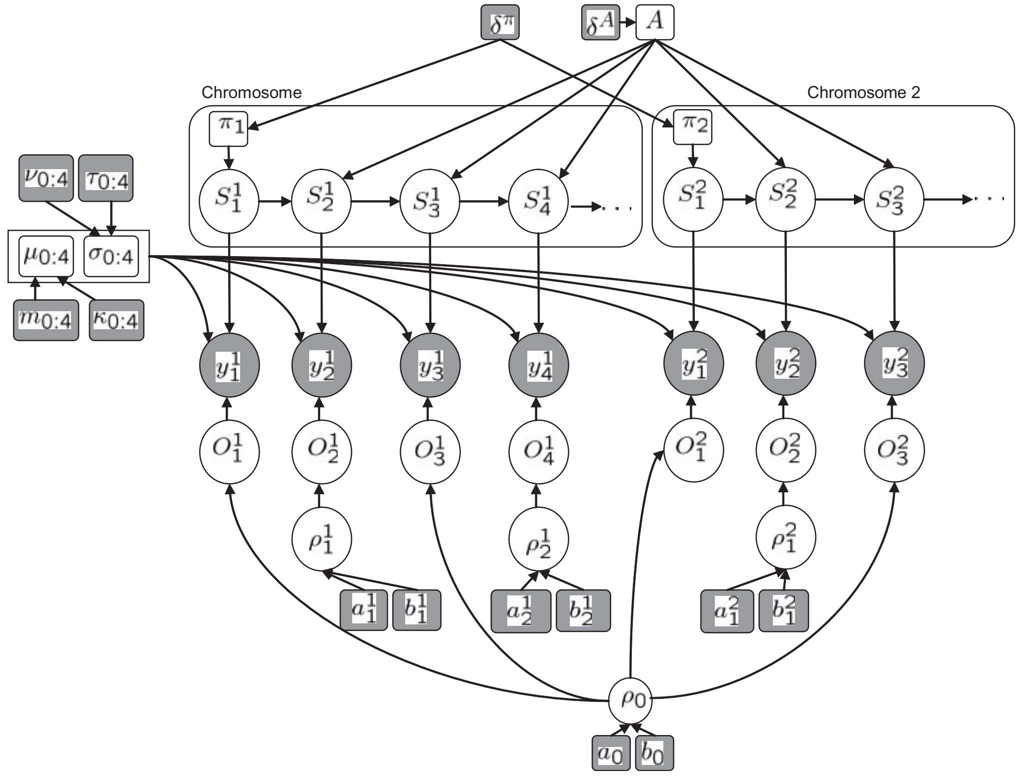


Fig. 1. Our model represented as a Bayesian network. Square nodes are parameters, round nodes are random variables. Shaded nodes are observed (known), unshaded nodes are hidden (unknown). We depict the start of 2 chromosomes (indicated by large rounded rectangles). Let c denote the chromosome, i represent the location along the chromosome and k represent the kth LSP on the chromosome; π_c is the initial state distribution of chromosome c; δ^{π} are hyperparameters for the π_c's; S_i^c is the state; y_i^c is the observation ($\log_2$ ratio); O_i^c indicates if this is an outlier or not; $\mu_{1:4}$ and $\sigma_{1:4}$ are the means and variances of states 1 to 4; μ_0 and σ_0 is the mean and variance of the outlier state; ρ_k^c is the probability of outlier for LSP locations; ρ_0 is the general background outlier probability; A is the Markov chain transition matrix; δ are the hyperparameters for A. For state j, m_j, τ_j are hyperparameters for μ_j; α_j, β_j are hyperparameters for σ_j; a_k^c, b_k^c are hyperparameters for ρ_k^c and are determined by LSPs; Hyper-parameters are shown shaded since they must be set by the user. In this example, we have assumed that locations 2 and 4 on chromosome 1 and location 2 on chromosome 2 correspond to known CNPs; other locations use the background outlier probability ρ_0. Hence the prior on O_1^1, O_3^1, O_1^2 and O_3^2 are all the same and equal to ρ_0.

Engler *et al*. [6] and Willenbrock and Fridlyand [24], it is much better to perform the segmentation and classification simultaneously, since the class labels can help with the segmentation as well as vice versa.

An obvious way to perform simultaneous segmentation and classification is to use an HMM. The first approach to do this was by Fridyland *et al*. [7]. However, in their approach, the states of the HMM do not have any intrinsic meaning (they are just indices to represent a discrete number of mean levels, typically $K = 5$). Hence post-processing was necessary to determine the labels. Guha *et al*. [9] modify this to use a ''supervised'' 4-state HMM, where the states are defined to mean loss, neutral, one-gain or multiple-gain. The advantage of this is two-fold: first, it is easy to perform simultaneous segmentation *and classification* using the Viterbi algorithm; secondly, we can impose informative priors on the parameters, since they now have biological meaning. This paper extends the model by adding robustness to outliers and location-specific priors (LSPs), which can be used to encode CNPs.

In addition to the work mentioned above, two recent papers have explored some interesting variations. Broet and Richardson [1] propose using a latent 1D Gaussian random field, as opposed to a latent 1D discrete random field (i.e., an HMM), to model spatial correlation between levels. However, this does not solve the classification problem. Engler *et al.* [6] introduce spatial dependence by breaking the data into overlapping triples, and then using a hierarchical random effects model. Unfortunately, because the triples are overlapping, the data is over-counted, so optimizing the likelihood turns out to be intractable. Instead, they compute a local maximum of the pseudolikelihood. We also use a hierarchical Bayesian model, but we are able to compute posterior estimates using an exact likelihood function.

2 METHODS

2.1 A mixture model HMM that considers outliers

Our model, sketched in Figure 1, is similar to the 4-state HMM in Guha *et al.* [9], where the states represent loss, neutral, one-gain and multiple-gain. (We also tried a 3-state model, where we combined the gain states, but results were not as good.) The main difference from Guha is that in our model the observation density is a mixture of 2 Gaussians, one representing inlier (clones belonging to one of the states) and the other representing outlier. We introduce binary indicator variables $O_i \in \{0, 1\}$ where $O_i = 1$ means

Table 1. User settable hyper parameters for our model, along with the values we used for the Willenbrock synthetic data and the real MCL data. f_i is the frequency of a known CNP at position i. In the synthetic data, we set this to 0.001. To help interpret these numbers, recall that the mean of a *beta*(a,b) random variable is $a/(a+b)$, and the mean of a $Ga(\alpha,\beta)$ random variable is α/β. In particular, this means $\sigma_1^2 = 0.1$ for synthetic and 0.07 for MCL, $\sigma_2^2 = 0.01$ for synthetic and 0.05 for MCL, etc

Parameter	Description	Synthetic	MCL
δ	Dirichlet prior on transition matrix A	1, 1, 1, 1	1, 1, 1, 1
$\alpha_{1:4}$	shape of gamma prior on inverse variances σ^{-2}	10, 100, 5, 5	15, 20, 10, 10
$\beta_{1:4}$	scale of gamma prior on inverse variances σ^{-2}	1, 1, 1, 1	1, 1, 1, 1
$\mathrm{m}_{1:4}$	prior mean on means μ	−0.1, 0, 0.58, 1	−0.4, 0, 0.3, 0.5
$\tau_{1:4}$	prior variance on means μ	0.5, 0.001, 1, 1	0.2, 0.1, 0.2, 0.2
a_0	a for beta prior for ρ_0 (prob of outlier)	0.01	0.00001
b_0	b for beta prior for ρ_0 (prob of outlier)	0.99	0.99999
a_i	a for beta prior for ρ_i (prob of outlier at CNP i)	0.001	f_i
b_i	b for beta prior for ρ_i (prob of outlier at CNP i)	0.999	$1 - f_i$

location i is an outlier, and $O_i = 0$ means it is an inlier. We model the outlier distribution with a Gaussian, μ_0, σ_0^2.

Using the mixture of Gaussians, the class-conditional density becomes

$$p(y_i \mid O_i, S_i = s) = \begin{cases} Gauss(y_i \mid \mu_0, \sigma_0) & \text{if } O_i = 1 \\ Gauss(y_i \mid \mu_s, \sigma_s) & \text{if } O_i = 0 \end{cases} \quad (1)$$

where y_i is the $\log_2$ ratio for clone i where the clones are ordered by their physical location on a chromosome. $S_i = s$ is the state label at i, where s is a discrete random variable $\in\{1,2,3,4\}$ with 1 corresponding to the loss state; 2 to the neutral state; 3 to the one-gain state; and 4 to the multiple-gain state. The unobserved sequence of states is governed by Markovian dynamics encoded in the transition matrix A. The transition matrix therefore models the spatial correlation expected to occur in the data. The clones labeled with a given state s are generated from a a common Gaussian distribution with μ_s and σ_s. The initial state distribution for $i = 1$ is multinomial random variable π that models the probability of starting in each state. In the observation density of our model, O_i acts like a ''switching parent'' variable, which selects between the outlier parameters μ_0, σ_0 or the inlier parameters, μ_s, σ_s. The O_i variables are modeled as conditionally independent. Hence there are no Markovian dynamics on the outliers. This allows the model to make temporary ''excursions'' to the outlier state, without incurring any ''penalty'' implicitly encoded by the state transition matrix.

Modeling the outliers as conditionally independent also allows us to encode CNPs. For each location that is known to be a CNP, we have an outlier probability, $\rho_i = p(O_i = 1)$; for all other locations, we have the ''background'' outlier probability, ρ_0.

2.2 Parameter estimation using 'pooling' across chromosomes

In addition to the outlier extension, we extend our model by estimating some of the parameters using pooled data across all chromosomes in the sample. Parameters A, μ, σ, ρ_0 are estimated by pooling. This assumes that the posterior distributions of these parameters are expected to be consistent across chromosomes, and therefore pooling is advantageous as their estimates are guided by more data. Moreover, the algorithm is more likely to 'visit' all the states by pooling the data, resulting in more robust estimates of the mean and variance of each state. However, not all the parameters can be estimated in this way. Sampling of the states S must be estimated on each chromosome separately as there is no physical interpretation for a state dependency between location 1 on chromosome c and the terminal location on chromosome $c - 1$. π must also be estimated independently for each chromosome since the telomeric regions of the chromosomes can have gains, losses or remain neutral and these initial states are not expected to be consistent across chromosomes. The model, showing pooling across chromosomes and the outlier parameters, is depicted as a Bayesian network in Figure 1. Since the figure shows pooling across chromosomes, (S,y,O,π and ρ) are indexed by both chromosome and location. The chromosome index was omitted above for notational clarity.

An obvious extension of pooling data across chromosomes is to pool data across samples, such as in Engler *et al*. [6]. However, due to numerous factors that are sample-specific such as ploidy of the tumour genome and proportion of tumour cells in the sample[24], we do not assume that mean levels of copy number change will be consistent across samples. Therefore we do not estimate mean levels of the states across samples. However, we suggest that jointly considering samples has considerable value for goals other than classification. Indeed, multiple aCGH samples of the same cancer subtype have something in common—this is precisely what scientists hope to discover! Presumably, multiple samples of the same cancer subtype will exhibit commonalities such as minimally overlapping aberrations (see [4] for examples in MCL cell lines) and locations of breakpoints. Detecting such features is the subject of future work, and for now, we limit our attention to modeling samples separately.

2.3 Priors

We use standard conjugate priors (see e.g., [8]) for all the parameters, as follows:

$$p(\mu_s \mid \sigma_s) = Gauss(m_s, \tau_s \sigma_s) \quad (2)$$

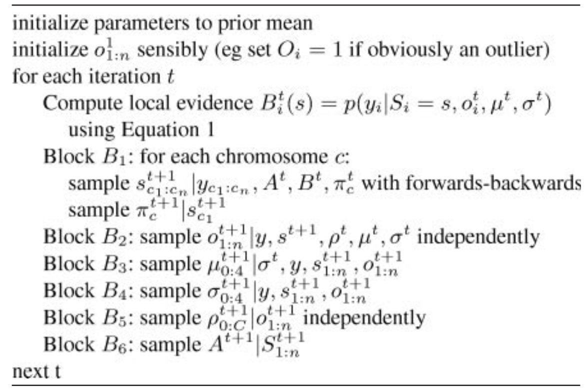
$$p(\sigma_s^{-2}) = Ga(\alpha_s, \beta_s) \quad (3)$$

$$p(A) = Dir(\delta) \quad (4)$$

$$p(\rho_i) = Beta(a_i, b_i) \quad (5)$$

As is apparent, these priors themselves have parameters, called hyper-parameters. We set these by hand. Specifically, we use a small fraction of validation data in order to estimate (by eye) the typical mean and variance of the loss, neutral and gain states. A more rigorous Bayesian approach would be to extend the hierarchy even further, and add priors to the hyperparameters. Previous work has shown that 3 levels of hierarchy (parameters, hyper-parameters, and hyper-hyper-parameters) is usually sufficient to obtain robustness to (hyper-hyper-)parameter settings. We plan to investigate this in the future. A summary of all the user-settable parameters is shown in Table 1.

Prior knowledge about CNPs is encoded as follows. Locations $i \in P$ which are known to come from CNPs get an adjustable parameter ρ_i which reflects the probability of outlier at that location. The parameters of the (*Beta*) prior on ρ_i is set so that the expected value of ρ_i is equal to the frequency of polymorphisms at that location in the population. Locations $i \notin P$, which are



```
initialize parameters to prior mean
initialize $o^1_{1:n}$ sensibly (eg set $O_i = 1$ if obviously an outlier)
for each iteration $t$
    Compute local evidence $B^t_i(s) = p(y_i|S_i = s, o^t_i, \mu^t, \sigma^t)$
        using Equation 1
    Block $B_1$: for each chromosome $c$:
        sample $s^{t+1}_{c_1:c_n}|y_{c_1:c_n}, A^t, B^t, \pi^t_c$ with forwards-backwards
        sample $\pi^{t+1}_c|s^{t+1}_{c_1}$
    Block $B_2$: sample $o^{t+1}_{1:n}|y, s^{t+1}, \rho^t, \mu^t, \sigma^t$ independently
    Block $B_3$: sample $\mu^{t+1}_{0:4}|\sigma^t, y, s^{t+1}_{1:n}, o^{t+1}_{1:n}$
    Block $B_4$: sample $\sigma^{t+1}_{0:4}|y, s^{t+1}_{1:n}, o^{t+1}_{1:n}$
    Block $B_5$: sample $\rho^{t+1}_{0:C}|o^{t+1}_{1:n}$ independently
    Block $B_6$: sample $A^{t+1}|S^{t+1}_{1:n}$
next t
```

Fig. 2. Pseudo code for the *pooled* algorithm. c_1 and c_n indicate the initial and terminal positions on chromosome c. n indicates the total number of clones in the sample.

not known to come from CNPs, share the same parameter ρ_0, which represents the background probability of outlier. The (*Beta*) prior on ρ_0 is set so that the expected value of ρ_0 is equal to the expected fraction of outliers, which we estimate by eye on a per dataset basis (once for synthetic data, once for MCL). We will let $C = |P|$ be the number of CNP locations, so ρ is a vector of length $C + 1$.

In order to ensure the model is identifiable (i.e., to avoid label switching), we enforce the following constraint on the mean parameters: $\mu_1 < \mu_2 < \mu_3 < \mu_4$, where the states represent loss, neutral, one-gain and multiple-gain. (We do this using a truncated Gaussian prior.) We impose a similar constraint on the state variances: $\sigma_{3,4} \geq \sigma_1 \geq \sigma_2$, which means that the gain states have higher variance than the loss state, which has higher variance than the neutral state (an empirical fact about most aCGH data). (We encode this using a truncated Gamma prior.) Note that handling truncated priors in EM (for MAP estimation) is harder than with MCMC, since it would require constrained optimization methods in the M step. Indeed, EM is usually only used to fit unidentifiable HMMs.

2.4 Algorithm

The algorithm is sketched in Figure 2. The output of the algorithm is the following: estimates of the states $\gamma_i(s) = p(S_i = s\,|\,y_{1:n})$ and outlier probabilities $\omega_i(o) = p(O_i = o\,|\,y_{1:n})$, as well as estimates of the parameters, $p(\theta\,|\,y_{1:n})$ where $\theta = (\pi, A, \mu, \pm\sigma, \rho)$. We use an MCMC algorithm called block Gibbs sampling to infer these quantities. The key to making this efficient is to use the forwards-filtering backwards-sampling algorithm for HMMs [21]. This is very similar to the more familiar forwards-backwards and Viterbi algorithms, except we sample state sequences from their posterior, rather than computing the most probable sequence or marginal state probabilities. Conditioned on knowing the states, it is easy to update the parameters of the model. The same intuition is used in EM, but the advantage of sampling is that we can model uncertainty in the parameters more easily.

To evaluate the effect of pooling across chromosomes, we implemented our model in two modes, one which models each chromosome of each sample independently (*single*), and the other which estimates the parameters of the model by summarizing over chromosomes in each sample (*pooled*), as shown in Figure 1. The algorithm for *pooled* mode is shown in Figure 2. To reduce the *pooled* model to the *single* model, we consider a single chromosome at a time and assign each chromosome its own set of private parameters A, μ, σ and ρ_0. We assess the relative performance of these two implementations in Section 3.

The running time is $O(NT)$ where N is the number of clones in the input and $T \sim 100$ is the number of MCMC iterations needed to obtain convergence (which we assess informally by monitoring quantities of interest by eye). The method is entirely standard except for the update of ρ. We update the ρ_i parameters (based on the sampled value of O_i) for those locations $i \in P$ known to correspond to CNPs; for all other locations, we update ρ_0 using the sufficient statistic $\sum_{i \notin P} O_i$. In *pooled* mode, the forward-filtering, backwards-sampling step (which samples a state sequence) is performed on each chromosome separately, as imposing dependencies on the terminal clone from one chromosome and the initial clone of the next chromosome is non-sensical. In both algorithm variants, we parameterize π separately with its own Dirichlet distribution; this allows the use of Gibbs sampling to update the hyperparameters by simple counting. In contrast, Guha *s*olve for π using the stationary distribution of A, which requires a Metropolis-Hastings step [9]. Currently we estimate $\hat{\gamma}_i(s) = p(S_i = s\,|\,y_{1:n})$ by counting the number of inlier samples for which $S_i = s$.

2.5 Evaluation methods

We evaluated our algorithm by calculating precision and recall for aberrations (gains and losses grouped together). Given a ground truth labeling and a predicted labeling of the clones (obtained by taking the max $p(S_i\,|\,y)$ probability), let *ntp* be the number of true positives (correctly predicted aberrations), let *nt* be the number of true aberrations, and let *np* be the number of predicted aberrations. Recall is defined as $\frac{ntp}{nt}$, meaning the proportion of true aberrations detected by the algorithm. Precision is defined as $\frac{ntp}{np}$, meaning the proportion of predicted aberrations that are true. By varying the threshold on the probabilities, we can vary the trade-off between precision and recall. To summarize the precision-recall curve in one number, we use the F-measure, which is the geometric mean:

$$F = 2 \times \frac{precision \times recall}{precision + recall} \qquad (6)$$

To summarize accuracy results over many samples or chromosomes, we use distributions of F-measures.

We now explain how we modify the above method to handle outliers. We first compute the posterior probability of outlier for each clone, $p(O_i = 1\,|\,y)$. We then rank these probabilities and take the top p_o% of them; finally, we select those whose absolute probability is above a threshold t_o. We then remove all those clones, which are deemed outliers, and compute precision-recall on the remaining locations in the usual way. Note that these parameters are not part of the algorithm. They are only used in the evaluation process. We use $p_o = 10\%$ and $t_o = 0.01$.

3 RESULTS

To systematically test our approach, we ran three variants of our algorithm on each data set:

- The baseline HMM (Base-HMM) which clamps the probability of outlier at each location to 0, $p(O_i = 1) = 0.0$. This reduces the model to an HMM with no outlier processing ability, as in [9].
- The robust HMM (Rob-HMM), which uses $C = 0$ CNPs but updates the global outlier probability $p(\rho_0\,|\,y)$ given data from all locations.
- The robust HMM augmented with location specific prior (LSP) knowledge (LSP-HMM). In particular, we allow all locations $i \in P$ to have their own prior probability of outlier, ρ_i.

For each of these variants, we also ran the algorithm in *single* and *pooled* mode. We also ran MergeLevels, considered to be the current best method.

3.1 Mantle cell lymphoma cell line data

To illustrate the performance of our method on real data, we used a set of 8 MCL cell lines (Granta-519, HBL-2, NCEB-1, Rec-1, SP49, UPN-1, Z138C and JVM-2) whose aCGH profiles were manually

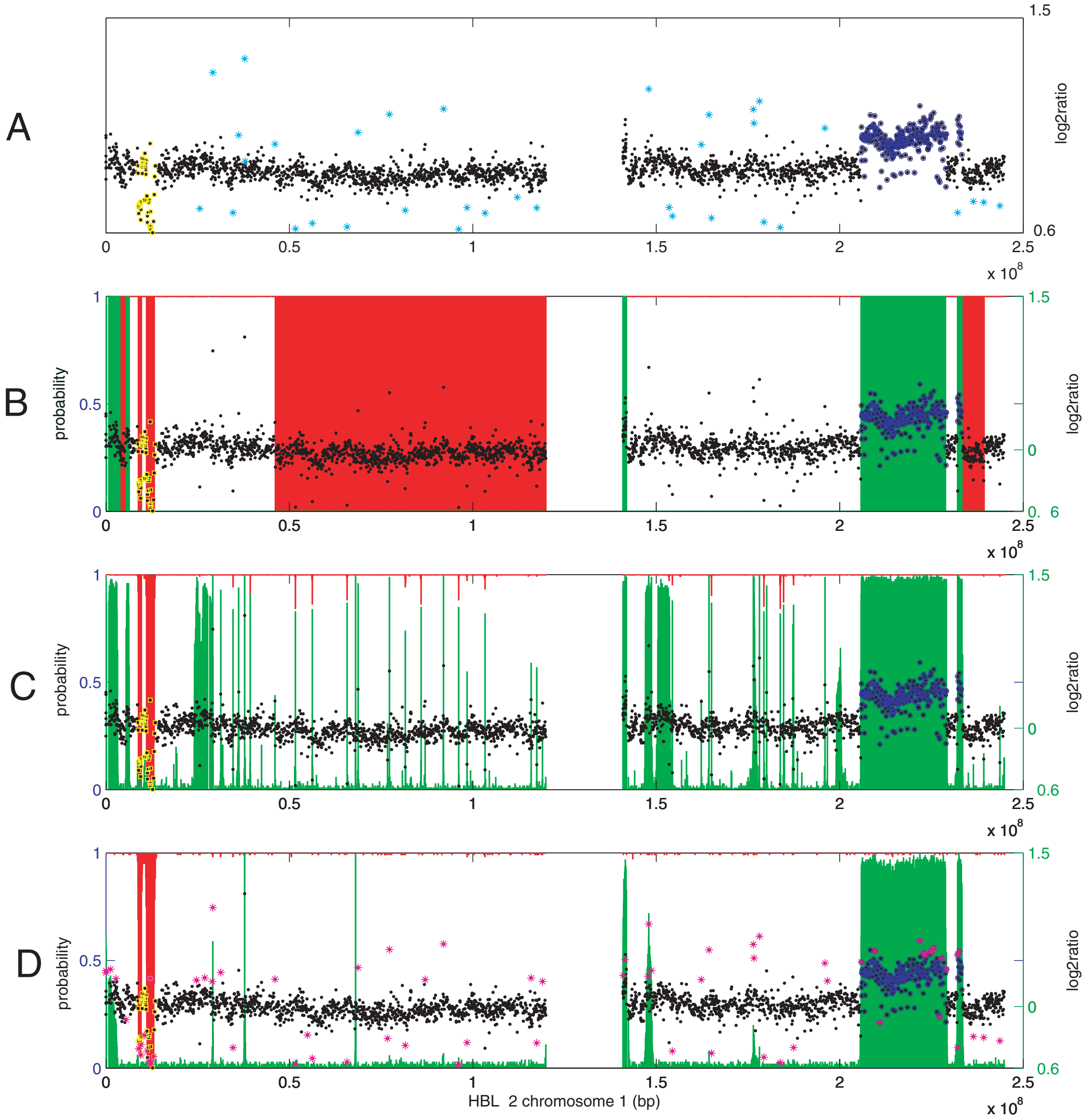


Fig. 3. Array CGH profile for chromosome 1 of the MCL cell line HBL2. The x axis for all panels indicate position in nucleotides (bp) along the chromosome. Panel **A** shows the $\log_2$ ratios (right axis) plotted against the position of each clone on the chromosome. The yellow squares indicate clones contained in a region labeled as a loss by an expert. The blue circles similarly indicate clones that are in a gain region. Clones marked with light blue stars indicate outliers. Panel **B** shows the predicted gains (vertical green bars) and losses (vertical red bars) output by MergeLevels. Note that while predicting all the ground truth aberrations correctly, MergeLevels predicts six additional aberrated regions, including two large loss regions near the ends of each chromosome arm. MergeLevels does not produce probabilistic output so we fix predicted aberrations at probability = 1 and all other locations at probability = 0 for comparative purposes. Panel **C** shows the output of the Base-HMM. The green curve indicates the marginal probability of gain at each location, the red curve indicates 1 minus the marginal probability of loss at each location (left axis). There are numerous false positive predictions with the Base-HMM, many of which are caused by single clone outliers. Panel **D** shows the output of the LSP-HMM (*pooled* mode) with green and red the same quantities as in panel **C** and purple stars indicating the set of predicted outliers. The LSP-HMM predicts all ground truth aberrations correctly and there are much fewer clones falsely predicted as aberrated compared to both MergeLevels and the Base-HMM. Note that the locations of the predicted outliers overlap many of the falsely predicted single clone aberrations by the Base-HMM. Notably, there are several outliers predicted in the leftmost loss region on the p-arm of the chromosome. These correspond to CNPs and therefore alert the user that the significance of this region of loss should be carefully considered.

analysed and published by deLeeuw *et al.* [4]. The data was generated using the Sub-Megabase Resolution Tiling (SMRT) arrays [13] using a set of approximately 32,000 clones that cover the human genome. We normalised the data published in deLeeuw *et al.* [4] according to the stepwise method described in Khojasteh *et al.* [15]. The normalized data was then manually labeled by identifying contiguous regions of gains and losses and then labeling the clones contained in the regions as gains or losses. This ‘ground truth’ labeling allowed us to test our model on high resolution real data, likely to contain CNPs. Only the autosomes (chromosomes 1

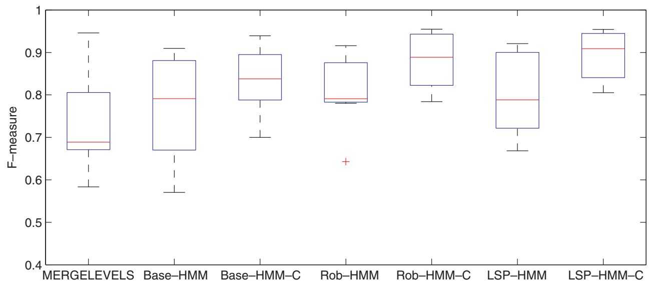


Fig. 4. Distributions of F-measures over eight MCL cell lines for MergeLevels, the Base-HMM, the Rob-HMM and the LSP-HMM with the CNP location prior in *single* and *pooled* mode (labeled with '-C'). All HMM variants performed better than MergeLevels (mean F-measure 0.73 ± 0.11). The LSP-HMM-C variant had the highest mean F-measure (0.89 ± 0.05), followed by the Rob-HMM-C (0.88 ± 0.06), followed by Base-HMM-C (0.84 ± 0.07). In all cases, *pooled* mode outperformed *single* mode.

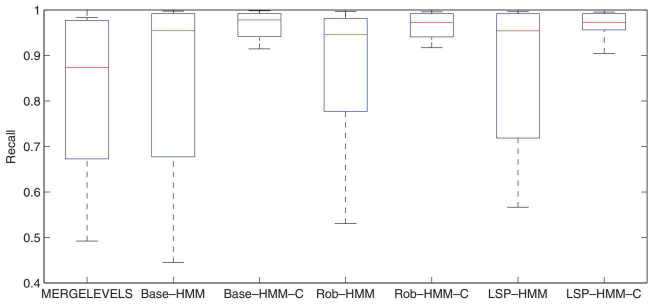


Fig. 5. Distributions of recall over eight MCL cell lines for MergeLevels, the Base-HMM, the Rob-HMM and the LSP-HMM with the CNP location prior in *single* and *pooled* mode (labeled with '-C') . All HMM variants had higher recall rates than MergeLevels (mean recall rate 0.82 ± 0.18). In all cases, *pooled* showed considerable improvement over *single* mode and showed very high recall rates. For Base-HMM-C, Rob-HMM-C and LSP-HMM-C the recall rates were the same at 0.97 ± 0.03.

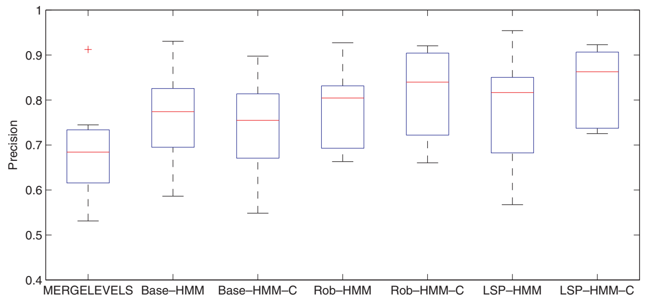


Fig. 6. Distributions of precision over eight MCL cell lines for MergeLevels, the Base-HMM, the Rob-HMM and the LSP-HMM with the CNP location prior in *single* and *pooled* mode (labeled with '-C'). Base-HMM had higher precision (0.76 ± 0.10) which was actually worsened slightly by pooling (Base-HMM-C: 0.74 ± 0.10). The Rob-HMM and Rob-HMM-C had precision of 0.78 ± 0.09 and 0.81 ± 0.09 indicating that robustness and robustness with pooling improves precision over the base model. Finally the LSP-HMM-C had the highest precision rates (0.83 ± 0.08). Pooling for the LSP-HMM showed the most benefit of all the HMM variants.

to 22) contained ground truth labeling therefore only these chromosomes were considered in our analysis. This reduced the number of clones per sample to 29,992. The data set contained a total of 195 aberrated regions: 123 losses and 72 gains covering approximately 1% of the human genome.

We used a list of CNPs (Wong *et al.*, unpublished) detected using SMRT arrays on a population of 95 normal individuals to set the LSP probability of an outlier. The list contains all the observed CNPs in the population. We discuss the potential use of other available CNP lists in Section 4.

Figure 3 shows chromosome 1 of HBL-2 with ground truth labels (A), with MergeLevels predictions (B) with Base-HMM predictions (C) and with LSP-HMM predictions (D). The Base-HMM and the LSP-HMM were both run in *pooled* mode. The LSP-HMM was given the complete list of CNPs described above that covered approximately 20% of the clones. We used $p_o = 10\%$ and $t_o = 0.01$ to determine outliers. Other parameters used for this data set are listed in Table 1. For MergeLevels, red bars indicate predicted regions of loss, green bars indicate predicted regions of gain. For the HMMs, red indicates 1 minus the probability of loss and green indicates probability of gain. These plots are similar in spirit to Engler *et al.* [6]. Figure 3 shows that the LSP-HMM predicts all of the aberrated clones with far fewer false positive predictions than both MergeLevels and the Base-HMM. Interestingly, MergeLevels and the Base-HMM are prone to different kinds of false positive predictions. MergeLevels tends to mis-label a small number of large segments with means slightly different than the neutral state mean, whereas the Base-HMM mis-labels a large number of very short segments usually corresponding to outliers. The LSP-HMM is relatively immune to both problems. In addition, the panel (D) depicts predicted outliers as purple stars, showing the qualitative advantage of providing additional information to the user in the output. This is particularly relevant to the left-most loss region in the p-arm. The ground truth labeling actually contains several clones that overlap CNPs. These clones are labeled as outliers by the LSP-HMM and therefore can instruct the user to treat the predicted loss with some degree of caution. In addition to this qualitative assessment of our algorithm, Figure 4 shows distributions of F-measures over the eight MCL cell lines. Distributions are shown as box-and-whisker plots where the line within the box indicates the median of the distribution, the top and bottom edges of the box indicate the third and first quartiles, the ends of the whiskers indicate the 95% confidence intervals of the distribution. The single point shown for Rob-HMM is outside the 95% confidence interval. The distributions show systematic improvement of the Base-HMM over MergeLevels, Rob-HMM over the Base-HMM and the LSP-HMM over the Rob-HMM. MergeLevels had a mean F-measure of 0.73 ± 0.10. Base-HMM had an F-measure of 0.77 ± 0.12 indicating that using an HMM framework improves accuracy over MergeLevels. Further gains were obtained by running the Base-HMM in *pooled* mode (F-measure for Base-HMM-C was 0.84 ± 0.07). Adding robustness in *pooled* mode (Rob-HMM-C) contributed additional improvement (F-measure was 0.88 ± 0.06). Finally using the robust model in *pooled* mode combined with prior knowledge on locations of CNPs (LSP-HMM-C) resulted in the highest accuracy (F-measure was 0.89 ± 0.05). In Figure 4, we can easily see from the boxplots that Base-HMM-C, Rob-HMM, Rob-HMM-C and LSP-HMM-C are all significantly better (at the 5% level) than MergeLevels. Base-HMM and LSP-HMM are not. We also performed a one way anova test (which is slightly less robust), and found that Rob-HMM-C and LSP-HMM-C are both significantly different (at the 5% level) to MergeLevels. Similar comments apply to the results on simulated data (see Section 3.2). Although the LSP-HMM-C is basically the same as the Rob-HMM-C, it is notable that it does not do worse despite being 'informed' by 20% of the locations in the LSP. This suggests that the model is robust to LSPs that are not supported by the data. This is significant given that our CNP list covers about 20% of the clones, yet in any one sample a much smaller portion of clones are expected to overlap a CNP (recall that the CNP list is made up of a union of all observed CNPs from a population of individuals). We note that the *pooled* mode worked considerably better for all HMM variants, demonstrating the advantage of ''borrowing statistical strength'' from all the data in the sample during parameter estimation.

To assess what was contributing to the differences in *F*-measure, we plotted precision and recall separately. The recall rates are shown in Figure 5 and demonstrate that pooling shows considerable improvement over *single* mode for the HMM variants. The recall rates were equally very high for the *pooled* HMM variants (0.97 ± 0.03). In contrast, differences in the HMM variants were observed for precision (see Figure 6). We observed improved precision of Rob-HMM-C over the Base-HMM-C indicating that considering outliers reduced the number of false positives. LSP-HMM-C had the highest precision (0.83 ± 0.09), therefore the CNP knowledge further reduced false positives (see Figure 6). The high recall rates for LSP-HMM-C suggests that any future effort to improve accuracy should first focus on reducing false positive predictions to improve precision. However, we noted numerous examples, such as at the centromeric end of the *q*-arm of HBL-2 chromosome 1 (Figure 3 **D**) where the falsely predicted aberrations could indeed be real.

3.2 Simulated data with outliers

To validate our model on additional data set with ground truth, we used the synthetic data created by Willenbrock and Fridlyand [24], downloaded from http://www.cbs.dtu.dk/~hanni/aCGH/. This data is fairly realistic, since it is generated by sampling segments from a large set of primary tumours [24]. To simulate CNPs, we modified this data by adding outliers planted randomly at 10% of the locations in the samples. The positions were sampled from a uniform distribution from 1 to 2000 (the number of clones in each sample). The $\log_2$ ratios for these outliers were sampled from a Gaussian distribution with mean 0 and variance 2. This gave us a data set with ground truth locations for the aberrated clones and for the positions of the outliers.

We chose 10% as the outlier fraction for the following reason. Our internally generated list of CNPs covers nearly 20% of the SMRT clones. However, publicly available CNPs represent approximately 1% of the SMRT clones. Therefore, we chose 10% as a reasonable compromise between these extremes. We also ran the Base-HMM and Rob-HMM on the original synthetic data and both performed extremely well (mean *F*-measure 0.95 ± 0.10 and 0.93 ± 0.12 respectively). This provided further justification to create a harder data set that contained the outliers.

In our experiments, we compared the effects of considering all the known outliers to adding additional locations to the prior which were not outliers. This simulated the effect of an incorrect prior. Note that we can choose the strength of the prior. We set the prior probability to 0.01.

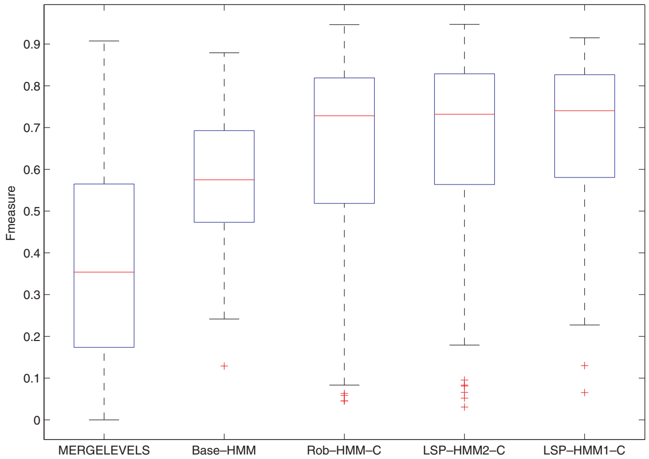


Fig. 7. *F*-measures for 100 samples of Willenbrock and Fridlyand's simulated data augmented with outliers. From left to right: MergeLevels had an *F*-measure of 0.37 ± 0.26. The Base-HMM had better accuracy (*F*-measure of 0.58 ± 0.16). Further improvement was gained with the Rob-HMM-C (*F*-measure = 0.64 ± 0.24). As expected, informing the LSP-HMM with the locations of the outliers (LSP-HMM1-C) resulted in the best performance. LSP-HMM2-C (*F*-measure = 0.66 ± 0.22) was informed with a superset of the outlier locations, and LSP-HMM1-C (*F*-measure 0.68 ± 0.19) was given all and only the outlier locations.

Figure 7 shows the distributions of accuracy on 100 samples for the three variants of our algorithm, including the LSP-HMM informed by a superset of the positions, and exactly all the positions of known outliers. Results on this data echo our results on MCL. MergeLevels performs considerably worse than all the HMMs: its *F*-measure was 0.37 ± 0.26 over 100 samples. The Base-HMM had a *F*-measure of 0.58 ± 0.16, validating that by using an HMM framework, significant improvement is attained over MergeLevels. As for MCL data, further improvement was attained by adding outlier detection. The Rob-HMM-C had a *F*-measure of 0.64 ± 0.24. Finally the versions of the LSP-HMM-C performed better when informed by a superset of the positions (*F*-measure=0.66 ± 0.22), and exactly all the positions (*F*-measure 0.68 ± 0.19) of the known outliers. This indicates that a weak prior, when supported by the data can help discover outliers, however contradictory evidence will usually overwhelm the prior when it is wrong.

4 DISCUSSION

We have presented a new model for classifying aberrated clones in aCGH data, which is robust to outliers and is able to leverage prior knowledge about CNP locations. We have demonstrated that on real and simulated data this model works better than a standard HMM and a state of the art method, DNAcopy+MergeLevels. We also determined that estimating parameters of the HMM using pooled data across chromosomes improves accuracy.

Our results showed that recall rates were very high for all HMM variants on the MCL data, and the differences in performance can be mainly attributed to precision rates. We showed that the LSP-HMM is immune to falsely predicting large regions that MergeLevels typically will mis-label and single clone outliers which the standard HMM falsely predicts as gains. We also showed qualitatively how the LSP-HMM enables the user to cross-reference predicted outliers with known CNPs and therefore allows for a more thorough interpretation of any predicted gains and losses.

As mentioned previously, the hyperparameters for both MCL and the synthetic data were set by hand. We believe that sensitivity to parameter settings (in particular with LSPs) are partially mitigated by pooling data across chromosomes. We showed in Figure 5 and Figure 6 how pooling improved both recall and precision rates for the LSP-HMM. We also noted that even though the CNP list for the MCL data consisted of 20% of the clones, the data overwhelms the prior at most locations. In *pooled* mode this phenomenon is significantly more pronounced as there is substantially more data available to help overwhelm the prior in locations where it is wrong. To further test this theory, our future work will involve accumulating a set of CNPs that is a union of numerous sets of previously published CNPs, for example Iafrate *et al*. [12], Sebat *et al*. [22], Tuzun *et al*. [23], Conrad *et al*. [3] and McCarroll *et al*. [17]. We anticipate that as long the the prior is not too strong a more comprehensive list of LSPs will further help aCGH analysis and the interpretation of results.

In addition to pooling, we plan to add levels to the hierarchy of the model to make it robust to parameter settings. We will put hyper-hyperparameters on the hyperparameters as discussed in Section 2.3. This increases the number of parameters to estimate, but the potential benefits of avoiding hand-tuning of parameters offset this additional cost. In addition, we also set the number of states of the HMM by hand. We noticed that the 4-state model performed better than the 3-state model, however the variance on the 4th state always converged to high values. This allowed the 4th state to ‘compete’ with the outlier process to explain the outliers, and therefore may have resulted in false positives. We are currently working on a new model that solves the ambiguities observed between high-variance states and the outlier process.

To evaluate the clinical applicability of our model, we plan to apply the method to samples extracted from a cohort of lymphoma patients. The aCGH profiles for these patients have been manually classified and numerous clinically relevant aberrations have been identified. We are also developing new models to identify locations of recurrent aberrations across samples, and to use other forms of prior knowledge, such as the locations of fragile sites. Combined with CNP information, we anticipate that such models will be extremely useful in profiling sub-types of cancer with aCGH.

ACKNOWLEDGEMENTS

This work is supported by a research grant from Genome Canada. SPS is supported by the Michael Smith Foundation for Health Research. We thank Kendi Wong from the British Columbia Cancer Research Centre for help with CNP determination and Kevin Leyton-Brown from the UBC Computer Science department for providing computing resources.

REFERENCES

[1] P Broët and S Richardson. Detection of gene copy number changes in cgh microarrays using a spatially correlated mixture model. *Bioinformatics*, Feb 2006.

[2] P G Buckley, K K Mantripragada, A. Piotrowski, T Diaz de Ståhl, and J P Dumanski. Copy-number polymorphisms: mining the tip of an iceberg. *Trends Genet*, 21(6):315–317, Jun 2005

[3] D F Conrad, T D Andrews, N P Carter, M E Hurles, and J K Pritchard. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*, 38(1): 75–81, Jan 2006.

[4] R J de Leeuw, J J Davies, A Rosenwald, G Bebb, R D Gascoyne, M J Dyer, L M Staudt, J A Martinez-Climent and W L Lam. Comprehensive whole genome array cgh profiling of mantle cell lymphoma model genomes. *Hum Mol Genet*, 13(17): 1827–1837, Sep 2004.

[5] P H Eilers and R X de Menezes. Quantile smoothing of array CGH data. *Bioinformatics*, 21(7): 1146–1153, Apr 2005.

[6] D A Engler, G Mohapatra, D N Louis, and R A Betensky. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations (acgh). *Biostatistics*, Jan 2006.

[7] J Fridlyand, A Snijders, D Pinkel, D Albertson, and A Jain. Hidden Markov Models approach to the analysis of array CGH data. *Journal of Multivariate Statistics*, 90: 132–153, 2004.

[8] A Gelman, J Carlin, H Stern, and D Rubin. *Bayesian data analysis*. Chapman and Hall, 2004. 2nd edition.

[9] S Guha, Y Li, and D Neuberg. Bayesian hidden markov modeling of array cgh data. Technical report, Harvard School of Public Health, 2006.

[10] L Hsu, S G Self, D Grove, T Randolph, K Wang, J J Delrow, L Loo, and P Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6(2): 211–226, Apr 2005.

[11] P Hupé, N Stransky, J Thiery, F Radvanyi, and E Barillot. Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422, Dec 2004.

[12] A J Iafrate, L Feuk, M N Rivera, M L Listewnik, P K Donahoe, Y Qi, S W Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nat Genet* 36(9):949–951, Sep 2004.

[13] A S Ishkanian, C A Malloff, S K Watson, R J DeLeeuw, B Chi, B P Coe, A Snijders, D Albertson, D Pinkel, M Marra, V Ling, C MacAulay, and W Lam. A tiling resolution DNA microarray with complete coverage of the human genome. *Nat Genet*, 36(3):299–303, Mar 2004.

[14] K Jong, E Marchiori, G Meijer, A V Vaart, and B Ylstra. Breakpoint identification and smoothing of array comparative genomic hybridization data. *Bioinformatics* 20(18):3636–3637, Dec 2004.

[15] M Khojasteh, W L Lam, R K Ward, and C MacAulay. A stepwise framework for the normalization of array cgh data. *BMC Bioinformatics*, 6:274–274, Nov 2005.

[16] W R Lai, M D Johnson, R Kucherlapati, and P J Park. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics* 21(19):3763–3770, Oct 2005.

[17] S A McCarroll *et al.* Common deletion polymorphisms in the human genome. *Nat Genet* 38(1):86–92, Jan 2006.

[18] A B Olshen, E S Venkatraman, R Lucito, and M Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5(4):557–572, Oct 2004.

[19] F Picard, S Robin, M Lavielle, C Vaisse, and J J Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics* 6(1):27–27, Feb 2005.

[20] D Pinkel and D G Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet* 37Suppl: 11–17, Jun 2005.

[21] S Scott. Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American Statistical Association*, 2002.

[22] J Sebat, B Lakshmi, J Troge, J Alexander, J Young, P Lundin, S Månér, H Massa, M Walker, M Chi, N Navin, R Lucito, J Healy, J Hicks, K Ye, A Reiner, T C Gilliam, B Trask, N Patterson, A Zetterberg, and M Wigler. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, Jul 2004.

[23] E Tuzun, A J Sharp, J A Bailey, R Kaul, V A Morrison, L M Pertz, E Haugen, H Hayden, D Albertson, D Pinkel, M V Olson, and E E Eichler. Fine-scale structural variation of the human genome. *Nat Genet*, 37(7):727–732, Jul 2005.

[24] H Willenbrock and J Fridlyand. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, Sep 2005.