Published in final edited form as:

Stat Med. 2010 July 10; 29(15): 1608–1621. doi:10.1002/sim.3866.

A Metastasis or a Second Independent Cancer? Evaluating the Clonal Origin of Tumors Using Array Copy Number Data

Irina Ostrovnaya¹, Adam B. Olshen¹, Venkatraman E. Seshan², Irene Orlow¹, Donna G. Albertson³, and Colin B. Begg¹

¹Memorial Sloan-Kettering Cancer Center, New York, NY 10021

²Columbia University, New York, NY 10032

³University of California, San Francisco, CA 94143

SUMMARY

When a cancer patient develops a new tumor it is necessary to determine if it is a recurrence (metastasis) of the original cancer, or an entirely new occurrence of the disease. This is accomplished by assessing the histo-pathology of the lesions. However, there are many clinical scenarios in which this pathological diagnosis is difficult. Since each tumor is characterized by a distinct pattern of somatic mutations, a more definitive diagnosis is possible in principle in these difficult clinical scenarios by comparing the two patterns. In this article we develop and evaluate a statistical strategy for this comparison when the data are derived from array copy number data, designed to identify all of the somatic allelic gains and losses across the genome. First a segmentation algorithm is used to estimate the regions of allelic gain and loss. The correlation in these patterns between the two tumors is assessed, and this is complemented with more precise quantitative comparisons of each plausibly clonal mutation within individual chromosome arms. The results are combined to determine a likelihood ratio to distinguish clonal tumor pairs (metastases) from independent second primaries. Our data analyses show that in many cases a strong clonal signal emerges. Sensitivity analyses show that most of the diagnoses are robust when the data are of high quality.

Keywords

Clonality; Diagnosis; Likelihood ratio; Second primary cancer; Metastasis

1. INTRODUCTION

The defining feature of cancer is metastasis, the ability of tumors to colonize distant sites of the body. Independent (second primary) cancers also occur frequently. Distinguishing a second primary from a metastasis or a local recurrence is often of great clinical relevance, as it can affect the appropriateness of local (surgical) versus systemic (medical) treatment. Historically pathologists have distinguished these on the basis of gross and microscopic pathologic criteria. In recent years cancer investigators have begun to explore new methods to accomplish this by comparing the molecular profiles of the two tumors [1–3]. These studies involve the side-by-side comparison of pairs of tumors (from the same patient) on the basis of patterns of somatic mutations. In this article we explore how to construct a formal statistical comparison of these mutational patterns in the setting in which the two

tumors have been evaluated using genome-wide array techniques to identify allelic gains and losses across the entire genome of a tumor. Note that we use the term somatic mutation generically to represent any somatic change, which could include point mutations, small insertions or deletions, or other genetic alterations. However, the methods discussed in this article pertain only to the longer allelic gains or losses that are detectable by current copy number micro-arrays.

In making this differential diagnosis our purpose is to determine whether or not the tumors share a clonal origin, i.e. one of the tumors is a metastasis or local recurrence. That is, one wishes to assess the evidence that both tumors are derived from a single “clonal” cell that experienced the pivotal mutations that led to tumor development. In practice, since additional independent mutations are likely to occur subsequently in the two tumors as they develop separately, two clonal tumors will consist of a mixture of identical (clonal) mutations and independent mutations, while tumors that arise independently will consist solely of independently occurring mutations. Our strategy involves comparing the mutational patterns in the two tumors to see whether they are sufficiently similar to conclude that the tumors are clonal.

In earlier work our group has examined this problem in the context where the mutational data consist of the presence or absence of loss of heterozygosity (LOH) at a set of candidate markers. We developed a simple statistical test for this setting, in which the number of concordant events on the two tumors is evaluated against a reference distribution in an analogous fashion to Fisher’s Exact Test. Although based on simplifying assumptions, we demonstrated that this test has good statistical properties when the marginal mutation probabilities at the candidate loci do not have exceptionally wide variation [4]. In subsequent work, we developed a likelihood based approach to formally account for variations in these marginal probabilities [5]. In this article we use the same likelihood structure as a building block for our approach to the formulation for examining array copy number data. By scanning the entire genome for copy number changes array technology has the potential to provide a comprehensive comparison of the two mutational profiles, and to provide insights beyond those available from studies using a pre-defined set of candidate markers. In particular, array copy number data can pinpoint the places in the genome where these allelic gains and losses begin and end, offering the potential for identifying the exact matches that are the hallmark of clonal allelic changes [6].

Other investigative groups have used array copy number data to examine clonal relatedness of tumor pairs. Several empirical strategies have been adopted. In some, allelic changes have been identified using either the chromosome arm [7] or the chromosome band [8] as the unit of analysis for statistical tests or clustering algorithms. Many investigators have evaluated the similarity between profiles visually or by merely listing the chromosome arms or bands that have similar and different alterations [9–17]. In contrast, other investigators have used the marker values on the array directly, either to determine if the paired tumors cluster together using hierarchical clustering [18], a technique used also by Ghazani et al. [19], Teixeira et al. [8] and Agelopoulos et al. [20], or to create a diagnostic measure, or similarity score, to characterize quantitatively the similarity of the two tumors using similarities of tumors paired from different patients as a benchmark. A similarity measure of this nature proposed by Waldman et al. [18] has also been used by Hwang et al. [21], Nyante, DeVries, Chen and Hwang [22] and Torres et al. [23]. Building on this idea, Bollet et al. [24] used a modified version of the similarity score proposed by Waldman et al. [18] to reflect the relative frequency of *exact* matches of estimated end points of detected allelic changes.

The method we propose in this article is similar in spirit to Bollet et al., in that we believe that examination of potential matching gains and losses provides the strongest evidence for

establishing clonality. That is, we believe that methods that use the chromosome arm or band as the unit of analysis are inefficient, and that selection of the similarity measure is crucial. However, our method differs from Bollet et al. in several important ways. Our method focuses on the closeness of potentially matching allelic changes on an individual basis, recognizing that the segmentation algorithm that determines the exact positions of the beginning and end of any given allelic change will be imprecise due to the noise level in the array. We also seek a normative approach that allows the diagnosis to be made for an individual patient without reference to a data set of patients with the same disease. That is, while we use the available data set to create a reference distribution for benchmarking the diagnosis of each patient in order to study and compare the diagnoses using different methods, our approach does not require such data for diagnosing a new patient in practice.

2. METHODS

Before defining notation and constructing our model, we outline the nature of the data using an illustrative example. In this example, involving two squamous cell tumors from a patient with cancer of the mouth, the evidence favoring the clonal origin of the pair of tumors is quite strong. The two tumors have been analyzed using a bacterial artificial chromosome (BAC) array [6,25], and the results are displayed in Figure 1. Each dot on the graphs is a marker value that represents the allelic copy number at a specific genetic locus (there are approximately 2400 such markers on this BAC array). The markers are displayed sequentially across the 22 chromosomes, with the two tumors aligned vertically. Chromosomes X and Y are excluded. The horizontal black lines at log ratio 0 represent the normal copy number (i.e. the expected 2 copies). If the markers in a region are significantly higher than the black line then we conclude that there has been an allelic gain, and these are represented by red lines. Allelic losses (below the line) are represented by blue lines. The locations of gains and losses are determined by a statistical “segmentation” algorithm that divides the genome into regions of estimated equal copy number, each of which can be classified as “gain”, “loss” or “no change”. We have used the circular binary segmentation (CBS) algorithm [26], a method that has been shown to have good statistical properties [27,28]. In the figure we have used a one-step CBS algorithm that picks the most prominent allelic change within a chromosome arm but does not search for more complex patterns of gains and/or losses (see later discussion). The algorithm examines every possible segment in a chromosome arm and selects the one that maximizes the mean difference between markers within the segment versus those markers outside it. A statistical test is then used to determine whether this difference is significant. In the absence of a significant result the chromosome arm is not segmented. We used a significance level of 0.01, and further considered a significant segment to be a true allelic change only if the mean marker value in the segmented band exceeded a distance of 1.25 median absolute deviations (1.25 MAD criterion) from the normal copy number benchmark. This criterion was also used to identify whole arm gains or losses.

In considering the plots in Figure 1 our goal is to determine whether or not the two tumors are of clonal origin. Clonal tumors are characterized by the identical somatic mutations that occur in the originating clonal cell. In Figure 1 we see a loss of the entire chromosome arm on 3p on both tumors. Other concordant whole arm changes are observed for 8q (gain), 16q (gain) and 20p (gain). In general, the losses and gains appear to be positively correlated, suggesting clonality. However, the real strength of the evidence favoring the clonal origin of these tumors lies in the precision of the matching of allelic changes that occur within chromosome arms. For example there is a common loss on 10q, and a magnified display of the results for this chromosome arm is provided in Figure 2. Here we see strong evidence of a region of loss in the middle of the arm that looks similar in both tumors. If this allelic loss is indeed “clonal”, then the true change must begin and end at exactly the same genetic

locations. Thus our goal is to judge whether relatively close matches of this nature could in fact represent exact matches for which the estimated endpoints differ due to statistical error in the estimation procedure. For 10q the regions of loss are closely but not exactly matched. Nonetheless, this does appear, visually, to be a plausible clonal event. Our challenge in this article is to assess the strength of evidence for and against the hypothesis that this event is indeed clonal. We then need to aggregate this evidence with the evidence from all of the other chromosome arms in order to obtain a diagnosis for the two tumors.

2.1 Model Formulation

Our goal is conceptually simple. In examining two tumors from a single patient we seek to distinguish two hypotheses: the hypothesis, denoted H_I , that the tumors arose independently, and the hypothesis H_M that the tumors are clonally related, i.e. one is a metastasis of the other. The key data are the allelic changes that have been identified by the segmentation algorithm. Among clonal tumors, we use c to represent the proportion of all identified allelic changes expected to have occurred in the originating, clonal cell. The parameter c is necessarily 0 for independent tumor pairs, and so under H_I $c=0$, while under H_M $c>0$. Our likelihood has two components. The first is a multinomial component that characterizes the correlation of observed allelic changes on different chromosome arms, i.e. using the chromosome arm as the unit of analysis. In the second component of the likelihood, we focus solely on arms in which concordant allelic changes were observed, i.e. a gain on both tumors or a loss on both tumors. In this component of the likelihood the closeness of the observed allelic changes are evaluated with respect to sampling distributions generated by permuting residuals of marker values under H_I and H_M separately. Details of how this is accomplished are described in Section 2.2 and in Appendix 2. Both portions of the likelihood are influenced by the parameter c .

The initial step is a segmentation analysis of each of the chromosome arms of the two tumors (see Figure 1). This analysis allows us to assign the arm as representing an allelic gain (represented by the horizontal red lines in the figure), a loss (represented by the horizontal blue lines), or no change. If the change is entirely contained within the chromosome arm it is characterized as a gain or loss by the middle segment. If there is only one change point the allelic change is characterized by the segment mean farthest from 0. Also, a “gain” (or loss) can represent either a whole arm gain or a partial arm gain.

Using the chromosome arm as the unit of analysis, we summarize the patterns of gains and losses on the two tumors using the following notation. Using the suffix 1 to represent a gain, 2 for a loss, and 3 for no change, we thus set $r_{11i} = 1$ if gains are observed on the i^{th} chromosome arm on both tumors (0 otherwise), $r_{22i} = 1$ if losses are observed on both tumors, $r_{12i} = 1$ if there is a gain on one tumor and a loss on the other, $r_{13i} = 1$ if there is a gain on one tumor and no change on the other, $r_{23i} = 1$ if there is a loss on one tumor and no change on the other, and $r_{33i} = 1$ if there is no change on either tumor. Further let $r_i = [r_{11i}, r_{22i}, r_{12i}, r_{13i}, r_{23i}, r_{33i}]$ summarize the combination of events on the two tumors for the i^{th} chromosome arm. The distribution of $\{r_i\}$ is multinomial, as described later in this section.

For the second component of the likelihood we restrict attention to those chromosome arms exhibiting either concordant gains ($r_{11i} = 1$) or concordant losses ($r_{22i} = 1$), i.e. those changes that are potentially clonal, excluding arms with whole arm gains or losses. This information is structurally independent of the frequencies that comprise the multinomial portion of the likelihood. For each of these arms we create a closeness statistic, t_i for the i^{th} arm, representing the closeness of the segmented bands on the two tumors. The closeness statistic captures the degree to which the evidence favors the hypothesis that these represent an identical (clonal) mutation. The definition of this closeness statistic and its distributions when the tumors are clonal versus independent are described in detail in Section 2.2. We use

Ψ_1 and Ψ_2 to represent the sets of arms containing concordant partial arm gains and losses, respectively, i.e. the arms for which the closeness statistics are calculated.

To construct the likelihood we need to know the marginal probabilities of allelic gains and losses for each chromosome arm. For the i^{th} chromosome arm of each of the tumors, let p_{1i} be the probability of a gain, p_{2i} for a loss, and p_{3i} for no change, with $p_{1i} + p_{2i} + p_{3i} = 1$. In our analyses we have calculated the empirical relative frequencies of gains and losses within the chromosome arms in each dataset using the cohort of pairs of tumors being analyzed, and have derived $\{p_{1i}, p_{2i}, p_{3i}\}$ accordingly. Further details are provided in Appendix 1. We assume throughout that these marginal probabilities are the same for both tumors.

It follows that the likelihood takes the form

$$L = \prod_i \left[cp_{1i} + \frac{(1-c)^2 p_{1i}^2}{1 - cp_{1i} - cp_{2i}} \right]^{r_{11i}} \left[cp_{2i} + \frac{(1-c)^2 p_{2i}^2}{1 - cp_{1i} - cp_{2i}} \right]^{r_{22i}} \left[\frac{2(1-c)^2 p_{1i} p_{2i}}{1 - cp_{1i} - cp_{2i}} \right]^{r_{12i}} \left[\frac{2(1-c) p_{1i} p_{3i}}{1 - cp_{1i} - cp_{2i}} \right]^{r_{13i}} \left[\frac{2(1-c) p_{2i} p_{3i}}{1 - cp_{1i} - cp_{2i}} \right]^{r_{23i}} \left[\frac{p_{3i}^2}{1 - cp_{1i} - cp_{2i}} \right]^{r_{33i}}$$

where $f_{Mi}(t_i)$ and $f_{Ii}(t_i)$ are the probability density functions of the closeness statistic t_i when the allelic changes are clonal and independent, respectively. These are derived in detail in Section 2.2.

The various terms are derived from the fact that observed concordant gains or losses could represent clonal events or subsequent independent events that are concordant by chance. For example, the probability of observing a concordant gain on the i^{th} arm is the probability of a clonal gain, i.e. a gain in the originating “clonal” cell, added to the conditional probability of an independently occurring (concordant) gain in the dominant clones of both the resulting primary tumor and the metastasis, given that there was no initial clonal gain or loss. That is

$$\Pr(\text{concordant gain}) = cp_{1i} + (1 - cp_{1i} - cp_{2i}) \left(\frac{(1-c)p_{1i}}{1 - cp_{1i} - cp_{2i}} \right)^2 = cp_{1i} + \frac{(1-c)^2 p_{1i}^2}{1 - cp_{1i} - cp_{2i}}.$$

Likewise the probability of a concordant loss is given by:-

$$\Pr(\text{concordant loss}) = cp_{2i} + (1 - cp_{1i} - cp_{2i}) \left(\frac{(1-c)p_{2i}}{1 - cp_{1i} - cp_{2i}} \right)^2 = cp_{2i} + \frac{(1-c)^2 p_{2i}^2}{1 - cp_{1i} - cp_{2i}}.$$

Among these concordant gains (losses) the conditional probability that an individual gain (loss) is a clonal gain (loss) is given by b_{1i} (b_{2i}), where

$$b_{1i} = \frac{cp_{1i}}{cp_{1i} + \frac{(1-c)^2 p_{1i}^2}{1 - cp_{1i} - cp_{2i}}}.$$

and

$$b_{2i} = \frac{cp_{2i}}{cp_{2i} + \frac{(1-c)^2 p_{2i}^2}{1 - cp_{1i} - cp_{2i}}}.$$

By specifying we obtain the likelihood ratio for the clonal versus independence diagnoses using $L(c)/L(c=0)$. In the analyses presented in Section 3 we use $c = 0.5$ to represent the

clonal hypothesis. We note that we could estimate c for any given dataset by maximum likelihood, but this would preclude the calculation of a likelihood ratio, due to the composite nature of H_M . That is, estimated likelihood ratios would always either be 1 or would favor the clonal hypothesis. In order to frame the problem in a diagnostic context it is necessary to specify c rather than estimate it. We present some reassuring sensitivity analyses that show that the results are not sensitive to the choice of c in the range 0.25–0.75 for the data sets analyzed.

Ultimately the differential diagnosis for the patient under investigation depends on the prior probabilities of these two diagnoses, reflecting the long-run relative frequencies with which pairs of tumors in the given clinical setting are clonal or independent, augmented if necessary with other relevant information extraneous to the mutational profiles. If the prior probability that the tumors are clonal is defined to be π , and the corresponding posterior probability is Π , then the posterior odds is given by

$$\frac{\Pi}{1 - \Pi} = \frac{\pi}{1 - \pi} \cdot \frac{L(c)}{L(c=0)}$$

In the absence of meaningful prior information in our present state of knowledge, we focus on likelihood ratios throughout, effectively assuming that $\pi = 0.5$.

2.2 Comparisons of Specific Concordant Mutations

The latter terms in the likelihood represent comparisons of potentially clonal changes, i.e. comparisons of chromosome arms where concordant allelic changes have been identified. These individual comparisons are potentially very informative, since an exact or very close match of both endpoints of the allelic change can provide strong evidence for clonal relatedness. For this reason we exclude from these specific comparisons whole arm gains or losses, since these events are relatively frequent, and so exact matching of the endpoints at the chromosome boundaries is not informative in the same way that it is for within-arm changes. Matching whole arm gains or losses do, however, contribute to the first part of the likelihood.

The precise comparison of within-arm changes is accomplished by specifying a closeness statistic, t_i and evaluating its distribution separately under H_I and H_M . This involves analysis of the individual markers on the array. In the following, for simplicity, we drop the suffix “i” that identifies the chromosome arm. Let x_{uk} represent the measurement of the u^{th} marker of the k^{th} tumor on a specific chromosome arm that has concordant allelic changes on the two tumors, where $u = 1, \dots, n$ and $k = 1, 2$, and where n represents the number of markers on the chromosome arm. Let the copy number change begin at marker i_k and end at marker j_k for the k^{th} tumor. That is, markers i_k through j_k , inclusive, represent the markers of allelic gain (or loss). If the mutation under investigation is clonal then $i_1 = i_2$ and $j_1 = j_2$. The CBS algorithm is used to obtain estimates of the endpoints, denoted $\hat{i}_k$ and $\hat{j}_k$. Our “closeness” statistic reflects the similarity of the length and positioning of the two changes:

$$t = |\hat{i}_1 - \hat{i}_2| + |\hat{j}_1 - \hat{j}_2|. \quad (2)$$

Small values of t are indicative of a possible clonal mutation. We note that using this statistic, longer chromosome arms with more markers have more power to detect clonal signals.

First consider the distribution of t when the allelic changes have arisen independently. To generate this reference distribution, denoted $f_I(t)$ we must recognize that while chromosomal breakpoints may occur randomly in cells, the alteration is more likely to be retained if it contains a gene (a hotspot) for which there is an advantage to having an abnormal number of copies, such as an oncogene (for a gain) or a tumor suppressor gene (for a loss) [29]. We first generate randomly a location for this hypothetical mutational hotspot where the observed regions of allelic loss or gain on the two tumors overlap. We then randomly generate new (true) regions of allelic change for the two tumors, restricted to the set of changes that overlap the hotspot. We add permuted residuals (differences between markers values and their corresponding segment means) to these generated allelic changes, and then estimate the start and stop points for the allelic changes. If concordant allelic changes are detected by CBS on both tumors then the estimated endpoints are used to calculate a value of t from the reference distribution. This process is then repeated a large number of times to determine $f_I(t)$. A similar algorithm is used to generate the reference distribution under the assumption that the change is clonal (identical), denoted $f_M(t)$ with the difference that a single “true” region of allelic change is assigned to both tumors. Details of these algorithms are set out in Appendix 2. Smoothed estimates of these two reference distributions are then obtained using kernel density estimation, with the default parameters for the R function *density* [30].

3. DATA ANALYSES

We analyze initially the illustrative case that was described in Section 2. The segmentation in Figure 1 reveals 8 allelic gains in tumor 1 and 7 allelic gains in tumor 2, with 4 of these occurring on the same arm (concordant gains). There are 8 losses on tumor 1 and 11 losses on tumor 2, and 7 of these are concordant losses. Of the 11 chromosome arms with concordant changes, several involve a whole arm gain or loss in at least one of the tumors. Thus, there are 6 arms remaining for which we can conduct the detailed comparison of the endpoints using the methods described in Section 2.2. One of these comparisons (10q) is plotted on Figure 2. The odds for this loss favor the clonal hypothesis by a factor of 3 to 1. Of the 5 remaining comparisons three favor the clonal hypothesis: 8q, 79 to 1; 11q, 120 to 1; 18p, 34 to 1. The remaining two comparisons appear to represent independent mutations: 5q, 6 to 1 in favor of independence; 13q, 5 to 1 in favor of independence. Combining all the data the likelihood ratio is 5.5×10^6 to 1, overwhelmingly favoring the common clonal origin of these two tumors. We have analyzed data in this way from 3 complete studies. For two of the studies this involves re-analysis of data that are publically available. The third study was conducted by one of us (DGA).

3.1 Analyses of Ipsilateral Breast Tumors

In the first study [24] the investigators have examined pairs of breast cancers that occurred separately within the same (ipsilateral) breast in 22 patients. Some of these tumor pairs are suspected to be independently occurring breast cancers on the basis of clinico-pathologic information, while in other cases the second tumors are suspected to be recurrences. Clinical diagnoses were determined based on the congruence of the histology and location of the tumors. Second tumors were classified as recurrences (i.e. clonal, C) if they had the same histologic subtype, a similar or increased growth rate, a similar or loss of dependence on either estradiol or progesterone, and a similar or increased differentiation compared with the initial primary [24]. On this basis, 9 of the 22 patients were classified as independent primaries (I), and the remaining 13 were classified as clonal (C) (see Table 1). The ACGH data were obtained from the xba chip of the Affymetrix 50K mapping array and are available through ACTuDB [31]. In order to magnify the signal, remove short germ-line copy number variations, and diminish array artifacts that lead to hypersegmentation, we are

using these data averaged over 15 adjacent markers in our analysis. This leads to a total number of markers of a similar order of magnitude to the illustrative case in Figure 1.

The log likelihood ratios for the 22 patients are plotted in the red histogram in Figure 3. In the figure we also show a “benchmark” histogram of the likelihood ratio distribution for independent tumors. This follows a strategy also used by Bollet et al. and others whereby we compare tumors from different patients, since these tumor pairs are of independent origin by definition. Our two sets of 22 tumors provide $22 \times 21 = 462$ such independent pairings (where each pair contains one 1st primary and one 2nd primary) and the likelihood ratios for these pairings are displayed in Figure 3 in the black histogram. Thirteen of the likelihood ratios for the actual tumor pairs (in red) are very large, providing strong evidence for clonal origin of the tumor pairs. There are a further 4 patients with marginally positive likelihood ratios, i.e. within the range of the reference histogram but above the 95th percentile, and the remainder are negative for clonality. It seems reasonable to consider the cases in the upper 95th percentile but below the maximum reference value observed to represent “equivocal” cases, where the evidence favors clonality but is not strongly conclusive.

The individual likelihood ratios and resulting classifications are presented in Table 1 (in the “LR” column). The results are mostly consistent with the clinical diagnoses. They are also broadly consistent with the results obtained using the authors’ “partial identity score” (“PIS”, see Section 4 for details). Note that Bollet et al. classified all patients with PIS scores above the 95th percentile of their reference distribution as clonal, but for consistency with our classifications we list those scores above the 95th percentile but within the range of reference distribution as “equivocal” in the table.

The most interesting case is #22, diagnosed as clonal by both the LR and the PIS, but considered to be clinically independent primaries. This case possesses two individual mutations that point strongly to clonality, 8p (80 to 1) and 11q (36 to 1). These individual mutations are plotted on the top two panels of Figure 4. Interestingly, this case also highlights some of the practical difficulties we face in accounting for the evidence in a fully algorithmic way. Although our method identifies most potentially clonal mutations, it will occasionally miss some possible candidates due to arbitrary features of the selection algorithm. For example, we elected to compare only mutations that are both designated as either gains or losses. In the lower two panels of Figure 4 we see highly plausible clonal mutations that were missed. For 6p, the short segment in the first tumor (top panel) is considered a loss, while for the second tumor the long segment is considered a gain. This is because we make the classification of gain versus loss on the basis of the distance from the normal copy number, itself estimated from the average of all the markers in the array. Yet, this looks clearly like a highly plausible clonal event. A similar pattern emerges in 13q. Thus the evidence for clonality in this patient may be substantially stronger than is represented by the formal analysis.

3.2. Analysis of Data to Diagnose Invasive Lobular Carcinoma

We have also analyzed publically available data from Hwang et al. [21] who studied tumors from women with an invasive lobular carcinoma (ILC) who had previously been diagnosed with lobular carcinoma in situ (LCIS). Here the investigators were interested in the scientific issue of whether LCIS is a precursor lesion for invasive breast cancer. This dataset involves 24 pairs of tumors, and the tumors were analyzed using BAC arrays with approximately 2400 markers. The results are displayed in Figure 5. That is, all possible pairings of LCIS and ILC tumors from different patients were analyzed and the resulting distribution of likelihood ratios is displayed in black. As in the previous example the juxtaposition of the 24 actual within-patient comparisons in red with the reference histogram in black again produces a group of patients with very strong evidence for clonal relatedness, others with

more equivocal, less convincing evidence, and others with results that are strongly consistent with independent origin of the tumors.

3.3. Analysis of Data to Diagnose Head and Neck Cancer

The illustrative case described at the beginning of this section came from a study of 21 head and neck tumors from 9 patients conducted at the University of California, San Francisco by one of us (DGA). Eight of the tumor pairings were considered clinically and pathologically to represent tumor recurrences, the remaining seven comparisons being diagnosed as new primaries. Only two of these pairings produce strongly clonal patterns, as shown in Figure 6, the remaining comparisons being consistent with the reference distribution of independent pairings. An interesting feature of these data is the fact that the reference distribution (in black) extends to log likelihood ratios in excess of 10, values that are nominally highly indicative of clonality. This appears to reflect the relatively poor quality of the arrays for this study, which used tumor specimens from formalin-fixed paraffin-embedded archival material, unlike the previous two studies which used fresh frozen tissue, known to produce much better quality array data.

3.4 Sensitivity to Selection of c

We have evaluated the sensitivity of our analyses to the arbitrary choice of $c = 0.5$, the parameter that represents the relative frequency of clonal mutations in tumor pairs that are genuinely clonal. We repeated all of our analyses with $c = 0.2$ and with $c = 0.8$. For the Bollet et al. [24] dataset all three analyses produce consistent diagnoses for 18 of the 22 patients (82%). Here we define consistency to represent likelihood ratios that are consistently greater than 1 or consistently less than 1. For all but one of the inconsistent cases the likelihood ratio was in the equivocal range (between the 95th percentile and the maximum observed value of the reference histogram) for the analyses with $c = 0.5$ shown in Figure 3. For the Hwang et al. [21] data 19 of the 24 patients (79%) were diagnosed consistently. Three of the 5 inconsistent cases were in the equivocal range for $c=0.5$. For the head and neck cancer dataset only 2 of the 15 comparisons had strong evidence for clonality at $c = 0.5$, and this pattern re-emerged for analyses at $c = 0.2$ and $c = 0.8$. These results suggest that when the analysis provides very strong evidence for either hypothesis we can be confident of the diagnosis despite the arbitrary choice of c . Conversely, log likelihood ratios in the equivocal range must be viewed with caution.

4. SIMULATIONS

We have explored the performance of our method (LR) in comparison with other measures that have been suggested. First, Waldman et al [18] have proposed a “similarity score” (SS) that characterizes the broad correlation of chromosomal gains and losses. In this method the chromosome arms are characterized as exhibiting either an allelic gain or loss or no change. Individual chromosome arms contribute positively to this score if they exhibit concordant changes, with the contributions from chromosome arms for which allelic changes are infrequent being weighted correspondingly higher. Bollet et al. [24], like us, used a strategy that focuses on individual break points within a chromosome arm that are identified by a segmentation algorithm. They propose a “partial identity score” (PIS) that is a weighted count of breakpoints that exhibit an exact match on the two tumors, including those at the ends of the chromosome arms, with weights determined by the observed relative frequencies of the breakpoints in the dataset. They judge the significance of this score for an individual tumor pair with reference to a distribution of independent pairings using pairs of tumors from different patients. Like our proposed method, these methods require an initial segmentation of the data to determine the allelic gains and losses and their locations. Although these authors have used different segmentation strategies in their published work,

in the following we have employed for consistency the same segmentation method, i.e. the CBS algorithm.

To compare the methods we have resimulated the data from each of the three datasets under sampling schemes in which the tumors are either clonal or independent. The resulting distributions of the measures are evaluated using the area under the ROC curve to characterize the degree of separation achieved. Higher AUCs indicate superior ability of the method to distinguish clonal from independent tumor pairs. Details of how the simulated distributions were constructed are provided in Appendix 3. The key results are provided in Table 2, which shows comparative AUCs in relation to the strength of the clonality signal (i.e. the parameter c), and the degree of “noise” in the marker values, characterized by “SD”, a proportionality factor for the mean standardized residual of the markers from the estimated segment means in the original dataset. The results show clearly that the SS of Waldman et al. has inferior properties, presumably due to the fact that this score does not utilize the information in the within-chromosome matching of observed allelic changes. The LR and the PIS have generally similar properties when the data are generated with the same degree of noise as in the original datasets, although the properties of the PIS relative to the LR seem to degrade as the noise level in the markers increases. Again this is plausibly explained by the fact that the PIS relies on exact allelic matches, and these are increasingly unlikely to be detected accurately as the noise level increases.

5. DISCUSSION

Our goal in this work has been to develop a formal statistical procedure to make the differential diagnosis of metastases from second independent primaries on the basis of somatic allelic changes obtained from array copy number data. This is difficult for many reasons. The first, and possibly the most difficult step, is to organize the voluminous data into a conceptual framework that facilitates formal statistical analysis. Because of the richness and complexity of the data, this process is necessarily somewhat ad hoc, following a growing tradition in statistical genomics [32]. Our belief is that the pivotal information for establishing the clonal origin of pairs of tumors lies in the precise comparison of the locations of specific allelic gains and losses that are potentially clonal events. These comparisons are then combined with the gross correlation patterns of losses and gains across all chromosome arms to determine an overall diagnosis for the patient. The data analyses of our various examples using this methodology suggest that the method can provide conclusive diagnoses for individual patients where the DNA is of high quality and the clonality signals are strong. Our simulation study suggests also that the statistical properties of the method compare favorably with simpler classification measures that have been proposed by other investigators.

The complexity of the data means that any proposed method, including our own, necessarily has limitations due to the analytic and modeling trade-offs required. A particularly difficult feature of the problem is the fact that the two alternative diagnoses that we are trying to distinguish are structured very differently. Under the hypothesis that the two tumors are independent the somatic mutational patterns are presumed to have arisen independently. However, we know that different genetic loci experience mutations with different frequencies in cancers, and so the method requires knowledge of these “marginal” mutation probabilities to effectively filter out the induced correlation that will necessarily occur in the mutational profiles of biologically independent tumors. Our knowledge at present of these marginal probabilities is limited, and we chose to estimate them from the relatively small data sets at our disposal. Under the hypothesis that the tumors share a clonal origin, tumors must contain allelic gains or losses that occurred in the original “clonal” cell that led to the cancers. However, even tumors of clonal origin may, and usually do, harbor numerous other

non-clonal mutations. Consequently, we recognize that a change on one tumor that is not observed on the other should be evidence against the clonal hypothesis. We have approached this issue by constructing a likelihood in which the relative frequency of clonal mutations in tumors that are clonal is assumed to be known (c), but in practice we have very limited knowledge of this parameter. The initial term in the likelihood captures the broad correlation of gains and losses across the chromosome arms. In other words, an allelic gain on one tumor that is not replicated on the other tumor will contribute to a negative correlation, and that in turn provides evidence against the clonal hypothesis.

Application of the method to our various examples demonstrates that it has the potential to provide convincing evidence that some tumor pairs are of clonal origin. However, the method has some arbitrary features. First, the method requires an initial segmentation analysis to identify the allelic gains and losses. This is influenced strongly by both the segmentation method used and by the parameters of this analysis, namely the significance level for detecting an allelic change, and the MAD criterion for ensuring that the signal detected is sufficiently strong. Segmentation methodology is an evolving area of research beyond the scope of this article, and higher resolution arrays can lead to increased sensitivity to array artifacts [33]. In practice, some judgment may need to be exercised when using segmentation to find the level of resolution that seems to be the most credible for identifying copy number changes in the data set under investigation. Second, we have restricted the entire testing strategy to the assumption that each chromosome arm possesses at most one allelic gain or loss. In practice, multiple changes may be observed within a single chromosome arm. If these more complex patterns match closely the evidence favoring clonality can be enhanced. Indeed we see such a pattern in Figure 7. This is from chromosome 5q on patient #13 in the Bollet et al. [24] data, a patient with strong overall evidence for clonality. The method could possibly benefit from further refinement to accommodate complex changes of this nature, although we acknowledge that it is not straightforward to generalize our approach to accommodate these complex changes. Our impression is that when the tumors are truly clonal, comparison of the observed single most prominent change on each chromosome arm will usually provide considerable strength of evidence. Our empirical results show that many of the tumor pairs that we analyze are diagnosed as clonal very convincingly.

Other authors have suggested alternative strategies for tackling the problem. We compared our results with two recently proposed methods, both based on the creation of arbitrarily constructed scoring schemes as the diagnostic classifier. Waldman's similarity score is a measure that essentially captures the broad correlation of allelic changes with the chromosome arm as the unit of analysis. This is similar in spirit to our construction of the first portion of the likelihood. Bollet's partial identity score, which counts the number of places where the endpoints of allelic changes match exactly on the two tumors is similar in spirit to our construction of the second portion of the likelihood. Our comparisons of these methods suggest that use of the partial identity score has similar diagnostic properties to our likelihood ratio approach, but that use of Waldman's similarity score lacks power. Our simulations suggest that a method (such as Bollet's) that relies on *exact* matches of allelic changes is unlikely to have good properties when the arrays are "noisy". Our method is also self-contained, in that one can use it to analyze data from a single patient without an available dataset of other patients to provide a benchmark for the score, as is required by the other proposed methods. Conversely, we do need information with which to estimate the marginal probabilities of allelic changes on each chromosome arm. It is important to note that all of the methods are dependent on the selection of segmentation algorithm, and that we need to be able to filter out small germ-line copy number variants. We have accomplished this by local averaging, but one could attempt to identify and exclude these based on prior knowledge, as Bollet et al. have done.

We have focused on the statistical issues, but there are numerous practical aspects of molecular testing that can influence the data and the resulting analyses. To accomplish array copy number testing, tumor cells must be isolated for analysis. The specimen may be substantially contaminated with normal stromal or interstitial cells, and this can radically reduce the detectable signal from any allelic change. The “quality” of the data can also be affected by whether the tumor samples are fresh frozen or obtained from formalin fixed paraffin-embedded archival material. This “quality” is reflected in the clarity of the signals that identify allelic changes. In poor quality data it is both harder to detect the changes, and also the endpoints of the changes are estimated with much greater variability. As indicated above, our analytic strategy depends on several “tuning” parameters. It also depends on further arbitrary choices, such as how to classify changes as gains versus losses, as indicated in our discussion of Figure 4, and on the extent to which we elect to reduce the total number of markers by averaging adjacent markers. We need further research to determine how to select these parameters to optimize the method, recognizing that the choices may be dependent at the outset on the overall degree of noise in the data. We view this entire methodology as a suggested framework for the task of differential diagnosis of metastases and second primaries, and recognize that additional work is needed to refine the methodological details.

Acknowledgments

We are grateful to Kevin Eng for programming work conducted early in the development of this project; Marc Bollet, Philippe Hupe and colleagues for supplying data from their study of ipsilateral breast cancer; and Brian Schmidt and Antoine Snijders for their work on the head and neck dataset.

Sponsors: This research was supported by the National Cancer Institute, awards CA098438, CA125829 and CA124504.

APPENDICES

The following appendices provide technical details about the method, and about how the simulations were generated. R code for conducting the likelihood ratio method is available at <http://www.mskcc.org/mskcc/html/13287.cfm>

Appendix 1

Estimation of Marginal Probabilities

Let p_{1i}^* , p_{2i}^* and p_{3i}^* be the empirical relative frequencies of gains, losses and no change, respectively, observed on chromosome arm i across all patients and tumors in the dataset under investigation. Rather than use these empirical relative frequencies in the likelihood calculations for an individual patient, we have found it to be expedient to rescale these marginal probabilities to reflect the overall frequencies of gains and losses in the patient under investigation. This is because these overall frequencies vary greatly from patient to patient and can create considerable instability in the likelihood ratio calculations in the

absence of rescaling. Let $h(\cdot)$ represent the logit function. Then $\sum_i h(p_{1i}^*)/m$ represents the average logit relative frequency of gains across all patients, while $(2r_{11} + r_{12} + r_{13}) / 2m$ represents the relative frequency of gains for the specific patient under investigation, where

$r_{11} = \sum_i r_{11i}$ etc., and where m is the number of distinct chromosome arms evaluated per tumor. The number of chromosome arms for which data are available depends on the platform and resolution. For BAC arrays there are 39 chromosome arms. Our rescaled marginal probabilities for the patient under investigation are then calculated using

$h(p_{1i})=k_1 h(p_{1i}^*)$, $h(p_{2i})=k_2 h(p_{2i}^*)$ and $p_{3i} = 1 - p_{1i} - p_{2i}$, where the rescaling parameters are $k_1 = \frac{2(2r_{11}+r_{12}+r_{13})}{\sum_i h(p_{1i}^*)}$ and $k_2 = \frac{2(2r_{22}+r_{12}+r_{23})}{\sum_i h(p_{2i}^*)}$. We note that if $p_{1i} + p_{2i} \geq 0.95$, then we arbitrarily set $p_{3i} = 0.05$ and rescale p_{1i} and p_{2i} accordingly.

Appendix 2

Algorithm for Generating Reference Distributions

Step 1. Segment the data using the CBS algorithm to obtain the position of the detected allelic gain or loss.

Step 2. Obtain residuals for the markers $r_{uk} = x_{uk} - \hat{\theta}_k$ for $u < \hat{i}_k$ or $u > \hat{j}_k$ and $r_{uk} = x_{uk} - \mu_k$ for $\hat{i}_k \leq u \leq \hat{j}_k$ where

$$\hat{\mu}_k = \sum_{u=\hat{i}_k}^{\hat{j}_k} x_{uk} / (\hat{j}_k - \hat{i}_k + 1) \text{ and } \hat{\theta}_k = \left[\sum_{u=1}^{\hat{i}_k-1} x_{uk} + \sum_{u=\hat{j}_k+1}^n x_{uk} \right] / (n - \hat{j}_k + \hat{i}_k - 1)$$

are the segmented means for the allelic change and the normal segments, respectively.

Step 3. Generate the location of the mutational hotspot h^* , where h^* is selected uniformly from the markers in the common interval, i.e. the interval between $\max(\hat{i}_1, \hat{i}_2)$ and $\min(\hat{j}_1, \hat{j}_2)$; if the intervals do not overlap separate hotspots are generated for each tumor.

Step 4. Generate the “true” endpoints of the allelic changes in the reference sample: i_1^* and i_2^* sampled from $U(1, h^*)$ and j_1^* and j_2^* sampled from $U(h^*, n)$ where $U(i, j)$ represents uniform sampling between the markers i and j , inclusive.

Step 5. Obtain $\{r_{uk}^*\}$, a permuted set of residuals, permuted separately for each tumor.

Step 6. Create permuted marker values using

$$\begin{aligned} x_{uk}^* &= \hat{\theta}_k + r_{uk}^* \text{ if } u < i_k^* \text{ or } u > j_k^* \text{ and } x_{uk}^* \\ &= \hat{\mu}_k + r_{uk}^* \text{ if } i_k^* \leq u \leq j_k^*, \\ &\text{where } \hat{\theta} \\ &= \frac{(n - \hat{j}_1 + \hat{i}_1 - 1)\hat{\theta}_1 + (n - \hat{j}_2 + \hat{i}_2 - 1)\hat{\theta}_2}{(n - \hat{j}_1 + \hat{i}_1 - 1) + (n - \hat{j}_2 + \hat{i}_2 - 1)} \text{ and } \hat{\mu} \\ &= \frac{(\hat{j}_1 - \hat{i}_1 + 1)\hat{\mu}_1 + (\hat{j}_2 - \hat{i}_2 + 1)\hat{\mu}_2}{(\hat{j}_1 - \hat{i}_1 + 1) + (\hat{j}_2 - \hat{i}_2 + 1)}. \end{aligned}$$

Step 7. Segment the new datasets to obtain the estimated endpoints of the regions of allelic change, denoted $(\hat{i}_1^*, \hat{j}_1^*)$ and $(\hat{i}_2^*, \hat{j}_2^*)$, and include the results only if these segments are both determined to be statistically significant.

Step 8. Calculate the reference value for the test statistic using $t^* = |\hat{i}_1^* - \hat{i}_2^*| + |\hat{j}_1^* - \hat{j}_2^*|$.

Step 9. Repeat steps 3–8 a large number of times to obtain the distribution of t^* .

The reference distribution for t under the hypothesis that the two observed allelic changes are clonal is generated in the same way, merely by changing Step 4. In this case we randomly generate the endpoints of the common allelic change below and above the hotspot, i^* from $U(1, h)$, and j^* from $U(1, h)$ and set $i_1^* = i_2^* = i^*$ and $j_1^* = j_2^* = j^*$. Smoothed estimates of

these two reference distributions (densities), denoted $f_I(t)$ and $f_C(t)$ are then obtained using kernel density estimation, with the default parameters for the R function *density* [26].

Appendix 3

Construction of Simulations

The simulated comparisons of the different discrimination measures were accomplished by re-simulating datasets consisting of 100 clonal and 100 independent tumor pairs, constructed randomly from the three original datasets in the following manner. Note that for the Bollet et al. dataset simulations were created from the original full resolution data, but analyses were conducted as outlined in the article, e.g. for the LR and SS approaches each full resolution dataset generated by the simulation was smoothed by local averaging prior to analysis. However, for the PIS method we retained the full resolution data since this method was seen to have superior accuracy with full resolution data (results not shown).

Specification of Dataset Characteristics

Each tumor in the dataset was segmented using the CBS algorithm as described in Section 2, with the exception that the algorithm was unrestricted, i.e. it was permitted to detect more than one allelic change per chromosome arm. Residuals were calculated for each marker relative to the segmented means, and a residual standard deviation was calculated for each tumor, denoted S_j for the j^{th} tumor. The total number of allelic changes detected is denoted N_j for the j^{th} tumor. The marginal probabilities of allelic changes (p_{1i} , p_{2i} , p_{3i}) were estimated as in Appendix 1.

Generation of Simulated Clonal Tumor Pairs

We first generate the signal function for each tumor in the pair. If the clonality signal is c , we want on average a proportion c of the allelic changes to be clonal changes, and the remaining proportion $1-c$ to be independent changes. For the first tumor in a simulated pair we first selected the total number of allelic changes N_j randomly from the empirical distribution of N_j . We then selected cN_j chromosome arms randomly (with replacement) to represent arms with clonal changes using a randomization that is weighted by the frequencies $p_{1i} + p_{2i}$. For each selected arm we randomly selected a tumor and then selected one of the allelic changes from the available set of changes on that arm. These cN_j designated clonal changes are then assigned to the segmented mean functions of both tumors. For each tumor we then selected the remaining $(1-c)N_j$ independent changes using the same methodology separately for each tumor. This process resulted in defined signal functions for each of the tumors, with overlapping functions for, on average, a proportion c of the allelic changes.

Generation of Random Marker Data and Analyses

For each tumor, we selected S_j^* randomly from the empirical distribution of the observed standard deviations. We generated marker residuals from the normal distribution $N(0, dS_j^*)$, where d represents the “noise” level ($d=1, 1.25$ and 1.5 in our simulations), and added these to the signal function to obtain the dataset of marker values. The resulting data from the tumor pair were analyzed to obtain the LR statistic as described in Section 2.1. Tumors generated using the data from Bollet et al. were smoothed by averaging consecutive markers in groups of 15. These data were then segmented using the one-step CBS algorithm. The resulting data from the tumor pair were analyzed to obtain the LR statistic. The SS of Waldman et al. was calculated using the same chromosome arm classifications as was used for the LR approach. To obtain the PIS of Bollet et al. the CBS algorithm was unrestricted, i.e. multiple significant allelic changes were allowed to be detected on the same

chromosome arm. Also, as noted above, for the Bollet et al. dataset the PIS was calculated using the full resolution dataset generated as above.

For each data set and each configuration of c and d the process was repeated 100 times to simulate 100 clonal tumor pairs, and 100 times with $c=0$ to simulate independent tumor pairs. The discrimination of these two samples was measured by calculating the area under the ROC curve. The standard errors of the resulting AUCs are approximately $\pm 4\%$.

REFERENCES

1. Ha PK, Califano JA. The molecular biology of mucosal field cancerization of the head and neck. *Critical Reviews in Oral Biology and Medicine*. 2002; 14:363–369. [PubMed: 14530304]
2. Hafner C, Knuechel R, Stoeck R, Hartmann A. Clonality of multifocal urothelial carcinomas: 10 years of molecular genetic studies. *International Journal of Cancer*. 2002; 101:1–6.
3. Huang J, Behrens C, Wistuba I, Gazdar AF, Jagirdar J. Molecular analysis of synchronous and metachronous tumors of the lung: impact on management and prognosis. *Annals of Diagnostic Pathology*. 2001; 5:321–329. [PubMed: 11745069]
4. Begg CB, Eng KH, Hummer AJ. Statistical tests for clonality. *Biometrics*. 2007; 63:522–530. [PubMed: 17688504]
5. Ostrovnaya I, Seshan VE, Begg CB. Comparison of properties of tests for assessing tumor clonality. *Biometrics*. 2008; 68:1018–1022. [PubMed: 18266893]
6. Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genetics*. 1998; 20:207–211. [PubMed: 9771718]
7. Jiang JK, Chen YJ, Lin CH, Yu IT, Lin JK. Genetic changes and clonality relationship between primary colorectal cancers and their pulmonary metastases--an analysis by comparative genomic hybridization. *Genes, Chromosomes and Cancer*. 2005; 43:25–36. [PubMed: 15723340]
8. Teixeira MR, Ribeiro FR, Torres L, Pandis N, Andersen JA, Lothe RA, Heim S. Assessment of clonal relationships in ipsilateral and bilateral multiple breast carcinomas by comparative genomic hybridisation and hierarchical clustering analysis. *British Journal of Cancer*. 2004; 91:775–782. [PubMed: 15266323]
9. Nishizaki T, Chew K, Chu L, Isola J, Kallioniemi A, Weidner N, Waldman FM. Genetic alterations in lobular breast cancer by comparative genomic hybridization. *International Journal of Cancer*. 1997; 74:513–517.
10. Weiss MM, Kuipers EJ, Meuwissen SG, van Diest PJ, Meijer GA. Comparative genomic hybridisation as a supportive tool in diagnostic pathology. *Journal of Clinical Pathology*. 2003; 56:522–527. [PubMed: 12835298]
11. Wa CV, DeVries S, Chen YY, Waldman FM, Hwang ES. Clinical application of array-based comparative genomic hybridization to define the relationship between multiple synchronous tumors. *Modern Pathology*. 2005; 18:591–597. [PubMed: 15696129]
12. Knösel T, Schlüns K, Dietel M, Petersen I. Chromosomal alterations in lung metastases of colorectal carcinomas: associations with tissue specific tumor dissemination. *Clinical and Experimental Metastasis*. 2005; 22:533–538. [PubMed: 16475023]
13. Gallegos Ruiz MI, van Cruijsen H, Smit EF, Grunberg K, Meijer GA, Rodriguez JA, Ylstra B, Giaccone G. Genetic heterogeneity in patients with multiple neoplastic lung lesions: a report of three cases. *Journal of Thoracic Oncology*. 2007; 2:12–21. [PubMed: 17410004]
14. Park SC, Hwang UK, Ahn SH, Gong GY, Yoon HS. Genetic changes in bilateral breast cancer by comparative genomic hybridisation. *Clinical and Experimental Medicine*. 2007; 7:1–5. [PubMed: 17380298]
15. Nestler U, Schmidinger A, Schulz C, Huegens-Penzel M, Gamberdinger UA, Koehler A, Kuchelmeister KW. Glioblastoma simultaneously present with meningioma--report of three cases. *Zentralblatt für Neurochirurgie*. 2007; 68:145–150. [PubMed: 17665342]

16. Haller F, Schulten HJ, Armbrust T, Langer C, Gunawan B, Fuzesi L. Multicentric sporadic gastrointestinal stromal tumors (GISTs) of the stomach with distinct clonal origin: differential diagnosis to familial and syndromal GIST variants and peritoneal metastasis. *American Journal of Surgical Pathology*. 2007; 31:933–937. [PubMed: 17527083]
17. Agaimy A, Pelz AF, Corless CL, Wünsch PH, Heinrich MC, Hofstaedter F, Dietmaier W, Blanke CD, Wieacker P, Roessner A, Hartmann A, Schneider-Stock R. Epithelioid gastric stromal tumours of the antrum in young females with the Carney triad: a report of three new cases with mutational analysis and comparative genomic hybridization. *Oncology Reports*. 2007; 18:9–15. [PubMed: 17549339]
18. Waldman FM, DeVries S, Chew KL, Moore DH 2nd, Kerlikowske K, Ljung BM. Chromosomal alterations in ductal carcinomas in situ and their in situ recurrences. *Journal of the National Cancer Institute*. 2000; 92:313–320. [PubMed: 10675380]
19. Ghazani AA, Arneson N, Warren K, Pintilie M, Bayani J, Squire JA, Done SJ. Genomic alterations in sporadic synchronous primary breast cancer using array and metaphase comparative genomic hybridization. *Neoplasia*. 2007; 9:511–520. [PubMed: 17603634]
20. Agelopoulos K, Tidow N, Korsching E, Voss R, Hinrichs B, Brandt B, Boecker W, Buerger H. Molecular cytogenetic investigations of synchronous bilateral breast cancer. *Journal of Clinical Pathology*. 2003; 56:660–665. [PubMed: 12944548]
21. Hwang ES, Nyante SJ, Yi Chen Y, Moore D, DeVries S, Korkola JE, Esserman LJ, Waldman FM. Clonality of lobular carcinoma in situ and synchronous invasive lobular carcinoma. *Cancer*. 2004; 100:2562–2572. [PubMed: 15197797]
22. Nyante SJ, DeVries S, Chen YY, Hwang ES. Array-based comparative genomic hybridization of ductal carcinoma in situ and synchronous invasive lobular cancer. *Human Pathology*. 2004; 35:759–763. [PubMed: 15188144]
23. Torres L, Ribeiro FR, Pandis N, Andersen JA, Heim S, Teixeira MR. Intratumor genomic heterogeneity in breast cancer with clonal divergence between primary carcinomas and lymph node metastases. *Breast Cancer Research and Treatment*. 2007; 102:143–155. [PubMed: 16906480]
24. Bollet MA, Servant N, Neuvial P, Decraene C, Lebigot I, Meyniel JP, De Rycke Y, Savignoni A, Rigai G, Hupe P, Fourquet A, Sigal-Zafrani B, Barillot E, Thiery JP. High-resolution mapping of DNA breakpoints to define true recurrences among ipsilateral breast cancers. *Journal of the National Cancer Institute*. 2008; 100:48–58. [PubMed: 18159071]
25. Snijders AM, Nowak N, Segraves R, Blackwood S, Brown N, Conroy J, Hamilton G, Hindle AK, Huey B, Kimura K, Law S, Myambo K, Palmer J, Ylstra B, Yue JP, Gray JW, Jain AN, Pinkel D, Albertson DG. Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*. 2001; 29:263–264. [PubMed: 11687795]
26. Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*. 2004; 5:557–572. [PubMed: 15475419]
27. Lai WR, Johnson MD, Kucherlapati R, Park PJ. Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*. 2005; 21:3763–3770. [PubMed: 16081473]
28. Willenbrock H, Fridlyand J. A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*. 2005; 21:4084–4091. [PubMed: 16159913]
29. Houldsworth J, Chaganti RSK. Comparative genomic hybridization: an overview. *American Journal of Pathology*. 1994; 145:1253–1260. [PubMed: 7992829]
30. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*. 1991; 53:683–690.
31. Hupe P, La Rosa P, Liva S, Lair S, Servant N, Barillot E. ACTuDB, a new database for the integrated analysis of array-CGH and clinical data for tumors. *Oncogene*. 2007; 26:6641–6652. [PubMed: 17496932]
32. Speed TP. Terence's Stuff: Statistics without probability. *Institute of Mathematical Statistics Bulletin*. 2008; 36:12.

33. van de Wiel M, Brosens R, Eilers PHC, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B. Smoothing waves in array CGH tumor profiles. *Bioinformatics*. 2009; 25:1099–1104. [PubMed: 19276148]

Figure 1.

Whole genome segmentation of tumors from the patient with oral cancer described in Section 2. Individual markers are plotted as log ratios. These are measures that are typically normalized in a standardized way relative to a sample of pooled normal DNA from multiple individuals in a manner recommended by the manufacturer. The numbers on the x-axis indicate chromosomes. The red (blue) lines represent allelic gains (losses) as determined by the segmentation algorithm.

Figure 2.
Detailed view of chromosome 10q segmentation of the patient with oral cancer described in Section 2.

Figure 3.
Likelihood ratios for patients in Bollet et al. data (red with cross hatching) superimposed on reference histogram from independent tumor pairings from different patients (black).

Figure 4.
Suspected clonal mutations from selected patient from Bollet et al. (2008)

Figure 5. Likelihood ratios for patients in Hwang et al. data (red with cross hatching) superimposed on reference histogram from independent tumor pairings from different patients (black).

Figure 6. Likelihood ratios for patients in the Albertson dataset (red with cross hatching) superimposed on reference histogram from independent tumor pairings from different patients (black).

Figure 7.
Example of a closely matching complex change, from 5q on patient #13 in Bollet et al. (2008).

Table 1

Results for Diagnoses of Ipsilateral Breast Cancer¹

Pt #	Diagnostic Measures		Diagnoses ⁵		
	LR ²	PIS ³	Clinical ⁴	LR	PIS ³
1	5.0×10^{-4}	0.02	I	I	I
2	1.9	0.00	I	E	I
3	2.0×10^5	0.47	C	C	C
4	6.6×10^4	0.28	C	C	C
5	3.3×10^{26}	0.56	C	C	C
6	9.4	0.10	C	E	I
10	2.6×10^{-2}	0.06	I	I	I
11	1.5×10^{-4}	0.03	I	I	I
12	5.5	0.12	I	E	E
13	1.9×10^{16}	0.24	C	C	E
14	1.2×10^8	0.31	C	C	C
15	2.7×10^6	0.13	C	C	E
16	4.1	0.32	I	E	C
18	6.3×10^{-4}	0.00	I	I	I
19	1.8×10^7	0.33	C	C	C
20	9.9×10^{-1}	0.14	I	E	E
21	1.3×10^{27}	0.36	C	C	C
22	3.6×10^3	0.39	I	C	C
23	5.5×10^{13}	0.34	C	C	C
24	1.8×10^9	0.31	C	C	C
25	2.3×10^{16}	0.38	C	C	C
26	7.5×10^{13}	0.52	C	C	C

¹Data courtesy of Bollet et al. [21]

²LR – likelihood ratio

³ PIS – Bollet et al.’s Partial Identity Score. Note that in their article Bollet et al. classified all patients above the 95th percentile of the reference distribution as clonal, but for comparability with the likelihood ratio strategy we classify here patients between the 95th percentile and the maximum reference value as equivocal (E).

⁴ Diagnoses based solely on clinical criteria (from Bollet et al. [24])

⁵ I – independent (multiple primaries); C – clonal (recurrence/metastasis); E – equivocal.

Table 2

Diagnostic AUCs from Simulations¹

Noise	Signal (c)	Data from Bollet et al.			Data from Hwang et al.			Data from Albertson		
		LR	PIS	SS	LR	PIS	SS	LR	PIS	SS
1 SD	0.25	0.80	0.80	0.62	0.82	0.80	0.73	0.80	0.77	0.68
	0.50	0.95	0.97	0.82	0.97	0.97	0.92	0.93	0.89	0.87
	0.75	0.99	0.95	0.93	0.99	0.99	0.99	0.98	0.95	0.97
	0.95	0.99	0.97	0.97	0.99	1.00	1.00	1.00	0.98	0.99
1.25 SD	0.25	0.81	0.77	0.63	0.80	0.77	0.70	0.71	0.67	0.60
	0.50	0.94	0.87	0.83	0.96	0.94	0.88	0.91	0.87	0.85
	0.75	0.98	0.92	0.93	0.97	0.97	0.95	0.97	0.94	0.93
	0.95	1.00	0.94	0.97	0.99	0.99	0.97	0.98	0.96	0.96
1.5 SD	0.25	0.79	0.71	0.65	0.73	0.69	0.64	0.69	0.66	0.58
	0.50	0.94	0.83	0.84	0.94	0.87	0.86	0.83	0.76	0.74
	0.75	0.98	0.86	0.93	0.97	0.92	0.93	0.94	0.90	0.86
	0.95	1.00	0.89	0.98	0.99	0.98	0.98	0.95	0.89	0.90

¹The table displays the area under the ROC curve comparing the distribution of the measure under clonal sampling versus independent sampling. An area of 0.5 represents no discriminatory power, while an AUC of 1 represents perfect discrimination.