Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology & Biostatistics Working Paper Series

Year 2008 *Paper* 15

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

Irina Ostrovnaya* Adam Olshen† Venkatraman E. Seshan‡

Irene Orlow** D G. Albertson†† Colin B. Begg‡‡

*Memorial Sloan-Kettering Cancer Center, ostrovni@mskcc.org

†Memorial Sloan-Kettering Cancer Center, Dept. of Epidemiology and Biostatistics, olshena@biostat.ucsf.edu

‡Columbia University, ves2111@columbia.edu

**Memorial Sloan-Kettering Cancer Center, orlowi@mskcc.org

††albertson@cc.ucsf.edu

‡‡Memorial Sloan-Kettering Cancer Center, beggc@mskcc.org

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

Irina Ostrovnaya, Adam Olshen, Venkatraman E. Seshan, Irene Orlow, D G. Albertson, and Colin B. Begg

Abstract

When a cancer patient develops a new tumor it is necessary to determine if this 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, and it is frequently relatively straightforward. However, there are many clinical scenarios in which this pathological diagnosis is difficult. Since each tumor is characterized by a genetic fingerprint of somatic mutations, a more definitive diagnosis is possible in principle in these difficult clinical scenarios by comparing the fingerprints. In this article we develop and evaluate a statistical strategy for this comparison when the data are derived from array comparative genomic hybridization, a technique designed to identify all of the somatic allelic gains and losses across the genome. Our method involves several stages. First a segmentation algorithm is used to estimate the regions of allelic gain and loss. Then the broad correlation in these patterns between the two tumors is assessed, leading to an initial likelihood ratio for the two diagnoses. This is then further refined by comparing in detail each plausibly clonal mutation within individual chromosome arms, and the results are aggregated to determine a final likelihood ratio. The method is employed to diagnose patients from several clinical scenarios, and the results show that in many cases a strong clonal signal emerges, occasionally contradicting the clinical diagnosis. The "quality" of the arrays can be summarized by a parameter that characterizes the clarity with which allelic changes are detected. Sensitivity analyses show that most of the diagnoses are robust when the data are of high quality.

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

Irina Ostrovnaya[1], Adam B. Olshen[1], Venkatraman E. Seshan[2], Irene Orlow[1], Donna G. Albertson[3], Colin B. Begg[1]

1. Memorial Sloan-Kettering Cancer Center, New York
2. Columbia University, New York
3. University of California, San Francisco

Correspondence to: Colin B. Begg, Department of Epidemiology and Biostatistics, 307 East 63rd Street, New York, NY 10021, beggc@mskcc.org

Abstract

When a cancer patient develops a new tumor it is necessary to determine if this 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, and it is frequently relatively straightforward. However, there are many clinical scenarios in which this pathological diagnosis is difficult. Since each tumor is characterized by a genetic fingerprint of somatic mutations, a more definitive diagnosis is possible in principle in these difficult clinical scenarios by comparing the fingerprints. In this article we develop and evaluate a statistical strategy for this comparison when the data are derived from array comparative genomic hybridization, a technique designed to identify all of the somatic allelic gains and losses across the genome. Our method involves several stages. First a segmentation algorithm is used to estimate the regions of allelic gain and loss. Then the broad correlation in these patterns between the two tumors is assessed, leading to an initial likelihood ratio for the two diagnoses. This is then further refined by comparing in detail each plausibly clonal mutation within individual chromosome arms, and the results are aggregated to determine a final likelihood ratio. The method is employed to diagnose patients from several clinical scenarios, and the results show that in many cases a strong clonal signal emerges, occasionally contradicting the clinical diagnosis. The “quality” of the arrays can be summarized by a parameter that characterizes the clarity with which allelic changes are detected. Sensitivity analyses show that most of the diagnoses are robust when the data are of high quality.

KEY WORDS: Statistical diagnosis; Likelihood ratio; Array CGH; Second primary cancer; 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 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. However, in recent years cancer investigators have begun to explore new methods to accomplish this by comparing the molecular profiles of the two tumors. These studies involve the side-by-side comparison of pairs of tumors (from the same patient) on the basis of patterns of somatic mutations, such as allelic gains or losses, micro-satellite instability, or point mutations in genes that frequently experience somatic mutations in tumors. In this article we explore how to construct a formal statistical comparison of the mutational patterns in the setting in which the two tumors have been evaluated using genome-wide array comparative genetic hybridization (ACGH), a molecular genetic technique designed to identify allelic gains and losses across the entire genome of a tumor.

These studies have potentially important clinical implications. For example, a patient treated effectively for a localized primary head and neck cancer may at a later date present with a solitary lung nodule. If the nodule is a localized second primary lung cancer it can be treated effectively by surgery, though lung surgery is risky and very invasive. On the other hand, if the tumor is a metastasis from the head/neck primary, the prognosis of the patient is necessarily poor, as the cancer will almost certainly have also metastasized to other parts of the body (even though these other metastases may not yet be detectable). In this case invasive surgery would impose needless risks and morbidity on a patient who will have relatively little time left to live. Yet if the two tumors have the same cell type the pathologist has essentially no direct evidence on which to base the diagnosis.

In making this differential diagnosis our fundamental purpose is to determine whether or not the tumors share a clonal origin. That is, one wishes to determine if both tumors are derived from a single “clonal” cell that experienced the pivotal mutations that led to tumor development. Many studies exploring the use of molecular profiling in this context have

been conducted in recent years. For example, investigators studying lung cancer have used microsatellite markers to distinguish patterns of microsatellite instability (Huang 2001, Dacic 2005, Geurts 2005, Leong 1998, Shin 2001) and several investigators have also used mutational analysis of the important cancer genes p53 and/or K-ras (Hiroshima 1998, Holst 1998, Lau 1997, Shimizu 2000, Shin 2001, Murase 2003, Matsuzoe 1999, Sozzi 1995, van Rens 2002). Similar studies have been conducted to distinguish contralateral breast cancers from metastases, and in other cancer sites (Imyanitov 2002, Regitnig 2004, Kollias 2000, Janschek 2001, Tse 2003, Schlechter 2004, Stenmark-Askmalm 2001, Chunder 2004). By studying the mutational pattern, one can establish a genetic fingerprint of the tumor. When the mutational profiles of two apparently independent primary tumors from the same patient are compared, it is possible in principle to see whether these genetic fingerprints are sufficiently similar that we can determine with confidence that they share a clonal origin, i.e. the second primary is really a metastasis from the first primary.

The comparison of mutational profiles of tumors to determine clonality is a challenging statistical problem, and a number of authors have proposed techniques for this purpose. In earlier work we examined two new statistical tests, based on the setting in which the mutational events at candidate genetic loci are assessed for correlation, with a view to determining if the correlation exceeds the level that is plausible on the basis of chance (Begg et al. 2007, Ostrovnaya et al. 2008). These tests have been shown to be reasonably powerful provided that information is available from a considerable number of candidate genetic loci that experience mutational events with reasonably high frequency in the cancer under study, and that the “signal” is relatively strong, i.e. the preponderance of the observed somatic mutations occur in the clonal phase of development. Other authors have approached this problem in different ways. For example Sieben et al. (2003) and Brinkmann et al. (2004) both construct likelihood ratios to distinguish the evidence favoring the two hypotheses, though the construction is somewhat different in each case. Another approach was advocated in earlier work by Kuukasjarvi et al. (1997), who proposed a measure of clonal relatedness based on the frequency of occurrence of concordant mutations in the tumors, and this measure has been used by other authors such as Jiang et al. (2005) and Goldstein et al. (2005a,b).

The preceding methods are all based on the setting in which we observe mutations in a pre-specified set of candidate markers in each tumor, and we evaluate the collective concordance of these mutational profiles. However, there are a limited number of genetic loci at which mutations are known to occur frequently in tumors, and these tend to differ between cancer sites. As a result, sometimes very few mutations are observed in a specific patient, even when a relatively comprehensive set of loci have been examined, and so there can be limited statistical power to distinguish the two diagnoses reliably (Orlow et al. 2008). Since the common somatic mutations in tumors are frequently losses or gains of segments of DNA, the issue of clonality can be studied for the entire genome using array technology, specifically array comparative genomic hybridization (ACGH) (Pinkel et al. 1998). By scanning the entire genome for copy number changes this 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, ACGH can pinpoint the places in the genome where these gains and losses begin and end, offering the potential for identifying the exact matches that are the hallmark of clonal mutations.

Statistical methods for comparing ACGH data in this context have typically employed strategies that simply count mutational events, as in the methods described above for studies based on markers at candidate genetic loci. For example, investigators have used data from the arrays to define the presence or absence of, say, loss of heterozygosity (LOH) at the level of the chromosome arm (Jiang et al. 2005) or chromosome band (Teixeira et al. 2004) in order to define the unit of analysis for statistical tests or clustering algorithms. Many investigators have evaluated the similarity between profiles only visually and through listing the chromosomes arms or bands that have similar and different alterations, for example Nishizaki (1997), Weiss (2003), Wa et al. (2005), Knosel (2005), Ruiz (2007) , Park (2007), Nestler (2007), Haller (2007) and Agaimy (2007). More specific approaches have been used by Waldman et al. (2000) who employed three distinct strategies for classifying pairs of tumors as clonal or independent. First, these investigators used hierarchical clustering of the marker values on the array, designating tumors as clonal if they cluster together in a pair. Hierarchical clustering has also been used by Ghazani et al (2007), Teixeira et al. (2004) and Agelopoulos (2003). Another strategy considered by Waldman et al. is to simply report the percentage of chromosome arms with concordant gains or losses. Finally, this group

has used a similarity score that characterizes the broad correlation of gains and losses across chromosome arms. The similarity score is then benchmarked against the distribution of this measure when tumors from different patients are compared. Some of these strategies were further used in Hwang (2004) and Nyante (2004), published by the same group, and Torres (2007).

None of these methods have taken advantage of the distinctive evidence available from ACGH data when compared with studies involving candidate genetic loci, namely the granularity of the information regarding the allelic gains and losses. In principle, this feature of the data provides the ability to pinpoint the start and stop regions of the allelic changes, with a view to determining an exact match between the mutations on the two tumors. An exception is the recent article by Bollet et al. (2008) where a modified version of the similarity score proposed by Waldman et al. (2000) was used to reflect the relative frequency of *exact* matches of estimated end points of detected allelic changes. In our experience the noise level in the arrays is usually too great to identify the exact endpoints of the allelic changes with confidence, and so matching algorithms need to address directly the statistical variation in the estimation of where allelic changes have occurred, and the positioning of the endpoints of the gain or loss. Furthermore, a comprehensive approach for making the diagnosis of the second tumor as clonal versus independent (of the first tumor) needs to take into account the broad correlation of the observed allelic gains and losses on the two tumors, as well as interrogating specific matching gains and losses to determine the probabilities that these matches represent clonal somatic events. In this article we outline a comprehensive statistical diagnostic strategy constructed along these lines, explore its performance on several available datasets, and describe a research agenda that will be needed to validate its statistical properties.

2. Examples

We utilize data from various sources to illustrate the challenges faced. These include two unpublished studies in which we are involved as co-investigators, and two studies from the literature for which the data are publicly available.

We introduce the problem in the context of an example in which the evidence favoring the clonal origin of the pair of tumors is quite strong. This involves two squamous cell tumors from a patient with cancer of the mouth. These were suspected of being related tumors by the pathologist, and indeed the molecular profiles support this diagnosis. The two tumors have been analyzed using a BAC array (Pinkel et al. 1998; Snijders et al. 2001), and the results are displayed in Figure 1. Each dot on the graph is a marker value that represents the allelic copy number at a specific genetic locus (there are approximately 2400 such markers on a 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 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. Many statistical techniques for ACGH segmentation are available. We have used the circular binary segmentation (CBS) algorithm (Olshen et al. 2004), a method that has been shown to have good statistical properties (see Lai et al. 2005; Willenbrook and Fridyland 2005). For Figure 1, and throughout this manuscript, 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). 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 further criterion is intended to eliminate experimental artifacts such as batch effects. Note that the thresholds for gain or loss are different for every array and depend on the noise level.

The plots in Figure 1 show a broad correlation between the patterns of allelic changes. For example there appears to be a loss of the entire chromosome arm on 3p on both tumors. Other concordant whole arm changes are observed for 8q(gain), 16q(gain),19p(gain) and 20p(gain). In general, the losses and gains appear to be fairly strongly correlated. That is there seem to be more concordances than we might expect by chance. 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. However, the noise in the marker values, and the resulting uncertainty surrounding the estimation of the region of loss, can lead to statistical error in the estimated regions of loss. 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. We note that this patient is 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), and we will present an analysis of this patient and a summary of the analyses for all tumor combinations later in Section 5.

Typically, the evidence for or against clonality is much less clear-cut than for this patient. Our second example involves two skin melanomas that have been diagnosed in the same patient. These melanomas were classified as independent primaries by the pathologist, and they occurred 2.4 years apart in distinct anatomic locations, one on an arm and the other on a leg. This time the arrays are from the 244K Agilent platform, an array with far more marker values than the BAC arrays featured in the first example. However, the data for this patient are quite noisy, and so we elected to perform our analyses using new marker values that represent averages of 49 adjacent markers. This averaging was accomplished to reduce the degree of scatter. It also leads to a total number of markers that is of the same order of magnitude as for the BAC arrays. This patient's data are plotted in Figure 3. For these two tumors there are some notable similarities. Indeed the patterns in the higher numbered chromosomes are visually similar, and there is a moderately strong overall correlation across the genome, suggesting that the clinico-pathological diagnosis that the two melanomas are independent may be wrong. However, comparison of concordant within-arm allelic changes reveals only one change that strongly favors a clonal origin (on 2p) while most of the other observed changes appear to represent independent somatic events. This

patient is from a study of clonality in 19 patients with double primary melanomas (Orlow et al. 2008).

We also analyze in detail publically-available datasets from two published studies. In the first study (Bollet et al. 2008) 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 metastases. The ACGH data were obtained from the Affymetrix Genechips Human Mapping 50K Array and are available through ACTuDB (Hupe et al. 2007). In order to magnify the signal and diminish the array artefacts we are using these data averaged over 15 adjacent markers in our analysis. Again this leads to a total number of markers of a similar order of magnitude as the other datasets. In a second study, also involving breast cancer, Hwang et al. (2004) have studied the 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 a total of approximately 2400 markers.

3. Conceptual Model

Our analytic goal is conceptually straightforward. We wish to determine whether the two tumors are biologically independent, or whether the tumors are clonally related, i.e. both originating from the same “clonal” cell in which the acquired pivotal mutations occurred that provoked the cell to proliferate uncontrollably, leading ultimately to cancer. Thus, in our hypothesis of independent origin of the tumors, the sets of somatic mutations on each tumor must have occurred independently of each other. Under the clonal hypothesis the two tumors must possess one or more mutations that are identical. The existence of these clonal mutations ensures that a positive correlation in the mutational profiles would be expected, and so examination of the strength of this correlation is a major aspect of our analysis. However, we note that correlation of the patterns of gains and losses is likely even in independent tumors. This is because allelic gains and losses

tend to be observed in tumors in genetic regions for which there is a selective advantage, such as in the neighborhood of oncogenes or tumor suppressor genes. Thus, even in the absence of clonal origin of the tumors, there will be a common tendency for gains and losses to occur on the same chromosome arms. Our methods adjust for this phenomenon using background data to estimate the probabilities of gains and losses for each chromosome arm for the cancer type under investigation. Also, in clonal tumors, we expect additional "independent" allelic changes in each cell colony to occur, thereby adding "noise" to the clonal signal. Statistical noise in the marker values can also be accentuated for various experimental reasons: the tumor sample may be contaminated with an unknown proportion of normal cells; the tumor itself may have developed considerable heterogeneity of cell clones with distinct somatic changes; there may be artefacts in the array technology; there may be copy number variants in the germ line that masquerade as clonal events.

After examining the broad pattern of correlation across the genome, we examine more carefully the specific chromosome arms on which concordant mutations have been observed, i.e. a loss on both tumors or a gain on both tumors. We examine the exact locations of these allelic gains or losses to determine the plausibility that the two changes are actually clonal, i.e. they represent the same change that occurred in the original clonal cell. We have developed new methodology for accomplishing this comparison. Our overall strategy is based on the premise that these precise within-chromosome comparisons provide the most compelling evidence for identifying clonal tumors.

We approach the problem from a "theoretical" perspective. That is, we construct a sampling model that we conjecture to be a realistic representation of the way in which the marker data are generated under the two competing diagnoses (independent origin of the tumors versus clonal origin). This model is then used to obtain statistical results that characterize the relative strength of the evidence favoring each of these hypotheses. The results are expressed as likelihood ratio statistics. Ultimately, a more satisfying (and better calibrated) strategy may be to generate an optimal discrimination measure, and then characterize the distribution of the measure in training data consisting of tumor pairs "known" to be clonal and pairs known to be independent, all derived from the relevant clinical scenario under investigation, e.g. cancers of a specific

anatomic site and/or cell type. However, at present there are very few data of this nature available, and indeed one cannot be sure that diagnoses based on classical pathology are correct. That is, our problem is akin to the creation and evaluation of a diagnostic test when there is no "gold standard" reference test (Begg 1987). Despite this problem our diagnostic setting is unusual in that we can construct plausible reference distributions for our diagnostic test statistics under the "independence" hypothesis. We can do this by pairing tumors from different patients. By definition, all such pairs of tumors must have occurred independently. We use this strategy to calibrate our results for each dataset.

4. Detailed Analytic Model

The initial step of the analysis is a segmentation analysis of each of the chromosome arms of the two tumors (see Figures 1 and 3). In our analyses we have used the CBS algorithm with the significance level and further constraint as defined in the previous section. This analysis allows us, for each chromosome arm of each tumor, to assign the arm as representing an allelic gain, a loss, or no change. Comparing the patterns from the two tumors, we identify arms in which gains occur in both tumors or losses occur in both tumors. We define the former as "concordant gains", and the set of such arms is represented by Ψ_g. Likewise the set of arms with concordant losses is denoted Ψ_l.

Correlation of Mutational Patterns

The arrays we have been using contain sufficient data for 39 autosomal chromosome arms that are considered to be statistically independent units of the genome. Let $r_{ggi} = 1$ if gains are observed on the ith chromosome arm on both tumors (0 otherwise), $r_{lli} = 1$ if losses are observed on both tumors, $r_{gli} = 1$ if there is a gain on one tumor and a loss on the other, $r_{gni} = 1$ if there is a gain on one tumor and no change on the other, $r_{\ln i} = 1$ if there is a loss on one tumor and no change on the other, and $r_{nni} = 1$ if there is no change on either tumor. In evaluating the correlation in these outcomes between the tumors we must recognize the fact that the probabilities of gains and losses will be

specific to each chromosome arm, in addition to being specific to the tumor type under investigation. For the i[th] chromosome arm let these probabilities be p_{gi} for a gain, p_{li} for a loss, and p_{ni} for no change, with $p_{gi} + p_{li} + p_{ni} = 1$. Our analytic strategy requires knowledge of these marginal probabilities, and there are growing data resources for this purpose. However, in our analyses we have calculated the empirical relative frequencies of gains and losses in each dataset using the cohort of pairs of tumors being analyzed, and have used these as estimates of p^*_{gi}, p^*_{li}, and p^*_{ni}. We then obtained patient-specific estimates of the marginal probabilities using $\log it(p_{gi}) = \log it(p^*_{gi}) \log it[(2r_{gg} + r_{gl} + r_{gn})/78] / \sum \log it(p^*_{gi})/39$, and analogous formulas for p_{li} and p_{ni}, where $r_{gg} = \sum_i r_{ggi}$, etc. We have used these rescaled probabilities to avoid the risk of creating extreme results merely because the overall mutation frequency is unusually low or high for the patient, since this overall frequency is in part determined by the "quality" of the array data (see Sections 5 and 6).

For our problem of differential diagnosis we have chosen to evaluate the evidence distinguishing the two diagnoses (H_I, the independence hypothesis, and H_C, the clonal hypothesis) using likelihood ratios. As our knowledge develops, it should be possible to refine the diagnostic strategy to accommodate the prior probabilities for each diagnosis, based on the long-term relative frequencies of the two diagnoses in the clinical scenario, adjusted also possibly using relevant clinical information, such as the concordance of cell type and other features that inform current pathologic diagnostic rules.

We construct a likelihood as follows:

$$P(r \mid c) = \prod_i \left[cp_{gi} + \frac{(1-c)^2 p_{gi}^2}{1 - cp_{gi} - cp_{li}} \right]^{r_{ggi}} \left[cp_{li} + \frac{(1-c)^2 p_{li}^2}{1 - cp_{gi} - cp_{li}} \right]^{r_{lli}} \left[\frac{2(1-c)^2 p_{gi} p_{li}}{1 - cp_{gi} - cp_{li}} \right]^{r_{gli}} \left[\frac{2(1-c) p_{gi} p_{ni}}{1 - cp_{gi} - cp_{li}} \right]^{r_{gni}} \left[\frac{2(1-c) p_{li} p_{ni}}{1 - cp_{gi} - cp_{li}} \right]^{r_{\ln i}} \left[\frac{p_{ni}^2}{1 - cp_{gi} - cp_{li}} \right]^{r_{nni}} \quad (1)$$

where $r = \{r_{ggi}, r_{lli}, r_{gli}, r_{gni}, r_{\ln i}, r_{nni}\}$ represents the pattern of gains and losses across all of the chromosome arms. The parameter c represents, in clonal pairs of tumors, the

proportion of observed mutations that are expected to be clonal, and we assume that this proportion applies to both gains and losses equally. By specifying a value for c we can obtain the likelihood ratio for the clonal versus the independence diagnoses using $P(r \mid c) / P(r \mid c = 0)$.

Comparisons of Specific Concordant Mutations

We augment the broad evaluation of correlation across the genome with specific comparisons for chromosome arms on which a common overlapping loss or gain spanning only a part of a chromosome has been observed on both tumors. The goal is to assess the evidence for and against the clonal origin *of each specific mutational change.* Let x_{uk} represent the measurement of the u^{th} marker of the k^{th} tumor on a specific chromosome arm which has concordant allelic changes on the two tumors, where $u = 1,..,n,$ 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 $\hat{j}_k$. We define a "closeness" statistic $t,$ representing the similarity of the length and positioning of the two changes:

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

Thus small values of *t* are indicative of a possible clonal mutation. Under H_I we assume that the allelic changes have arisen independently, and so the reference distribution for *t* under H_I should thus reflect the distribution of *t* when independent allelic gains or losses have been generated on each tumor. To generate an appropriate reference distribution 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 or genes for which there is an advantage to having an abnormal number of copies, such as an oncogene or a tumor suppressor gene. To address this phenomenon we first generate a location for a hypothetical mutational hotspot, which we presume to be located 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 permute the data (as described below) and use the CBS algorithm on the permuted data for each tumor to estimate the start and stop points for the allelic changes. If concordant allelic changes are detected by CBS on both tumors then the data set is considered to be "admissible", and the estimated endpoints are used to calculate the reference test statistic. Re-applying the same segmentation algorithm (CBS) to the data simulated in the reference distribution automatically adjusts the procedure for the segmentation error. This process is then repeated a large number of times to establish the reference distribution for t under H_1.

Let the sample means of the segmented marker values be

$\hat{\mu}_k = \sum_{u=\hat{i}_k} x_{uk} / (\hat{j}_k - \hat{i}_k + 1)$ for the mutated portion and

$\hat{\theta}_k = \left[\sum_{u=1}^{\hat{i}_k - 1} x_{uk} + \sum_{u=\hat{j}_k + 1}^{n} x_{uk} \right] / (n - \hat{j}_k + \hat{i}_k - 1)$ for the normal copy number portion. These are used to obtain residuals for each of the marker values:-

$$r_{uk} = x_{uk} - \hat{\theta}_k \text{ for } u < \hat{i}_k \text{ or } u > \hat{j}_k$$

$$= x_{uk} - \hat{\mu}_k \text{ for } \hat{i}_k \le u \le \hat{j}_k .$$

The reference distribution is constructed using the following steps. [An asterisk denotes terms representing the reference distribution.]

(1) Generate the location of the mutational hotspot h^*, where h^* is selected uniformly from 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. [For simplicity we assume that the hotspot occurs at a marker value, and define $U(i, j)$ to represent uniform sampling of the markers between i and j, inclusive.]

(2) 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).

(3) Obtain $\{r_{uk}^*\}$, a permuted set of the residuals $\{r_{uk}\}$, permuted separately for each tumor.

(4) Create the permuted marker values $\{x^*_{uk}\}$ using

$$x^*_{uk} = \hat{\theta} + r^*_{uk} \text{ if } u < i^*_k \text{ or } u > j^*_k$$
$$= \hat{\mu} + r^*_{uk} \text{ if } i^*_k \le u \le j^*_k,$$

where

$$\hat{\theta} = \frac{(n - \hat{j}_1 + \hat{i}_1 - 1)\hat{\theta}_1 + (n - \hat{j}_2 + \hat{i}_2 - 1)\hat{\theta}_2}{(n - \hat{j}_1 + \hat{i}_1 - 1) + (n - \hat{j}_2 + \hat{i}_2 - 1)}$$
$$\hat{\mu} = \frac{(\hat{j}_1 - \hat{i}_1 + 1)\hat{\mu}_1 + (\hat{j}_2 - \hat{i}_2 + 1)\hat{\mu}_2}{(\hat{j}_1 - \hat{i}_1 + 1) + (\hat{j}_2 - \hat{i}_2 + 1)}.$$

(5) Segment the new datasets to obtain the estimated endpoints of the regions of allelic change, denoted $(\hat{i}^*_1, \hat{j}^*_1)$ and $(\hat{i}^*_2, \hat{j}^*_2)$. Include the results only if these changes are both determined to be significant by the CBS segmentation method.

(6) Calculate the reference value for the test statistic using

$$t^* = \left|\hat{i}^*_1 - \hat{i}^*_2\right| + \left|\hat{j}^*_1 - \hat{j}^*_2\right|.$$

(7) Repeat the process a large number of times to obtain the distribution of t^*.

A reference distribution for t under the clonal hypothesis H_C can be generated in exactly the same manner, merely by changing step 2. Here we randomly generated the endpoints of the allelic change below and above the hotspot, i^* from U(1, h), and j^* from U(h, n), and set $i^*_1 = i^*_2 = i^*$ and $j^*_1 = j^*_2 = j.^*$ Also, in step 2, if the intervals do not overlap, a single hotspot is generated between the two intervals.

Smoothed estimates of these two reference distributions (densities), denoted $f_I(t)$ and $f_C(t)$, are then obtained using kernel density estimation, with a standard default R bandwidth selection and kernel (Sheather and Jones 1991). The ratio $f_C(t) / f_I(t)$ is then used as the likelihood ratio to characterize the evidence for and against the hypothesis that the mutation under investigation is clonal.

In an effort to assess the validity of this strategy from a purely statistical perspective we have evaluated its frequentist properties by performing simulations in which the reference distribution of t is evaluated under a model in which the two mutations are

generated independently, and the noise in the marker values is generated by a normal distribution. Specifically, we determined the mean value for markers at normal copy number, denoted by θ, and the mean in the region of allelic change, denoted by μ, with common variance σ^2. These were chosen to specify the signal strength, represented by $|\mu - \theta| / \sigma$, and one of the means was set to 0 and the variance set to 1 without loss of generality. For each simulation we first selected a true mutational hotspot at marker h. This was randomly generated from the n markers for each data set. We then generated a data set as follows. First the "true" endpoints of the allelic changes were randomly generated, i_1 and i_2 as U(1, h), and j_1 and j_2 as U(h, n). Observed marker values were generated as normal random variables. That is, x_{uk} was generated as N(θ, σ^2) for $u < i_k$ or $u > j_k$ and as N(μ, σ^2) for $i_k < u < j_k$ The CBS algorithm was used on these data to estimate the endpoints, denoted $\hat{i}_1, \hat{j}_1, \hat{i}_2, \hat{j}_2$, and the test statistic t was calculated using (2).

Following the procedure outlined above, the tail area probability (p-value) was calculated as the relative number of times that $t^* \leq t$ based on 1000 replicates from the reference distribution. The entire process was then repeated 1000 times to determine the relative frequency matching the tail-area probabilities generated by the algorithm. The simulation standard error is about $\pm$0.02. The procedure was allowed as many attempts as necessary to complete the 1000 replicates required, and likewise it was allowed as many attempts as necessary to generate a significant, concordant data set. In configurations with a signal strength ranging from 0.5 standard deviation units to 3, and numbers of markers from 65 to 140 (the typical numbers of markers on a chromosome arm of the arrays used in some of our examples, after data averaging) the observed relative frequencies from the simulation ranged from 0.02 to 0.06 for tail-area probabilities less than 0.05 as determined by our permutation-based algorithm. This exercise gives us confidence that our permutation-based procedure produces tail-area probabilities that are approximately accurate when data are generated using normal errors in the marker values.

Global Analysis and Patient Diagnosis

The final step in the analysis is the aggregation of the evidence obtained from the correlation of the broad mutational patterns and the similarity analyses of specific concordant mutations. This provides a final assessment of the strength of the evidence favoring H_I versus H_C for the two tumors. We create an augmented likelihood that combines the evidence from these two sources. To do this we need to recognize that even for clonal tumors not all observed mutations are expected to be clonal. However, since our likelihood only involves comparison of potentially clonal concordant mutations, we need a mixing parameter that represents, under the clonal hypothesis, the proportion of "concordant" mutations that are expected to be clonal (as opposed to the proportion of all observed mutations that are clonal, denoted by the parameter c). Setting

$$b_{gi} = \frac{cp_{gi}}{cp_{gi} + \dfrac{(1-c)^2 p_{gi}^2}{1 - cp_{gi} - cp_{li}}}$$

and

$$b_{li} = \frac{cp_{li}}{cp_{li} + \dfrac{(1-c)^2 p_{li}^2}{1 - cp_{gi} - cp_{li}}}$$

the full likelihoods under the two hypotheses can be expressed as follows:

$$L_I = \prod_i P(r \mid c = 0) \prod_{i \varepsilon \Psi_g} f_{Ii}(t_i) \prod_{i \varepsilon \Psi_l} f_{Ii}(t_i)$$

$$L_C = \prod_i P(r \mid c) \prod_{i \varepsilon \Psi_g} (b_{gi} f_{Ci}(t_i) + (1 - b_{gi}) f_{Ii}(t_i)) \prod_{i \varepsilon \Psi_l} (b_{li} f_{Ci}(t_i) + (1 - b_{li}) f_{Ii}(t_i)), \qquad (3)$$

where $f_{Ii}(t_i)$ and $f_{Ci}(t_i)$ represent the reference distributions of the similarity statistic t_i for the comparison on the i^{th} chromosome arm.

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

$$\frac{\Pi}{1-\Pi} = \frac{\pi}{1-\pi} \bullet \frac{L_C}{L_I}.$$

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

5. Data Analyses

We analyze initially the illustrative cases that were described earlier in Section 2. Data from the first of these, involving two squamous cell tumors from a patient with cancer of the mouth are presented in Figures 1 and 2. The segmentation analysis 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. The resulting likelihood ratio statistic using (1) is 214 to 1 in favor of clonality versus independence. In other words, the degree of broad correlation in allelic gains and losses is quite strongly supportive of the clonal hypothesis. 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 of the changes. One of these comparisons (10q) is plotted on Figure 2. The odds for this loss favor the clonal hypothesis by a factor of 3 to1. 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. When these comparisons are augmented with the broad comparisons using (2), the odds for clonality are 5.5 X 10^6 to 1, overwhelmingly favoring the common clonal origin of these two tumors.

Our second example from Section 2 comes from a study of 19 patients with double primary melanomas that were assembled to examine the possible relationship of second primary melanomas with their initial primaries. These samples were examined for LOH at a set of candidate markers, and the results seem to confirm generally that most if not all of the tumors are independent (Orlow et al. 2008). However, for two of the 19 patients, the comparison of the LOH profiles was marginally statistically significant, and for one of these we had sufficient tumor tissue to obtain ACGH on both tumors (note that most primary cutaneous melanomas are too small for CGH analysis using current technology). This case is displayed in Figure 3. The likelihood ratio from the broad

correlation of the gains and losses favors independence with odds of 31 to 1. There are 3 concordant mutations amenable to a comparison of the specific changes with the following results: 2p, 15 to 1 favoring of clonality; 17q, 5 to 1 favoring independence; 22q, 1.3 to 1 favoring independence. Thus the aggregate likelihood ratio is 13 to 1 in favor of independence.

These previous examples lead us to likelihood ratios that ideally represent the strengths of evidence favoring each of the hypotheses/diagnoses for the patient, H_I versus H_C. For the first patient we arrived at odds for H_C of 5.5 x 10^6 to 1. But do these seemingly overwhelming odds really supply the certainty of the diagnosis of H_C that the numbers imply? All of the examples we present involve clinical scenarios where the "correct" diagnosis is uncertain. That is, clinical and pathological data do not provide us with a "gold standard" reference diagnosis, and indeed a goal of research into the use of molecular techniques such as ACGH in this setting is to provide a more accurate standard. However, when we have at our disposal a more complete dataset of patients from the clinical scenario under investigation, we can create a plausible reference distribution for our diagnostic statistics under H_I by comparing pairs of tumors from different patients, tumors which necessarily arose independently. In the following more comprehensive analyses we use this strategy to add further insights into the properties of our method.

First we examine the 22 patients from the study by Bollet et al. (2008). Clinical details are provided in Table 1, along with the diagnostic classifications based on our analyses. The goal for each patient is to determine if a second ipsilateral breast cancer is a new primary or a recurrence of the initial primary cancer. 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 (see Bollet et al. 2008). On this basis, 9 of the 22 patients were classified as independent primaries (I), and the remaining 13 were classified as clonal (C). The final 3 columns of Table 1 show the broad likelihood ratio calculations using (1), the likelihood ratio augmented with results from specific within-chromosome comparisons using (3),

and the diagnoses based on the latter statistics. For this dataset we classified cases as independent if ln(LR2) < -0.5, equivocal if -0.5 < ln(LR2) < 6, and clonal if ln(LR2) > 6, for reasons further described below. Our classifications are mostly in agreement with the clinical classifications, with the notable exception of case #22. For this case, considered clinically to represent two independent tumors, the broad correlation (LR1) modestly favors the clonal hypothesis but there are individual mutations that point strongly to clonality on 8p (80 to 1) and 11q (36 to 1), leading to a final likelihood ratio in favor of clonality of 3.6×10^3 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 only compare 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 clearly looks 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.

This patient represents an example of a case in which the molecular evidence seems to clearly contradict the diagnosis based on standard clinical criteria. However, the data are not always so clear-cut, and it is also much harder to be convinced that two tumors are independent, since independence is characterized (visually) merely by the absence of strikingly clonal features such as the allelic changes highlighted in Figure 4. One way to judge the credibility of our calculated likelihood ratios is to create a benchmark reference distribution for independent comparisons by conducting analyses on all comparisons formed by pairing tumors from different patients, a strategy that has also been used by Bollet et al. and others. Our two sets of 22 tumors provide 22X21=462 such independent pairings (where each pair contains one 1^{st} primary and one 2^{nd} primary) and the likelihood ratios (using (3)) for these pairings are displayed in Figure 5 in the black histogram. Superimposed in red with cross-hatching is a histogram of the results from the 22 actual within-patient comparisons (from Table 1). The results show that a

likelihood ratio of 3.8 to 1 (ln(LR2)=1.3) corresponds to the upper 99th percentile of the likelihood ratio distribution for independent tumors, and so values considerably in excess of this are unlikely by chance. In this and subsequent analyses we define the region from the 95th percentile of the reference distribution to the maximum value recorded as an "equivocal" diagnostic region. It can be seen from Figure 5 that for this dataset the equivocal region spans ln(LR2) values between -0.9 and 5. Consequently the 13 patients with likelihood ratios in excess of this region (including case #22) would appear to be definitively clonal. For the 4 patients with log LR2 values below -0.9 the evidence strongly favors independence. Five patients fall into the "equivocal" zone, with odds favoring clonality of 1.9 to 1 (case #2), 9.4 to 1 (case #6), 5.5 to 1 (case #12), 4.1 to 1 (case #16), and 1 to 1 (case #20). These results give confidence that with good quality data the method has the potential to provide definitive classifications for the majority of patients.

We have applied the same series of analyses to another published example, this time a comparison of LCIS and ILC breast tumors from each of 24 patients (Hwang et al. 2004). The purpose of this study was to examine the general hypothesis that LCIS is a precursor lesion to invasive breast cancer (ILC), and so the authors were interested in the frequency with which clonal relatedness could be identified or proved. The results are characterized in the two histograms (Figure 6), calculated in a similar way to Figure 5 above. 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. This distribution has slightly greater spread than for the Bollet et al. data. In fact, the equivocal region stretches from a log LR2 value of 0.3 to a value of 8. The juxtaposition of the 24 actual within-patient comparisons in red with the reference histogram again produces a group of patients with very strong evidence for clonal relatedness (8 of the 24 patients). The remaining cases are spread through the "equivocal" (5 patients) and "independent" (11 patients) regions.

We performed a similar analysis on our dataset of 21 tumors from 9 patients with multiple head and neck cancers, from which our illustrative patient in Figures 1 and 2 was drawn. Eight of the tumor pairings were considered clinically and pathologically to represent tumor recurrences. Only two of these pairings produce strongly clonal patterns. These are the two extreme observations on the right of Figure 7 in red. One

case, considered clinically to be an independent primary, has odds in favor of clonality of 78 to 1. However, this falls in the equivocal range of the independent reference distribution for this dataset, which spans likelihood ratios that nominally favor the clonal hypothesis by large factors, with ln(LR2) values ranging from 4 to 12.

It is noticeable from Figure 7 that the reference distribution of likelihood ratios from independent pairings is much broader than for the other two datasets in Figures 5 and 6 and includes likelihood ratios whose nominal values strongly favor the clonal hypothesis. This appears to reflect the fact that the datasets differ with respect to the clarity with which allelic changes are detected. Defining the signal strength to be the 90^{th} percentile of the absolute values of the detected segment means divided by the standard deviation of the residuals, reflecting how separated the larger segment means are from the rest of the array values, we find that the mean signal strengths are 4.2 for the Bollet et al. data, 3.6 for the Hwang et al. data, and 2.4 for the head and neck dataset. High signal strength would appear to translate into a tighter reference distribution, and to clearer separation of tumor pairs into clusters representing independent pairs and clonal pairs. Signal strength also appears to affect the normative values of the likelihood ratios, which should generally be less than 1 for independent tumors. The upper 95^{th} percentile of the reference distribution for independent pairings is 0.4 to 1 for the Bollet et al. data, 1.3 to 1 for the Hwang et al. data, but it is 59 to 1 for the head and neck cancer dataset. While the arrays were performed on fresh frozen tissue in the studiy by Bollet et al.,, the head and neck study used formalin fixed paraffin embedded tissue, and it is well known that this source of tissue produces ACGH arrays of much poorer quality.

Finally, we have evaluated the sensitivity of our analyses to the arbitrary choice of $c = 0.5$ as our parameter representing 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. 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 for the analyses with $c = 0.5$ shown in Figure 5. For the Hwang et al. 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 H_C or H_I 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. The results also support the use of a "signal strength" measure of the clarity of the allelic changes observed, as suggested in the previous paragraph, to characterize the quality of the array data and the consequent conclusiveness of the resulting diagnoses.

6. Discussion

Cancer pathology is in a period of fast evolution at present, stimulated by the knowledge gained from the sequencing of the human genome, and from related developments (Triche 2006). Historically, pathologists have diagnosed cancer on the basis of histologic and cytologic features observed by macro- and micro-scopic examination, in recent years complemented by various laboratory tests. They make differential diagnoses of metastases from second independent primaries on the basis of the comparability of these pathologic features, along with relevant clinical information and common sense rules regarding this information, such as the expectation that a metastasis would be unlikely to have cells that have better differentiation, or an *in situ* component. However, ultimately, it is generally accepted that the crucial features of a cancer that determine its behavior and ancestry are the somatic mutations that have accumulated in the tumor cells. Thus, examination of these mutational patterns holds the definitive key to the accurate differential diagnosis of a metastasis versus a second independent primary cancer.

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 genetic fingerprints obtained from ACGH data. However, this is difficult for many reasons. In this article we have focussed on the statistical challenges. 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 (Speed 2008). After considering numerous options for summarizing the data, 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. Our strategy thus inevitably involves multiple stages. We must use segmentation methods to first identify the allelic gains and losses, and then we must use the new methods presented in this article to assess the closeness of their estimated locations. These comparisons are building blocks of information that are then combined with the gross correlation patterns of the losses and gains across the genome to determine an overall diagnosis for the patient. In our limited efforts to date to validate this strategy we observe that the method has good statistical properties in an ideal setting in which there is at most a single allelic loss or gain in each chromosome arm, and where the random errors in the marker values in the arrays are normally distributed. 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.

A difficult feature of the problem is the fact that the two hypotheses that we are trying to distinguish are structured very differently. Under the independence hypothesis, H_I, the somatic mutational patterns are presumed to have arisen independently. However, we know that different genetic loci experience mutations with very 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, which are different for different cancer types, is limited, and we chose to estimate them from the relatively small data sets at our disposal. Under the clonal hypothesis, H_C, the tumors are linked by allelic gains or losses that occurred in the original “clonal” cell that led to the cancers, and are thus identical. Therefore H_C is characterized by tumors that share some (at least one) clonal mutations, but these tumors may, and usually do, harbor numerous other non-clonal mutations. Consequently, we need a method that appropriately weighs the negative evidence of the presence of clearly non-clonal mutations against the positive evidence of closely matching mutations. We have approached the problem by constructing a likelihood in

which the relative frequency of clonal mutations in tumors that are clonal is assumed known (c), but in practice we have very limited knowledge of this parameter.

Because of the preceding features, we have leaned heavily in interpreting our analyses on the use of a "null" distribution of our likelihood ratio statistic, created by comparing tumors from different patients, tumors which are necessarily independent. Thus, despite the fact that the purpose of our analysis is differential classification of patients into H_C and H_I our analysis ultimately has a significance testing flavor in which we rely on the null distribution of the statistic under H_I to help define the appropriate diagnostic classifications. We note that a simple strategy for analyzing the data would be to formulate the problem as a significance test, with the diagnosis of independence as the null hypothesis, denoted H_I. The broad correlation of gains and losses could then be viewed as a set of independent, non-identically distributed multinomials with one outcome for each multinomial. Dale (1986) has proposed tests for independent non-identically distributed multinomials with sparse data, and has studied their properties. In our notation her test statistic would be $\sum_i \sum_j \sum_k (r_{jki} - q_{jki})^2 / q_{jki}$ for $j = g, l, n$ and $k = g, l, n$, where i represents the chromosome arm, and where $q_{ijk} = p_{ji} p_{ki}$. However, we examined this test in our context where each outcome $r_{...}$ takes the value 1 or 0. We found that it does not appear to have good small sample properties, and so we did not pursue this approach further. Formulation of the problem as a significance test of H_I would have followed the strategy we have used previously for the comparison of the mutational profiles at candidate markers (Begg et al. 2007, Ostrovnaya et al. 2008).

Application of the method to our various examples demonstrates clearly that it has the potential to convincingly establish the clinico-pathological diagnosis, and to change it in some patients. However, there are many limitations, and much additional research is needed to refine it and to better understand its statistical properties. The key areas for further investigation are as follows. First, since we seek a "better" diagnosis than the current standard, there is no gold standard benchmark against which to evaluate the classifications of the new method. Ultimately, clinical follow-up studies of patients may help to determine the gold standard, in that the clinical courses of patients with

metastases will generally be much worse than those of patients with new primaries. The absence of a gold standard diagnosis also inhibits our ability to calibrate the magnitudes of the likelihood ratios produced by the method. Second, the method requires that an initial segmentation analysis be performed to identify the allelic gains and losses. This is a statistical analysis in and of itself and it 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. Third, our method requires specification of marginal mutation rates in each chromosome arm, and a specification of the parameter c that characterizes the strength of the clonality signal. Although we need further research to understand the sensitivity of the method to errors in the specification of these parameters, our sensitivity analyses provide us with some confidence that diagnoses with high likelihood ratios are insensitive to the choice of c. Fourth, we have restricted the entire testing strategy to the assumption that each chromosome arm possesses at most one allelic gain or loss. In practice, sometimes multiple changes may be observed within a single chromosome arm. If these more complex patterns match closely on the two tumors the evidence favoring clonality can be greatly enhanced. Indeed we see such a pattern in Figure 8. This is from chromosome 5q on patient #13 in the Bollet et al. data, a patient with strong overall evidence for clonality. The segmentation for this plot is not restricted to the first detected allelic change, as in our previous analyses. We restricted our method to one-step changes for analytical simplicity, but the method could benefit from further refinement to accommodate complex changes of this nature which would seem to provide very strong evidence for clonal relatedness. Finally, we have focussed on the statistical issues, but in practice there are numerous practical aspects of molecular testing that can greatly influence the data and the resulting analyses. To accomplish ACGH testing tumor cells must be isolated for analysis. The tumor cells may be substantially contaminated with normal stromal or interstitial cells, and this can radically reduce the detectable signal in the allelic changes. As we have seen in our examples, the “quality” of the array data can also be affected by whether the tumor samples are fresh frozen or obtained from formalin fixed paraffin-embedded archival material.

The “quality” of the array data 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. Our analytic

strategy depends on several “tuning” parameters, including the significance level of the segmentation algorithm, the MAD criterion used to try to eliminate artifactual signals, and the choice of c to reflect the clonality signal. 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 elected 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 much additional research is needed to refine the methodological details.

Acknowledgements:

We thank 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. The research was supported by the National Cancer Institute, awards CA098438, CA125829 and CA124504.

References

Agaimy, A., Pelz, A. F., Corless, C. L., Wünsch, P. H., Heinrich, M. C., Hofstaedter, F., Dietmaier, W., Blanke, C. D., Wieacker, P., Roessner, A., Hartmann, A., and Schneider-Stock, R. (2007), "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*, 18, 9-15.

Agelopoulos, K., Tidow, N., Korsching, E., Voss, R., Hinrichs, B., Brandt, B., Boecker, W., and Buerger, H. (2003), "Molecular Cytogenetic Investigations of Synchronous Bilateral Breast Cancer," *Journal of Clinical Pathology*, 56, 660-665.

Begg, C. B. (1987), "Biases in the Assessment of Diagnostic Tests," *Statistics in Medicine,* 6, 411-423

Begg, C. B., Eng, K., and Hummer, A. J. (2007), "Statistical Tests for Clonality," *Biometrics*, 63, 522-530.

Bollet, M. A., Servant, N., Neuvial, P., Decraene, C., Lebigot, I., Meyniel, J-P., De Rycke, Y., Savignoni, A., Rigaill, G., Hupe, P., Fourquet, A., Sigal-Zafrani, B., Barillot, E., and Thiery, J-P. (2008), "High-Resolution Mapping of DNA Breakpoints to Define True Recurrences Among Ipsilateral Breast Cancers," *Journal of the National Cancer Institute*, 100, 48-58.

Brinkmann, D., Ryan, A., Ayhan, A., McCluggage, W. G., Feakins, R., Santibanez-Korf, M. F., et al. (2004), "A Molecular Genetic and Statistical Approach for the Diagnosis of Dual-Site Cancers," *Journal of the National Cancer Institute*, 96, 1441-1446.

Chunder, N., Roy, A., Roychoudhury, S., and Panda, C. K. (2004), "Molecular Study of Clonality in Multifocal and Bilateral Breast Tumors," *Pathology, Research and Practice*, 200, 735-741.

Dacic, S., Ionescu, D. N., Finkelstein, S., and Yousem, S. A. (2005), "Patterns of Allelic Loss of Synchronous Adenocarcinomas of the Lung," *American Journal of Surgical*

Pathology, 29, 897-902.

Dale, J. R. (1986), "Asymptotic Normality of Goodness-of-Fit Statistics for Sparse Product Multinomials," *J. R. Statist. Soc B,* 48, 48-59.

Ghazani, A. A., Arneson, N., Warren, K., Pintilie, M., Bayani, J., Squire, J. A., and Done, S. J. (2007), " Genomic Alterations in Sporadic Synchronous Primary Breast Cancer Using Array and Metaphase Comparative Genomic Hybridization," *Neoplasia*, 9, 511-520.

Goldstein, N. S., Vicini, F. A., Hunter, S., Odish, E., Forbes, S., and Kestin, L. L. (2005), "Molecular Clonality Relationships in Initial Carcinomas, Ipsilateral Breast Failures, and Distant Metastases in Patients Treated with Breast-Conserving Therapy: Evidence Suggesting that Some Distant Metastases are Derived From Ipsilateral Breast Failures and that Metastases Can Metastasize," *American Journal of Clinical Pathology*, 124, 49-57.

Goldstein, N. S., Vicini, F. A., Hunter, S., et al. (2005), "Molecular Clonality Determination of Ipsilateral Recurrence of Invasive Breast Carcinomas After Breast-Conserving Therapy: Comparison With Clinical and Biologic Factors," *American Journal of Clinical Pathology*, 123, 679-689.

Geurts, T. W., Nederlof, P. M., van den Brekel, M. W., et al. (2005), "Pulmonary Squamous Cell Carcinoma Following Head and Neck Squamous Cell Carcinoma: Metastasis or Second Primary?," *Clinical Cancer Research*,11, 6608-6614.

Haller, F., Schulten, H. J., Armbrust, T., Langer, C., Gunawan, B., and Füzesi, L. (2007), "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*, 31, 933-937.

Hiroshima, K., Toyozaki, T., Kohno, H., Ohwada, H., and Fujisawa, T. (1998), "Synchronous and Metachronous Lung Carcinomas: Molecular Evidence for Multicentricity," *Pathology International*, 48, 869-876.

Holst, V. A., Finkelstein, S., and Yousem, S. A. (1998), "Bronchioloalveolar

Adenocarcinoma of Lung: Monoclonal Origin for Multifocal Disease," *American Journal of Surgical Pathology*, 22, 1343-1350.

Huang, J., Behrens, C., Wistuba, I., Gazdar, A. F., and Jagirdar, J. (2001), "Molecular Analysis of Synchronous and Metachronous Tumors of the Lung: Impact on Management and Prognosis," *Annals of Diagnostic Pathology*, 5, 321-329.

Hupe, P., La Rosa, P., Liva, S., Lair, S., Servant, N., and Barillot, E. (2007) ACTuDB, a New Database for the Integrated Analysis of Array-CGH and Clinical Data for Tumors. *Oncogene*, 26, 6641-52.

Hwang, E., Nyante, S. J., Yi Chen, Y., Moore, D., DeVries, S., Korkola, J. E., Esserman, L. J., and Waldman, F. M. (2004), "Clonality of Lobular Carcinoma In Situ and Synchronous Invasive Lobular Carcinoma," *Cancer*, 100, 2562-2572.

Imyanitov, E. N., Suspitsin, E. N., Grigoriev, M. Y., Togo, A. V., Kuligina, E. S., Belogubova, E. V, et al. (2002), "Concordance of Allelic Imbalance Profiles in Synchronous and Metachronous Bilateral Breast Carcinomas," *International Journal of Cancer*, 100, 557-564.

Janschek, E., Kandioler-Eckersberger, D., Ludwig, C., et al. (2001), "Contralateral Breast Cancer: Molecular Differentiation Between Metastasis and Second Primary Cancer," *Breast Cancer Research and Treatment*, 67, 1-8.

Jiang, J-K., Chen, Y-J., Lin, C-H., Yu, I-T., and Lin, J-K. (2005), "Genetic Changes and Clonality Relationship Between Primary Colorectal Cancers and Their Pulmonary Metastases – An Analysis by Comparative Genomic Hybridization," *Genes, Chromosomes and Cancer*, 43, 25-36.

Knösel, T., Schlüns, K., Dietel, M., and Petersen, I. (2005), "Chromosomal Alterations in Lung Metastases of Colorectal Carcinomas: Associations With Tissue Specific Tumor Dissemination," *Clinical and Experimental Metastasis*, 22, 533-538.

Kollias, J., Man, S., Marafie, M., et al. (2000), “Loss of Heterozygosity in Bilateral Breast Cancer,” *Breast Cancer Research and Treatment*, 64, 241-251.

Kuukasjarvi, T., Karhu, R., Tanner, M., Kahkonen, M., Schaffer, A., Nupponen, N., et al. (1997), “Genetic Heterogeneity and Clonal Evolution Underlying Development of Asynchronous Metastasis in Human Breast Cancer,” *Cancer Research*, 57, 1597-1604.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005), “Comparative Analysis of Algorithms for Identifying Amplifications and Deletions in Array CGH Data,” *Bioinformatics*, 21, 3763-3770.

Lau, D. H., Yang, B., Hu, R., Benfield, J. R. (1997), “Clonal Origin of Multiple Lung Cancers: K-ras and p53 Mutations Determined by Nonradioisotopic Single-Strand Conformation Polymorphism Analysis,” *Diagnostic Molecular Pathology*, 6, 179-184.

Leong, P. P., Rezai, B., Koch, W. M., et al. (1998), “Distinguishing Second Primary Tumors From Lung Metastases in Patients With Head and Neck Squamous Cell Carcinoma,” *Journal of the National Cancer Institute*, 90, 972-977.

Matsuzoe, D., Hideshima, T., Ohshima, K., Kawahara, K., Shirakusa, T., and Kimura, A. (1999), “Discrimination of Double Primary Lung Cancer From Intrapulmonary Metastasis By p53 Gene Mutation,” *British Journal of Cancer*, 79, 1549-1552.

Murase, T., Takino, H., Shimizu, S., et al. (2003), “Clonality Analysis of Different Histological Components in Combined Small Cell and Non-Small Cell Carcinoma of the Lung,” *Human Pathology*, 34, 1178-1184.

Nestler, U., Schmidinger, A., Schulz, C., Huegens-Penzel, M., Gamerdinger, U. A., Koehler, A., Kuchelmeister, K. W. (2007), “Glioblastoma Simultaneously Present With Meningioma--Report of Three Cases,” *Zentralbl Neurochir*, 68, 145-150. Epub 2007 Jul 30.

Nishizaki, T., Chew, K., Chu, L., Isola, J., Kallioniemi, A., Weidner, N., and Waldman, F. M. (1997), “Genetic Alterations in Lobular Breast Cancer by Comparative Genomic Hybridization,” *International Journal of Cancer*, 74, 513-517.

Nyante, S. J., Devries, S., Chen, Y. Y., and Hwang, E. S. (2004), “Array-Based Comparative Genomic Hybridization of Ductal Carcinoma In Situ and Synchronous Invasive Lobular Cancer,” *Human Pathology*, 35, 759-763.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004), “Circular Binary Segmentation for the Analysis of Array-Based DNA Copy Number Data,” *Biostatistics*, 5, 557-572.

Orlow, I., Tommasi, D., Bloom, B., Ostrovnaya, I., Cotignola, J., Mujumdar, U., Busam, K. J., Jungbluth, A. A., Scolyer, R. A., Thompson, J. F., Armstrong, B. K., Berwick, M., Thomas, N., and Begg, C. B. (2008), “Molecular Profiling to Distinguish Multiple Independent Primary Melanomas from Melanoma Metastases,” *Cancer Research*, submitted.

Ostrovnaya, I., Seshan, V. E., Begg, C. B. (2008), “Comparison of Properties of Tests for Assessing Tumor Clonality,” *Biometrics*, in press.

Park, S. C., Hwang, U. K., Ahn, S. H., Gong, G. Y., and Yoon, H. S. (2007), “Genetic Changes in Bilateral Breast Cancer by Comparative Genomic Hybridisation,” *Clinical and Experimental Medicine*, 7, 1-5.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., et al. (1998), “High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Micro-Arrays,” *Nature Genetics*, 20, 207-11.

Regitnig, P., Ploner, F., Maderbacher, M., and Lax, S. F. (2004), “Bilateral Carcinomas of the Breast With Local Recurrence: Analysis of Genetic Relationship of the Tumors,” *Modern Pathology*,17, 597-602.

Schlechter, B. L., Yang, Q., Larson, P. S., et al. (2004), “Quantitative DNA Fingerprinting May Distinguish New Primary Breast Cancer From Disease Recurrence,” J*ournal of Clinical Oncology*, 22, 1830-1838.

Sheather, S. J. and Jones M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc.* B, 683–690.

Shelley Hwang, E., Nyante, S. J., Yi Chen, Y., Moore, D., DeVries, S., Korkola, J. E., Esserman, L. J., and Waldman, F. M. (2004), "Clonality of Lobular Carcinoma In Situ and Synchronous Invasive Lobular Carcinoma," *Cancer*, 100, 2562-2572.

Shimizu, S., Yatabe, Y., Koshikawa, T., et al. (2000), "High Frequency of Clonally Related Tumors in Cases of Multiple Synchronous Lung Cancers as Revealed by Molecular Diagnosis," *Clinical Cancer Research*, 6, 3994-3999.

Shin, S. W., Breathnach, O. S., Linnoila, R. I., et al. (2001), "Genetic Changes in Contralateral Bronchioloalveolar Carcinomas of the Lung," *Oncology*, 2001, 60, 81-87.

Sieben, N. L. G., Kolkman-Uljee, S. M., Flanagan, A. M., le Cessie, S., Cleton-Jansen, A. M., Cornelisse, C. J., and Fleuren, G. J. (2003), "Molecular Genetic Evidence for Monoclonal Origin of Bilateral Ovarian Serous Borderline Tumors," *American Journal of Pathology*, 162, 1095-1101.

Snijders, A. M., et al. (2001), "Assembly of Microarrays for Genomewide Measurement of DNA Copy Number," *Nature Genetics*, 29, 263-264.

Sozzi, G., Miozzo, M., Pastorino, U., et al. (1995), "Genetic Evidence for an Independent Origin of Multiple Preneoplastic and Neoplastic Lung Lesions," *Cancer Research*, 1995, 55, 135-140.

Speed, T. P. (2008), "Terence's Stuff: Statistics Without Probability," *Institute of Mathematical Statistics Bulletin*, 36, 12.

Stenmark-Askmalm, M., Gentile, M., Wingren, S., and Stahl, O. (2001), "Protein Accumulation and Gene Mutation of p53 in Bilateral Breast Cancer," South-East Sweden Breast Cancer Group. *Acta Oncologica*, 40, 56-62.

Teixeira, M. R., Ribeiro, F. R., Torres, L., Pandis, N., Anderson, J. A., Lothe, R. A., and

Heim, S. (2004), “Assessment of Clonal Relationships in Ipsilateral and Bilateral Multiple Breast Carcinomas By Comparative Genomic Hybridization and Hierarchical Clustering Analysis,” *British Journal of Cancer*, 91, 775-782.

Torres, L., Ribeiro, F. R., Pandis, N., Andersen, J. A., Heim, S., and Teixeira, M. R. (2007), “Intratumor Genomic Heterogeneity in Breast Cancer With Clonal Divergence Between Primary Carcinomas and Lymph Node Metastases,” *Breast Cancer Research and Treatment*, 102, 143-155.

Triche, T.J. (2006), “Technologies in Molecular Biology: Diagnostic Applications”, In Oncology: An Evidence-Based Approach, eds. A.E. Chang et al., Springer, New York, pp269-284.

Tse, G. M., Kung, F. Y., Chan, A. B., Law, B. K., Chang, A. R., and Lo, K. W. (2003), “Clonal Analysis of Bilateral Mammary Carcinomas By Clinical Evaluation and Partial Allelotyping,” *American Journal Clinical Pathology*, 120, 168-174.

van Rens, M. T., Eijken, E. J., Elbers, J. R., Lammers, J. W., Tilanus, M. G., and Slootweg, P. J. (2002), “P53 Mutation Analysis for Definite Diagnosis of Multiple Primary Lung Carcinoma,” *Cancer*, 94,188-196.

Venkatraman, E. S., and Olshen, A. B. (2006), “A Faster Circular Binary Segmentation Algorithm for the Analysis of Array CGH Data,” *Bioinformatics*, 23, 657-663.

Wa, C. V., DeVries, S., Chen, Y. Y., Waldman, F. M., and Hwang, E. S. (2005), “Clinical Application of Array-Based Comparative Genomic Hybridization to Define the Relationship Between Multiple Synchronous Tumors,” *Modern Pathology*, 18, 591-597.

Waldman, F. M., DeVries, S., Chew, K. L., Moore, D. H. 2nd, Kerlikowske, K., and Ljung, B. M. (2000), “Chromosomal Alterations in Ductal Carcinomas In Situ and Their In Situ Recurrences,” *Journal of the National Cancer Institute*, 92, 313-320.

Weiss, M. M., Kuipers, E. J., Meuwissen, S. G., van Diest, P. J., and Meijer, G. A. (2003), “Comparative Genomic Hybridisation as a Supportive Tool in Diagnostic Pathology,”

Journal of Clinical Pathology, 56, 522-527.

Willenbrock, H., and Fridlyand, J. (2005), “A Comparison Study: Applying Segmentation to Array CGH Data for Downstream Analyses,” *Bioinformatics*, 21, 4084-4091.

Table 1								
Clinical Data and Results for Diagnoses of Ipsilateral Breast Cancer (Bollet et al. 2008)								
Pt #	**Histology[1] 1st**	**2nd**	**Time Interval[2]**	**Quadrant[3]**	**Clinical Diagnosis**	**ACGH Results LR1**	**LR2**	**Diagnosis[4]**
1	Ductal	Ductal	6.5	Same	I	4.3×10^{-4}	5.0×10^{-4}	I
2	Ductal	Lobular	5.3	Same	I	1.9	1.9	E
3	Ductal	Ductal	3.1	Same	C	1.1×10^{4}	2.0×10^{5}	C
4	Lobular	Lobular	3.5	Same	C	7.1×10^{1}	6.6×10^{4}	C
5	Ductal	Ductal	2.0	Same	C	1.1×10^{6}	3.3×10^{26}	C
6	Lobular	Lobular	3.1	Same	C	9.4	9.4	E
10	Lobular	Ductal	5.0	Different	I	2.6×10^{-2}	2.6×10^{-2}	I
11	Lobular	Ductal	6.3	Same	I	1.5×10^{-4}	1.5×10^{-4}	I
12	Lobular	Lobular	2.9	Different	I	5.5	5.5	E
13	Ductal	Ductal	4.6	Same	C	1.4×10^{3}	1.9×10^{16}	C
14	Lobular	Lobular	2.5	Same	C	3.7×10^{3}	1.2×10^{8}	C
15	Ductal	Ductal	3.3	Same	C	3.9×10^{2}	2.7×10^{6}	C
16	Ductal	Ductal	3.8	Same	I	2.5×10^{1}	4.1	E
18	Ductal	Ductal	2.2	Same	I	8.7×10^{-3}	6.3×10^{-4}	I
19	Ductal	Ductal	3.0	Same	C	2.8×10^{-1}	1.8×10^{7}	C
20	Ductal	Ductal	1.4	Different	I	2.2	9.9×10^{-1}	E
21	Ductal	Ductal	4.2	Same	C	4.8×10^{3}	1.3×10^{27}	C
22	Ductal	Micro-Pap	3.5	Same	I	1.3	3.6×10^{3}	C
23	Ductal	Ductal	0.8	Same	C	3.6×10^{2}	5.5×10^{13}	C
24	Ductal	Ductal	1.0	Same	C	5.7×10^{3}	1.8×10^{9}	C
25	Ductal	Ductal	2.2	Same	C	3.5×10^{5}	2.3×10^{16}	C
26	Ductal	Ductal	1.8	Same	C	1.8×10^{4}	7.5×10^{13}	C

1. It is presumed generally that tumors must have the same histology to be clonally related.
2. Time interval between tumor diagnoses in years: the longer the interval, the less likely it is that the second tumor is a metastasis.
3. A closer anatomical relationship (same quadrant) is believed to increase the probability of clonal relatedness.
4. I – Independent Primary; C – Clonal (metastasis); E – Equivocal (diagnosis uncertain).

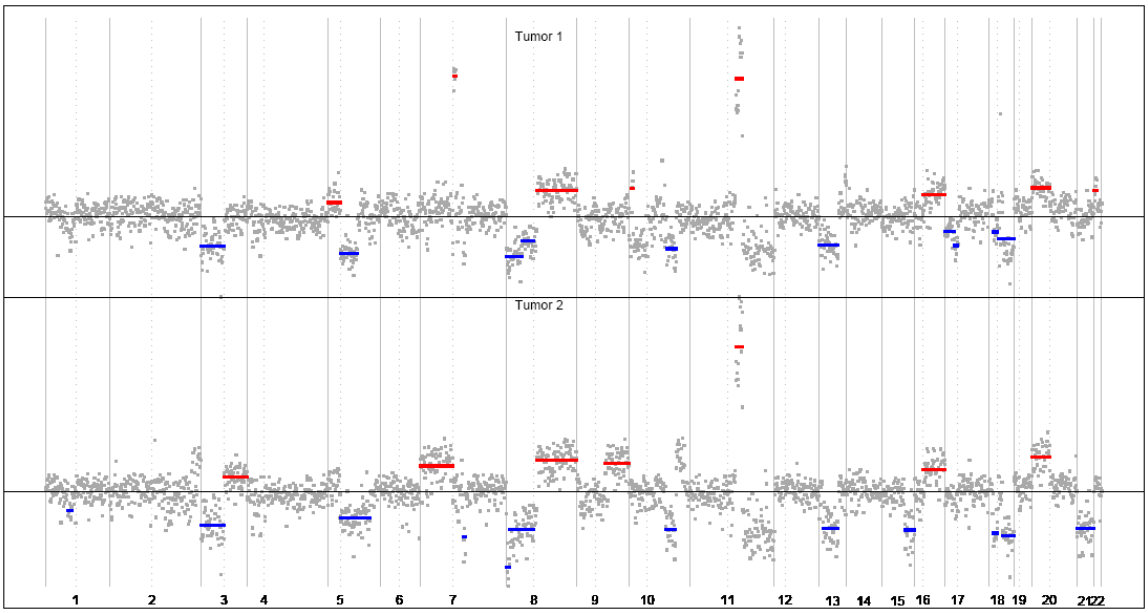

Figure 1. Whole genome segmentation of tumors from the patient with cancer of the mouth described in Sections 2 and 5. The red (blue) lines represent allelic gains (losses) as determined by the segmentation algorithm.

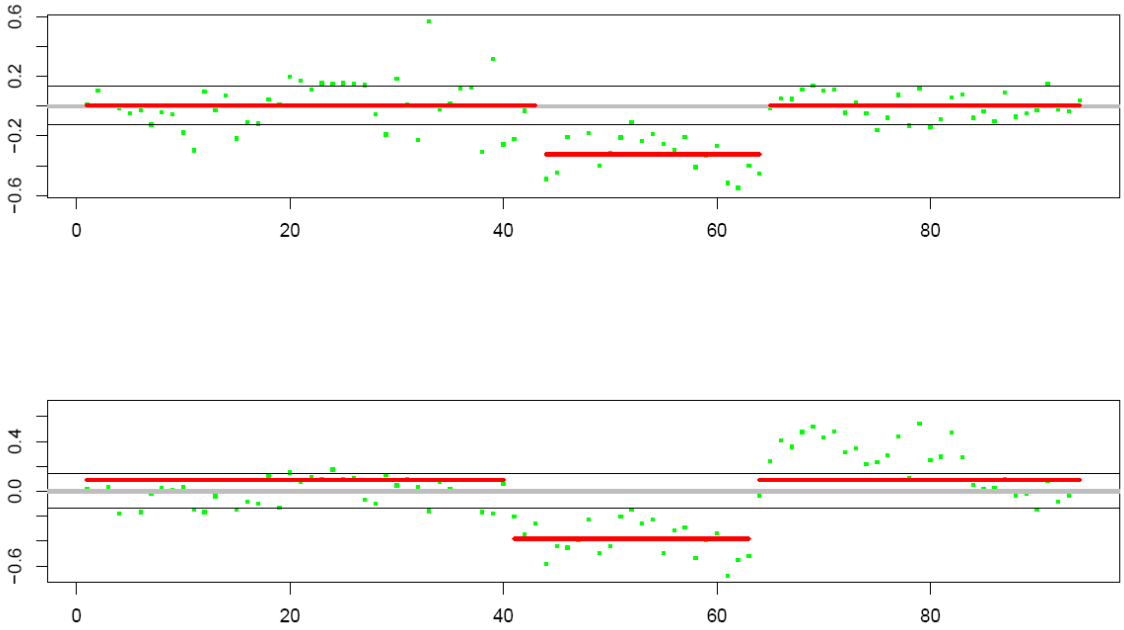

Figure 2. Detailed view of chromosome 10q segmentation of the patient with cancer of the mouth described in Sections 2 and 5.

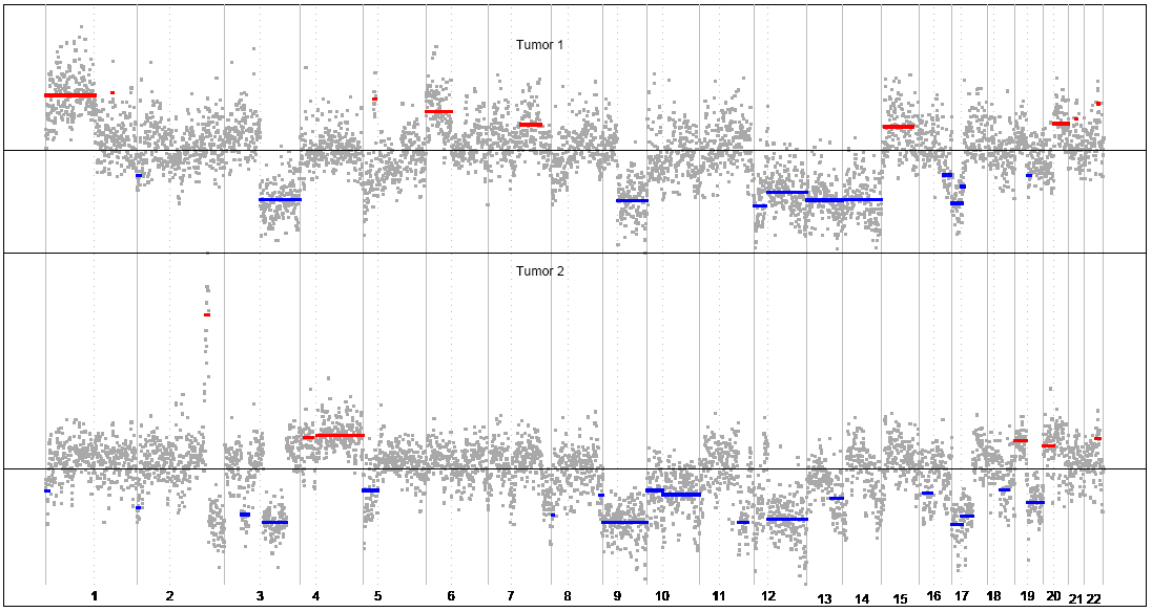

Figure 3. Whole genome segmentation of tumors from the patient with two melanomas described in Sections 2 and 5. The red (blue) lines represent allelic gains (losses) as determined by the segmentation algorithm.

.

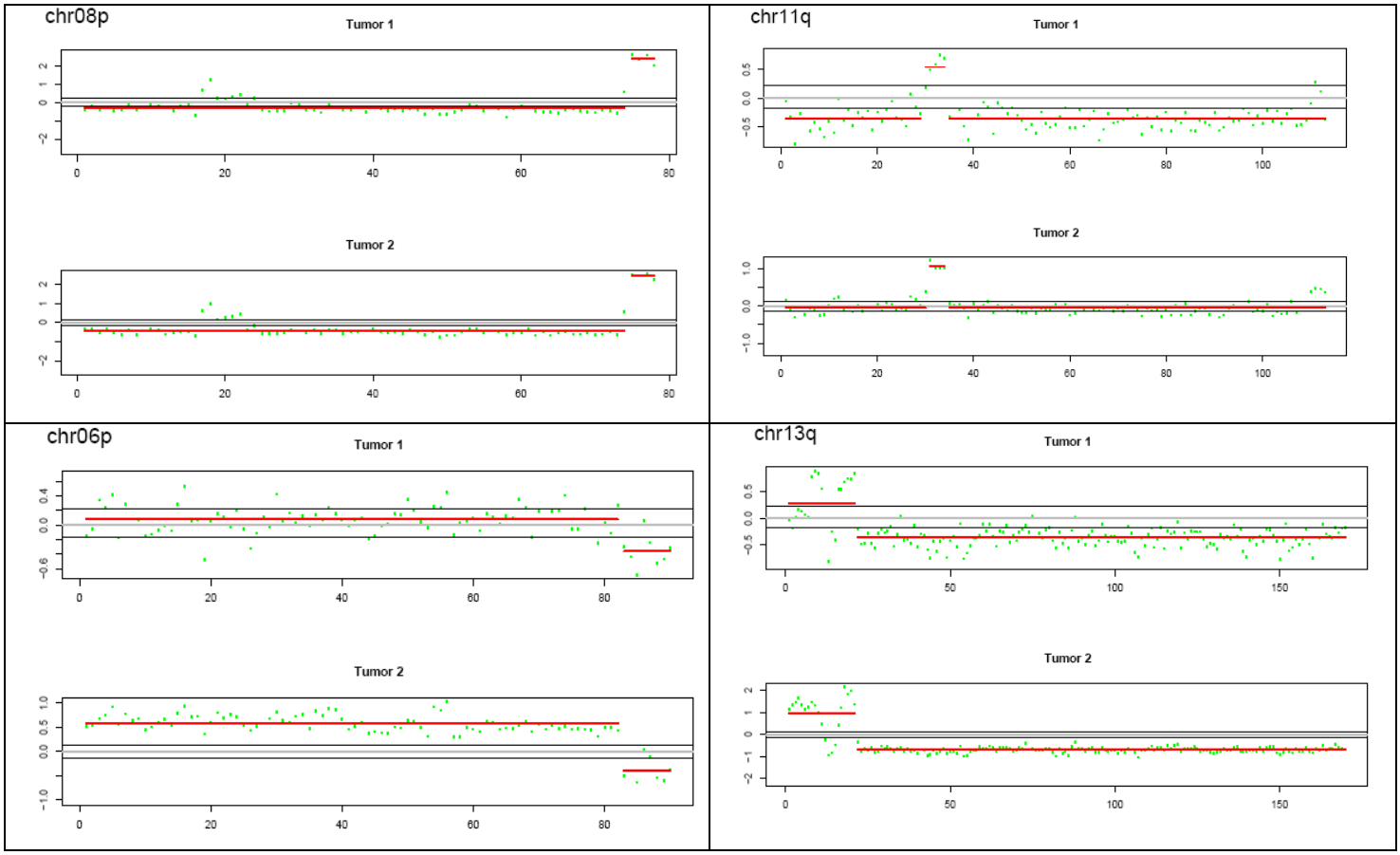

Figure 4. Clonal Mutations from Patient #22 from Bollet et al. (2008)

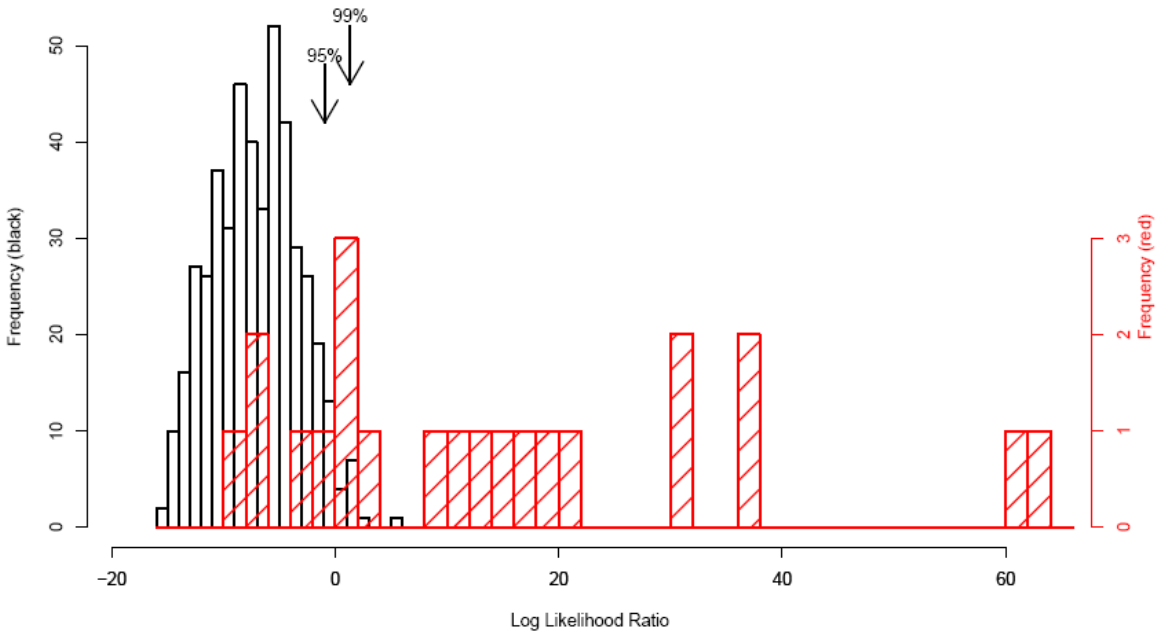

Figure 5. Likelihood ratios for patients in Bollet et al. data (blue) superimposed on reference histogram from independent tumor pairings from different patients (black).

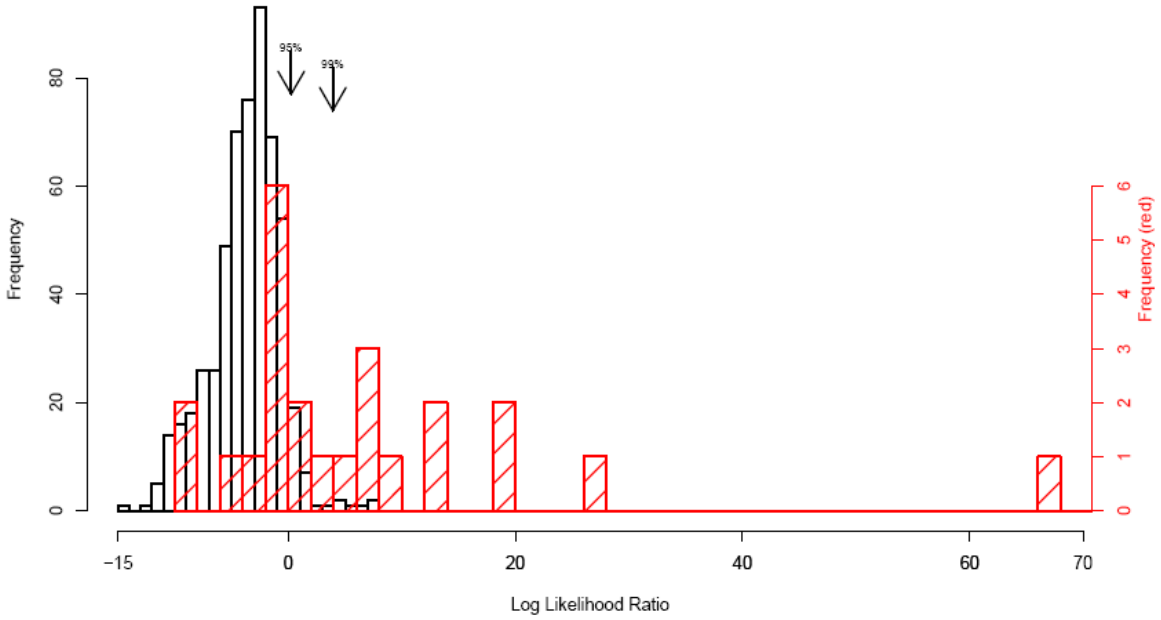

Figure 6. Likelihood ratios for patients in Hwang et al. data (blue) superimposed on reference histogram from independent tumor pairings from different patients (black).

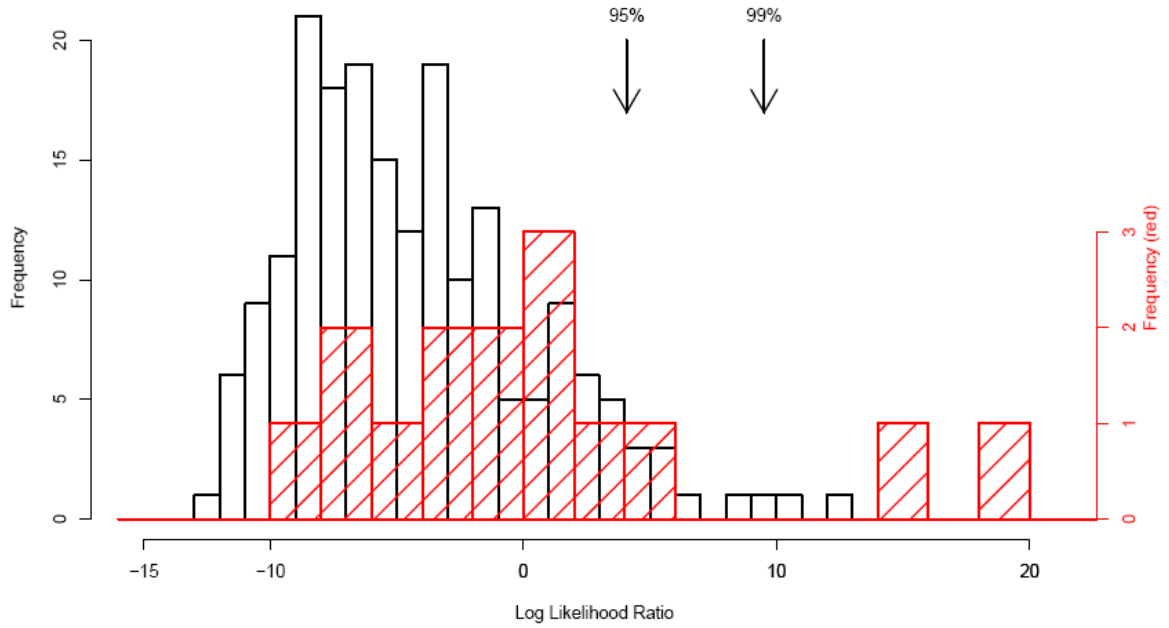

Figure 7. Likelihood ratios for patients in head and neck cancer dataset (blue) superimposed on reference histogram from independent tumor pairings from different patients (black).

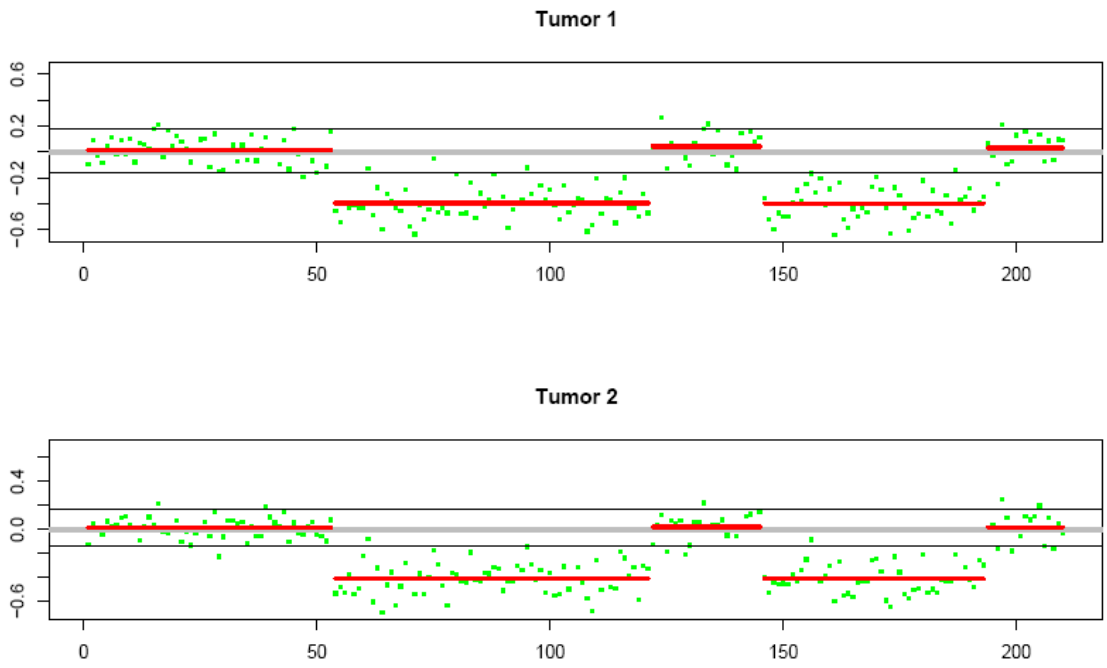

Figure 8. Example of a closely matching complex change, from 5q on patient #13 in Bollet et al. (2008).