Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2012, Article ID 797812, 16 pages
doi:10.1155/2012/797812

Research Article

Inference of Tumor Phylogenies from Genomic Assays on Heterogeneous Samples

Ayshwarya Subramanian,[1] Stanley Shackney,[2] and Russell Schwartz[3]

[1] *Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA*
[2] *Intelligent Oncotherapeutics, Pittsburgh, PA 15243, USA*
[3] *Department of Biological Sciences and Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA*

Correspondence should be addressed to Russell Schwartz, russells@andrew.cmu.edu

Received 2 November 2011; Accepted 21 February 2012

Academic Editor: Ali Khraibi

Tumorigenesis can in principle result from many combinations of mutations, but only a few roughly equivalent sequences of mutations, or “progression pathways,” seem to account for most human tumors. Phylogenetics provides a promising way to identify common progression pathways and markers of those pathways. This approach, however, can be confounded by the high heterogeneity within and between tumors, which makes it difficult to identify conserved progression stages or organize them into robust progression pathways. To tackle this problem, we previously developed methods for inferring progression stages from heterogeneous tumor profiles through computational unmixing. In this paper, we develop a novel pipeline for building trees of tumor evolution from the unmixed tumor data. The pipeline implements a statistical approach for identifying robust progression markers from unmixed tumor data and calling those markers in inferred cell states. The result is a set of phylogenetic characters and their assignments in progression states to which we apply maximum parsimony phylogenetic inference to infer tumor progression pathways. We demonstrate the full pipeline on simulated and real comparative genomic hybridization (CGH) data, validating its effectiveness and making novel predictions of major progression pathways and ancestral cell states in breast cancers.

1. Introduction

The application of genomic technologies to cancers has revealed that patients with tumors that appear indistinguishable to the clinician may have completely different causes at the molecular level [1, 2] resulting in very different prognoses [3] and responses to possible treatments [4]. Nonetheless, most human cancers seem to follow a relatively small number of progression pathways [1, 2, 5], each characterized by an approximately equivalent sequence of mutations. This observation is key to the success of targeted therapeutics, a groundbreaking approach to cancer treatment in which drugs are developed to treat specific molecular abnormalities shared by large subgroups of patients [6]. By identifying common progression pathways and characterizing their conserved features, it is hoped that we can find new subgroups of patients who will respond to a common treatment, identify the specific abnormalities that will provide effective therapeutic targets for those subgroups, and develop clinically useful diagnostic tests to identify new patients in those subgroups. There are considerable practical challenges to each of these steps, however.

One of the significant challenges to identifying and characterizing progression pathways is the heterogeneity of cancers both within and between patients [7]. Any two patients, even with a common progression pathway, will exhibit many differences in the details of the causal mutations along that pathway, as well as in the assortment of random passenger mutations distinct to each patient that do not contribute to their pathology [8]. Even within a single patient, a tumor will generally be highly heterogeneous, with genetically distinct cell populations corresponding to different stages along the progression of their tumor and possibly even different branches along those progression pathways within a single tumor [9]. This heterogeneity is problematic for methods for profiling tumor states, since there is at present no technology to determine the genetic states of single cells at a genomic scale. Genome-wide methods for tumor profiling—such as expression microarrays, RNA-seq, or array comparative genomic hybridization (aCGH)—necessarily mix contributions from many discrete cell types. This mixing would be expected to result in a conflation of distinct states along a progression pathway, obscuring characteristics of individual subpopulations of cells and hiding the discrete steps in progression that may provide clinically valuable markers of early stages in progression or important clues to major decision points in a tumor's evolution. This heterogeneity is particularly challenging to phylogenetic approaches to inferring tumor progression [10], which depend on our ability to at least approximately identify discrete steps in tumor evolution and can benefit greatly from information about ancestral states and the combinations of states present in distinct tumor samples [11].

There are various ways to approach the problem of heterogeneity in tumor phylogeny inference. One approach is to use alternative technologies designed to profile single cells as a way of directly observing discrete states within tumors. This approach has been successfully used for tumor phylogeny inference from single cell fluorescent in situ hybridization (FISH) data [12, 13]. Using single-cell assays has substantial drawbacks, however, because single-cell technologies can profile only a few preselected markers per cell. An alternative is to separate cells into approximately homogeneous populations prior to applying genomic methods, as was done recently by [14], who used a combination of microdissection and post-dissection cell sorting to separate discrete sub-populations of cells prior to whole-genome DNA copy number profiling by aCGH. A third alternative, used in the present work, is to apply genomic technologies to heterogeneous samples but attempt to computationally separate distinct cell populations from the outputs of these samples. Such computational unmixing methods have been previously used in tumor analysis to correct for stromal contamination of tumor cells [15] and have been useful to similar applications of evolutionary inference from heterogeneous samples, such as in reconstructing evolutionary steps in viral quasispecies [16].

In previous work, we proposed the use of such unmixing methods for identifying cell states for phylogeny inference [11] and demonstrated their ability to separate biologically meaningful tumor cell populations from expression microarray data [11] and aCGH data [17]. In this paper, we build on that prior work by developing a pipeline for converting inferred cell profiles into phylogenetic trees describing likely stages of tumor progression and common progression pathways by which they evolve. This pipeline implements four distinct steps. The first applies our prior unmixing model [17] to infer profiles of major progression steps from heterogeneous tumor data. The second step uses a novel statistical test to identify amplified genomic regions that can serve as markers of progression. The third step then uses a second statistical approach to call these markers as amplified or nonamplified in individual inferred cell states, creating a matrix of phylogenetic states suitable for character-based phylogenetic inference. The fourth step then applies maximum parsimony phylogeny inference to the resulting data to identify likely progression trees, labeled by changes in the marker set inferred in step two. These progression trees establish a model of tumor evolution identifying discrete steps of progression among these markers and possible ancestral stages of tumor progression not directly apparent from the identified components. Validation on simulated data demonstrates the effectiveness of the method at identifying markers, assigning them to progression states, and inferring trees from those states. Application to real breast cancer CGH data results in a phylogeny that recapitulates key features of our current understanding of major breast cancer progression pathways while elaborating in several potentially significant ways. The work represents, to our knowledge, the first use of character-based phylogenetic inference for similar whole-genome tumor profiles, providing advantages over prior distance-based approaches in identifying likely markers and describing specific mutations that may underlie key steps in tumor progression.

In the remainder of this paper, we present our method and an application to a publicly available aCGH data set. In Section 2.1, we describe our overall phylogenetic inference pipeline and the novel computational and statistical methods developed for it. In Section 2.2, we provide details on specific use of the methods developed here and their application to the analysis of the breast tumor aCGH data of [14]. In Section 3, we present the results, identifying a set of phylogenetic markers and a resulting tumor phylogeny. In Section 3, we also discuss the biological significance of the results, examining both their concordance with prior literature and interesting novel predictions of the methods. Finally, in Section 4, we consider avenues for future work.

2. Materials and Methods

2.1. Algorithms. At a high level, our method consists of an analysis pipeline to convert raw data on profiles of heterogeneous tumor samples into phylogenetic inferences on computationally inferred profiles of discrete cell states. While the method can in principle work with any technology for profiling tumor state, we assume in the present work that we are specifically using aCGH data describing DNA copy numbers at a discrete genome-wide probe set. The data are assumed to be in the form of copy numbers of n probes in m tumors or tumor sections. These data are assumed to be raw or baseline normalized raw input, rather than the conventional log ratios.

The overall analysis pipeline is summarized in Figure 1. The pipeline consists of the following steps:

(1) computational unmixing of raw aCGH data to infer aCGH profiles of well-populated tumor states,

(2) identification of significantly amplified marker regions of the genome from the component aCGH data,

(3) assignment of marker states to components,

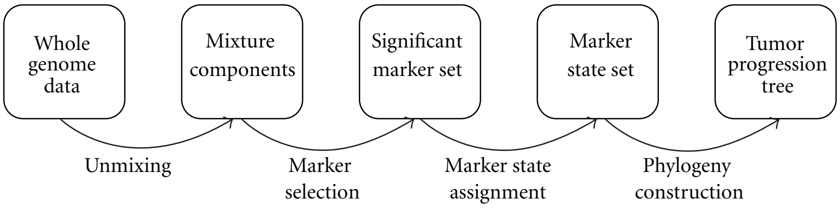


Figure 1: Workflow diagram summarizing the major steps in our unmixing-based phylogenetic analysis pipeline.

(4) phylogenetic inference on cell states to produce an inferred progression tree.

The individual steps of this analysis are as described below.

2.1.1. Unmixing Analysis. Our phylogenetic approach assumes data that has been separated into mixture components. We initially accomplished this assignment using an unmixing method previously developed by our group [17] based on an interpretation of the problem as that of fitting a simplex to an observed set of data points, where simplex vertices will then correspond to inferred components of the mixture. The method is based on prior work by Ehrlich and Full [18] adapted to better handle the high dimension and noise level characteristic of genomic data. We have since updated that method to use nonnegative matrix factorization (NNMF) [19] to eliminate the possibility of negative copy number values and other artifacts that can induce in the code.

Figure 2 illustrates the unmixing procedure. We first preprocess the data by applying L1-L1 total variation denoising to the raw aCGH profiles. In the initial method, we then use principal components analysis (PCA) to convert aCGH profiles of tumor samples to points in a low-dimensional space. The aCGH profiles are then explained as mixtures drawn from a set of common cell types by fitting a simplex to the point set, with some allowance for noise in the data. Any point in the simplex can then be explained as a linear combination of the vertices of the simplex. These vertex points are interpreted as the cell types from which each tumor sample is generated and can be projected back into the original dimension of the aCGH array to construct virtual aCGH profiles of the inferred cell types. The outputs of the method are an inferred set of mixture components, identifying a projected copy number of each cell type at each probe, and a set of mixture fractions, explaining each observed tumor sample as a sum of fractional contributions of cell types. The mixture components can be represented as a matrix C in which each entry c_{ij} describes the inferred copy number of component or cell type i at aCGH probe j. For the present pipeline, we use only the component matrix C and discard the mixture fractions. Space does not permit a detailed description of the method, so we refer the reader to [11] for a more thorough description of our general unmixing strategy for tumor phylogenetics and to [17] for a detailed discussion of the specific noise-tolerant unmixing algorithm used in our primary results here. Our most recent algorithm functions identically except that initially dimensionality reduction is accomplished by NNMF rather than PCA and an additional nonnegativity constraint is imposed during the optimization of components C.

The primary results below are based on components previously determined in Tolliver et al. [17] by the PCA-based method, although the improved method is applied to develop components from simulated data and from a secondary breast cancer data set to provide additional points of comparison.

2.1.2. Identification of Amplified Genomic Regions. Once we have the inferred components, it is next necessary to identify markers for tracking phylogenetic state. For aCGH data, we seek genomic regions that are amplified in subsets of tumors. We focus on amplifications due to a technical limitation of the unmixing approach. Unmixing is performed in the linear, rather than log, domain, and a deletion represents only a small linear change in copy number, so we expect the method to have poor sensitivity to deletions. Given the high variability from probe to probe in the data, it is necessary to use a statistically robust test for amplification. To accomplish this, we developed a test designed to test for significant amplification of a window of w contiguous probes across the m components.

We assume Gaussian noise in the data, thus modeling each individual probe as drawn from a Gaussian distribution with mean 1 (corresponding to diploid DNA). The variance is assumed to be the empirically measured variance, σ^2, across all probes in all components. We then seek to reject the hypothesis that the collection of $w \times m$ probes under consideration were drawn from the corresponding Gaussian. For this purpose, we take as our statistic the sum of squares of Z-scores of the probe values:

$$X_k = \sum_{i=1}^{m} \sum_{j=k}^{k+w-1} \left(\frac{c_{ij} - 1}{\sigma} \right)^2, \tag{1}$$

where k is the index of the first probe in the window. Under the null hypothesis, this statistic would be expected to be chi-square distributed with $w \times m$ degrees of freedom. We thus test for significant amplification with a one-sided chi-square significance test for the appropriate degrees of freedom.

We apply this test to sliding windows of probes of fixed width w across the genome. After identification of discrete amplified windows, we apply a postprocessing step to collapse any overlapping amplified windows into a single larger window and treated the union of probes in all overlapping significant windows as the marker for subsequent analysis.

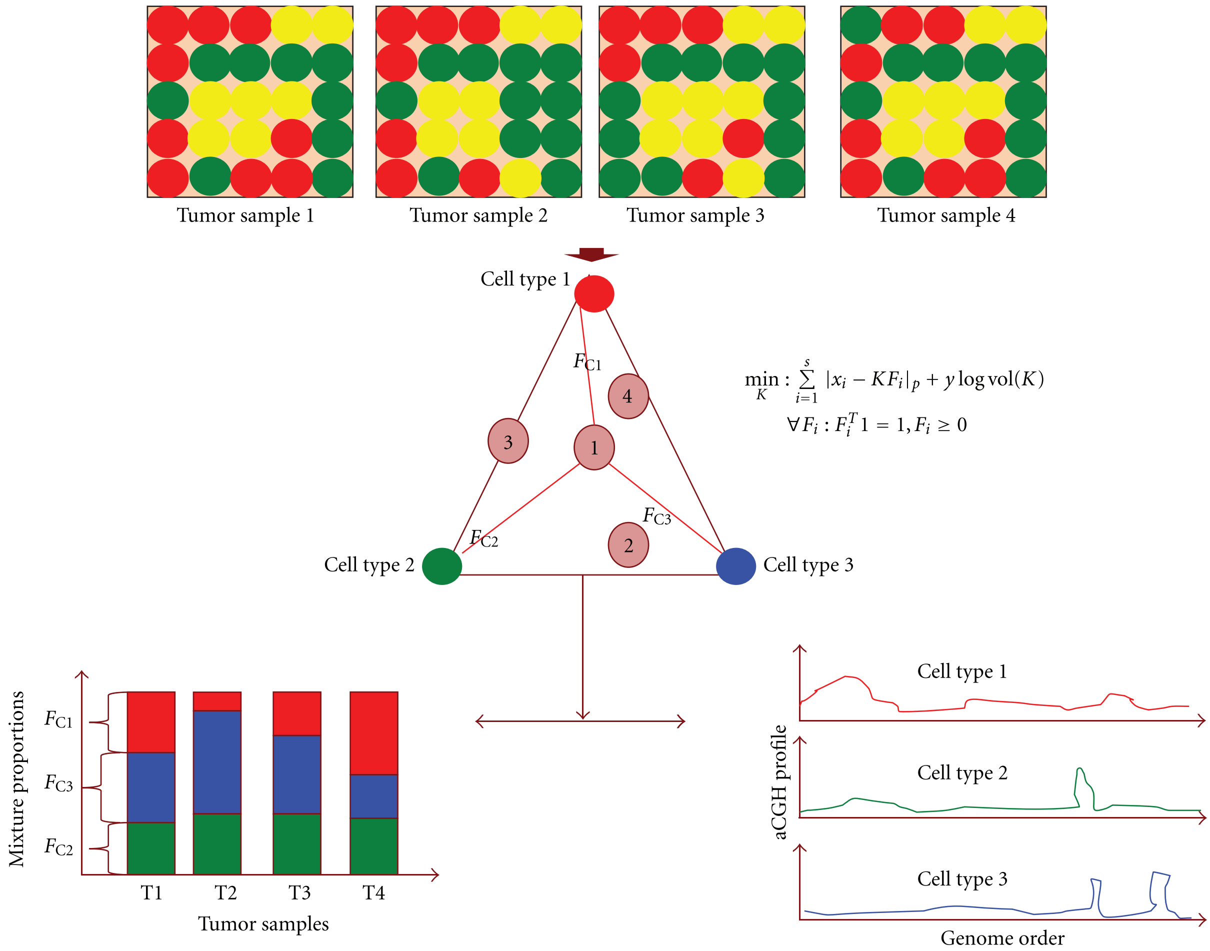


FIGURE 2: Illustration of the unmixing approach. Tumor samples T1–T4 are assayed by aCGH, generating genome-wide copy number profiles. The aCGH profiles are interpreted as points in a space (two-dimensional in the example) and are unmixed by fitting a simplex to the point set (a 3-simplex, or triangle, in the example). The vertices of the simplex represent inferences of three cell types (1, 2, and 3) from which T1–T4 can be explained. These vertices are then projected back to the dimension of the aCGH arrays to construct virtual aCGH profiles of the inferred cell types. The outputs are these virtual aCGH profiles and the inferred fractional amount of each cell type in each tumor sample.

We would normally expect the detected regions to be a subset of those one would find by performing a comparable statistical test on the raw aCGH measurements rather than the inferred components, as we would expect that features that are not robust to a significant fraction of samples will be interpreted as noise and suppressed at the unmixing step.

The scan for significant windows was done through custom Matlab code using the `chi2cdf` function for chi-square significance testing.

2.1.3. Assignment of Marker States to Components. After identifying a set of markers, we next need to determine the states of those markers in each inferred cell component. For this purpose, we again treat the problem as that of attempting to reject the hypothesis that the individual copy numbers are drawn from a Gaussian of mean 1 and variance corresponding to the empirically measured variance across all probes. For each component i and marker j, we compute the mean copy number over all probes in the given marker for the given component:

$$\mu_{ij} = \frac{1}{b_j - a_j + 1} \sum_{k=a_j}^{b_j} c_{ik}, \tag{2}$$

where a_j is the leftmost probe index and b_j the rightmost probe index for marker j. We then evaluate the single-sided P value for the hypothesis that μ_{ij} is drawn from a Gaussian with mean 1 and variance $\sigma^2/(b_j - a_j + 1)$, where σ^2 is again the empirical variance across all probes in all components. We implicitly build in a prior probability that any given marker is not amplified in any given component by using a P value cutoff of 0.001 for calling a probe amplified. The result of this analysis is an assigned state (amplified or not amplified) for each component at each phylogenetic marker. These values can be represented as an $m + 1 \times k$ matrix P of phylogenetic markers, where element p_{ij} is a binary value indicating whether marker j is amplified or not amplified in component i.

Custom Matlab code was used to assign phylogenetic states to each component at each marker using the normcdf function.

2.1.4. Phylogeny Construction. The matrix of phylogenetic marker states P produced in the previous step serves as the input to a character-based phylogenetic inference. Given the lack of any sound empirical basis for setting parameters for a Bayesian or maximum likelihood method, we favor use of a simpler parsimony method and therefore treat tumor phylogeny inference as the problem of finding a maximum parsimony Steiner tree [20] in which the observed components are leaves of the tree. For similar reasons, we do not weight markers, treating gain or loss of any marker as equally likely and seeking a minimum weight Steiner tree capable of explaining the data. The actual phylogeny construction is accomplished with PAUP [21] (Portable version 4.0b10 for Unix). The program was run with the maximum parsimony optimality criterion using heuristic search for 10 repetitions, random sequence addition, and the tree bisection reconnection option for swapping. Trees were visualized with GraphViz [22].

2.2. Computational Analysis

2.2.1. Simulated Data. As a first validation, we applied our methods to a set of simulated aCGH data to specifically test the effectiveness of our method at identifying markers, grouping them into components, and properly placing the components in a phylogenetic tree. We simulated data for a single hypothetical chromosome of 1000 probes, assuming cell states evolve according to a binary tree from an initially diploid root state. We then assumed each of the edges would contribute a single mutation, represented as a segment of 11 consecutive probes with amplification level 20 placed uniformly at random on the simulated chromosome, rejecting placements that would place segments less than 10 probes away from another segment. We then drew 200 simulated tumor samples from this tree of components by choosing a single node at random from the tree and using all nodes on the path from the root to chosen node as the mixture components of that sample. We chose mixture fractions for the components in each simulated tumor sample by choosing uniform random weights for each component assigned to the sample and normalizing by the sum of these weights to derive fractional contributions of each component to each mixture fraction. Finally, we add simulated Gaussian noise to each probe value for each simulated tumor sample with mean zero and standard deviation set to 0.05, 0.10, 0.15, and 0.20 in separate experiments. We then applied the NNMF-based unmixing algorithm with regularization parameter 100 and the analysis pipeline described above using a P value cutoff of 10^{-6} and window size of 5 for marker identification.

We measured accuracy based on amplified segments correctly identified, components correctly identified, and tree edges correctly identified. We first assessed the fraction of the amplified segments correctly identified during marker selection for each scenario. Next, we computed the fraction of components correctly identified, with an assignment judged correct if it was assigned the same state as the true component for all markers that were correctly identified in the previous step. Finally, we assessed the fraction of tree edges correctly identified among those subdividing nodes correctly identified in the previous step. A tree edge was considered correct if it subdivided the node set identically in the inferred tree and in the true tree when collapsed to the subset of nodes identified correctly during the marker assignment step. All three analyses were repeated for $k = 4$–7 components for each of the four noise levels.

2.2.2. Real Data. Our primary analysis consisted of application of our method to a set of previously identified mixture components derived in [17] using a publicly available set of aCGH data from sectioned primary ductal breast tumors [14]. This dataset was selected because the sectioning and cell sorting approach developed by Navin et al. was specifically chosen to facilitate phylogenetic inference and provides additional data on intratumor heterogeneity useful in validating the methods. 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.

The raw aCGH data was preprocessed and unmixed as described in our prior work [17]. As before, data was converted from log to linear domain, denoised with a total variation denoising, and unmixed to generate components. Six components were chosen, as described in the prior work, based on an analysis of the eigen-decomposition of the data. The resulting components are the same as those described in that prior paper and we refer the reader there for detailed information on the unmixing method and its application to this data set.

Phylogenetic markers were determined from the resulting component matrix as described in Approach. We used a window size of $w = 20$ for the initial sliding-window scan of the genome. The P value threshold for each window in isolation was set to 10^{-8} to account for Bonferroni correction for the 78,855 sliding windows of size 20 possible for the 78,874 probes. This threshold corresponds to a corrected P value threshold of 7.9×10^{-4}. After collapsing overlapping windows,we found a total of 27 phylogenetic marker regions significantly amplified across samples. In order to investigate the possible biological significance of these markers, we identified all genes overlapping the probe set for each marker region using the UCSC Genome Browser [23] applied to the human reference genome build 17 (NCBI35). We use NCBI build 35, rather than a more recent build, to conform to the aCGH platform specifications. We further attempted to identify any genes with a known association with cancer by manually examining Online Mendelian Inheritance in Man (OMIM) [24] entries for all genes overlapping the probes, specifically noting those with a prior association with cancers in general or breast cancers specifically.

2.2.3. Application to an Independent Data Set. As a secondary validation of our approach, we applied it to a second set of mixture components derived from a second publicly available second breast cancer aCGH dataset [25] consisting of 44 predominantly advanced primary breast tumors and 10 breast cancer cell lines. The dataset consists of 59 samples and 6691 probes each corresponding to a single gene, making it substantially lower in resolution than the Navin et al. dataset. We ran our recent NNMF-based unmixing method with TV denoising regularization parameter 6 and, unmixing parameters $k = 6$ components and $\gamma = 100$ regularization, with window size 20 and bonferroni corrected P-value cutoff 1.7×10^{-7}. While the lower resolution of the data prevented direct comparison to the Navin et al. results, we evaluated the method based on its ability to identify four markers (on 1q, 8q, 17q and 20q) specifically cited by the authors of the study as well as others that showed up as important markers in the analysis of the Navin et al. data.

3. Results and Discussion

3.1. Results

3.1.1. Simulated Data. Figure 3 summarizes results on the simulated data. Surprisingly, marker-level accuracy generally improves with increasing component numbers but appears relatively insensitive to noise level over the ranges examined here. The average accuracy across all scenarios is 79.2%. No false positive markers were detected in any of the simulations. Component-level accuracy shows a more complicated profile, with generally worse performance for larger numbers of components at any given noise level. Analysis of specific identified components suggests a common error is the identification of more than one inferred component closely corresponding to a single true component, leading to other true components getting omitted from the data. The overall average accuracy in component assignment is 72.8% over all scenarios. The accuracy of tree edges in partitioning the identified components is 100% across most noise levels and component numbers, except for 20% noise and 15% noise for $k = 6$ components and 20% noise for $k = 7$ components. The overall accuracy in inferring tree edges is 94.8%. It is important to note, though, that we defined these error measures so that the method would not be penalized for failed marker detection in assessing component or tree edge detection nor be penalized for failed component detection in assessing tree edge detection. This decision was motivated by a desire to assess the accuracy of each step independent of the others. The reported accuracies would appear more pessimistic if we counted components correct only if all markers were detected or counted tree edges incorrect if the components they separate were not detected.

3.1.2. Real Tumor Data. Application of our analysis to the [14] data yielded six components corresponding to inferred cell states, in addition to a seventh normal cell type added to root the subsequent tree. The components themselves and a detailed analysis of those components and the associated mixture fractions are provided in our prior work [17] and we therefore refer the reader to that prior literature for a detailed discussion of the mixture components by themselves.

We next analyzed the components to find significantly amplified marker regions. The analysis yielded a total of 27 nonoverlapping regions at which the components collectively showed significant amplification. The full set of marker regions is provided in Table 1. In addition, we provide a list of genes overlapping the regions that have some known association to cancers. Most of the regions contain at least one gene known to have some prior association with cancers, including several genes specifically associated with breast cancers (CD55, MDM4, WNT2, ERBB2, GRB7, BCAS, CCNE, CTTN, AURKA, BCL2, MYC, TNFRSF11A, ZNF217, CYP24A1). In several other cases, a region lacking known cancer-associated genes is found adjacent to one with a known association and might be presumed to be part of a common amplicon (e.g., 18q22.2-18q22.3).

These regions overlap a total of 343 genes, of which 56 (16.3%) were manually found to be associated with cancers in OMIM. It is difficult to rigorously establish a global frequency with which genes are cancer related, but we can derive an estimate by reference to the work Bajdik et al. [26], who used a text-mining approach to determine that 1,943 genes as of the time of their work were annotated as cancer-related in OMIM. Comparing this number to the number of Refseq transcripts, 27,704 (NCBI genome build 35), provides an estimate that 7.01% of all genes are annotated as cancer-associated in OMIM. The comparison suggests that the marker regions identified by our study are strongly enriched for known cancer-related genes. A chi-squared statistical test shows this difference in frequencies to be highly significant (chi-square score 43.2, P value <0.0001).

We would expect the unmixing to screen out amplifications that occur in only a small fraction of samples, leading to the discovery of fewer but more robust markers than would be found from the raw aCGH data. To test that assumption, we also ran the marker selection method on the raw aCGH data. This process yielded 47 marker regions, including 24 of the 27 found from the unmixed data. Three markers (Markers 6, 22, and 23) are found only from the unmixed data. Due to space limitations, we do not provide the complete list of markers obtained from the raw data.

We next assigned states to each of the identified marker regions in each component. Table 2 shows the full assignment of marker states to components. We further manually examined the copy number profiles for the predicted components in each marker region. Figure 4 provides two illustrative examples, showing the inferred copy number data for the six components and identifying those components determined to be amplified versus nonamplified. Figure 4(a) shows the inferred profile for marker 1, corresponding to locus 1q32.1-1q32.2. C1, C3, C4, and C5 are determined to be amplified, which appears to provide a good correspondence to those with copy numbers significantly above one. It is worth noting, however, that there is a finer resolution of amplification apparent in the Figure 4(a): C1 shows broad but low amplification across the region, C3 shows a more specific amplification of the subregion approximately from probes 5250 to 5300, and C4 shows a distinct pattern of multiple amplicons across the region. These observations suggest the marker-identification method is performing well at a coarse resolution but that there is considerable finer-scale structure that could in principle exploited by a more sophisticated marker selection strategy, particularly where contiguous regions show distinct patterns of amplification.

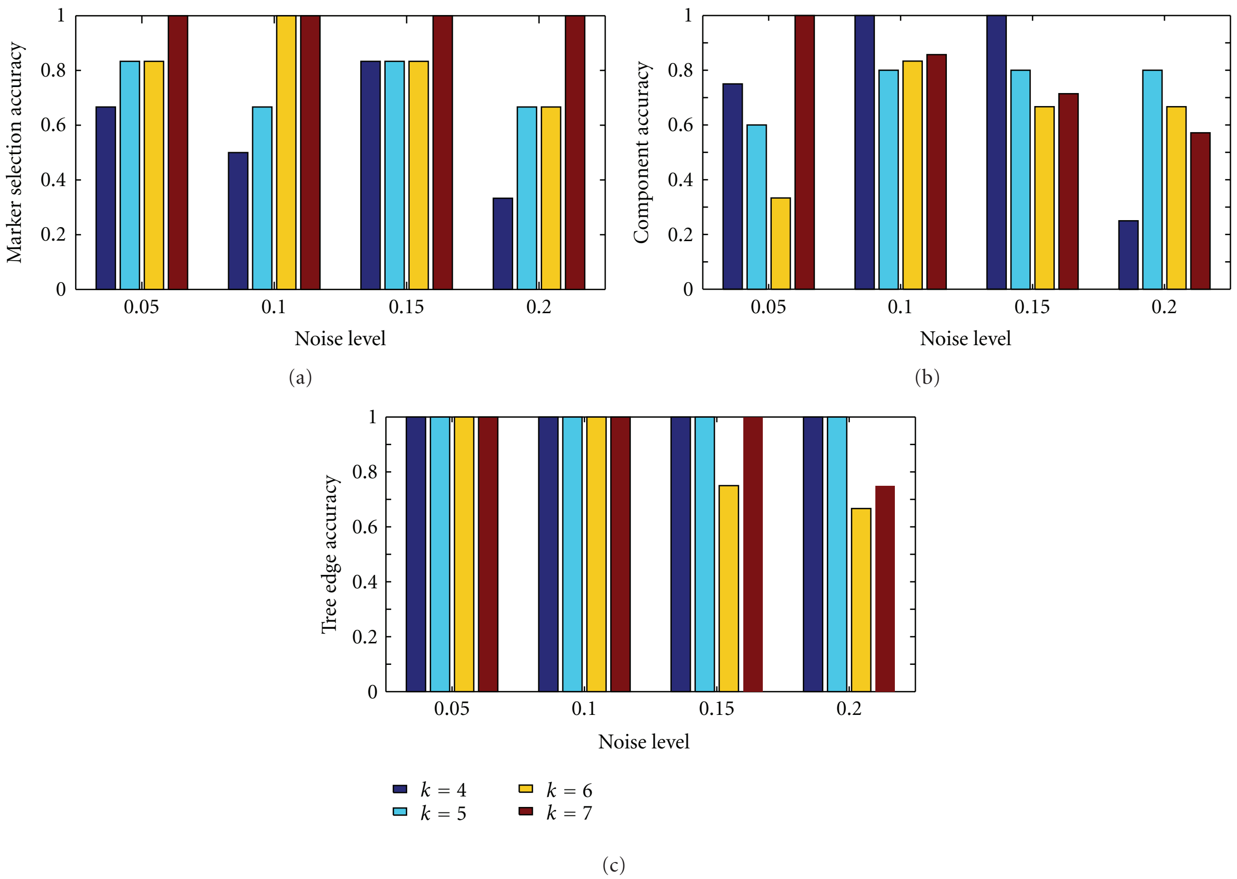


Figure 3: Quantification of accuracy on simulated data from $k = 4$–7 components and noise levels 0.05–0.20. (a) Fraction of markers correctly predicted in each experiment. (b) Fraction of components correctly identified on all identified markers in each experiment. (c) Fraction of tree edges correctly identified for the components and markers identified in each experiment.

Figure 4(b) shows a second example, the inferred copy number profile for marker 20, corresponding to an amplicon at 17q12-17q21.2. We would expect this site to be picked up as a marker and to show high amplification, since it is the site of the Her-2 locus. The region again shows a strong but selective amplification, with C5 and C6 highly amplified (although with distinct fine-scale structures), C4 slightly amplified, and others showing no amplification. The result again confirms that the method produces correct answers at a coarse resolution, although there may be a finer-scale structure that could exploited by a more sophisticated method.

Using the resulting probes, we then performed phylogenetic inference. Figure 5 shows the phylogenetic tree produced from the six inferred progression components and the additional normal component manually added to the analysis. The majority of markers are gained at a unique point in the tree and never subsequently lost. Marker 9 (8q12.1) is lost in the tree in the transition to component C4. In addition, some markers are inferred to be gained more than once in the tree. Most notable of these is the collection of 17q markers, which are gained separately in the subtree leading to component 6 and that leading to Steiner node 8 and then to components 4 and 5.

3.1.3. Application to an Independent Data Set. Application to a second component set derived from the lower-resolution data of Pollack et al. [25] provides a secondary validation of the reproducibility of the results on distinct datasets, aCGH platforms, and unmixing methods for a common tumor type. The method identified 20 markers, shown in Table 3. The lower resolution of the data leads to substantially more possible genes per amplicon than were found with the Navin et al. data, making it infeasible to conduct a similar analysis of the genes identified. We therefore must compare the two results more indirectly based on markers reported by Pollack et al. in their own analysis of their data as well as known breast cancer markers found in the primary analysis of the Navin et al. data above. Pollack et al. described finding 1q, 8q, 17q, and 20q as predominantly amplified regions in the data, and our method did find sizeable amplicons on each of these regions. Other amplicons appear to correspond to several important tumor markers, including the HER2, CCND1, c-myc, and CCNE1 loci noted in the analysis of the Navin et al. data as well as the FGFR1 locus that is conspicuously absent from our analysis of the Navin et al. data. Of note, the CCNE1 locus is found as a significant marker when analyzing the unmixed components but is not detected by a similar marker analysis of the raw data without unmixing. All other

Table 1: Marker regions determined to be significantly amplified across components for the data of Navin et al. [14]. The table provides, for each marker region, a unique identifier, cytogenetic coordinates, probe positions along the genomic axis, and gene IDs for genes identified as having some known association with cancers.

Marker ID	Cytogenetic coordinates	Chromosome positions	Annotated cancer-related genes
1	1q32.1-1q32.2	196117366-206330147	CD55, MDM4, NR5A2,PTPN7, IL10, CNTN2,CD34
2	1q44	242649493-245131380	SMYD3
3	2p12	76777788-78642108	None
4	3q25.1-3q25.2	151037467-154216571	None
5	5p15.33-5p14.2	3485419-24119655	PAPD7, TAG, CDH18
6	5q21.1-5q21.3	100224934-106834646	None
7	5q22.3-5q23.1	115172420-118711133	TNFAIP8, ATG12, SEMA6A
8	7q31.2-7q31.31	116016939-120372452	ING3,ASZ1, WNT2, ST7
9	8q12.1	55969808-58018737	None
10	8q12.3-8q13.2	63931435-69387571	MYBL1
11	8q13.2-8q13.3	69634776-74092165	TRPA1
12	8q21.11-8q24.3	77351432-143296089	MYC
13	11q13.2-11q13.4	67830873 -70354248	CCND1, CTTN, FGF4, FGF3
14	11q14.1-11q13.4	74383378-82935709	None
15	11q23.3	115171785-116542726	None
16	12p11.22-12p11.21	28901065-33207415	ERGIC2
17	15q25.2-15q25.3	82525637-85513682	None
18	15q26.3	96434691-99661839	None
19	17q11.2	23392447-25127504	RAB34, NEK8, TRAF4, FOXN1
20	17q12-17q21.2	32705491-37628927	STAT5, ERBB2, GRB7
21	17q21.33	45403785-47282174	SPAG9, UTP18, CA10, ANKRD40, CACNA1G, PPP1R9B
22	18q21.32-18q22.2	56806538-66527883	TNFRSF11A, BCL2, SERPINB5, SERPINB13, SERPINB4, SERPINB3, CDH19
23	18q22.2-18q22.3	66607283-71314138	None
24	19q12	34017456-36812510	CCNE1
25	20q13.12	44249187-45563781	None
26	20q13.2-20q13.32	50440150-57022263	ZNF217, CYP24A1, BCAS1, AURKA, CTCFL, ZBP1, RAB22A, GNAS, SDX16
27	20q13.33	57624055-58571221	None

markers found in the unmixed data are also found in the raw data, as was observed with the Navin et al. data. Figure 6 shows the inferred phylogenetic tree. For these data, it was not necessary to add a normal root component C0, as was done with the Navin et al. data, because the method directly inferred component C1 to be nonamplified at all markers and thus to serve as the expected normal root.

3.2. Discussion. Analysis on simulated data shows the method to have generally good accuracy at identifying amplified markers, identifying complete components with defined patterns of marker amplification, and grouping these components into phylogenies. The dependence of accuracy on various model parameters is difficult to analyze, with generally better marker-level accuracy but worse component-level and tree-edge-level accuracy as greater numbers of components are modeled. Examination of different noise levels, chosen to roughly approximate noise levels observed on the real data, shows no strong dependence within a range of 5–20% noise. Overall, the results suggest that methods show good although far from perfect performance, picking out 79.2% of true markers and greater than 72.8% of true components in most scenarios and correctly identifying 94.8% of tree edges dividing the identified components. The high specificity of the marker assignment, with no false positives observed in any of the tests, suggests that there

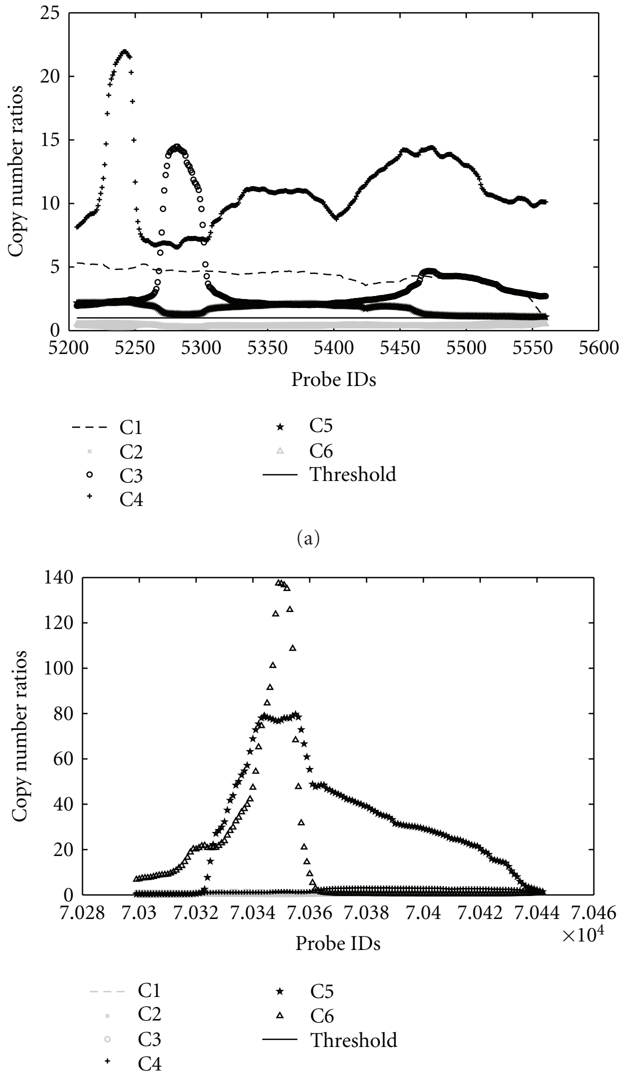
Table 2: Phylogenetic states of all components at all identified progression markers for the data of Navin et al. [14]. Columns show the states for the six inferred components (C1–C6). The additional normal component (C0) used to root the tree is included for completeness. "1" corresponds to an amplified region and "0" to nonamplified.

Marker ID	C0	C1	C2	C3	C4	C5	C6
1	0	1	0	1	1	1	0
2	0	1	0	1	1	1	0
3	0	0	0	1	0	0	0
4	0	0	0	1	0	0	0
5	0	0	0	1	1	1	0
6	0	0	0	0	1	0	1
7	0	0	0	0	1	0	1
8	0	0	0	1	0	0	0
9	0	1	0	1	0	1	0
10	0	1	0	1	1	1	0
11	0	1	0	1	1	1	0
12	0	1	0	1	1	1	0
13	0	0	0	0	1	1	0
14	0	0	1	0	1	0	0
15	0	0	0	0	0	0	1
16	0	1	0	0	1	0	0
17	0	0	0	0	1	0	0
18	0	0	0	0	1	0	1
19	0	0	0	0	1	1	1
20	0	0	0	0	1	1	1
21	0	0	0	0	1	1	1
22	0	0	1	0	0	0	0
23	0	0	1	0	0	0	0
24	0	1	0	0	0	0	1
25	0	0	0	1	1	1	0
26	0	0	0	1	1	1	0
27	0	0	0	1	1	1	0

may be room to tune the methods to improve accuracy by trading off sensitivity for a somewhat higher rate of false positives. While simulated data provides some assessment of the effectiveness of the method, however, there are many features of tumor evolution that are not yet well enough understood to permit a faithful simulation of real tumor data. In assessing our methods, we must therefore rely primarily on more indirect validation on real data.

There is no closely comparable method to ours of which we are aware that we could use as a basis for comparison and we therefore validate the results on the Navin et al. data primarily by considering whether they are consistent with prior knowledge about breast tumors. One could in principle validate our results against recent work of Navin et al. [27] using single-cell analysis of the subsections of Tumor 10 analyzed here. Navin's phylogenetic approach, however, leads to progression trees dominated by changes in overall ploidy, which is not examined in our trees and precludes any direct comparison. As noted previously, a majority of the markers we find correspond to some genes with known cancer associations. These include well-characterized breast cancer amplicons at 17q, 11q, and 20q [28–30]. The most notable absence among well-known breast cancer markers would be the 8p locus associated with the gene FGFR1. A majority of the markers (16 of 27) include genes with some annotated relationship with cancers, although only 7 of those (markers 1, 12, 13, 20, 22, 24, and 26) are annotated in OMIM as specifically associated with breast cancers.

Of those markers lacking an annotated association with breast cancers, many are in close proximity to and inherited with breast-cancer-associated markers and might plausibly be assumed to contain distinct portions of common amplicons. Table 4 identifies those proximal markers that are coinherited in the tree and likely reflect common amplicons. For example, 17q is interpreted as three distinct markers (markers 19–21), and although only marker 20 contains genes with an annotated breast cancer association (ERBB2/Her-2/neu, STAT5, and GRB7), all are inherited together apparently as a common amplicon. Similar explanations can account for markers 2 on 1q, which is coinherited with marker 1 (MDM4); markers 10 and 11 on 8q, which are coinherited with marker 12 (MYC); marker 25 and 27 on 20q, which are



(a)

(b)

Figure 4: Inferred copy number profiles for mixture components in the vicinity of three markers from the data of Navin et al. [14]. The x-axis of each figure corresponds to probes within a specific marker region and the y-axis to copy number relative to the diploid control in that region for each component. The thin solid line in each plot at value 1 shows the diploid threshold. Amplified components appear in black and nonamplified in grey. (a) Marker 1, corresponding to the amplicon at 1q32.1-1q32.2. (b) Marker 20, corresponding to the amplicon at 17q12-17q21.2.

Table 3: Amplified markers with probe boundaries and corresponding cytogenetic coordinates for the data of Pollack et al. [25].

Marker ID	Start probe ID	End probe ID	Cytogenetic coordinate
1	1	37	1p36
2	136	272	1p34-1p22
3	330	649	1p13-1q44
4	671	790	2p24-2p13
5	1170	1210	3p25-3p21
6	1810	1889	5p15-5q11
7	2056	2229	5q23-6p21
8	2253	2331	6p21
9	2532	2865	7p22-7q36
10	2935	3106	8p12-8q24
11	3235	3264	9q22-9q31
12	3800	3926	11q12-11q14
13	4194	4255	12q12-12q14
14	4478	4522	13q22-14q11
15	4523	4566	14q11-14q12
16	4968	5367	16p13.3-17q11
17	5384	5448	17q11.2-17q21
18	5478	6056	17q21-19q13.4
19	6057	6230	20p13-20q13.33
20	6231	6312	21q11-21q22.3

Table 4: Marker regions amplified simultaneously during tumor evolution. The table provides, for each such set of marker regions, a unique identifier, cytogenetic coordinates, and corresponding specific edges or paths in the phylogenetic tree.

Coamplified markers	Phylogeny edges
18q21.32-18q22.2,18q21.2-18q21.3	12→ C2
1q32.1-1q32.2,1q44	11 → 10
5q21.1-5q21.3, 5q22.3-5q23.1	11 → C6, 8 → C4
8q12.3-8q13.2,8q13.2-8q13.3,8q21.11-8q24.3	11 → 10
20q13.12,20q13.2-20q13.32,20q13.33	10 → 9
7q31.2-7q31.31	9 → C3
15q25.2-15q25.3,15q26.3	8 → C4
17q11.2,17q12-17q21.2,17q21.33	11→ C6, 9 → 8

coinherited with marker 26 (ZNF217, CYP24A1, BCAS1, and AURKA). In other cases, however, we observe coinherited markers for which no specific explanation is available for any of the markers. It is impossible to say purely from a computational analysis whether these represent false positives, discoveries not annotated specifically in OMIM, or even novel but significant associations with breast cancer progression.

Examining the phylogeny itself allows us to further examine the possible biological significance of the data and its concordance with current knowledge about breast cancer progression. In this regard, it is helpful to interpret the tree as a set of possible progression pathways from the healthy root cell type (C0). As the tree implies, however, different progression pathways do not function in isolation but rather may share some common features in early progression.

The first internal node, Steiner node 12, is inferred to be identical to the root, but diverges at the top level into two pathways. The first such progression pathway (C0 →

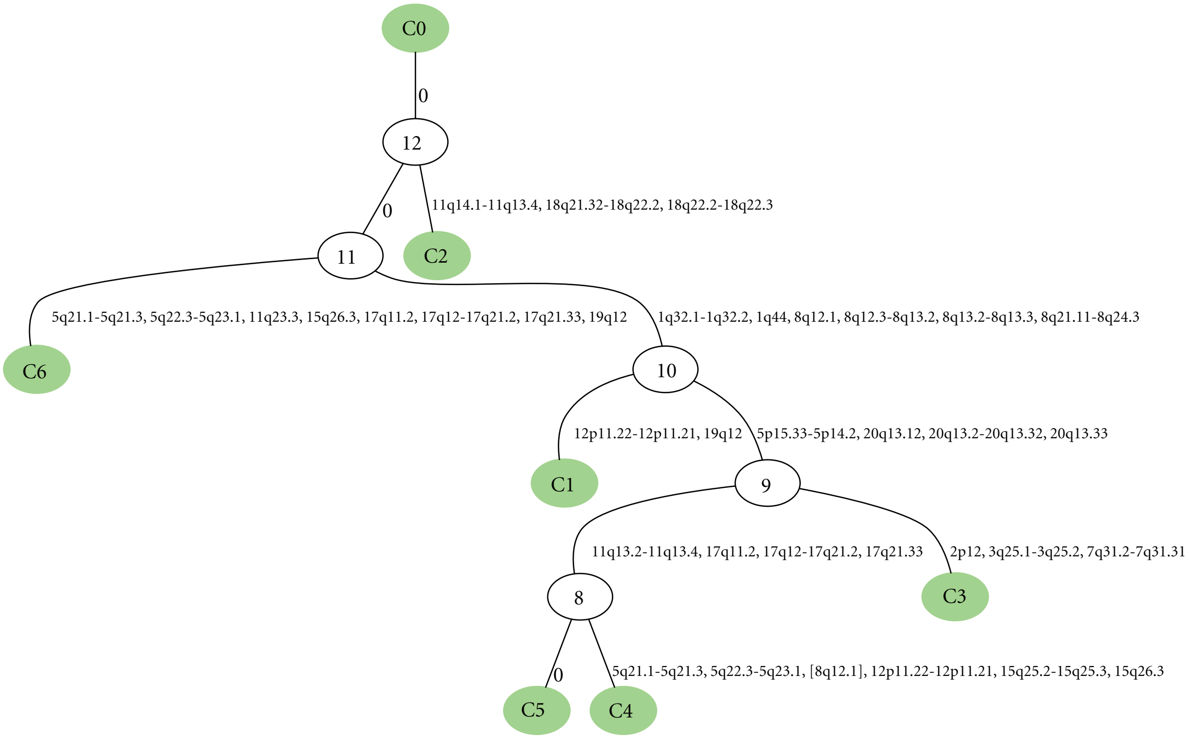

C0
0
12
0
11q14.1-11q13.4, 18q21.32-18q22.2, 18q22.2-18q22.3
11
C2
5q21.1-5q21.3, 5q22.3-5q23.1, 11q23.3, 15q26.3, 17q11.2, 17q12-17q21.2, 17q21.33, 19q12
1q32.1-1q32.2, 1q44, 8q12.1, 8q12.3-8q13.2, 8q13.2-8q13.3, 8q21.11-8q24.3
C6
10
12p11.22-12p11.21, 19q12
5p15.33-5p14.2, 20q13.12, 20q13.2-20q13.32, 20q13.33
C1
9
11q13.2-11q13.4, 17q11.2, 17q12-17q21.2, 17q21.33
2p12, 3q25.1-3q25.2, 7q31.2-7q31.31
8
C3
0
5q21.1-5q21.3, 5q22.3-5q23.1, [8q12.1], 12p11.22-12p11.21, 15q25.2-15q25.3, 15q26.3
C5
C4

FIGURE 5: Inferred phylogenetic tree for the mixture components from the data of Navin et al. [14]. Nodes are labeled by component for the six inferred components C1–C6 and the normal component C0. Internal nodes are inferred ancestral states (Steiner nodes) and are each labeled by a unique identifier (8–12). Tree edges are labeled with the markers inferred to be amplified across each. Markers inferred to be lost along a given edge are shown in brackets and edges with no markers gained or lost are labeled "0."

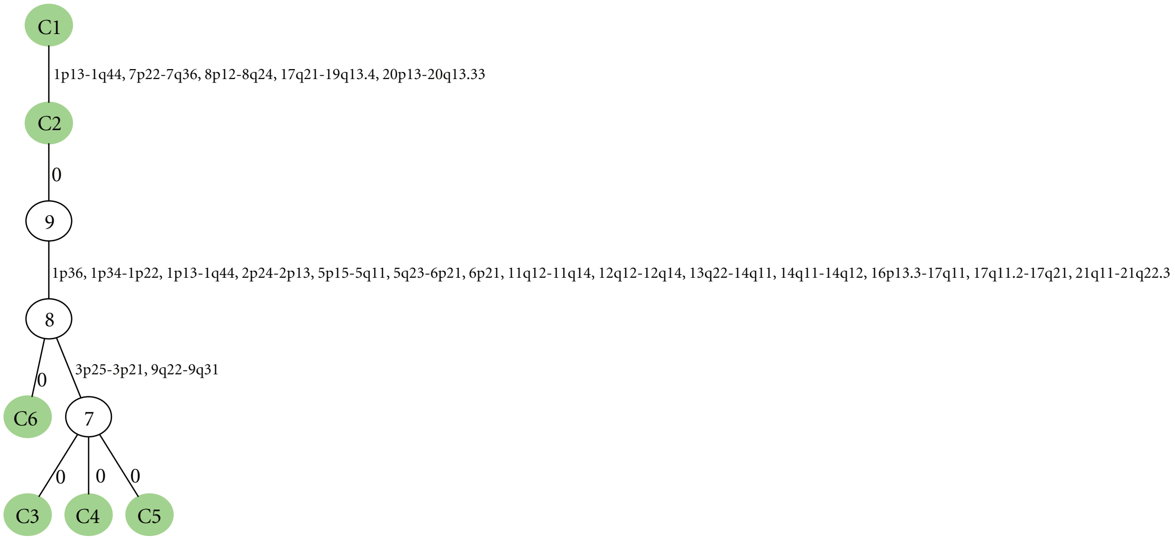

C1
1p13-1q44, 7p22-7q36, 8p12-8q24, 17q21-19q13.4, 20p13-20q13.33
C2
0
9
1p36, 1p34-1p22, 1p13-1q44, 2p24-2p13, 5p15-5q11, 5q23-6p21, 6p21, 11q12-11q14, 12q12-12q14, 13q22-14q11, 14q11-14q12, 16p13.3-17q11, 17q11.2-17q21, 21q11-21q22.3
8
0
3p25-3p21, 9q22-9q31
C6
7
0
0
0
C3
C4
C5

FIGURE 6: Inferred phylogenetic tree for components derived from the data of Pollack et al. [25]. Nodes are labeled by component for the six inferred components C1–C6. Internal nodes are inferred ancestral states (Steiner nodes) and are each labeled by a unique identifier (7–9). Tree edges are labeled with the markers inferred to be amplified across each. Markers inferred to be lost along a given edge are shown in brackets and edges with no markers gained or lost are labeled "0."

12 → C2) describes a short terminal progression pathway isolated from the rest of the tree. The progression pathway is resolved only to a single step of mutation corresponding to amplification of 11q14.1–11q13.4, 18q21.32-18q22.2, 18q22.2–18q22.3. 11q is a known breast cancer amplicon [29, 30] and harbors CCND1, which has been found to be amplified in breast cancers [31]; FGF3 and FGF4, which are known oncogenes [32]; and CTTN, which is frequently overexpressed in breast cancers [33]. The region also contains other genes, such as NPAT, with functions in cell cycle regulation that might be considered candidates for an oncogenic function. 18q21.32-18q22.2 harbors the oncogene BCL2, which is involved in the MYC pathway [34] and TNFRSF11A, which is frequently expressed in late-stage breast cancers [35, 36]. The marker also harbors several SERPIN genes known to be tumor associated. 18q22.2–18q22.3 does not carry any currently known cancer-related genes but may be gained due to proximity to 18q21.32-18q22.2 as part of a common amplicon. Together, these abnormalities appear to define a distinct subclass of breast tumor cells with early divergence from all other cell types.

Within the subbranch rooted at Steiner node 11, one branch leads directly to a terminal node characterizing a second progression pathway (C0 → 11 → C6). This progression pathway is characterized by amplification of 5q21.1-5q21.3, 5q22.3-5q23.1, 11q23.3, 15q26.3, and 19q12 and is one of two subtrees characterized by amplification of 17q11.2, 17q12-17q21.2, and 17q21.33. The 17q region is a well-established breast cancer hotspot [28, 30], including genes ERBB2 (Her-2/neu), GRB7, and STAT5. 19q12 contains CCNE1, an important prognostic marker for breast cancer progression [37, 38]. CCNE1 amplification has been specifically associated with basal-like breast cancers [39], but has been previously identified as coassociated with particularly aggressive Her-2 positive breast tumors [40]. Our phylogeny is consistent with the notion that 17q/19q coamplification defines a distinct subtype of Her-2 positive tumors. Region 15q26.3 has no genes specifically noted to be breast-cancer associated in OMIM, although amplification of the locus was identified as predictive of recurrence in systematic breast cancers [41] and the region contains IGF1R, an antiapoptotic gene broadly amplified in cancers [42]. The biological significance of the 5q amplicon is not apparent. While 5q22.3-5q23.1 has several genes associated with cancers (e.g., ATG12, TNFAIP8, SEMA6A, which are associated with lung cancer), they are predominantly tumor suppressors. Likewise, there is no obvious relevance to the 15q amplicon, although it is close to other known 15q markers.

The next major division in the tree corresponds to the branch from Steiner nodes 11 to 10, characterized by gains in 1q32.1-1q32.2, 1q44, 8q12.1, 8q12.3-8q13.2, 8q13.2-8q13.3, and 8q21.11-8q24.3. Both 1q and 8q are rich in tumor-associated genes. 1q32.1 includes the breast cancer associated gene MDM4, a putative oncogene involved in apoptosis regulation of p53 activity [43], in addition to various genes associated with cancers more generally. 8q21.11-8q24.3 includes the MYC locus, another well known breast cancer amplicon [30]. We can suggest, then, that the 11 → 10 branch corresponds to a specific subset of progression pathways characterized by MYC amplification and suppression of apoptosis.

A third progression pathway can be identified within this branch through progression into C1 (C0 → 12 → 11 → 10 → C1). The final step on this pathway is characterized by amplifications on 12p11.22-12p11.21 and 19q12. 19q12 is the locus of CCNE1 suggesting a generic connection to cell cycle control on this pathway. 12p11.22-12p11.21 has no known cancer-related genes but carries the apoptosis-related gene DNM1L and the telomerase-related gene DDX11 [44].

Further progression pathways diverge from Steiner node 10 through Steiner node 9 with gains on 5p15.33-5p14.2, 20q13.12, 20q13.2-20q13.32, and 20q13.33. The 5p amplicon contains two genes with known cancer associations, CDH18 [45] and PAPD7 [46], although neither appears to have a known role in breast cancers specifically. 20q13.2-20q13.32 contains several genes associated with breast cancers, including ZNF217, CYP24A1, BCAS1, and AURKA [30], making it difficult to ascribe a particular mechanism to this branch.

Within the Steiner node 9 subtree, we can characterize a fourth progression pathway terminating in C3 (C0 → 11 → 10 → 9 → C3). The final step on this progression pathway corresponds to gains on 2p12, 3q25.1-3q25.2, and 7q31.31-7q31.32. The 7q31.32 marker contains the WNT2 gene associated with many cancer types, including breast cancer [47]. 7q31.31 has no known cancer-related genes and is perhaps gained due to its proximity to 7q31.2. 3q25.1-3q25.2 has been previously detected as an amplicon in fraction of breast cancers [48], although we can offer no mechanistic explanation for its presence. We are not aware of any prior suggestion of an association between 2p12 and cancers.

The remaining two terminal nodes of the tree, C4 and C5, appear likely to represent two steps on a common progression pathway. Both branchs from Steiner node 9 through 8 by acquisition of 17q11.2, 17q12-17q21.2, 17q21.33 (the Her-2 locus) along with 11q13.2. This subtree might thus be characterized primarily as a second Her-2 positive progression group associated with gain of CCND1, distinct from the Her-2 positive progression group terminating at C6 and associated with gain of CCNE1. C5 branches from Steiner node 8 with no changes, indicating a single progression pathway corresponding to C0 → 11 → 10 → 9 → C5 → C4. The final step in this pathway is then characterized by a series of amplifications on 5q21.1, 5q22.3, 12p11.22, 15q25.2, 15q26.3, and loss of 8q12.1. We would not expect loss of a previously gained marker, and can suggest that this apparent loss might be better explained as a miscall of the state of that marker. Most of these loci have no annotated association with any cancers, with the only specific annotated breast cancer association being to 11q13.2, described above. This lack of associations may again represent false positive inferences specifically associated with this component. We can suggest, however, that such markers might be have been missed if they are specific only to late progression of one subtype of Her-2 positive breast tumor. Summarizing across the tree, we can note that there is clear support in the prior literature for many of the specific markers, although there is little evidence one way or the other supporting the specific sequences of mutations suggested by our phylogeny analysis. Nonetheless, these pathways make several novel predictions that may warrant further investigation. Chief among these would be the identification of two apparently distinct pathways to Her-2/neu amplification that separate relatively early in progression and exhibit distinct sets of co-occurring amplifications.

The tree suggests several distinct patterns of coamplification that may be useful in identifying or classifying novel subtypes, particularly with respect to Her-2 amplifying tumors. Of particular interest are the observation of two distinct Her-2 amplifying subtrees, one showing coamplification with CCND1 and c-myc and the other with CCNE1. Loden et al. have previously reported separate Cyclin-D amplified and Cyclin-E amplified subgroups of breast cancer following separate pathways of oncogenesis, with Her-2/neu overexpression and c-myc amplification accompanying both subgroups. Coamplification of Her-2, CCND1, and c-myc is supported by additional literature, with this particular coamplification associated with later or more advanced stages of breast cancer [40, 49, 50]. Janocko et al. [49], however, do suggest that c-myc amplification should occur late in this sequence, a finding not supported by our phylogeny. Other more recent work has supported the idea of Her-2 and CCNE1 coamplification in breast cancers [51, 52] with Scaltriti et al. specifically suggesting this coamplification as a possible mechanism for Herceptin resistance in Her+ breast tumors. Other patterns of complication are apparent in the tree although not to our knowledge supported by prior literature or any obvious functional interpretation, for example, the observation of coamplification of loci on 5q and 15q in both Her-2 amplifying subtrees.

Additional analysis of the Pollack et al. [25] provides little additional insight into breast tumor development, although it does provide some independent validation of our method. While the lower resolution of those data prevents an analysis of specific amplified breast tumor genes comparable to that done with the Navin et al. data, we can nonetheless observe that the method is effective at picking out those amplicons noted by the authors of that study. Furthermore, the additional markers it detects beyond those four include several of those also inferred to be important progression markers on the Navin et al. data and supported by extensive prior literature, most prominently the loci of Her-2, CCND1, and CCNE1. These results show that the method can robustly find at least some prominent known tumor markers across two distinct sets of tumor samples using very different aCGH platforms and distinct unmixing methods. The tree itself provides no obvious new insights into breast tumor progression, as the method detected only four components that were actually distinct at the level of assigned markers, with three components determined to be amplified at all markers. Furthermore, all identified components were inferred to lie along a single progression pathway. It is notable that the tree implies amplification of most of the identified markers in a majority of components, perhaps because of the late clinical stages of the tumor samples and the presence of cell lines that would provide reasonably homogeneous representations of advanced states of breast tumor progression.

4. Conclusion

In this paper, we have developed a computational pipeline for tumor phylogeny inference from genome-scale profiles of tumor state, specifically to test the feasibility of using computational unmixing methods to circumvent the problem of cell type heterogeneity in tumor phylogeny in-ference. We have developed a set of statistical tests to allow us to analyze computationally inferred mixture components—representing inferred profiles of well-populated cell types from which heterogeneous tumor samples can be explained—to identify phylogenetic markers, assign them to specific inferred cell types, and use them in phylogenetic inference of tumor progression. We have demonstrated the approach with specific application to aCGH DNA copy number data, applied to a breast cancer data set [14], showing that the method is effective at locating biologically meaningful markers of tumor progression and assembling a biologically plausible model of breast tumor progression pathways. The inferred progression pathways provide several novel suggestions about possible steps in tumor evolution and key molecular abnormalities associated with progression. These inferences may provide useful guidance into the basic biology of tumor development as well as suggestions of possible targets for future diagnostics and therapeutics. Further application to a secondary lower-resolution breast tumor data set [25] and to a series of simulated aCGH data sets provides additional evidence for the effectiveness of the method at identifying markers of tumor progression, grouping them correctly into well-represented progression states, and accurately placing these states in phylogenetic trees.

Validation remains a challenge for tumor phylogeny inference, as there is no alternative method by which we can determine progression pathways with certainty for any real tumor data set. Simulated data can lend some confidence that the method works effectively relative to a model of the real data, as has been done here, but real tumor progression mechanisms are likely to be far more complex than our simulation models can capture. Comparison to single-cell approaches like FISH [13, 49, 53] and single-cell sequencing efforts [27] can help to verify the pure cell states determined by the unmixing within a single sample and potentially validate some ancestral states predicted by the phylogenetic inference. FISH data provides only a few markers per cell, making it infeasible for a comprehensive validation of the results of our method, but could be used prospectively on targeted markers selected from an inferred phylogeny. Single-cell sequencing approaches could in principle eventually overcome this limitation given sufficient volumes and quality of data. Other sources of data in which more information is available about the true pathways of progression might also be useful. While we know of no such data currently available, one might in principle construct such a data set by, for example, studying discrete passages of cell lines or through the use of animal models in which one can monitor tumor development and progression over time. While gathering such a data set would be beyond the scope of the present work, it could in principle provide a basis for a more thorough future assessment of the accuracy of the pipeline implemented here or other methods for the problem of tumor phylogeny inference.

While this pilot study was intended to establish the feasibility of an unmixing approach to tumor phylogenetics, there are many ways by which the work might be advanced in the future. It will be important to further establish the reproducibility of the specific markers and phylogenetic pathways in additional breast tumor datasets. Novel markers found to be robustly predictive of particular progression pathways will ultimately need to be experimentally verified. In addition, it will be important to establish that the approach is applicable to other forms of tumors. Each of the individual steps of analysis also might benefit from improvement. The approach developed here depends on use of an unmixing method for identifying progression states, a problem which itself might benefit from improvements in the model and algorithms to more precisely fit the kind of sparse noisy data characteristic of tumor data sets. Adapting the methods to more reliable data types, such as next generation sequencing data, may also prove valuable in that regard. The results on marker detection suggest there is room for improvement in more precisely determining the fine-scale structure of specific amplicons, especially when contiguous regions show distinct patterns of amplification across components. Likewise, there would appear to be room for improvement in better discriminating between normal and slightly elevated copy numbers. It is a weakness of the general approach that, because the unmixing models must work in linear rather than log space, they have difficulty distinguishing the relatively small linear change between normal and deleted regions. Improving sensitivity for deletions, or for subtler variations among amplification levels, may provide additional data for phylogeny construction. Finally, the phylogeny construction itself used a standard parsimony method not specifically tailored to tumor progression. This parsimony model has advantages in not requiring parameters for which there is currently no empirical basis and in allowing us to test for unexpected behavior, such as loss of previously amplified regions, that can help to validate the method. Nonetheless, there is now sufficient data that one might in principle learn more sophisticated probabilistic models of cancer progression or of the behavior of particular amplicons and build these models into the phylogeny inference.

Acknowledgments

A. Subramanian, S. Shackney, and R. Schwartz were supported in this work by US National Institutes of Health award 1R01CA140214 and R. Schwartz by US National Science Foundation award 0612099 and US National Institutes of Health award 1R01AI076318.

References

[1] C. M. Perou, T. Sørile, M. B. Eisen et al., “Molecular portraits of human breast tumours,” *Nature*, vol. 406, no. 6797, pp. 747–752, 2000.

[2] T. R. Golub, D. K. Slonim, P. Tamayo et al., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[3] C. Sotiriou, S. Y. Neo, L. M. McShane et al., “Breast cancer classification and prognosis based on gene expression profiles from a population-based study,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 18, pp. 10393–10398, 2003.

[4] T. Sørlie, C. M. Perou, R. Tibshirani et al., “Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 19, pp. 10869–10874, 2001.

[5] D. T. Ross, U. Scherf, M. B. Eisen et al., “Systematic variation in gene expression patterns in human cancer cell lines,” *Nature Genetics*, vol. 24, no. 3, pp. 227–235, 2000.

[6] M. D. Pegram, G. Konecny, and D. J. Slamon, “The molecular and cellular biology of HER2/neu gene amplification/overexpression and the clinical development of herceptin (trastuzumab) therapy for breast cancer,” *Cancer Treatment and Research*, vol. 103, pp. 57–75, 2000.

[7] A. Kamb, S. Wee, and C. Lengauer, “Why is cancer drug discovery so difficult?” *Nature Reviews Drug Discovery*, vol. 6, no. 2, pp. 115–120, 2007.

[8] I. Bozic, T. Antal, H. Ohtsuki et al., “Accumulation of driver and passenger mutations during tumor progression,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 43, pp. 18545–18550, 2010.

[9] C. A. Smith, A. A. Pollice, L. P. Gu et al., “Correlations among p53, Her-2/neu, and ras overexpression and aneuploidy by multiparameter flow cytometry in human breast cancer: evidence for a common phenotypic evolutionary pattern in infiltrating ductal carcinomas,” *Clinical Cancer Research*, vol. 6, no. 1, pp. 112–126, 2000.

[10] R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer, “Inferring tree models for oncogenesis from comparative genome hybridization data,” *Journal of Computational Biology*, vol. 6, no. 1, pp. 37–51, 1999.

[11] R. Schwartz and S. E. Shackney, “Applying unmixing to gene expression data for tumor phylogeny inference,” *BMC Bioinformatics*, vol. 11, article no. 42, 2010.

[12] G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz, “Reconstructing tumor phylogenies from heterogeneous single-cell data.,” *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 407–427, 2007.

[13] G. Pennington, C. A. Smith, S. Shackney, and R. Schwartz, “Expectation-maximization method for reconstructing tumor phylogenies from single-cell data,” in *Computational Systems Bioinformatics Conference (CSB ’06)*, pp. 371–380, 2006.

[14] N. Navin, A. Krasnitz, L. Rodgers et al., “Inferring tumor progression from genomic heterogeneity,” *Genome Research*, vol. 20, no. 1, pp. 68–80, 2010.

[15] R. Etzioni, S. Hawley, D. Billheimer, L. D. True, and B. Knudsen, “Analyzing patterns of staining in immunohistochemical studies: application to a study of prostate cancer recurrence,” *Cancer Epidemiology Biomarkers and Prevention*, vol. 14, no. 5, pp. 1040–1046, 2005.

[16] N. Beerenwinkel, M. Däumer, T. Sing et al., “Estimating

HIV evolutionary pathways and the genetic barrier to drug resistance," *Journal of Infectious Diseases*, vol. 191, no. 11, pp. 1953–1960, 2005.

[17] D. Tolliver, C. Tsourakakis, A. Subramanian, S. Shackney, and R. Schwartz, "Robust unmixing of tumor states in array comparative genomic hybridization data," *Bioinformatics*, vol. 26, no. 12, pp. i106–i114, 2010.

[18] R. Ehrlich and W. Full, "Sorting out geology—unmixing mixtures," in *Use and Abuse of Statistical Methods in the Earth Sciences*, pp. 33–46, Oxford University Press, 1987.

[19] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[20] A. W. F. Edwards and C. L. L. Sforza, "The reconstruction of evolution," *Heredity*, vol. 18, 1963.

[21] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods)*, Version 4. Sinauer Associates, Sunderland, Mass, USA, 2003.

[22] J. Ellson, E. Gansner, L. Koutsofios, S. North, and G. Woodhull, "Graphviz— open source graph drawing tools," in *Graph Drawing*, P. Mutzel, M. Jünger, and S. Leipert, Eds., vol. 2265 of *Lecture Notes in Computer Science*, pp. 594–597, Springer, Berlin, Germany, 2002.

[23] W. James Kent, C. W. Sugnet, T. S. Furey et al., "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, 2002.

[24] McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). Online Mendelian Inheritance in Man, OMIM (tm), 2010, http://www.ncbi.nlm.nih.gov/omim/.

[25] J. R. Pollack, T. Sørlie, C. M. Perou et al., "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 20, pp. 12963–12968, 2002.

[26] C. D. Bajdik, B. Kuo, S. Rusaw, S. Jones, and A. Brooks-Wilson, "CGMIM: automated text-mining of Online Mendelian Inheritance in Man (OMIM) to identify genetically-associated cancers and candidate genes," *BMC Bioinformatics*, vol. 6, article no. 78, 2005.

[27] N. Navin, J. Kendall, J. Troge et al., "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, no. 7341, pp. 90–95, 2011.

[28] J. S. Ross and J. A. Fletcher, "The HER-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy," *Stem Cells*, vol. 16, no. 6, pp. 413–428, 1998.

[29] S. Saito, K. Morita, and T. Hirano, "High frequency of common DNA copy number abnormalities detected by bacterial artificial chromosome array comparative genomic hybridization in 24 breast cancer cell lines," *Human Cell*, vol. 22, no. 1, pp. 1–10, 2009.

[30] I. Bièche and R. Lidereau, "Genome-based and transcriptome-based molecular classification of breast cancer," *Current Opinion in Oncology*, vol. 23, no. 1, pp. 93–99, 2011.

[31] K. Mu, L. Li, Q. Yang et al., "Detection of CHK1 and CCND1 gene copy number changes in breast cancer with dual-colour fluorescence in-situ hybridization," *Histopathology*, vol. 58, no. 4, pp. 601–607, 2011.

[32] B. M. Zaharieva, R. Simon, P. A. Diener et al., "High-throughput tissue microarray analysis of 11qI3 gene amplification (CCND1, FGF3, FGF4, EMS1) in urinary bladder cancer," *Journal of Pathology*, vol. 201, no. 4, pp. 603–608, 2003.

[33] E. Schuuring, E. Verhoeven, W. J. Mooi, and R. J. A. M. Michalides, "Identification and cloning of two overexpressed genes, U21B31/PRAD1 and EMS1, within the amplified chromosome 11q13 region in human carcinomas," *Oncogene*, vol. 7, no. 2, pp. 355–361, 1992.

[34] S. M. Aukema, R. Siebert, E. Schuuring et al., "Double-hit B-cell lymphomas," *Blood*, vol. 117, no. 8, pp. 2319–2331, 2011.

[35] H. R. Park, S. K. Min, H. D. Cho, D. H. Kim, H. S. Shin, and Y. E. Park, "Expression of Osteoprotegerin and RANK Ligand in Breast Cancer Bone Metastasis," *Journal of Korean Medical Science*, vol. 18, no. 4, pp. 541–546, 2003.

[36] D. H. Jones, T. Nakashima, O. H. Sanchez et al., "Regulation of cancer cell migration and bone metastasis by RANKL," *Nature*, vol. 440, no. 7084, pp. 692–696, 2006.

[37] A. M. Sieuwerts, M. P. Look, M. E. Meijer-Van Gelder et al., "Which cyclin E prevails as prognostic marker for breast cancer? Results from a retrospective study involving 635 lymph node-negative breast cancer patients," *Clinical Cancer Research*, vol. 12, no. 11, pp. 3319–3328, 2006.

[38] C. Sotiriou, M. Paesmans, A. Harris et al., "Cyclin E1 (CCNE1) and E2 (CCNE2) as prognostic and predictive markers for endocrine therapy (ET) in early breast cancer," *Journal of Clinical Oncology*, vol. 22, no. 14S, 2004.

[39] R. Agarwal, A. M. Gonzalez-Angulo, S. Myhre et al., "Integrative analysis of cyclin protein levels identifies cyclin B1 as a classifier and predictor of outcomes in breast cancer," *Clinical Cancer Research*, vol. 15, no. 11, pp. 3654–3662, 2009.

[40] C. B. Moelans, R. A. De Weger, H. N. Monsuur, R. Vijzelaar, and P. J. Van Diest, "Molecular profiling of invasive breast cancer by multiplex ligation-dependent probe amplification-based copy number analysis of tumor suppressor and oncogenes," *Modern Pathology*, vol. 23, no. 7, pp. 1029–1039, 2010.

[41] K. T. Hwang, W. Han, J. Cho et al., "Genomic copy number alterations as predictive markers of systemic recurrence in breast cancer," *International Journal of Cancer*, vol. 123, no. 8, pp. 1807–1815, 2008.

[42] R. Nahta, D. Yu, M. C. Hung, G. N. Hortobagyi, and F. J. Esteva, "Mechanisms of disease: Understanding resistance to HER2-targeted therapy in human breast cancer," *Nature Clinical Practice Oncology*, vol. 3, no. 5, pp. 269–280, 2006.

[43] F. Toledo and G. M. Wahl, "MDM2 and MDM4: p53 regulators as targets in anticancer therapy," *International Journal of Biochemistry and Cell Biology*, vol. 39, no. 7-8, pp. 1476–1482, 2007.

[44] P. van der Lelij, K. H. Chrzanowska, B. C. Godthelp et al., "Warsaw breakage syndrome, a cohesinopathy associated with mutations in the XPD helicase family member DDX11/ChlR1," *American Journal of Human Genetics*, vol. 86, no. 2, pp. 262–266, 2010.

[45] R. Venkatachalam, E. T. P. Verwiel, E. J. Kamping et al., "Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients," *International Journal of Cancer*, vol. 129, no. 7, pp. 1635–1642, 2011.

[46] C. Walowsky, D. J. Fitzhugh, I. B. Castaño, J. Y. Ju, N. A. Levin, and M. F. Christman, "The topoisomerase-related function gene TRF4 affects cellular sensitivity to the antitumor agent camptothecin," *Journal of Biological Chemistry*, vol. 274, no. 11, pp. 7302–7308, 1999.

[47] A. M. Brown, "Wnt signaling in breast cancer: have we come full circle?" *Breast Cancer Research*, vol. 3, no. 6, pp. 351–355, 2001.

[48] F. Forozan, E. H. Mahlamäki, O. Monni et al., "Comparative genomic hybridization analysis of 38 breast cancer cell lines: a basis for interpreting complementary DNA microarray data,"

Cancer Research, vol. 60, no. 16, pp. 4519–4525, 2000.

[49] L. E. Janocko, K. A. Brown, C. A. Smith et al., "Distinctive patterns of Her-2/Neu, c-myc, and cyclin D1 gene amplification by fluorescence in situ hybridization in primary human breast cancers," *Communications in Clinical Cytometry*, vol. 46, no. 3, pp. 136–149, 2001.

[50] K. Al-Kuraya, P. Schraml, J. Torhorst et al., "Prognostic relevance of gene amplifications and coamplifications in breast cancer," *Cancer Research*, vol. 64, no. 23, pp. 8534–8540, 2004.

[51] E. A. Mittendorf, Y. Liu, S. L. Tucker et al., "A novel interaction between HER2/neu and cyclin e in breast cancer," *Oncogene*, vol. 29, no. 27, pp. 3896–3907, 2010.

[52] M. Scaltriti, P. J. Eichhorn, J. Cortés et al., "Cyclin E amplification/overexpression is a mechanism of trastuzumab resistance in HER2$^+$ breast cancer patients," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 9, pp. 3761–3766, 2011.

[53] D. Wangsa, K. Heselmeyer-Haddad, P. Ried et al., "Fluorescence in situ hybridization markers for prediction of cervical lymph node metastases," *American Journal of Pathology*, vol. 175, no. 6, pp. 2637–2645, 2009.