Published in final edited form as:
IEEE/ACM Trans Comput Biol Bioinform. 2013 ; 10(6): 1422–1431.

Novel multi-sample scheme for inferring phylogenetic markers from whole genome tumor profiles

Ayshwarya Subramanian,
Graduate student at the Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, 15213. ayshwarya@cmu.edu

Stanley Shackney, and
President of Oncotherapeutics, Pittsburgh PA 15243. sshackney@verizon.net

Russell Schwartz
Professor of Biological Sciences at Carnegie Mellon University. russells@andrew.cmu.edu

Abstract

Computational cancer phylogenetics seeks to enumerate the temporal sequences of aberrations in tumor evolution, thereby delineating the evolution of possible tumor progression pathways, molecular subtypes and mechanisms of action. We previously developed a pipeline for constructing phylogenies describing evolution between major recurring cell types computationally inferred from whole-genome tumor profiles. The accuracy and detail of the phylogenies, however, depends on the identification of accurate, high-resolution molecular markers of progression, i.e., reproducible regions of aberration that robustly differentiate different subtypes and stages of progression. Here we present a novel hidden Markov model (HMM) scheme for the problem of inferring such phylogenetically significant markers through joint segmentation and calling of multi-sample tumor data. Our method classifies sets of genome-wide DNA copy number measurements into a partitioning of samples into normal (diploid) or amplified at each probe. It differs from other similar HMM methods in its design specifically for the needs of tumor phylogenetics, by seeking to identify robust markers of progression conserved across a set of copy number profiles. We show an analysis of our method in comparison to other methods on both synthetic and real tumor data, which confirms its effectiveness for tumor phylogeny inference and suggests avenues for future advances.

Index Terms

Biology and genetics; Health; Trees; Segmentation

1 Introduction

Analysis of cancer genomes using high-throughput genomic methods has revealed a high degree of variability at the genetic level [1], [2] in otherwise histopathologically indistinguishable tumors. In the process, such analyses have identified specific molecular markers and pathways associated with the onset and progression of cancers in specific tissue types, classification of molecular subtypes and patient sub-populations [3], [4]. This information can inform the design of targeted therapeutics and diagnostic strategies [5]. Analysis of tumor progression has nonetheless been hindered by high heterogeneity in tumor progression pathways, even in tumors impacting similar regulatory pathways and biological processes [6], [7]. Heterogeneity can occur between patients or within a single patient, where sub-populations of cells may correspond to different states or even pathways of tumor progression. Computational cancer phylogenetics provides a strategy for making sense of

the complexity of tumor evolution by identifying recurring pathways of tumor evolution both within and across patients through the use of phylogenetic inference algorithms. Such methods, however, require some mechanism for identifying discrete states of progression and estimating evolutionary distances among them. In the case of character-based phylogeny approaches, this process involves identifying robust markers of progression whose presence or absence can be used to track tumor evolution. In the present work, we focus specifically on the problem of marker inference from array comparative genomic hybridization (aCGH) data providing genome-scale DNA copy number measurements. For these data, the problem corresponds to finding discrete genomic regions of DNA gain or loss that can serve as markers of tumor progression.

Existing methods for aCGH analysis include algorithms for smoothing, segmentation and combined segmentation and classification of both single- [8], [9], [10], [11], [12], [13] and multi-sample data [14], [15], [16], [17], [18], [19], [20]. Such methods can be highly effective at identifying discrete copy number variations in such data, but are poorly suited to the problem of phylogenetic inference because they do not constrain solutions to common markers across tumor samples. They thus provide no straightforward way to infer a set of robust markers with defined boundaries across patients and progression states for use in phylogenetic inference. A similar objective was considered by Picard et al. for their method, CGHSeg [21], which addresses the problem of joint segmentation and calling of multiple samples primarily as a way of improving accuracy of assignment using similarities between data. This method, though, was also not designed for the purpose of phylogenetic inference, and is inefficient for the data characteristics needed for these purposes, especially the combination of large numbers of markers with defined boundaries across a modest number of discrete samples characteristic of whole-genome datasets.

Our method is distinguished from other methodologically similar segmentation methods for CGH data primarily in that it is designed specifically to facilitate phylogenetic inference from tumor samples. We favor character-based phylogenetic methods, which allow us to intepret evolution of tumors in terms of gain or loss of specific discrete amplicons. For such inferences, we must interpret raw copy number data as sets of phylogenetic characters for which we can assign discrete states to each sample in a data set. To be useful for phylogenetic inference, such characters must describe common regions of copy number change that are shared across multiple samples. Hence, it is essential for our purposes to have a joint segmentation and calling algorithm that can output discrete phylogenetic character data. Typical segmentation algorithms, which seek only to find most plausible explanations of the raw data in terms of regions of amplification or loss, are unlikely to produce segmentations that yield common shared regions of gain or loss across samples. As detailed in Approach, our method involves a variety of innovations designed to improve its ability to find shared markers across samples useful for phylogenetic inference. In Results and Discussion, we show using simulated aCGH data that these innovations lead to an improved ability over prior methods to find markers and call them accurately in individual samples and that these improvements in marker detection translate to improved ability to reconstruct phylogenetic trees.

In previous work, we developed an approach to the problem of tumor phylogenetics based on the use of mixture models to infer discrete states of progression recurrent across tumor samples [22], [23]. We subsequently used this mixture modeling approach as the basis for a pipeline for tumor phylogeny inference [24]. For this pipeline, we developed a multi-sample segmentation method based on a simple statistical test applied to fixed-length windows of probes heuristically merged to identify amplicons from a set of inferred mixture components. The unmixing procedure in its present formulation can only reliably infer amplifications and, hence, we focus only on copy number amplifications in this work. An

additional statistical test would then call presence or absence of each amplicon in each component, converting the components into discrete character arrays suitable for character-based phylogenetic inference. Validation on a set of components derived from real breast tumor data [25] showed the marker selection method to be reasonably effective at finding known breast cancer amplicons suitable for use as phylogenetic markers. The segmentation step, however, showed a poor ability to resolve fine-scale structure within amplicons, limiting the number of phylogenetic markers and the ability of the method to discriminate between subtle changes in nearby markers. In addition, separating segmentation from calling left no way to guarantee that amplicons detected in the segmentation stage would in fact be called differently in different components and thus become useful markers for phylogenetics.

The present work is aimed at developing an improved marker detection method designed to maintain the advantages of our prior work in using multi-sample segmentation from mixture components to identify a robust set of common markers usable across samples, while adapting ideas from prior single-sample methods to improve fine-scale resolution of amplicon structure. The method uses a novel HMM scheme to do joint segmentation and calling of markers simultaneously from a set of mixture components. It is thus similar in character to the method of Picard et al. [21] although with fewer assumptions about shared features of amplicons across samples. Both FLLat [20] and the HMM-mix model in [15] deal with the issue of heterogeneity inference in multi-sample aCGH data through mixture modeling. The outputs are not directly suited for phylogeny analysis of a set of input samples as they consist of representative driver aberration profiles, similar to the outputs of our mixture models, rather than phylogenetic characters derived from those aberration profiles as in the present work. Other HMM-based methods [12], [13] are either single-sample based, primarily platform-specific or focus on other issues of multi-sample analysis. Our new approach allows joint segmentation and thus detection of phylogenetically useful markers across mixture components. In contrast to our prior work, the use of the HMM scheme also allows the method to detect changes in assortments of amplicons across components within regions of amplification. We analyze the method on both simulated and real data and compare it to related methods heuristically adapted to the problem of phylogenetic calling. The results show the method to give superior performance at both marker inference and phylogenetic reconstruction for biologically reasonable levels of experimental noise.

2 Approach

Our model is based on a generalization of the use of HMMs to multi-sample data for the purpose of finding a common marker set across a set of samples. It accomplishes this task by treating states of the HMM as tuples of amplification states across samples, with each copy number probe assigned one state. Any contiguous region of common state in which at least one component is called amplified can then serve as a single marker for phylogenetic inference.

2.1 The HMM model

2.1.1 Notation—Let the data $\mathbf{X}$ consist of m samples, each sample being a vector of log copy number intensity ratios at n genomic coordinates. We assume each of the m copy number profiles are ordered in genome coordinates starting from chromosome 1 to chromosome 22 and potentially X and Y. Thus $\mathbf{X}$ is a $m \times n$ data matrix where each element x_{ij} is a copy number ratio in the log domain where $i \subset \{1,2,\ldots m|$ and $j \subset \{1,2,\ldots n\}$. A Hidden Markov model defines the joint probability distribution of the sequence of x_{ij} in the observed matrix $\mathbf{X}$ by using another latent or hidden sequential state set. The HMM divides $\mathbf{X}$ into k distinct segments $\mathbf{S}$ where $k << n$ and each segment s_t is assigned one of the

possible hidden copy number states defined below and $t \subset \{1,2,\ldots k\}$. Each s_t is made up of as many members x_{ij} as its length. We denote by s_{at} an element x_{ij} that belongs to segment t of length l and is at position a in the segment where $t \subset \{1,2,\ldots k\}$ and $a \subset \{1,2,\ldots l\}$. An illustration of our model is shown in Fig. 1

We assume no linkage disequilibrium between the x_{ij}s and they are hence assumed mutually independent for all j. Further, we do not take into account whether the individuals are heterozygous or homozygous at each x_{ij}. We also note that as a preprocessing step, we smooth input data by replacing each probe value with the average over a window of five consecutive probes centered on that value.

2.1.2 Hidden State Space—We assume two possible copy number states for each x_{ij} : normal or aberrated (loss/gain). The normal state is indicated by 0 and aberrated by 1. The copy number states can be further assigned ploidy defintions whereby the normal state is thought of as being diploid and the aberrated state is aneuploid. Then for any position i, the hidden state is a binary vector $\mathbf{H}_i$ of size m where each element h_i is either 0 or 1 and $i \subset \{1,2,\ldots m\}$. Each $\mathbf{H}_i$ is thus one of 2^m possible state vectors in this 2-state paradigm. We, however, believe that the optimum segmentation of a dataset will normally be defined by fewer than 2^m combinations of unique state vectors. The assumption of n-tuples over {0,1} for n samples is particularly useful for character-based phylogenetic methods where the data must be represented as discrete states across markers.

2.1.3 Parameters—By definition, the sequence of states in the HMM follows a Markov model with transition probabilities defined between each pair of states. We assume the Markov model to be ergodic. Because our goal is to produce a phylogenetically useful set of amplicons rather than to infer the true amplicon structure per se, we do not learn model parameters directly from the data. Rather, we seek a model that will favor a simpler representation of the amplicon structure specifically preferring fewer and longer amplicons and preferentially finding amplicons with shared boundaries across samples. For this reason, we build into the model a prior expectation of the approximate frequency and length of amplicon expected, encoded in the HMM transition probabilities as follows:

1. Transition Probabilities (**A**)

   The Markov model underlying the HMM is described in Figure 1.

   As explained above, the basic Markov model has two possible states for each x_j : normal or 0 (N) and aberrated or 1 (A). We define four possible transitions:

   a. p_{NN}: The probability of staying in the normal state.

   b. p_{NA}: The probability of going from the normal state to an aberrant state.

$$p_{NA} = \left(\frac{p}{n*m}\right)\left(\frac{1}{2^m - 1}\right)$$

   where p is a penalty set to 0.001 in the present work, effectively penalizing the model for assigning large numbers of amplicons by creating a prior expectation of 0.001 amplicons occurring by chance across the entire data set. The value of 0.001 was chosen to act comparably to a p-value of 0.001 used in statistical approaches to this problem, effectively requiring a 1000-fold excess in likelihood for amplicon versus no amplicon to identify a region as amplified.

c. p_{AA}: The probability of going from an aberrant state to another aberrant state (or to itself; the possibilities are assumed to have the same transition rates). We set $p_{AA}=\frac{w-1}{w}$ to enforce an average amplicon width *w*, where we assume in the present work that *w* = 20. The other two transition probabilities are then fixed by p_{AA} and p_{NA}.

d. p_{AN}: The probability of going from an aberrant state to normal.

$$p_{AN}=1-(2^m-1)*p_{AA}$$

and

$$p_{NN}=1-(2^m-1)*p_{NA}$$

which is derived by subtracting the probability of going to all other $2^m - 1$ aberrant states.

2. Emission Probabilities (**O**)

Estimating Empirical Noise Levels: Before we define the emission probabilities, we introduce a measure to determine noise in copy number data that exploits the spatial dependence of the data. Empirical results on real aCGH datasets show that the data is log-Laplacian distributed [23], but we can adopt the approximation of this distribution as log-normal, modeling log copy number data as a true signal with additive Gaussian noise:

$$X_{ij}=S_{ij}+\mathcal{N}(0,\sigma^2)$$

where S is the signal. This log-normal model is commonly used for modeling aCGH data [20]. We introduce a non-standard formulation for inferring the noise in this framework that takes into consideration the spatial distribution of the probes. We developed an estimator of variance or, equivalently, standard deviation σ based on the average difference between adjacent probe values. We can pose this estimate in terms of the expectation of the difference between two normal random variables:

$$\begin{aligned}
&\sum_{i,j}\frac{|X_{i,j}-X_{i,j+1}|}{m*(n-1)}\\
&=E[|N_{i,j}(\mu,\sigma^2)\\
&\quad -N_{i,j+1}(\mu,\sigma^2)|]\\
&=E[|\mu\\
&\quad -\mu|]\\
&+E[|N_{i,j}(0,\sigma^2)\\
&-N_{i,j+1}(0,\sigma^2)|]=\sigma E[|N_{i,j}(0,1)\\
&\quad -N_{i,j+1}(0,1)|]\\
&=\sigma E[|N(0,2)|]\\
&=\sqrt{2}\sigma E[|N(0,1)|]\\
&=2\sqrt{2}\sigma\int_0^\infty\frac{x}{\sqrt{2\pi}}\exp\frac{-x^2}{2}dx\\
&=\frac{2\sigma}{\sqrt{\pi}}\int_0^\infty\exp(\\
&-u)du(u\\
&\quad =x^2/2)\\
&=\frac{2\sigma}{\sqrt{\pi}}-\exp(-u)|_0^\infty=\frac{2\sigma}{\sqrt{\pi}}
\end{aligned}$$

Therefore:

$$\sigma=\sum_{i,j}\frac{\sqrt{\pi}|X_{i,j}-X_{i,j+1}|}{2m(n-1)}$$

This non-standard formula is used, rather than the conventional estimate of standard deviation, $\sqrt{E[X^2]-E[X]^2}$, in order to better separate variance due to measurement noise, which we wish to model, and true variance in the signal due to different amplicon copy numbers, which we do not want included in the noise model.

To illustrate the difference between the two measurements, we can use a model of DNA drawn from a genome with amplified segments, where we assume for illustration a fixed segment length *L* with alternating amplification levels of 0 and *K* for some *K*, here simplifying by assuming no true measurement noise. In the limit of an infinite number of segments, the standard estimator would measure variance to be:

$$E[X^2]-E[X]^2=\frac{K^2}{2}-\left(\frac{K}{2}\right)^2=\frac{K^2}{2}-\frac{K^2}{4}=\frac{K^2}{4}$$

and thus standard deviation to be *K*/2.

Our estimator, on the other hand, would add contributions to the estimate only at boundaries between segments, giving for a single genome of infinite length an

estimated standard deviation of $\frac{K\sqrt{\pi}}{2L}$ variance of $\frac{K^2\pi}{4L^2}$. In general, then, our estimator will suppress spurious estimates of standard deviation of the noise due to true amplification by a factor proportional to the average amplicon length. Our expectation is that this will lead to more accurate estimates of the parameter σ of our noise model for real data than will a straightforward measurement of standard deviation of the data.

We can bound variance of the noise estimator under the assumption that the input is a stream of *n* i.i.d. normal random variables, corresponding to consecutive probes, by noting that the estimator would then be described by a random variable of the form

$$\frac{\sqrt{\pi}}{2n}\left(\sum |Z_j - Z_{j+1}|\right)$$

where each Z_j is assumed to be an independent $N(0,\sigma^2)$ random variable. The variance in the estimator would then be given by:

$$\frac{\pi}{4n^2}\mathrm{Var}\left(\sum |Z_j - Z_{j+1}|\right)$$

This in turn is given by

$$\frac{\pi}{4n^2}((n-1)\mathrm{Var}(|Z_j - Z_{j+1}|) - (n-2)\mathrm{Cov}(|Z_j - Z_{j+1}|, |Z_{j+1} - Z_{j+2}|))$$

for some arbitrary $1 < j < n$. That expression can be bounded as follows:

$$\frac{\pi}{4n^2}((n-1)\mathrm{Var}(|Z_j - Z_{j+1}|) - (n-2)\mathrm{Cov}(|Z_{j+1} - Z_j|, |Z_{j+1} - Z_{j+2}|)) \leq \frac{\pi}{4n^2}(n-1)\mathrm{Var}(|Z_j - Z_{j+1}|)$$
$$= \frac{\pi}{4n^2}(n-1)\mathrm{Var}(|N(0,\sigma^2) - N(0,\sigma^2)|)$$
$$= \frac{\pi\sigma^2}{4n^2}(n-1)\mathrm{Var}(|N(0,1) - N(0,1)|)$$
$$= \frac{2\pi\sigma^2}{4n^2}(n-1)\mathrm{Var}(|N(0,1)|)$$
$$= \frac{2\pi\sigma^2}{4n^2}(n-1)(E[|N(0,1)|^2] - E[|N(0,1)|]^2) = \frac{2\pi\sigma^2}{4n^2}(n-1)\left(E[\chi_1^2] - \left(\sqrt{\frac{2}{\pi}}\right)^2\right) = \frac{\pi\sigma^2}{2n^2}(n-1)\left(1 - \frac{2}{\pi}\right) \approx \frac{0.6\sigma^2}{n}$$

The variance of our estimator can thus be bounded by a term that falls approximately linearly with the number of probes, *n*, which can be expected to yield accurate estimates of σ for genome-scale data. We empirically validate the performance of the estimator in the Results and Discussion below. *Defining Emission Probabilities:* Once we have an estimate of the noise level, we define emission probabilities *O* by assuming each measured copy number x_{ij} comes from either a normal diploid distribution or an aberrant aneuploid distribution:

$$P(O_d|H) = \phi(x;\mu_d,\sigma) \text{ and } P(O_a|H) = \phi(x;\mu_a,\sigma)$$

where we assume here that diploid data has a mean $\mu_d = 0 + \mu$, where 0 corresponds to a mean ratio of one between observed data and a diploid control in the log-domain, and aneuploid data is modeled as having a mean ratio $\mu_a = 1 + \mu$ relative to a diploid control. The additive term μ is an empirically estimated mean of the data,

used to control for overall background amplification that may arise due to overall signal aneuploidy or as an artifact of the unmixing process.

3. Initial State Probabilities (π)

   The initial state probability π for all aberrated states is assumed to be $q = (p/(2^m - 1)/n)$ leaving an initial probability of the normal state of $1 - (2^m - 1)q$.

2.2 Selection of Optimal States

We employ an extension of the Viterbi algorithm to determine the optimal sequence of copy number states for a given multisample copy number data set, assigning amplification or normal condition to each sample at each probe. A state here is defined, as above, as a tuple of binary normal/amplification assignments for all samples at a single probe. Our method differs from the generic Viterbi algorithm only in that our outputs are real-valued copy number measurements, rather than a discrete set of output characters, and our emission probabilities are thus drawn from log normal distributions to allow for continuous values. This extension still allows for optimal solution of the log likelihood via dynamic programming, as with Viterbi over a discrete state set. More specifically, we find a maximum likelihood solution H of hidden state assignments by optimizing for the subproblem $\hat{H}(i,j)$, defined to be the maximum likelihood assignment of amplification states to the first i probes terminating in state b_j, for some canonical ordering of amplification vectors $b_0,\ldots,b_{2m}-1$ where b_0 is defined to be the all-diploid vector.

We solve this problem using the recurrence:

$$\hat{H}(i,j) = \max_k \begin{cases} \hat{H}(i-1,0)_{p_{nn}} \prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j=0 \\ \hat{H}(i-1,0)_{p_{an}} \prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j \neq 0 \\ \hat{H}(i-1,k)_{p_{na}} \prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j=0, k \neq 0 \\ \hat{H}(i-1,k)_{p_{aa}} \prod_{l=1}^{m} P(x_{il}|b_{jl}) & : \quad j \neq 0, k \neq 0 \end{cases}$$

where x_{il} is the observed copy number of probe i in sample l and b_{jl} is the binary amplification state of sample l in state j. The optimal assignment is then derivable by identifying $\max_k \hat{H}(n, k)$ and backtracking to reconstruct the full state assignment.

The above recurrence relation admits a dynamic programming algorithm with runtime $\mathcal{O}(2^{2m}n)$. The resulting algorithm was implemented in MATLAB.

3 Experimental Methods

3.1 Synthetic Data

To assess accuracy on data of known ground truth, we simulated a series of aCGH data sets across a range of assumed experimental noise levels. We assumed a log-normal noise model $Y_{ij} = Mij + 7V(0, \sigma)$ for each sample i (i= 1, 2,…, m) and aCGH probe position j (j= 1, 2,…, n). Here, Yij is the simulated copy number ratio in the log domain, M_{ij} is the amplification model and $7V(0, \sigma)$ is Gaussian noise. We modeled the distribution of copy numbers in tumor data by an exponential distribution $M_{ij} = 1 + \mathbb{1}\ (j \subset\subset Si)\text{Exp}(\lambda)$ where $\mathbb{1}$ is the indicator function for the presence of site j in an amplicon Si. We estimated the exponential rate λ from the real component data in Sec. 3.2 using the mean of observed probe values above 5, to minimize contamination by non-amplified probes. We then simulated a series of components to model tumor evolution over a complete binary tree of depth three. Beginning from an all-diploid root, we simulated amplicons of fixed width w = 20 in a hypothetical data set of 1161 probes (to match the proportion of amplifications in the real data) in 6

components, adding one new amplicon per non-root node to those present in the node's parent to model acquisition of successive amplicons over succeeding generations of progression. Amplicons were placed uniformly at random within the genome, rejecting and rerunning any placement that resulted in two amplicons within *w* probes of one another. We then generated observed signal values for amplified and non-amplified sites by the log-normal noise model described above. This process was repeated for 200 replicates each at noise levels $\sigma = 0$ to 1.8 in increments of 0.1.

Because our method uses an estimate of noise level derived from the data, we perform a preliminary validation of our estimates of noise level on the simulated data. Specifically, at each noise level, we apply our estimator of noise standard deviation σ to the data and evaluate its inferred value and standard deviation of that value by our estimator and a generic standard deviation computation.

For each replicate, we ran the HMM algorithm as described in Sec. 2. For comparison, we tested the same data on two alternatives: the single-sample method Circular Binary Segmentation (CBS) [26] using the MATLAB function *cghcbs* and the multisample *multiseg* function in the R package CGHSeg [21]. While there is no comparative method developed specifically for phylogenetics, we chose to compare with one single-sample and one multi-sample copy number segmentation algorithm. The CBS output was called at a threshold of $log_2(1.5)$ as amplified or normal. CGHSeg returns called values for each sample. Downstream analysis was performed to extract and merge probes called amplified in at least one sample to yield recurrent markers with common boundaries, each of which serves as a character for the phylogeny inference. Our choice of the algorithms CBS and CGHSeg was based on the accessibility to code, platform non-specificity and popularity of use. We have compared our method on some major usability and functionality criteria in Table 1.

Phylogenetic trees were inferred by adding an all-diploid root to the set of character states and then running unweighted maximum parsimony inference using PAUP [27].

Given an accurate phylogeny reconstruction algorithm, the accuracy of the phylogenies will depend on the quality of input markers or characters. The estimated markers must first be truly representative of changes in copy number. Second, normal regions of the genome must not be assigned amplification states. Third, for each sample, the markers must only be assigned amplification states if they are indeed present in the sample and represent the correct character state assignment for that sample. Quality of the methods by these criteria was measured on three tasks. First, accuracy of amplicon detection across samples was quantified by the sensitivity, defined as fraction of genuinely amplified markers assigned to an amplicon, and specificity, defined as the fraction of markers assigned to an amplicon that were in fact amplified. Second, accuracy of marker assignment to amplicons was measured, quantified by the fraction of amplicons correctly called as amplified or non-amplified for all components. Finally, accuracy of phylogeny inference was assessed, quantified by the Branch Score Distance [28] using the *treedist* function of PHYLIP [29], a measure of agreement between the true and inferred phylogenies.

3.2 Experiments : Real Data

3.2.1 Unmixed Data—We further demonstrated our methods on real data derived from a publicly available (NCBI GEO GSE16672) primary ductal breast carcinoma aCGH dataset [25]. This data set was chosen because the cell sorting and sectioning methods underlying the tumor data extraction were developed specifically to aid phylogenetic analysis, making them well suited to our purposes, and because the data contains multiple samples per tumor, making them especially useful for studies of tumor heterogeneity and mixture analysis. The raw data comprises 87 tumor sectors obtained from 14 ductal breast cancer tumors run on a

high-density ROMA platform with 83,055 probes. We confined our analysis to the twenty-two autosomal chromosomes, reducing the dataset to 78,874 probes. We converted the raw aCGH data from log to linear domain, denoised it with a total variation denoising and then subjected it to an unmixing analysis to infer 6 components, or putative tumor progression states, as described in [23]. We next converted the data back to the log domain after recentering around a mean of 1. We then ran our method as described in Sec. 2 using PAUP for maximum parsimony tree building as with the simulated data.

4 Results and Discussion

4.1 Synthetic Data

Because our method relies on an estimate of noise level in its input data, we begin by verifying the accuracy of our estimator. Fig. 2 shows a comparison of the proposed data noise estimator with the estimated standard deviation of the data. The results show that our estimator gives a highly accurate estimate of the noise level on our simulated data sets. We note that the 1161 probes used in each simulated data set is low compared to a typical genome-scale aCGH data set and the accuracy of the estimator would therefore be expected to be greater for typical real data sets. By contrast, the standard deviation of the data provides a highly biased estimate of noise, especially at lower noise levels, because it conflates noise in the data with variance due to true amplicons.

We next examined the effectiveness of our HMMCNA method in comparison to the available competing methods and our own prior work on the simulated data. The results are summarized in Figure 3. Fig. 3(a,b) shows accuracy at the level of amplicon assignment. Fig. 3(a) shows that our method has a higher sensitivity than either of the comparative methods or our own prior method [24] at low to medium noise levels (up to about 0.6). Anecdotally, we have found that the noise inference computation described earlier yields values in the range of 0.1–0.5 on a selection of real datasets. At higher noise levels, the sensitivity drops sharply. Fig. 3(b) shows that all three methods have a high specificity for amplicon calling, with no false positive calls until relatively high levels of noise. At high noise levels, CGHseg is most prone to false positive calls, CBS least prone, and our own method intermediate between the two. At lower noise levels (< 0.2), our method has the least specificity in comparison, a result expected due our method's windowing approach, which raises the likelihood of incor- rectly grouping normal probes adjacent to an amplicon into the amplicon.

Fig. 3(c) shows accuracy of calling amplification states within detected amplicons. All three methods closely track the sensitivity plot of Fig. 3(a) up to a noise level of about 1.0, suggesting that each is highly accurate in calling states given the amplicons at low to moderate noise levels. Again, our method shows a drop in calling accuracy at higher noise levels in comparison to the competitors.

Fig. 3(d) shows the accuracy at inferring phylogenetic trees, which is the specific goal of our method. Here, our method shows superior performance in comparison to CBS and CGHSeg across all noise levels. This result may be attributed to high calling accuracy in general combined with a specific bias of our method for finding amplicons with shared boundaries across samples, which are especially useful for phylogenetic inference. It is interesting to note that while CBS has better calling accuracy at higher noise levels, its phylogenetic performance is not commensurate. This observation can be explained either at the marker inference step, where inconsistencies in boundary detection between samples may create problems for phylogenetic inference, or at the phylogeny-building stage itself, in that the order of phylogenetic markers can influence the topology of the resulting trees. We can thus

conclude that our method does provide an advantage over the existing methods in accurate phylogeny reconstruction in the presence of moderate but biologically realistic noise levels.

4.2 Real Data

4.2.1 Results on Unmixed Data—We next applied our method to mixture components derived from the real breast cancer data set of Navin et al. [25] both for further validation and to illustrate its value in predicting progression on real tumor samples. The HMM method found 315 marker amplicons, more than a 10-fold increase compared to the 27 detected by our prior method [24]. There are, on average, 91 am-plicons per component with markers spanning 74.81% of the genome. Analysis is complicated by the fact that some inferred amplicons are quite large and include many genes, which might be presumed to be predominantly passenger genes irrelevant to the progression process. It has been observed that small amplicons, in the range of a few megabases, are a distinct phenomenon from the large chromosome-scale amplifications produced by aneuploidy and translocations [30], which we believe account for the bulk of the total genome coverage. We therefore screened out inferred amplicons covering more than 148 probes (approximately 2.5 Mb) and examined enrichment of the shorter amplicons alone for known breast cancer markers. This reduced the portion of the genome found in some amplicon to 16% of the autosomal probes. We used the UCSC genome Table browser NCBI build 35 (corresponding to the aCGH array platform build) to find 3869 unique genes within the remaining small amplicons (versus 15869 for the set of all detected amplicons). We then used the Catalogue Of Somatic Mutations In Cancer (COSMIC) Database v. 57 [31] to specifically identify those associated with breast cancer, identifying 1014 breast cancer associated genes covered by short amplicons (versus 4126 in the full amplicon set) out of a total of 6973 breast cancer associated genes in COSMIC. To test whether these numbers suggest an enrichment for breast cancer-associated genes in our amplicons, we performed a chi-square test of significance of enrichment of our gene set for breast cancer markers relative to the full 23307 unique Refseq-curated human genes in NCBI build 35. The short ampli-cons were found to be significantly enriched for breast cancer associated genes (chi-square score 30.24, p-value < 0.0001). The set of both large and small amplicons was also strongly enriched (chi-square score 363.41, p-value < 0.0001). Anecdotally, this set of amplicons carries several important markers not identified by our earlier method, notable among them being JUN, BRAF, KRAS, FGFR1, ESR1 and JAK2.

Figure 4 provides a visual comparison of results of our method to those of CBS and CGHSeg, using chromosome 17. Our method and CGHSeg produce similar results, although with some additional fine-scale amplicon structure identified by our method. CBS produced considerably more breakpoints than either other method. Over the entire genome, CBS produced 1425 distinct marker segments, a much higher number than our own method spanning 93.8% of the genome. We cannot definitely say to what degree these extra breakpoints reflect better sensitivity to true variations versus spurious breaks due to experimental noise. CGHSeg has substantially higher computational cost and could not complete analysis of the full genome in more than a month of processing and we therefore do not provide a full comparison to that method. It should be noted, though, that neither of these methods are designed to work with mixture components of the sort for which our method was developed, which might be expected to conform poorly to their error models.

Next, we analyzed the phylogenetic tree obtained from the markers, summarized in Figure 5. Nodes correspond to putative stages of progression and edges to ampli-cons gained during discrete steps of progression. For purposes of annotation of the phylogeny, we identified specific genes for the short amplicons, favoring those in the COSMIC breast cancer set when a short amplicon covered multiple genes and using genes cited by Navin et al. [25] in

their own analysis of their data to break ties. We annotated only a subset of large amplicons manually chosen because they carry genes we expect to be particularly important to breast cancer progression.

The resulting tree is shown in Figure 5. The tree exhibits homosplasy (recurrent mutation) but no reversion of markers, a result we believe to improve upon that of our prior method [24], which exhibited both homo-plasy and reversions. While the homoplasy might reflect genuine convergence of distinct progression pathways, it could also be explained by false positive calling errors or errors in phylogeny inference due to the maximum parsimony assumption.

Analyzing the tree in more detail reveals several features of note. The progression pathway to C5 occurs with the gain of HER2 (ERBB2) and CCND1 suggesting a distinct arm of HER2/CCND1 co-amplification. There are two other progression pathways leading to C6 and C1 that also show HER2 amplification. The pathway leading to C6 has an amplicon housing CCNE1, consistent with a notion of two distinct forms of HER2-amplifying tumors. It has been reported recently that cooccurrence of HER2 and CCNE1 leads to Herceptin therapy resistance in HER2 overexpressing breast cancer [32], [33]. The phylogeny supports this idea of distinct pathways of evolution of HER2-amplifying breast cancers, specifically including one pathway co-amplifying with CCND1 and one co-amplifying with CCNE1. We also observe late co-amplification of HER2 and a large am-plicon containing MYC in both CCND1-amplifying and CCNE1-amplifying variants, as well as a CCNE1/HER2-amplifying pathway that does not co-amplify MYC.

4.3 Runtime Analysis

We also compared the computation run-time for all three methods. The results are shown in Table 2. The results show HMMCNA to be by far the most efficient method, requiring seconds per chromosome. CGH-Seg was the least time-efficient. CBS gave intermediate values. These results illustrate a secondary advantage of our method in scaling efficiently to many more probes than the alternatives, a key advantage for a method designed for working on whole-genome data.

5 Conclusion

We have developed a novel method for joint segmentation and calling of multi-sample genome-scale DNA copy number data, designed specifically for use in tumor phylogenetics. The method uses a novel multi-sample HMM approach to identify consistent markers across a set of samples, typically mixture components inferred from raw tumor data, for use as markers for phylogenetic inference. Comparison with a state-of-the-art multi-sample scheme and a leading single-sample scheme shows that our method has superior performance at levels of experimental noise typical of real aCGH data for the specific task of tumor phylogenetics, as well as for the more general task of tumor marker inference. Further, the method substantially improves on our own prior work for the problem of phylogenetic inference from inferred mixture components through a novel HMM approach for multi-sample amplicon detection and improved methods for modeling noise in the data. In particular, our method outperforms the alternatives, and substantially outperforms our own prior method, in the noise range of 0.0–0.6, a region that subsumes the noise range of approximately 0.1–0.5 we have estimated for real aCGH data. These methodological improvements lead to a more than ten-fold increase in the number of markers available for phylogeny inference and detection of several important progression markers not previously found from these data.

While there is no obvious direct way to validate the results obtained from running HMMCNA on real data, we have shown indirect support for our results through comparison to established marker sets and anecdotally supported features of the inferred trees based on previously published research. The issue of assessing the true validity of our results remains a challenge since there is no known ground truth for either the quality of inferred amplicons or the reconstructed phylogeny from the amplicons.

In future work, we hope to improve on the current approach through a more realistic model of amplification distributions including handling of genomic deletions, algorithmic improvements to avoid combinatorial increase in state size with components, and improvements in the upstream unmixing and downstream phylogenetic inference steps. We further hope to explore how one might better tune the method to specifically detect markers most likely to be informative for phylogenetic inference. In addition, the method may have value for other applications of copy number data in phylogenetics and related problems.

Acknowledgments

A.S., S.S., and R.S. were supported in this work by U.S. National Institutes of Health awards 1R01CA140214 and 1R01AI076318.

References

1. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumors. Nature. 2000; vol. 406:747–752. [PubMed: 10963602]
2. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999; vol. 286:531–537. [PubMed: 10521349]
3. Sorlie T, Perrou CM, Tibshirani R, et al. Gene expression profiles of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA. 2001; vol. 98:10869–10864. [PubMed: 11553815]
4. Sotiriou C, Neo SY, McShane LM, et al. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proc Natl Acad Sci USA. 2003; vol. 100:10393–10398. [PubMed: 12917485]
5. Ashworth A, de Bono J. Translating cancer research into targeted therapeutics. Nature. 2010
6. van't Veer LJ, Dai H, van de Vivjer M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; vol. 415:530–536.
7. Miller L, Smeds J, George J, et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. Proceedings of the National Academy of Sciences of the United States of America. 2005; vol. 102(no. 38):13550–13555. [Online]. Available: http://www.pnas.org/content/102/38/13550.abstract. [PubMed: 16141321]
8. Olshen AB, Venkatraman ES, Lucito R, et al. Circular binary segmentation for the analysis of array-based DNA copy number data.". Biostatistics. vol. 5(no. 4):557–572. October 2004. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/15475419. [PubMed: 15475419]
9. Picard F, Robin S, Lavielle M, et al. A statistical approach for array CGH data analysis.". BMC Bioinformatics. 2005; vol. 6 [Online]. Available: http://dx.doi.org/10.1186/1471-2105-6-27.
10. Hsu L, Self S, Grove D, et al. Denoising array-based comparative genomic hybridization data using wavelets, *Biostatistics*. vol. 6. no. 2005; 2:211–226. [Online]. Available: http://biostatistics.oxfordjournals.org/content/6/2/211.abstract.
11. Eilers P, de Menezes R. Quantile smoothing of array CGH data. Bioinformatics. 2005; vol. 21(no. 7):1146–1153. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/21/7/1146.abstract. [PubMed: 15572474]
12. Wang K, Li M, Hadley D, et al. Penncnv: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome snp genotyping data, *Genome*

Research. vol. 17. no. 2007; 11:1665–1674. [Online]. Available: http://genome.cshlp.org/content/17/11/1665.abstract.

13. Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, Futreal PA, Stratton MR. Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. Biostatistics. 2010; vol. 11(no. 1):164–175. [Online]. Available: http://biostatistics.oxfordjournals.org/content/11/1/164.abstract. [PubMed: 19837654]
14. Pique-Regi R, Ortega A, Asgharzadeh S. Joint estimation of copy number variation and reference intensities on multiple DNA arrays using GADA. Bioinformatics. 2009; vol. 25(no. 10):1223–1230. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/25/10/1223.abstract. [PubMed: 19276152]
15. Shah S, Cheung K, Johnson NA, et al. Model-based clustering of array cgh data. Bioinformatics. 2009; vol. 25(no. 12):i30–i38. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/25/12/i30.abstract. [PubMed: 19478003]
16. Wiel V, Mark A, Brosens R, et al. Smoothing waves in array CGH tumor profiles. Bioinformatics. 2009; vol. 25(no. 9):1099–1104. [PubMed: 19276148]
17. Wu L, Chipman H, Bull S, Briollais L, Wang K. A Bayesian segmentation approach to ascertain copy number variations at the population level. Bioinformatics. 2009; vol. 25(no. 13):1669–1679. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/25/13/1669.abstract. [PubMed: 19389735]
18. Zhang N, Senbabaoglu Y, Li J. Joint estimation of DNA copy number from multiple platforms. Bioinformatics. 2010; vol. 26(no. 2):153–160. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/26/2/153.abstract. [PubMed: 19933593]
19. Beroukhim R, Getz G, Nghiemphu L, et al. Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. Proceedings of the National Academy of Sciences. 2007; vol. 104(no. 50):20007–20012. [Online]. Available: http://www.pnas.org/content/104/50/20007.abstract.
20. Nowak G, Hastie T, Pollack J, Tibshirani R. A fused lasso latent feature model for analyzing multi-sample aCGH data. Biostatistics. 2011; vol. 12(no. 4):776–791. [Online]. Available: http://biostatistics.oxfordjournals.org/content/12/4/776.abstract. [PubMed: 21642389]
21. Picard F, Lebarbier E, Hoebeke M, Rigaill G, Thiam B, Robin S. Joint segmentation, calling, and normalization of multiple CGH profiles. Biostatistics. 2011; vol. 12(no. 3):413–428. [Online]. Available: http://biostatistics.oxfordjournals.org/content/12/3/413.abstract. [PubMed: 21209153]
22. Schwartz R, Shackney S. Applying unmixing to gene expression data for tumor phylogeny inference. BMC Bioinformatics. 2010; vol. 11:42. [PubMed: 20089185]
23. Tolliver D, Tsourakakis C, Subramanian A, et al. Robust unmixing of tumor states in array comparative genomic hybridization data. Bioinformaticsno. 2010; vol. 26(no. 12):106–i114. [Online]. Available: http://bioinformatics.oxfordjournals.org/content/26/12/i106.abstract.
24. Subramanian A, Shackney S, Schwartz R. Inference of tumor phylogenies from genomic assays on heterogeneous samples. Proc. ACM-BCB'. 2011; 11
25. Navin N, Krasnitz A, Rodgers L, et al. Inferring tumor progression from genomic heterogeneity. Genome Research. 2010 Mar.vol. 20:68–80. [PubMed: 19903760]
26. Olshen A, Venkatraman E, Lucito R, Wigler M. Circular binary segmentation for the analysis of arraybased DNA copy number data. Biostatistics. 2004; vol. 5(no. 4):557–572. [Online]. Available: http://biostatistics.oxfordjournals.org/content/5/4/557.abstract. [PubMed: 15475419]
27. Swafford, D. PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sunderland: Massachussets; 2002.
28. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.". Molecular Biology and Evolution. 1994; vol. 11(no. 3):459–468. [Online]. Available: http://mbe.oxfordjournals.org/content/11/3/459.abstract. [PubMed: 8015439]
29. Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 1989; vol. 5:164–166.
30. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature. 1998; vol. 396:643–649. [PubMed: 9872311]

31. Bamford S, Dawson E, Forbes S, et al. The COSMIC (catalogue of somatic mutations in cancer) database and website. Br J Cancer. 2004

32. Mittendorf EA, Liu Y, Tucker SL, et al. A novel interaction between HER2/neu and cyclin E in breast cancer. Oncogene. 2010 Jul.vol. 29:3896–3907. [PubMed: 20453888]

33. Scaltriti, M.; Eichhorn, P.; Cortes, J., et al. Cyclin E amplification/overexpression is a mechanism of trastuzumab resistance in HER2+ breast cancer patients. Proceedings of the National Academy of Sciences. 2011. [Online]. Available: http://www.pnas.org/content/early/2011/02/09/1014835108.abstract

Biographies

Ayshwarya Subramanian Ayshwarya Subra-manian received her undergraduate honors degree in Biological Sciences from the Birla Institute of Technology and Science(BITS)-Pilani, India in 2007. Since then, she has been a Ph.D candidate at the Department of Biological Sciences at Carnegie Mellon University where she studies tumor progression using computational phylogenetics.

Stanley Shackney Dr. Shackney received his M.D. degree from the Harvard Medical School in 1964. He later worked at the National Cancer Institute,where was Head of the Section of Cell Kinetics. In 1984 he moved to Allegheny General Hospital where he became Professor of Cancer Cell Biology and Genetics Drexel University Medical School.

Russell Schwartz Russell Schwartz received his BS, MEng, and PhD degrees from the Department of Electrical Engineering and Computer Science at the Massachusetts of Technology, the last in 2000. He later worked in the Informatics Research group at Celera Genomics. He joined the faculty of Carnegie Mellon University in 2002, where he is currently a Professor of Biological Sciences.

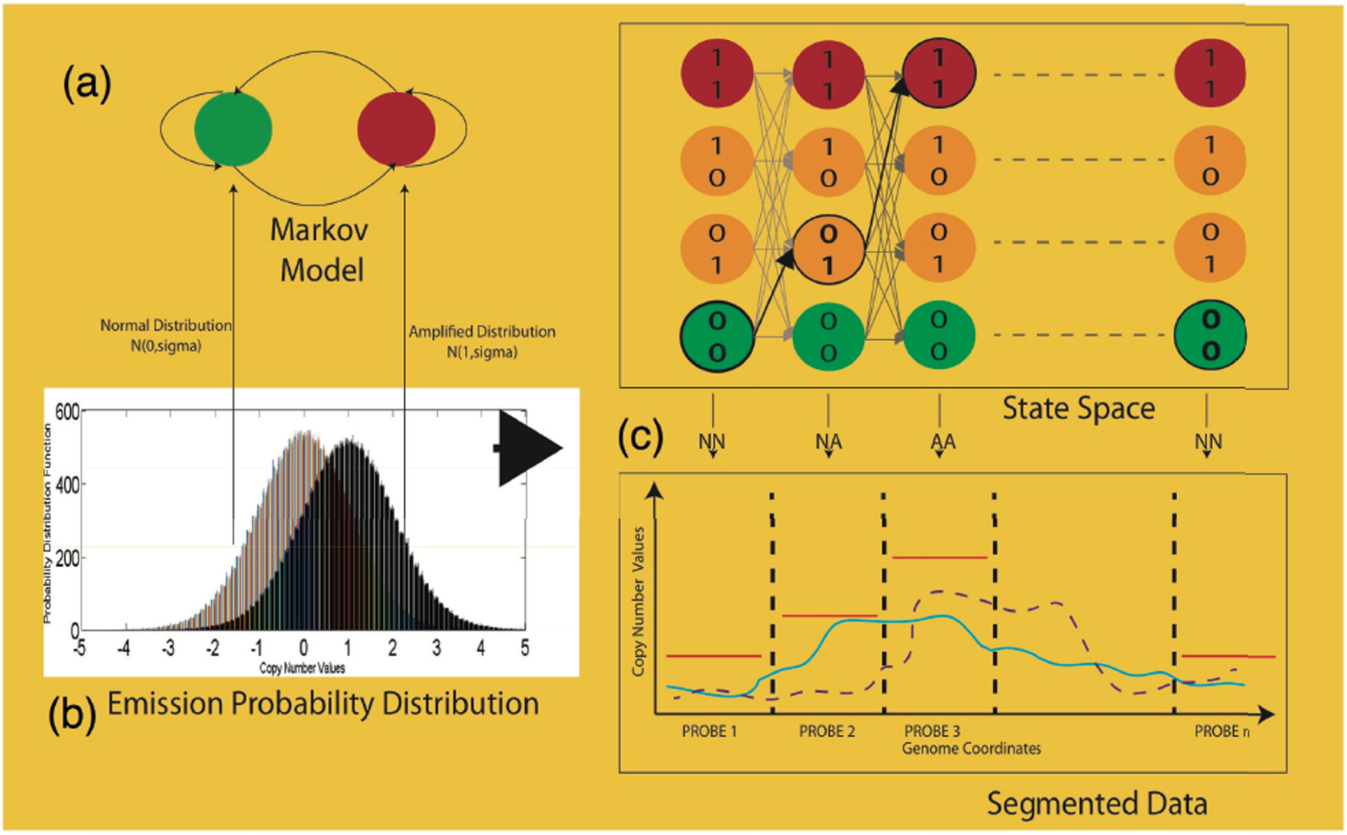

Fig. 1.

Representation of our HMM model, HMMCNA. The amplicon model (a) seeks to explain each probe in each progression state as either normal (green) or amplified (red) based on its fit to one of two copy number distributions (b). The HMM model (c) allows simultaneous maximization of the likelihood of these assignments across all probes and progression states, in the process segmenting the data and producing markers suitable for phylogenetic analysis. In the two-sample HMM example of (c), nodes labeled “1 1” (red) correspond to positions at which both samples are amplified, those labeled “0 0” (green) to positions at which neither sample is amplified, and those labeled “1 0” or “0 1” (orange) to positions at which exactly one of the two samples is amplified.

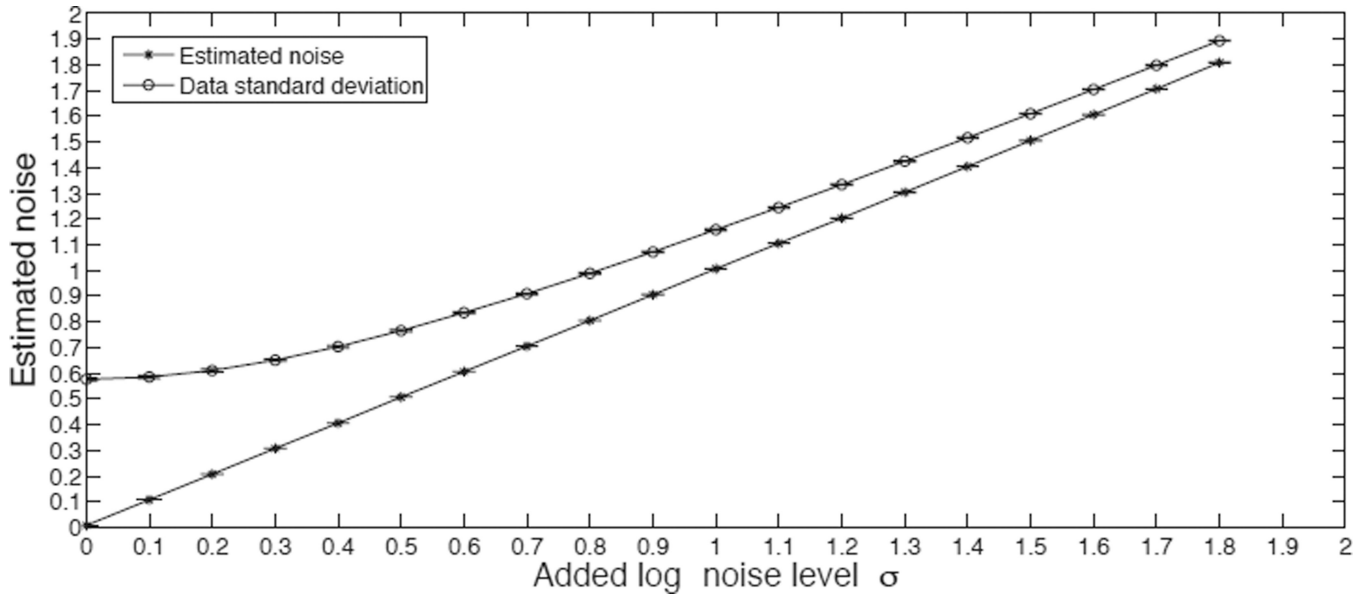

Fig. 2.

Comparison of noise estimates on simulated data derived from our method with those derived using the standard deviation of the data versus the true noise levels simulated for the data. Error bars show standard error of the estimates for each method.

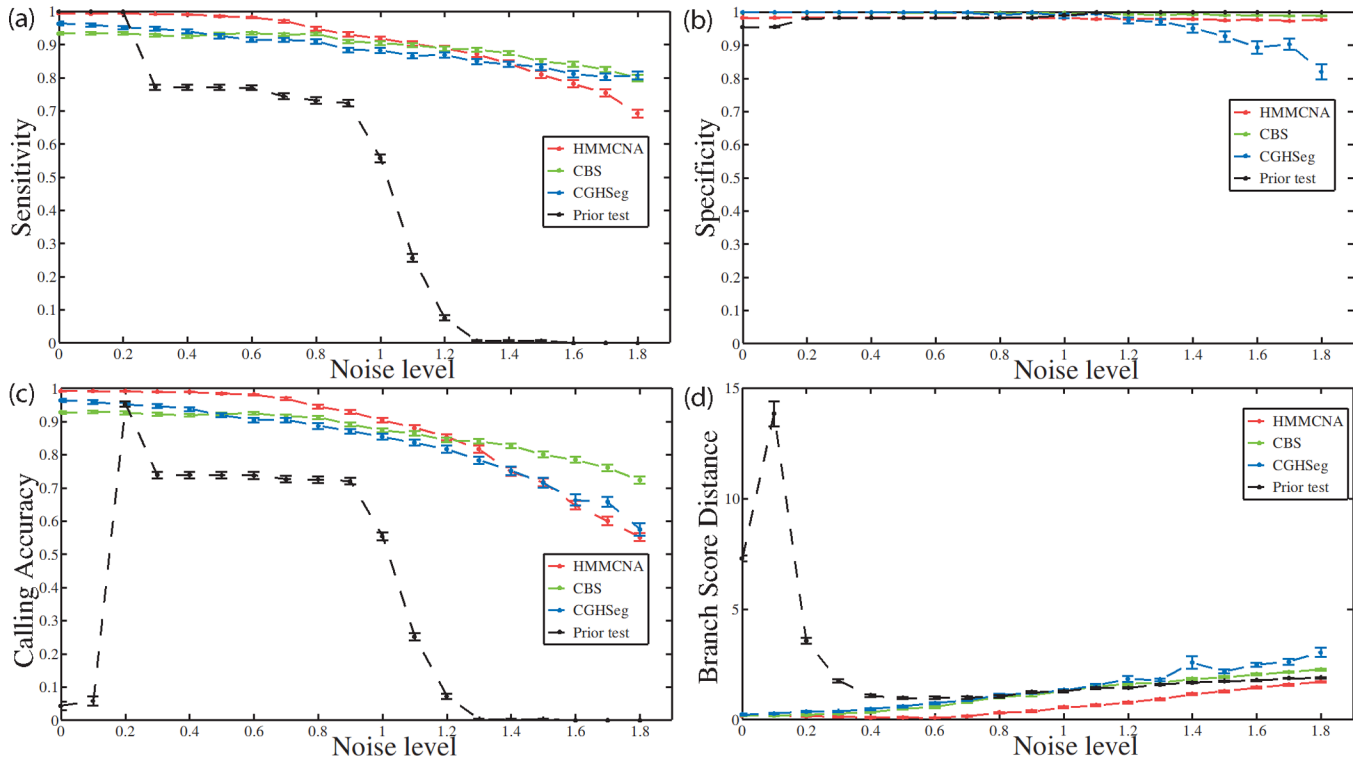

Fig. 3.

Accuracy of our method (HMMCNA), CBS, CGHseg, and our prior method on simulated data. (a, b) Accuracy in amplicon assignment, classified by the sensitivity (a) and specificity (b) of correctly assigning markers. (c) Calling accuracy, measured by the fraction of amplified markers assigned the correct amplification state. (d) Tree-building accuracy, quantified by the branch-score distance between the true and observed tree. All measures are reported as functions of the log-normal noise level σ, averaged over 200 independent runs per noise level.

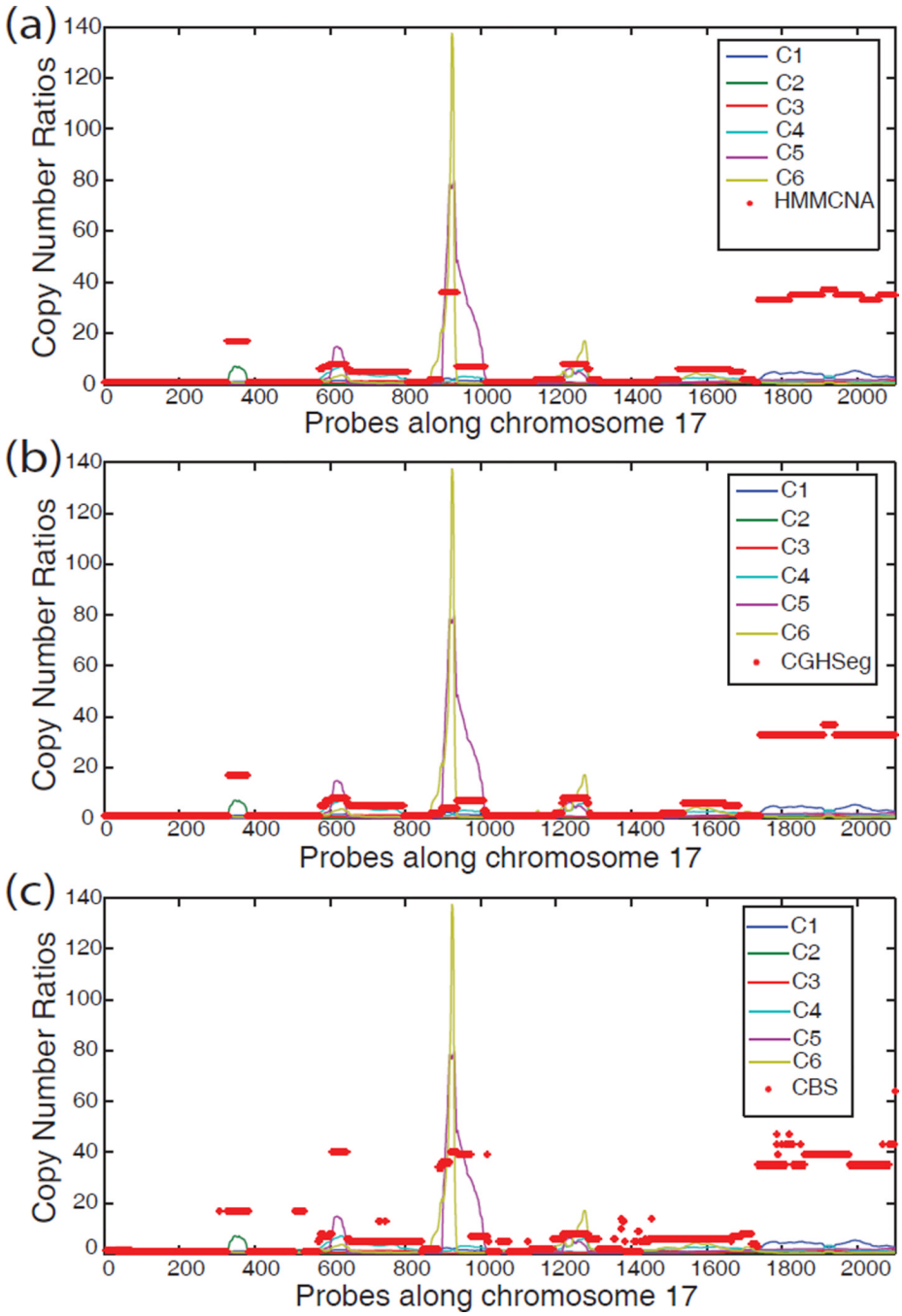

Fig. 4.

Segmentation of chromosome 17 using mixture components of Navin et al. (a) Our method, HMMCNA. (b) CGHSeg. (c) CBS.

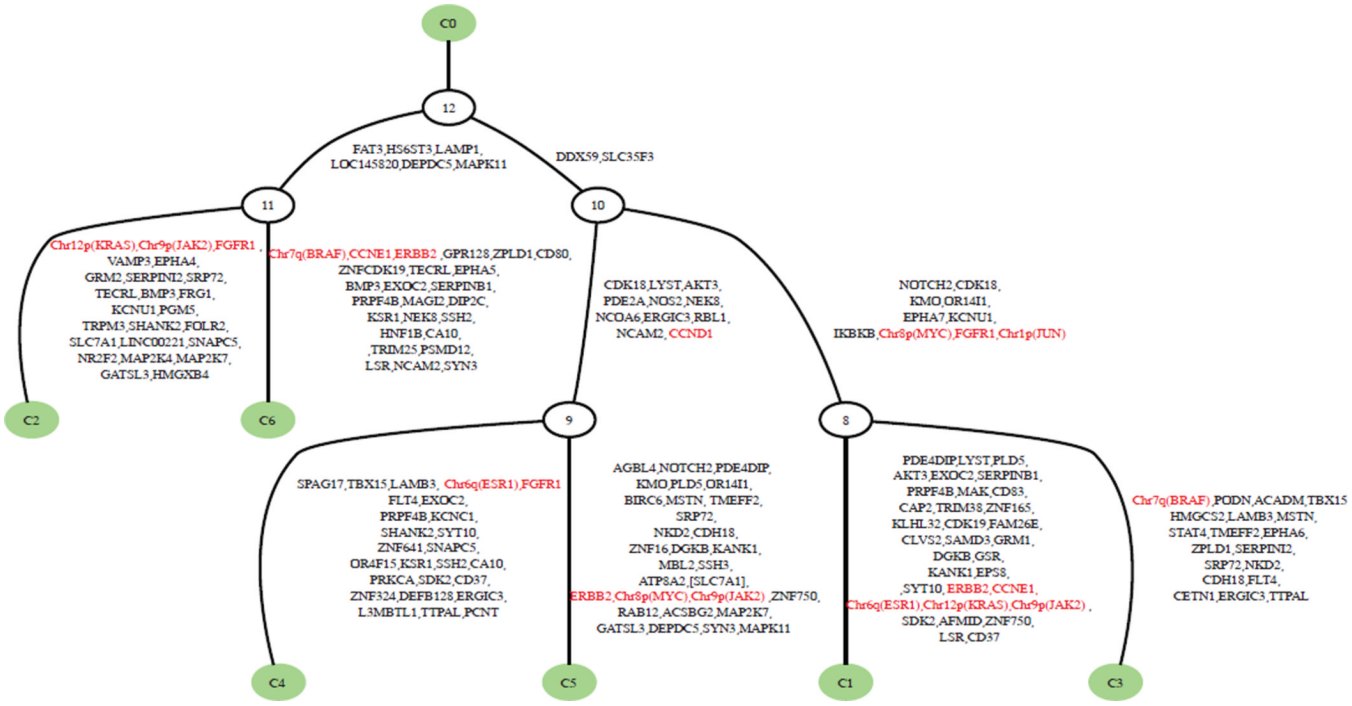

Fig. 5.

Maximum parsimony tree inferred from mixture components derived from real breast cancer data of Navin et al. [25]. Edges are labeled with putative driver genes, with those of particular note as breast cancer progression markers highlighted in red. Amplicons of 148 or fewer probes (approximately 2.5 Mb on average) are listed by gene while selected larger amplicons are listed by chromosome arm with genes of interest in parentheses. Green nodes are observed components and white are inferred ancestral states, also known as Steiner nodes.

TABLE 1

Qualitative comparison of HMMCNA with other state-of-the-art copy number segmentation methods. The table distinguishes methods based on whether they perform marker calling, whether they work on single- or multi-sample data, and whether they are generic with respect to input data or speciific to a particular data platform.

Method	Segmentation	Calling	Data	Platform specificity
CBS[8]	Yes	No	Single-sample	No
PennCNV[12]	Yes	Yes	Single-sample	SNP-Array
PICNIC[13]	Yes	Yes	Single-sample	SNP-Array
CGHSeg[21]	Yes	Yes	Multi-sample	No
GISTIC[19]	Yes	No	Multi-sample	Yes
HMMCNA	Yes	Yes	Multi-sample	No

TABLE 2

Computation run-time on real data for CBS, CGHSeg and our method, HMMCNA over the entire genome.

Method	Runtime
CBS[8]	1.0395h
CGHSeg[21]	41 days
HMMCNA	20.1s