Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Multivariate Analysis 90 (2004) 132–153

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

Hidden Markov models approach to the analysis of array CGH data

Jane Fridlyand,* Antoine M. Snijders, Dan Pinkel, Donna G. Albertson, and Ajay N. Jain

UCSF Comprehensive Cancer Center, 2340 Sutter Str. N412, San Francisco, CA 94143-0128, USA

Received 15 July 2003

Abstract

The development of solid tumors is associated with acquisition of complex genetic alterations, indicating that failures in the mechanisms that maintain the integrity of the genome contribute to tumor evolution. Thus, one expects that the particular types of genomic alterations seen in tumors reflect underlying failures in maintenance of genetic stability, as well as selection for changes that provide growth advantage. In order to investigate genomic alterations we are using microarray-based comparative genomic hybridization (array CGH). The computational task is to map and characterize the number and types of copy number alterations present in the tumors, and so define copy number phenotypes and associate them with known biological markers.

To utilize the spatial coherence between nearby clones, we use an unsupervised hidden Markov models approach. The clones are partitioned into the states which represent the underlying copy number of the group of clones. The method is demonstrated on the two cell line datasets, one with known copy number alterations. The biological conclusions drawn from the analyses are discussed.

AMS 2000 subject classifications: 92-08

Keywords: HMM; Array CGH; MMR; Chromosomal aberrations; Genomic profile; Mutation; Cancer

*Corresponding author. Fax: +415-502-3179.
E-mail address: janef@cc.ucsf.edu (J. Fridlyand).
URL: http://www.cc.ucsf.edu/jainlab/people.

0047-259X/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2004.02.008

1. Introduction

Tumors are driven by an accumulation of genetic and epigenetic changes resulting in altered levels of expression of certain genes, thereby modifying normal cell growth and survival. Many of these changes involve gains and/or losses of parts of the genome and are frequently observed in many different types of cancers. In order for a cell to become malignant, one or more mechanisms that normally maintain genome integrity and/or regulate cell division must be compromised, presumably through mutations that occur early in tumorogenesis. The relationship between DNA copy number and the transcriptional activity of the genes mapped to altered regions has been demonstrated in [14]. Moreover, gains and losses of chromosomes and chromosomal segments are involved in developmental abnormalities.

A variety of cytogenetic and more recently array-based [6,13,19] analytic methods have revealed a wide range in the number and types of DNA copy number aberrations present in human and murine tumors. This is likely to be a reflection of not only selection for increased growth advantage and survival, but also the different number and types of DNA copy number aberrations could point to underlying defective mechanism(s) permitting these aberrations.

1.1. Array CGH

Microarray-based comparative genomic hybridization (array CGH) provides a means to quantitatively measure DNA copy number aberrations and map them directly onto genome sequences. Typically, a test genomic DNA pool (e.g. tumor genomic DNA) is labeled with Cy3 and a reference genomic DNA pool (e.g. normal genomic DNA) is labeled with Cy5. The differentially labeled genomic DNA pools are then combined with unlabeled Cot-1 DNA, which blocks repetitive sequences in the genome, denatured and hybridized onto an array containing genomic clones. After hybridization, digital images are captured for each of the fluorescent dyes used in the hybridization. Image analysis software is used to calculate a ratio of the fluorescence intensities for each of the array targets. This ratio of test to reference intensity then reflects the relative DNA copy number between the two hybridized specimens for a certain locus. Arrays comprised of large-insert genomic clones, such as bacterial artificial chromosomes (BACs) provide reliable and quantitative measurements of DNA copy number aberrations, which can range from single copy gains and losses to homozygous deletions and high-level amplifications. The schematic representation of array CGH is shown in Fig. 1.

Analyses of array CGH data have shown that the genomes of established tumors are remarkably stable, as evidenced by similarity of tumor recurrences to primary tumors [2,20]. These observations indicate that the set of aberrations is maintained because of continued selective advantage. The dependence of the type of changes in a tumor on its genetic background has been previously demonstrated using both gene expression [5] and chromosomal data [18]. Desai et al . [5] demonstrated in murine models that initiating oncogenic events determine gene expression patterns of mammary tumors. Snijders et al. [18] showed that tumors with defects in mismatch

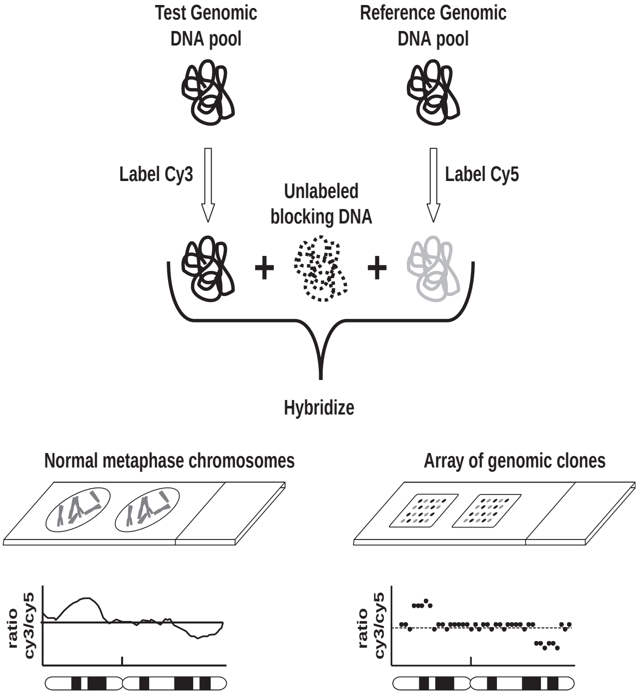

Fig. 1. Schematic representation of array CGH. The test and reference DNA are labeled with two different dyes and competitively hybridized to the array together with COT-1 DNA in order to block repeat sequences. The ratio of the fluorescence intensity for each spot is indicative of the relative copy number of the corresponding DNA sequence in the tumor sample. For display purposes the clones are ordered according to their mapping positions along the genome. The density of the clones on the array is high enough (approximately every 1.4 Mb) to be able to approximate the true copy number distribution.

repair (MMR) genes take a different genetic route to develop methotrexate resistance than MMR proficient tumors. These observations suggest that the spectrum of alterations that one sees in fully developed tumors is a composite of selection acting on the variation that is permitted to arise by the particular failures in genomic surveillance mechanisms(s) present in the tumor. The relation of mechanistic defect to aberration type has not been established for most sporadic tumor types; however, it is likely that some of the variety and complexity associated with these tumor genomes might be rationalized if associations could be developed between particular aberrations and specific effects in maintenance of genome stability. One’s hope is to eventually identify specific mechanistic failures and corresponding drug targets depending on the gene expression or copy number tumor profile.

The *karyotype* of a cell is defined as the chromosomal makeup of a somatic cell including the number, arrangement, size and structure of the chromosomes. Array CGH approach allows us to assess the number and types of dosage changes of chromosomal segments relative to a reference which generally has a normal diploid karyotype. We present an automated method for identifying and characterizing DNA copy number changes in a given sample. We distinguish four types of genomic changes: *copy number transitions* indicating *low-level* changes within a chromosome, *whole chromosomal gains and losses*, *focal aberrations* due to low-level alterations of very narrow regions (<1 Mb with the current resolution) and *high-level focal amplifications*. Alterations that can occur by only one biological event, for example a single copy gain or loss, are defined as "low-level". These result in small changes in observed ratios. "Amplifications" are defined as an increase in copy number of small regions of the genome that require multiple biological events to accomplish. Typically, these result in large ratio changes. More details on the underlying biology of the four types of changes are given in Section 2.5. Previously, it was possible to evaluate results of FISH and chromosomal CGH for overall number of genomic aberrations or to search for recurrent changes. The more detailed taxonomy has emerged only with the use of genome-wide array CGH [19] which allows one to more accurately determine the number, types and locations of the transitions of DNA copy number alterations throughout the genome.

The manual process of characterizing individual genomic profiles is time-consuming, prone to human error and non-reproducible. Fig. 2 demonstrates the possible range for the tumor profile complexity. While manual identification of the genomic changes for the profile (a) is straightforward, it becomes more complicated for the profile (b) and virtually impossible for (c).

With the coming abundance of array CGH data, the need for an automated reliable algorithm is substantial. Section 2 gives a brief overview of the available approaches and describes the application of the unsupervised hidden Markov model (HMM) method to array CGH data. Two cell-line data sets are presented in Section 2.1 and the results are given in Section 3. Finally, Section 4 summarizes our findings and outlines open questions.

2. Methods

For a given genomic profile, the goal is to partition the clones into sets with the same copy number and thus to characterize the copy number of the genomic segments. The biological model underlying this approach is that genomic rearrangements lead to gains or losses of segments of the genome, possibly spanning entire chromosomes, or, alternatively, to focal high-level amplifications. In particular, it is desirable to make use of the physical dependence of the nearby clones, which translates into copy number dependence. Currently, we are aware of two attempts at addressing this problem which have been developed in parallel with our method. In Olshen and Venkatraman [12] the authors develop a novel modification of binary segmentation referred to as *circular binary segmentation* to look for the change points along each chromosome that represent the copy number transitions. Jong et al. [10] apply a genetic local search algorithm to best segment the clones into clusters. Both methods operate on individual chromosomes.

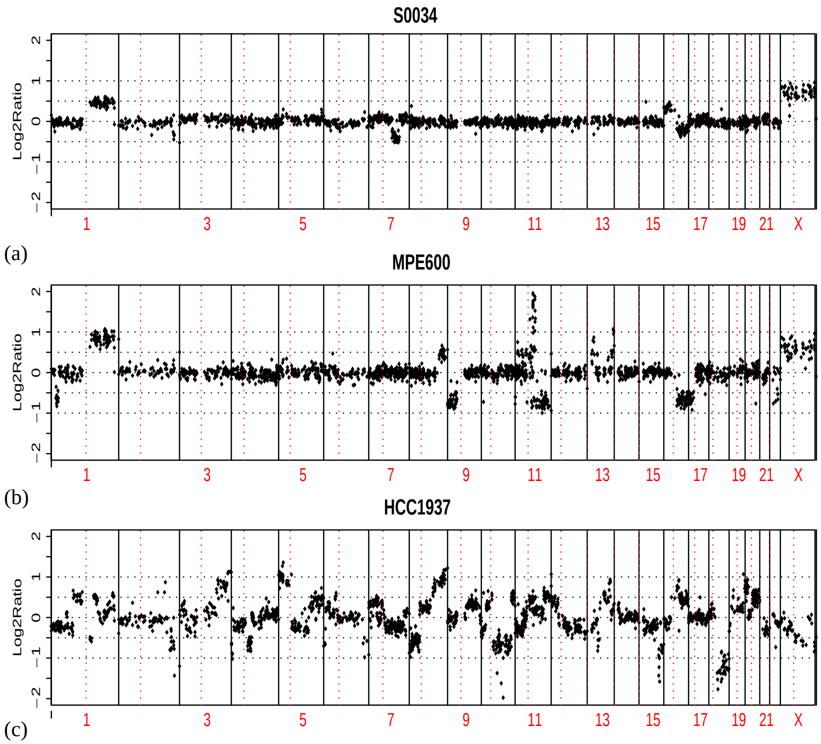

Fig. 2. Example of varying genomic complexity in breast tumors. Profiles from samples may range from those that have no genomic changes (a) to those with every single chromosome containing multiple aberrations (c). Note that amplifications (high-level focal gains) do not have to occur in the presence of many chromosomal changes, (c), but may arise in relatively quiet profiles, (b).

In this paper, we develop an algorithm which consists of two parts. In the first part, we partition clones on individual chromosomes into sets with the same underlying copy number. Often the copy number can be inferred by using thresholding of median value of $\log_2$ of the ratio of test (T) and reference intensities (R) or $\log_2 T/R$ in a given set. In the second part we look to characterize individual chromosomes according to whether there have been any copy number transitions or whole chromosome gains or losses. We also look for focal aberrations corresponding to the individual clones such as low level gains and losses or high level amplifications. The parameters of the algorithm have been derived using unpublished primary tumor data by cross-checking the results with an expert in array CGH.

2.1. Data

2.2. Cell lines

We demonstrate our approach on two publicly available cell-line data sets.

The first data set was also used by Olshen and Venkataraman [12] to show performance of their method.

2.2.1. Coriel cell lines

To test our ability to detect the low-level gains and losses, we use data featured in Snijders et al. [19]. (These data are freely available for download at http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html.) The data consists of single experiments on 15 fibroblast cell lines containing cytogenetically mapped partial or whole-chromosome *aneuploidy* (non-diploid copy number), and each array contained 2276 mapped BACs spotted in triplicate.

This data set is used as a proof of principle as it consists of very pure diploid cell lines with previously characterized genomic aberrations which are easily detectable by manual inspection of the profiles. To complicate the detection task in order to be able to optimize the parameters of the procedure and to compare the HMM approach to the more traditional clustering partitioning methods, we generate artificial chromosomes by adding Gaussian noise to the observed $\log_2 T/R$ ratios on the selected chromosomes.

2.2.2. MMR cell lines

The interplay between selection and genetic instability in shaping tumor genomes is currently most clearly established in tumors with defects in MMR. These tumors show a high level of microsatellite instability because of failures in MMR genes (generally MSH2 or MLH1). Cytogenetic analyses have shown that tumors with defects in MMR have fewer chromosomal changes than MMR competent solid tumors, suggesting that a greater proportion of the alterations required for malignancy occur in genes whose nucleotide sequences are susceptible to errors normally corrected by this system. Array CGH was done on 10 MMR deficient and 10 proficient cell lines to perform a high-resolution analysis of the effect of MMR competence on the number and types of the genomic alterations [18]. (Complete data set is available at http://cc.ucsf.edu/albertson/public.) Our aim was to confirm the previous findings and to investigate whether differences in specific MMR genes translate to differences in the types of genomic instability.

2.3. Data pre-processing

In an ideal measurement, the copy number of a given clone can be inferred by considering its $\log_2 T/R$ ratio. For instance, $\log_2 \frac{3}{2} = 1.58$ should correspond to one copy gain and $\log_2 \frac{1}{2} = -1$ would mean one copy loss. However, the primary task of estimating true copy number for a given clone is complicated by many experimental and biological factors such as purity and ploidy of a sample. The most frequent phenomenon arising in the analysis of the primary tumors is imperfect dissection leading to normal cell contamination. Generally, pathologists make sure that each tumor sample contains no more than 50% or even 30% of normal cells. (The purity of the tumor sample increases as the contamination proportion decreases.)

Additionally, not all tumor cells may have acquired a given aberration. Finally, the tumor cells may not be diploid, i.e. a large number of whole chromosomes are present in more than two copies. Here we define ploidy of the sample as the copy number of the majority of the genome. On our arrays we approximate the ploidy by the median copy number of the loci represented on the array, typically normalized to 0 as described below. All of these factors reduce the expected magnitude of copy number ratios and often make estimation of a true underlying copy number for a given clone impossible. Here we only partition clones into sets of the same underlying copy number and do not undertake the task of estimating the true copy number corresponding to a given state.

Formally, the observed unnormalized $\log_2 T/R$ ratio for a clone j in a sample i, x_{ij}, is determined by the true copy number of that clone in a tumor cell, c_{ij}^{T}, ploidy of the reference sample, pl_{ref}, normal cell admixture, a_i^{N}, ploidy of the normal (non-tumor) cells, c_i^{N}, and fraction of the tumor cells which have not acquired a given aberration, t_{ij}^{N}. Then, the proportion of the cells with a given aberration is $p_{ij}^{\mathrm{aber}} = (1 - a_i^{\mathrm{N}})(1 - t_{ij}^{\mathrm{N}})$ and

$$x_{ij} = \log_2\left(\frac{c_{ij}^{\mathrm{T}} p_{ij}^{\mathrm{aber}} + c_i^{\mathrm{N}}(1 - p_{ij}^{\mathrm{aber}})}{pl_{\mathrm{ref}}}\right) + \varepsilon_{ij}^{(1)},$$

where $\varepsilon_{ij}^{(1)}$ is experimental error. In addition, we allow observed $\log_2 T/R$ ratios on a chromosome to be dependent, i.e. the copy number for a clone is probabilistically related to the copy number of the physically proximal clones. Generally, the reference sample is derived from a normal (diploid) tissue and hence has ploidy $pl_{\mathrm{ref}} = 2$.

Note that as with any comparative technology, proper array normalization is necessary for meaningful multi-array analysis. We have found that no intensity or subarray-dependent normalization [21] is necessary with arrays described in Snijders et al. [19] and quantified using custom image analysis software *Spot* [9]. Instead we normalize by the global shift of the median of the observed $\log_2$ ratios for a given sample. The normalized observed $\log_2$ ratio is

$$\begin{aligned} y_{ij} = {} & \log_2\left(\frac{c_{ij}^{\mathrm{T}} p_{ij}^{\mathrm{aber}} + c_i^{\mathrm{N}}(1 - p_{ij}^{\mathrm{aber}})}{pl_{\mathrm{ref}}}\right) - \\ & \mathit{median}_j\left(\log_2\left(\frac{c_{ij}^{\mathrm{T}} p_{ij}^{\mathrm{aber}} + c_i^{\mathrm{N}}(1 - p_{ij}^{\mathrm{aber}})}{pl_{\mathrm{ref}}}\right)\right) + \varepsilon_{ij}. \end{aligned}$$

Recall that numerator of the second term approximates the ploidy of the tumor sample, pl_{T}, and, thus,

$$y_{ij} = \log_2\left(\frac{c_{ij}^{\mathrm{T}} p_{ij}^{\mathrm{aber}} + c_i^{\mathrm{N}}(1 - p_{ij}^{\mathrm{aber}})}{pl_{\mathrm{T}}}\right) + \varepsilon_{ij},$$

where ε_{ij} is assumed to be distributed as *iid Normal*$(0, \sigma_i^2)$ with a constant variance for all clones in a given hybridization. The distributional assumption is used in the likelihood calculation when fitting HMMs in Section 2.4 and is supported by the experimental data on self–self hybridizations (i.e. when a sample is compared to itself) where all the departures of the observed values from 0 are assumed to be experimental noise. Error variability has previously been shown to be independent of the copy number using the experiments on the cell lines with known genetic alterations. Note that the requirement that tumor sample ploidy is well approximated by the median of the unnormalized $\log_2$ ratios is necessary to perform a meaningful normalization. Also, with normalization the ploidy of the reference sample no longer plays a role. In the rest of the paper we partition normalized $\log_2 T/R$ ratios, y_{ij} in the sets of clones with the same copy number and determine alteration types for individual clones and chromosomal regions.

2.4. Unsupervised HMM partitioning

The HMM approach is a natural framework for the task at hand as the hidden states represent the underlying copy number of the clones. The data going into the model are the observed normalized $\log_2 T/R$ ratios of the clones. For each chromosome j and sample i, we independently fit an HMM, determine the number of states and allocate the clones to the derived states.

HMMs have been extensively applied to a wide range of problems [15]. A discrete time HMM with continuous output is characterized by the following:

(1) K, the number of states in the model. The states are hidden and, generally, physically meaningful. Typically, the states are interconnected in a way that any state can be reached from any other state. We denote the individual states as $S = S_1, \ldots, S_K$ and the state at the location l as $s_l, 1 \leqslant l \leqslant L$.

(2) The initial state distribution $\pi = \{\pi_k\}$ where

$$\pi_k = P\{s_1 = S_k\}, \quad 1 \leqslant k \leqslant K.$$

(3) The state transition probability distribution $A = \{a_{mp}\}$ where

$$a_{mp} = P\{s_{l+1} = S_p | s_l = S_m\}, \quad 1 \leqslant m, \ p \leqslant K.$$

For the special case of a model in which all states are connected, $a_{mp} > 0$ for all m, p.

(4) The emission distribution or probability density function $B = \{b_k(\mathbf{O})\}$ where

$$\{b_k(\mathbf{O})\} = \mathcal{G}(\mathbf{O}, \boldsymbol{\mu}_k, \mathbf{U}_k), \quad 1 \leqslant k \leqslant K,$$

where $\mathbf{O}$ is the vector being modeled, $\mathcal{G}$ is Gaussian density with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{U}_k$. More generally, $\mathcal{G}$ is any log-concave or elliptically symmetric density and the probability density function $\{b_k(\mathbf{O})\}$ is a finite mixture (see [15]).

Thus, an HMM with the fixed number of hidden states K can be characterized in terms of three parameters: (i) the initial state probabilities, π, (ii) the transition probability matrix, A, and (iii) the collection of Gaussian emission probability functions defined within each state, B. The parameters of the model may be represented in a compact way as $\lambda = (A, B, \pi)$ and the sequence of values on a given chromosome is written as $\mathbf{O} = (o_1, \ldots, o_T)$.

When fitting an HMM of size K to the vector of the observed $\log_2 T/R$ ratios of the clones that physically map to a given chromosome, we use the *Forward-Backward Procedure* to calculate $Lik(\lambda|\mathbf{O})$ or the likelihood of the parameters given the vector of values. To identify the *optimal* state sequence associated with a given vector, for each observation o_t we choose the state s_l which is *individually* most likely. Finally we re-estimate model parameters $\lambda = (A, B, \pi)$ to maximize $Lik(\lambda|\mathbf{O})$ using the *Baum-Welch method* or, equivalently, the *EM algorithm*. It has been shown (Rabiner [15]) that while the initial estimates for π and A may be fairly arbitrary, good initial estimates for B may be essential in the HMM with a continuous distribution output. We set parameters for π by placing a majority of the weight on the state corresponding to the "normal" (or median for a given sample) copy number (i.e. expected value of a state output is 0) and distributing the remaining probability uniformly among all other states. Similarly, to initialize A, we assign a high probability of remaining in the same state and low non-zero probabilities to transitioning between states. This results in an HMM model where all states are connected. Finally, to estimate initial emission probabilities, we segment the observations in K states using *partitioning among medoids* or *PAM* [11] and estimate the mean for each state by the median of the $\log_2 T/R$ ratios of the clones that were allocated to that state. Similarly, we estimate the common initial variance for the states in a given sample.

It remains to choose the number of states $K, 1 \leqslant K \leqslant K_{\max}$. We take a heuristic approach which minimizes

$$\psi(K) = -\log(Lik(\lambda|\mathbf{O})) + q_K D(L)/L, \quad K = 1, \ldots, K_{\max},$$

where q_K is the number of the parameters corresponding to the number of states, K, and $D(L)$ is a function of the number of L clones on a chromosome. Note that $D(L) = 2$ gives AIC or Akaike's information criterion [1] and $D(L) = \log(L)$ one obtains the Schwartz BIC or Bayesian information criterion [17].

Algorithm 1. *Segment clones into sets with the same underlying copy number*:

```
• For $k = 1 \ldots K_{\max}$ states:
    (1) Fit $k$-state HMM.
    (2) Calculate penalized negative log-likelihood $\psi(k)$.
• Choose the model corresponding to the number of states with the smallest $\psi(k)$,
  $K = \mathrm{argmin}_k\, \psi(k)$.
• If $K = 1$, then STOP.
• Calculate the median for each state and identify the two states whose medians
  are the closest. Compute $d = \min_{k_1 \neq k_2}^{K} |med_{k_1} - med_{k_2}|$ .
• While $d < threshold$ and $K > 1$.
    (1) Merge the two closest states.
    (2) Set $K = K - 1$ and recompute $d$.
```

The parameters of the procedure are the maximum size of an HMM model, $K_{\max}$, the model selection criterion $D(L)$ and the threshold for the state merge, *threshold*. The maximum number of states is limited by the number of clones on a given chromosome and the computing time, so we have used $K_{\max} = 5$. We have found that this was generally sufficient to recover the structure of even very complicated tumor profiles. In Section 3.1 we describe a small simulation study comparing the performance of *AIC* and *BIC* model size selection criteria. We conclude that *AIC* followed by the merging step leads to the most satisfying results. Finally, the *threshold* to merge the states should be determined by taking into account the specific problem at hand: i.e., what the investigator believes to be the level of purity of the populations and whether type I or type II error is more undesirable. For example, it is acceptable to have a number of false positives when structurally characterizing tumor profiles. On the other hand, one needs to be very conservative in calling the aberrations for medical diagnostic purposes [3]. In the past, we have allowed the threshold to be as low as 0.25 when analyzing primary tumors and as high as 0.5 when looking for genetic abnormalities homogeneously present in all cells. Generally, lower thresholds should be used with more heterogeneous samples.

It is possible to fit HMMs to all chromosomes simultaneously by constraining state means and variances to be the same across the chromosomes. The difficulty of doing this is that some copy number changes may be present in all cells in the population, while others are present only in some of the cells. Thus aberrations at different locations in the genome may have different ratios. Attempting to constraint the transition matrix to be the same across chromosomes is probably biologically even less plausible due to varying propensity for genomic instability on different chromosomes.

2.5. Characterizing genomic aberrations

We characterize the genomic profiles using four types of genomic alterations: *copy number transitions*, *whole chromosomal gains and losses*, *focal aberrations* and *high-level focal amplifications*. Different genomic alterations are likely to be initiated by failures in different types of mechanisms that normally maintain genome integrity and/or regulate proper cell division. For example, gains or losses of whole chromosomes are expected to occur following failures of karyokinesis or cytokinesis (divisions of nucleus or cell), and copy number transitions within a chromosome are likely to be initiated by DNA double-stranded breaks. The two types of focal aberrations reflect low-level gains or losses of DNA sequence spanning less than 1 Mb (one or two clones at the current array resolution) and gene amplifications, which we defined as focal high-level copy number changes. The focal aberrations are also a consequence of double-stranded breaks; however, there is likely to be a mechanistic difference between the breaks leading to focal aberrations and low-level copy gains and losses. It is believed that amplifications arise under selective pressure to replicate the same localized event more than once. The exact mechanisms of how different types of genomic instability arise are currently unknown; however, we have found so far that the above taxonomy is useful and is associated with existing clinical

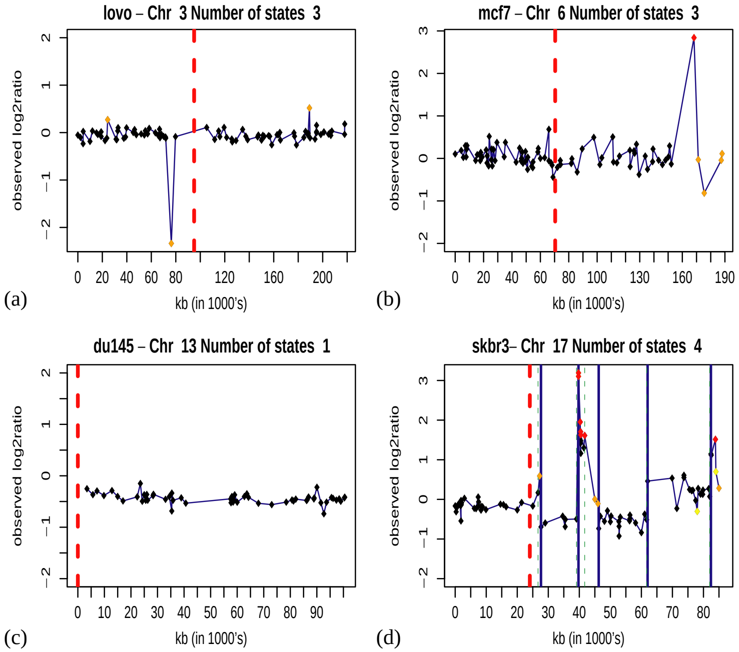

Fig. 3. Example of types of genomic aberrations. The dotted red line indicates the centromere of each chromosome. Solid blue and dotted green lines show the first clone after the transition and last clone before the transition, respectively, separating the regions with constant copy number. Red and orange dots indicate amplifications and focal aberrations. Yellow dots indicate outliers. Fig. (a) shows homozygous deletion and 1 copy focal gain (focal aberration). Fig. (b) shows an amplicon consisting of a single clone and 1 copy focal deletion (focal aberration). Note that the three neighboring clones are classified as aberrations as they are placed in alternating states and the last two clones fall into the same state but are telomeric clones. Fig. (c) shows whole chromosome loss. The structure of the chromosome in (d) is complicated: it contains transitions (normal, one copy gain and one copy loss states), an amplicon with complicated structure also separated by transitions and a single clone amplicon; outliers and focal aberrations.

phenotypes (work in progress). In particular, we demonstrate in Section 3.2 that specific types of genomic alterations are associated with different genetic mutations.

Once the clones on each chromosome are partitioned into the states, there are still parameters to be set by a user to assign the alteration types. We have defined the parameters in accordance with our experience with array CGH data. Examples for each type of alteration are shown in Fig. 3. The following algorithm outlines the procedure we use for categorizing the genomic alterations.

Algorithm 2. *Assign alteration types to genomic regions and to individual clones*:

(1) Estimate the sample standard deviation: Compute the median absolute deviation (MAD) of the clones in the states containing at least 20 clones located on the chromosomes partitioned in $\leqslant 3$ states. The standard deviation σ

is estimated as median of MADs for all such states. High σ indicates poor hybridization quality.

(2) Find outliers: Identify a clone as an *outlier* if its value differs from the median value of their state by $\geqslant 5\sigma$. The clones that are indicated as outliers over a series of hybridizations may represent natural copy number polymorphisms such as repeats or length polymorphisms [3] or mismapped clones. Outliers do not represent an alteration type but rather an auxiliary quantity used in finding amplifications and detecting array problems.

(3) Identify focal aberrations: A clone is called a *focal aberration* if it is (i) a single clone (two clones mapped to the *telomere* or to the ends of a chromosome) assigned to a state different from the state of both of the neighboring clones (one neighboring clone if at a telomere) or (ii) two or more clones mapping within 1 Mb whose states are different from the states of both of the neighboring clones. Focal aberrations may indicate true narrow gains or deletions or, alternatively, mismapped clones or natural copy number polymorphisms.

(4) Find transitions: Exclude clones marked as focal aberrations and place the *copy number transitions* between the two regions whose states differ. The transitions indicate the number of double-stranded breaks that occurred. The end of a previous region and the start of a new region are placed at the last and first clones of the regions immediately to the left and to the right of the transition.

(5) Identify whole chromosomal changes: A chromosome is called gained (lost) if it does not contain transitions and, after exclusion of clones that are outliers or focal aberrations all of the following conditions hold: (i) at least 95% of the clones have $\log_2 T/R$ greater (less) than 0, (ii) the null hypothesis H_0 that the mean $\mu = 0$ is rejected at p-value < 0.0001 using a t-test, and (iii) the median of the $\log_2 T/R$ ratios of the clones is greater (less) than a threshold $= 0.1$ (-0.1). Note that all the quantities here are ad hoc and the threshold is used to protect against rare but possible sporadic behavior of the chromosomes with sparse coverage.

(6) Find focal amplifications: The magnitude, focality and relative ratio difference are used to identify focal amplifications. A clone is amplified if any of the following is true:
   - It is an outlier and its $\log_2 T/R$ is (i) greater by $\mathit{diffVal}_1 = 1$ than the median value of the state containing the clone, or (ii) at least $\mathit{absValSingle} = 1$ and is greater by $\mathit{diffVal}_2 = 0.5$ than the median value of the state containing the clone;
   - It is an aberration and its $\log_2 T/R$ is (i) greater by $\mathit{diffVal}_1 = 1$ than the maximum of the median values of the two surrounding states, or (ii) greater than $\mathit{absValSingle} = 1$ and is greater by $\mathit{diffVal}_2 = 0.5$ than the maximum of the median values of the two surrounding states;
   - It belongs to a narrow (< 5 Mb) region and the median value of the state to which the clones in the region belong is greater than $\mathit{absValRegion} > 1.5$. Note, that the conditions above describe a focal amplification as a clone whose $\log_2 T/R$ is sufficiently high relative to its neighboring clones. In addition, we allow multiple amplicons in the narrow regions with very high $\log_2 T/R$ (last condition).

3. Results

3.1. Coriel cell lines

To assess the performance of our algorithm on the Coriel cell lines, we used the table of agreement between known karyotypes and manual segmentation of the array CGH profiles published by Snijders et al. [19]. There were 15 chromosomes with partial changes and 8 whole chromosomal monosomies (1 copy) or trisomies (three copies). In two chromosomes, the known regions of alterations contained only one or two clones and were located at the telomeric ends. In their analysis, Snijders et al. [19] confirmed all but one of the known partial and whole chromosomal changes. In addition they found a number of single clone aberrations that have not been confirmed by an alternative technique, such as FISH. Such localized aberrations may be real and not previously seen because of the crude resolution of the cytogenetic analysis. Alternatively, they could correspond to the mismapped clones or array artifacts.

In our analysis we have detected all of the whole and partial chromosomal changes confirmed by Snijders et al. [19], and all of the transitions except the two that occurred in the narrow telomeric regions described above. Our definition of a copy number transition does not allow us to identify such narrow regions near the telomere, and the clones in those regions were placed in a different state from the rest of the clones and identified as focal aberrations. Thus, all of the known alterations were found. Additionally, there were a number of single-clone aberrations (mean number of aberrations per sample is 3 with range of 0 to 8) found by the algorithm. These alterations were not evaluated by independent means. Fig. 4 shows four examples of application of our method to the Coriel cell lines.

In this analysis we used AIC as a model selection criterion followed by a merging step with the conservative threshold of 0.35 since we are dealing with pure diploid cell lines and hence the expected size of one copy gain is $\log_2(3/2) = 0.58$ and of one copy loss is $\log_2(1/2) = -1$. Interestingly, if merging was not used, additional transitions occurring at the same chromosomal locations emerge on chromosome 11 in many samples that separate regions with a small median $\log_2 T/R$ (see Fig. 5). Those transitions were not reproduced when the same samples were hybridized to a newer BAC array (data not shown), i.e. we conclude that they represent an array artifact.

In comparison with our approach, the change-point method of Olshen and Venkatraman [12] does not detect focal aberrations and it has missed a partial chromosomal change represented by 2 clones only, but it detected all but one of the longer partial changes that were called in [19]. One of the deletions was not detected because of the presence of an outlier on that chromosome. The same transitions on chromosome 11 that we have detected in the absence of merging were also found by Olshen and Venkatraman [12].

Simulations: In choosing a model selection criterion, the question of type I error versus type II error always arises. It is also clear that any criterion choice is confounded by the subsequent state merge step. Here we briefly address the choice of the criterion by using a simulation approach. In addition, we investigate whether

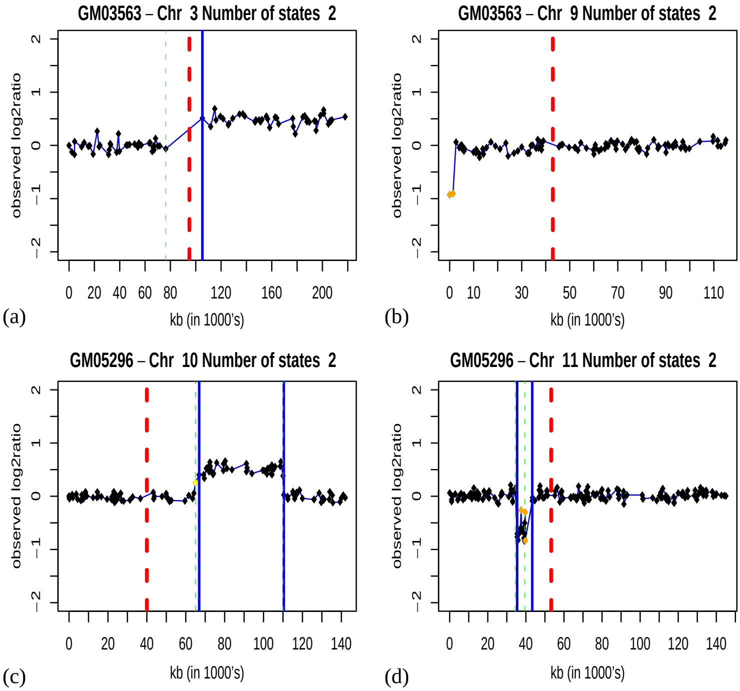

Fig. 4. Application of the HMM to Coriel cell lines. Fig. (a) shows one copy gain with one transition at the centromere. Fig. (b) contains monosomy at the telomeric end. The loss is coded as focal aberration since there are two clones only at the telomere. Figure in lower left: trisomy created by two transitions. The clone near the first transition point is coded as an outlier. In Fig. (d) narrow monosomy created by two transitions. The aberrations in the deletion region are likely to be mismapped clones.

treating array CGH data as dependent is indeed an improvement upon simply segmenting clones into a pre-assigned or optimized number of clusters.

Optimizing the parameters using the Coriel cell line data is difficult because on that data set any reasonable technique will do well. Instead, we generate artificial chromosomes by repeatedly adding Gaussian noise to the chromosomes that contain at least one transition. We investigate whether an AIC or BIC penalty gives better results with or without the state merging step and compare the performance of partitioning observations with HMM to PAM clustering results.

Three chromosomes each containing a monosomy or trisomy identified by a transition were chosen for the study. Since the true karyotype is known for each of the chromosomes, we record the “true” states, P^{true}, by placing the clones known to have the same copy number in the same cluster. Each clone is assigned state “1” if it has a lower copy number than the clones belonging to the other state or “2” otherwise. For each chromosome, we generate 150 artificial chromosomes by adding Gaussian noise distributed as $N(0, \sigma^2)$ to each $\log_2 T/R$ where $\sigma = 0.1, 0.2$ and 0.3 and repeating this 50 times for each σ. We then partition each artificial chromosome

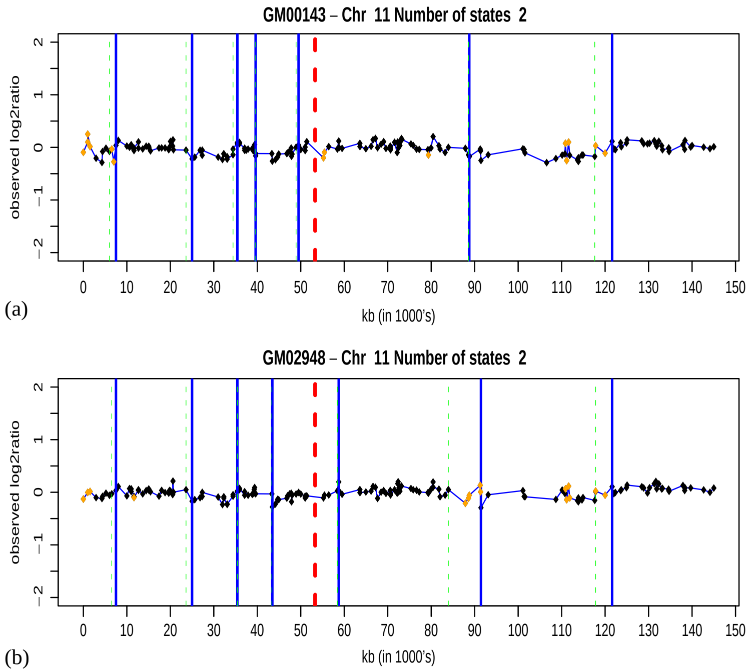

Fig. 5. Coriel cell lines: chromosome 11 artifact. Note the striking similarity between the locations of transitions in the two samples. The pattern is clearly present; however, by repeating the experiment with the same samples on a newer version of the array with expanded clone content did not reproduce this pattern. Thus, this pattern represents an array artifact, and motivates merging step. Note that the difference in median $\log_2 T/R$ ratios between states is less than 0.2, i.e. very small.

into states using the HMM method and choose the number of states with the AIC or BIC criterion with or without merging. For each chromosome and σ we record the number of times that the correct number of states $(= 2)$ was chosen. In addition we calculate the proportion of the clones assigned to a wrong state relative to the true states for that chromosome. The comparison is done by permuting the labels on the assigned states, P^{hmm}, so that there is a maximum overlap with the true states of these clones, P^{true}. More specifically, let R_I denote the set of all permutations of the integers $1, \ldots, I$ (here $I = 2$) and assume that there are L clones on a given chromosome. Find the permutation $\tau \in R_I$ such that

$$\sum_{i=1}^{L} I(\tau(P^{hmm}(\mathbf{x}_i)) = P^{\text{true}}(\mathbf{x}_i)) = \max_{\tau \in R_I} \sum_{i=1}^{L} I(\tau(P^{hmm}(\mathbf{x}_i)) = P^{\text{true}}(\mathbf{x}_i)),$$

where $I(\cdot)$ is the indicator function, equaling 1 if the condition in parentheses is true, and 0 otherwise. Then the misallocation proportion is computed as

$$p^{hmm} = 1 - \sum_{i=1}^{L} I(\tau(P^{hmm}(\mathbf{x}_i)) = P^{\text{true}}(\mathbf{x}_i^b))/L.$$

Finally, we apply the PAM algorithm, with $K = 2$ clusters to the artificial chromosome and in the same way calculate the proportion of the clones misallocated, p^{pam}.

Fig. 6 shows the number of times each HMM-based method picked the correct number of states for various noise levels. There is no difference between the methods when the amount of noise in the data is small ($\sigma = 0.1$); however, the superiority of AIC criterion combined with merging is apparent as the data becomes noisier. It is intuitively clear that it is advantageous to first choose a larger model and then reduce the number of states; i.e., we first choose a model according to a model selection criterion with a tendency to overfit, such as AIC, and then follow it with the backwards deletion step.

The result that AIC combined with merging tends to choose the correct number of states most frequently does not ensure that the clones are allocated to the correct states. We stratify the misallocation rate by the number of states picked by an HMM model: 2 or >2. In each stratum we compare misallocation rates of the HMM models and PAM.

Fig. 7 shows the disagreement rate for all chromosomes and noise levels combinations. The advantage of utilizing dependency in the data becomes apparent when comparing results of PAM with $K = 2$ clusters and HMM methods: PAM has a much higher misallocation rate as evident by the distributions shown with boxplots. Also, misallocation rate is very similar for all HMM methods given the number of states chosen. Thus, we can conclude that, indeed, AIC criterion with merge is more accurate than the other methods evaluated.

3.2. *MMR cell lines*

Genomes of tumors with defects in MMR have few chromosomal changes compared to most solid tumors, which are aneuploid with numerous chromosomal aberrations. These mainly cytogenetic observations agree with the expectation that MMR-deficient cells exhibit a higher frequency of nucleotide aberrations, resulting in a higher probability that growth promoting alterations will occur by mutations in genes whose nucleotide sequences are susceptible to errors normally corrected by MMR. Here, we indirectly tested our HMM approach to the array CGH analysis by confirming known cytogenetic results on MMR cell lines.

For each genomic profile we counted the number of gains or losses of whole chromosomes, copy number transitions within a chromosome, focal aberrations and number of chromosomes containing high level focal amplifications (see Fig. 8). In agreement with previous studies, we confirmed that MMR-deficient cells exhibited significantly fewer alterations relative to MMR-proficient cells in accord with earlier observations. However, we observed a substantial number of alterations in some MMR-deficient lines. We performed two-sided Wilcoxon rank tests to compare the number of alterations by type between MMR deficient and MMR proficient cell lines and obtained significant differences (wilcoxon rank test p-value <0.05) for comparison of copy number transitions, focal aberrations and number of

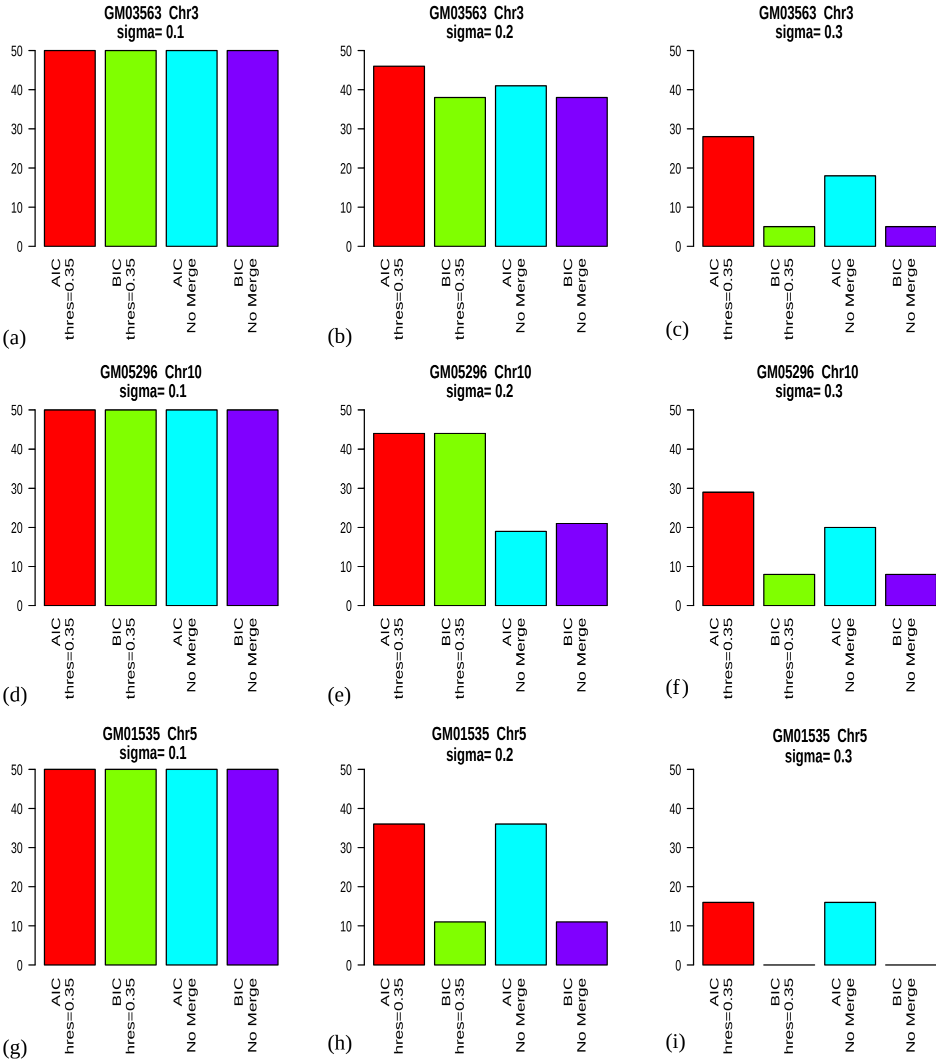

Fig. 6. Number of times the correct number of states is picked by each model. Fifty artificial chromosomes were generated for each of 3 × 3 combination. The performance of all four methods (AIC or BIC criterion with or without state merge) is equivalent (perfect) with small amount of noise ($\sigma = 0.1$). As the amount of noise increases, *AIC* criterion with state merge appears as a uniform winner with the largest number of times that correct number of states are picked.

chromosomes containing one or more amplifications. The distribution of whole chromosomal changes was the same in all MMR subtypes. We also found a possible dependency of aberration type on the specific MMR defect. Cells deficient in MLH1 had a higher frequency of transitions, focal aberrations and amplifications than MSH2 deficient cells.

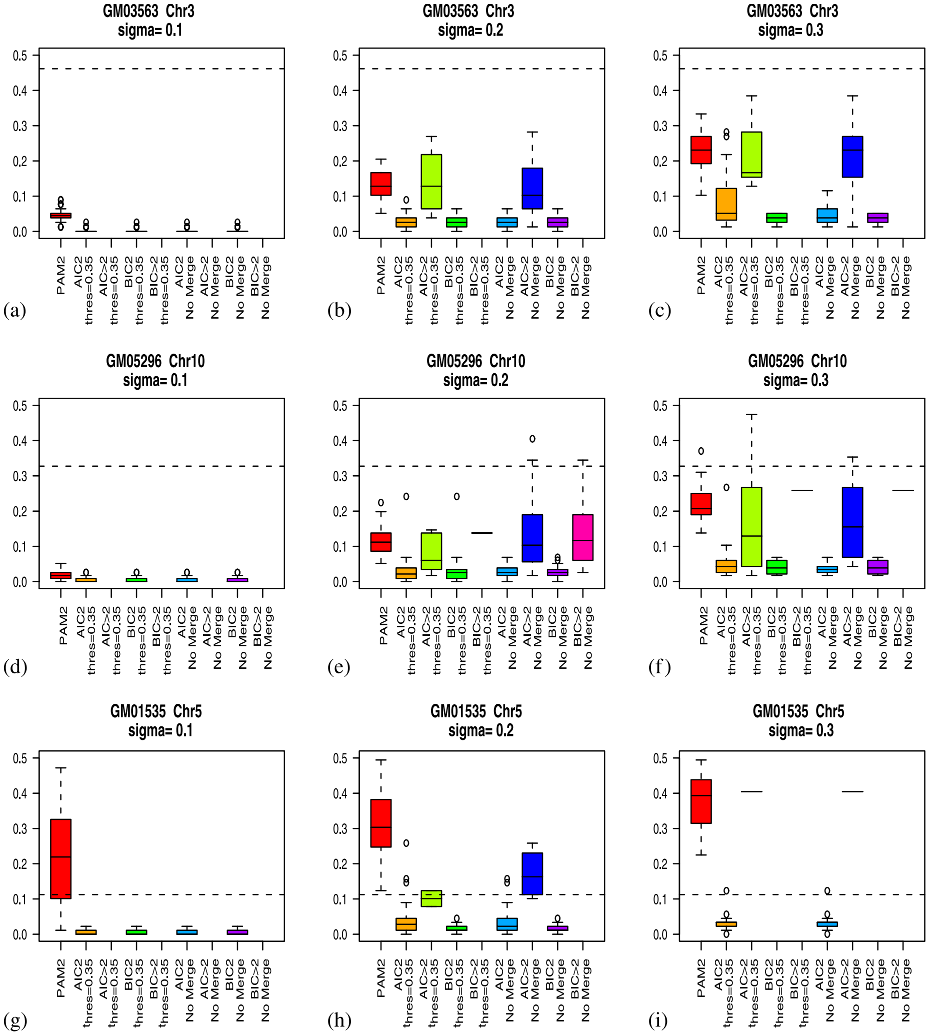

Fig. 7. Distribution of misallocation rates produced by different methods. The dotted horizontal line indicates misallocation rate if all clones were placed in the same state. We show distributions of all 3 × 3 combinations and all models stratifying by whether 2 or 3+ states were chosen by a corresponding HMM model. The performance of all HMM methods is very close given number of states chosen. The misallocation rate for the PAM method with 2 clusters underperforms all HMM methods and is frequently higher than the misallocation rate would be if all clones were placed in the same state.

4. Discussion, conclusions and work in progress

We have developed a method for automatic identification of structural abnormalities in tumor genomes using array CGH data. Here, we apply this method to two data sets and obtain results that clearly show that our approach is

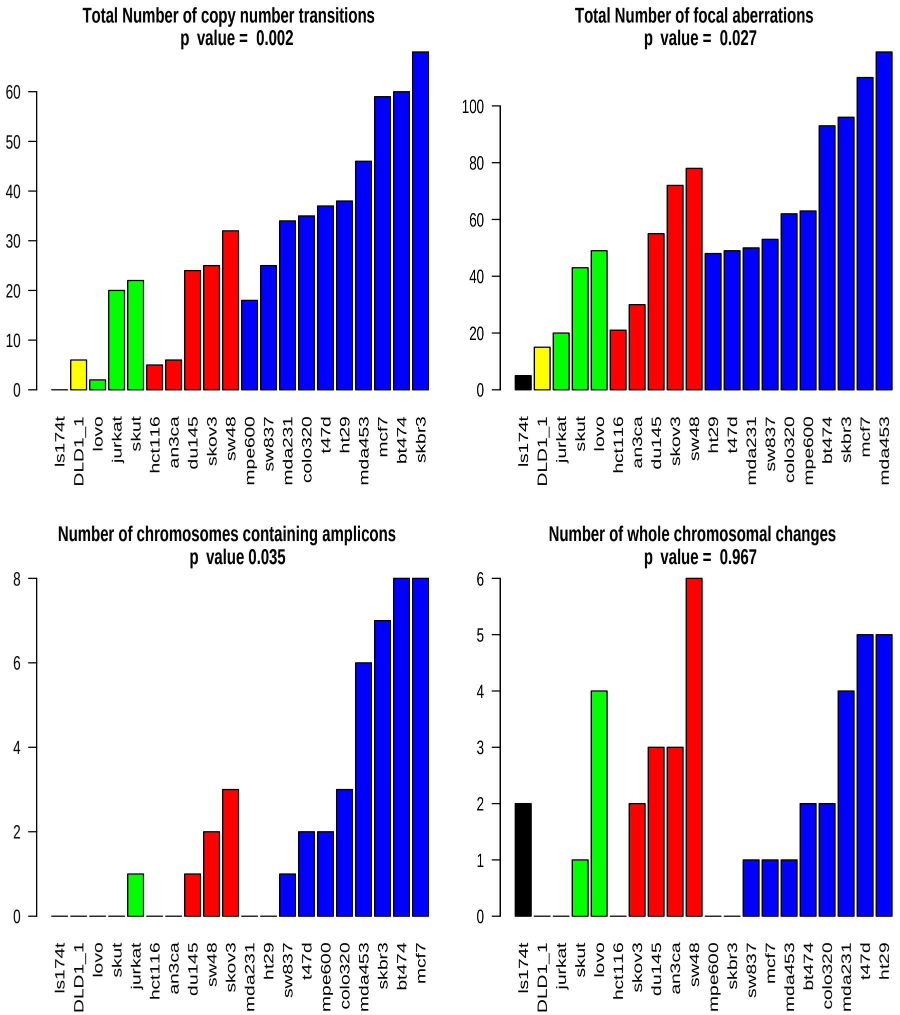

Fig. 8. Distribution of aberrations in MMR cell lines. Black refers to unknown MMR deficiency, yellow is MSH6, green is MSH2, red is MLH1 deficiency; and, finally, blue shows MMR proficient cell lines. The *p*-value is computed using the 2-sided wilcoxon rank sum test for equality between deficient and proficient cell lines. Proficient cell lines contain significantly more aberrations of each type except for whole chromosomal changes. MLH1 deficient cell lines have more alterations that MSH2/MSH6 deficient ones.

finding real alterations and is useful for classifying genomic profiles. Currently we are applying this methodology to primary breast cancer data to investigate the relationship of the frequencies and types of copy number alterations with defects in mechanisms that preserve genomic integrity.

Quantitative analysis of microarray data needs to include knowledge of the underlying biological processes. One of the main difficulties of developing quantitative methods for microarray data in general and array CGH data in particular is the lack of data sets with available ground truth. Thus it is not possible to make decisions only on a strictly mathematical basis. The appropriateness of the developed methodology can only be known in the long run on the grounds that the conclusions demonstrate utility for improving biological understanding and clinical decisions. For example, identification of an amplicon at 20q13.2 has led to a discovery of the novel oncogene, ZNF217 [4]. With the abundance of incoming array CGH data, the need for an automatic reliable algorithm is substantial, and the unsupervised HMM approach described in this paper is one possible algorithm for meeting the growing demand for interpreting genomic profiles. While the definition of alteration types is subjective and is based on the suggestions and experience of biologists, this summary approach to the profiles does correlate with the existing clinical phenotypes. Note that single clone aberrations need to be interpreted with extreme caution and whenever possible confirmed using a replicate hybridization or an alternative technique.

One interesting question related to biological mechanism is whether transitions occur at the identical locations in multiple samples. Recurrent locations may indicate *fragile sites* in the genome (heritable sensitive regions of chromosomes associated with chromosome breakage and other aberrations) or point to the mechanism of chromosomal abnormality. We are currently investigating recurrent locations of the transitions in primary breast tumors.

General genomic instability has been suggested to be associated with bad prognosis for cancer patients [7]. The overall number of genomic aberrations, possibly stratified by subtypes, can serve as a good measure of genomic complexity of a given tumor. We have already seen in prostate and breast primary tumor data sets that there is indeed a relation between the number of aberrations and time to recurrence or death (unpublished data).

We hope to further develop our approach by incorporating distance between clones in the transition probability and, thus, allowing for the analysis of arrays with uneven distribution of clones across the genome. We are also working to reduce reliance on user-defined thresholds and develop more objective methods for merging states and for declaring chromosomal gain or loss.

We would also like to adapt the output of the HMM partitioning method to more standard discrimination tasks such as clustering and discrimination. An interesting application of HMMs to the analysis of gene expression time course data was proposed by Schliep et al. [16]. There they used a model-based approach to clustering the time series where each cluster is represented by an HMM. A similar idea may work with the array CGH profiles.

The HMM approach may also help with identifying individual clones or regions with differential copy number between groups of interest by reducing the test multiplicity. For instance, each clone can be assigned the predicted value of its state thus reducing the dimensionality of the problem and possibly making it easier to discover interacting clusters of clones located physically apart on the genome.

5. Software

All the analyses in the manuscript were performed using open source R packages [8] which may be downloaded from the Comprehensive R Archive Network (http://cran.r-project.org) or the Bioconductor Web site (http://www.bioconductor.org).

Hidden Markov models: The `hidden` function from the `repeated` package can be used for fitting of discrete HMMs. The package can be downloaded from J. Lindsey website www.luc.ac.be/˜jlindsey/rcode.html.

Partitioning around medoids: The `pam` function from the `cluster` package is used for partitioning around medoids clustering.

Characterizing genomic alterations: The set of utility functions for finding alterations is available from the authors on request.

Acknowledgments

The authors are grateful to Terry Speed for his suggestion of utilizing hidden Markov models to exploit spatial dependence of the observations, to Adam Olshen for introducing the first author (JF) to the issues arising in the analysis of the array CGH data and to David Nelson for providing comments for the manuscript.

References

[1] H. Akaike, Fitting autoregressive models for prediction, in: Annals of the Institute of Statistical Mathematics, Kluwer Academic Publishers, Dordrecht, 1969, pp. 243–247.

[2] D.G. Albertson, Profiling breast cancer by array CGH, Breast Cancer Res. Treat. 78 (2003) 289–298.

[3] D.G. Albertson, D. Pinkel, Genomic microarrays in human genetic disease and cancer, Hum. Mol. Genet. 12 (2003) 145–152.

[4] C. Collins, J.M. Rommens, D. Kowbel, T. Godfrey, M. Tanner, S. Hwang, D. Polikoff, G. Nonet, J. Cochran, K. Myambo, K.E. Jay, J. Froula, T. Cloutier, W.-L. Kuo, P. Yaswen, S. Dairkee, J. Giovanola, G.B. Hutchinson, J. Isola, O.-P. Kallioniemi, M. Palazzolo, C. Martin, C. Ericsson, D. Pinkel, D. Albertson, W.-B. Li, J.W. Gray, Positional cloning of znf217 and nabcl: genes amplified at 20q13.2 and overexpressed in breast carcinoma, Proc. Natl. Acad. Sci. USA 95 (1998) 8703–8708.

[5] K.V. Desai, N. Xiao, W. Wang, L. Gangi, J. Greene, J.I. Powell, R. Dickson, P. Furth, K. Hunter, R. Kucherlapati, R. Simon, E.T. Liu, J.E. Green, Initiating oncogenic event determines gene-expression patterns of human breast cancer models, Proc. Natl. Acad. Sci. USA 10 (2002) 6967–6972.

[6] G. Hodgson, J.H. Hager, S. Volik, S. Hariono, M. Wernick, D. Moore, D.G. Albertson, D. Pinkel, C. Collins, D. Hanahan, J.W. Gray, Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas, Natur. Genet. 929 (2001) 459–464.

[7] J. Hu, V. Khanna, M.W. Jones, U. Surti, Comparative study of primary and recurrent ovarian serous carcinomas: comparative genomic hybridization analysis with a potential application for prognosis, Gynecol. Oncol. 89 (3) (2003) 369–375.

[8] R. Ihaka, R. Gentleman, R: a language for data analysis and graphics, J. Comput. Graphical Statist. 5 (1996) 299–314.

[9] A.N. Jain, T.A. Tokuyasu, A.M. Snijders, R. Segraves, D.G. Albertson, D. Pinkel, Fully automatic quantification of microarray image data, Genome Res. 12 (2002) 325–332.

[10] K. Jong, E. Marchiori, A. van der Vaart, B. Ylstra, G. Meijer, M. Weiss, Chromosomal breakpoint detection in array comparative genomic hybridization data, in: In Applications of Evolutionary Computing: Evolutionary Computation and Boinformatics, Vol. 2611, Springer, Berlin, 2003, pp. 54–65.

[11] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley, New York, 1990.

[12] A.B. Olshen, E.S. Venkatraman, Change-point analysis of array-based comparative genomic hybridization data, in: Proceedings of the Joint Statistical Meetings, 2002, pp. 2530–2535.

[13] D. Pinkel, R.D. Segraves, S.C. Sudar, I.P.D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B.M. Ljung, J.W. Gray, D.G. Albertson, High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarray, Natur. Genet. 20 (1998) 207–211.

[14] J.R. Pollack, T. Sorlie, C.M. Perou, C.A. Rees, S.S. Jeffrey, P.E. Lonning, R. Tibshirani, D. Botstein, A.L. Borresen-Dale, P.O. Brown, Microarray analysis reveals a major direct role of DNA copy number alteration in the transriptional program of human breast cancers, Proc. Natl. Acad. Sci. USA 99 (2002) 12963–12968.

[15] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, in: Proceedings of the IEEE, Vol. 77, February 1989, pp. 257–285.

[16] A. Schliep, A. Schonhuth, C. Steinhoff, Using hidden Markov models to analyze gene expression time course data, Bioinformatics 19 (2003) 255–263.

[17] Schwarz, G., Estimating the dimension of a model, Ann. Statist. (1978) 461–464.

[18] A.M. Snijders, J. Fridlyand, D. Mans, R. Segraves, A.N. Jain, D. Pinkel, D.G. Albertson, Shaping of tumors and drug-resistant genomes by instability and selection, Oncogene 22 (2003) 4370–4379.

[19] A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, N. Conroy, G. Hamilton, A.K. Hindle, B. Huey, K. Kimura, S. Law, K. Myambo, J. Palmer, B. Ylstra, J.P. Yue, J.W. Gray, A.N. Jain, D. Pinkel, D.G. Albertson, Assembly of microarrays for genome-wide measurement of DNA copy number, Natur. Genet. 29 (2001) 4281–4286.

[20] F.M. Waldman, S. DeVries, K.L. Chew, D.H. Moore, K. Kerlikowske, B.M. Ljung, Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences, J. Natl. Cancer Instit. 92 (2000) 313–320.

[21] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed, Normalization for CDNA microarray data : a robust composite method addressing single and multiple slide systematic variation Nucleic Acids Res. 30 (2002) e15.