BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 24 2013, pages 3113–3120
doi:10.1093/bioinformatics/btt546

Genome analysis

Advance Access publication September 23, 2013

Methods and challenges in timing chromosomal abnormalities within cancer samples

Elizabeth Purdom[1,*], Christine Ho[1], Catherine S. Grasso[2], Michael J. Quist[2], Raymond J. Cho[3] and Paul Spellman[2]

[1]Department of Statistics, University of California, Berkeley, 367 Evans Hall Berkeley, CA 94720-3860, USA, [2]Department of Molecular and Medical Genetics, Oregon Health & Science University, Portland, OR 97239, USA and [3]Department of Dermatology, University of California, San Francisco, CA 94115, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Tumors acquire many chromosomal amplifications, and those acquired early in the lifespan of the tumor may be not only important for tumor growth but also can be used for diagnostic purposes. Many methods infer the order of the accumulation of abnormalities based on their occurrence in a large cohort of patients. Recently, Durinck *et al.* (2011) and Greenman *et al.* (2012) developed methods to order a single tumor's chromosomal amplifications based on the patterns of mutations accumulated within those regions. This method offers an unprecedented opportunity to assess the etiology of a single tumor sample, but has not been widely evaluated.

Results: We show that the model for timing chromosomal amplifications is limited in scope, particularly for regions with high levels of amplification. We also show that the estimation of the order of events can be sensitive for events that occur early in the progression of the tumor and that the partial maximum likelihood method of Greenman *et al.* (2012) can give biased estimates, particularly for moderate read coverage or normal contamination. We propose a maximum-likelihood estimation procedure that fully accounts for sequencing variability and show that it outperforms the partial maximum-likelihood estimation method. We also propose a Bayesian estimation procedure that stabilizes the estimates in certain settings. We implement these methods on a small number of ovarian tumors, and the results suggest possible differences in how the tumors acquired amplifications.

Availability and implementation: We provide implementation of these methods in an R package `cancerTiming`, which is available from the Comprehensive R Archive Network (CRAN) at http://CRAN.R-project.org/.

Contact: epurdom@stat.Berkeley.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 29, 2012; revised on April 26, 2013; accepted on September 18, 2013

1 INTRODUCTION

Tumors accumulate large numbers of mutations and other chromosomal abnormalities due to defects in the genomic repair mechanisms of tumor cells. Not all of these abnormalities are believed to be crucial for tumor growth and progression, and a question of great importance is to try to identify the critical abnormalities. One possible indicator of the importance of an abnormality is when it occurred, relative to other abnormalities. The most straightforward approach for determining the progression of these abnormalities is to evaluate multiple samples from the same individual, such as primary and metastatic samples (Frumkin *et al.*, 2008; Gerlinger *et al.*, 2012; Nishizaki *et al.*, 1997; Sasatomi *et al.*, 2002), sub-clonal populations (Campbell *et al.*, 2008), or different portions of the same tumor (Navin and Hicks, 2010; Siegmund *et al.*, 2009).

It is usually difficult to have data on multiple time points in the progression of an individual tumor; rather it is more common to have cross-sectional data with a single time point from multiple individuals. In this case, we cannot directly observe the accumulation of genomic abnormalities and must infer it. There has been a great deal of interest in identifying driver mutations and events based on the frequency of their occurrence across patients (e.g. Beroukhim *et al.*, 2007; Cancer Genome Atlas Research Network, 2008, 2011; Taylor *et al.*, 2008). Many statistical methodologies rigorously analyze frequencies of aberrations to determine those that are significantly represented in the population, with specific statistical methods developed for mutations, copy-number abnormalities and others types of genomic profiles (Beroukhim *et al.*, 2007; Brodeur *et al.*, 1982; Huang *et al.*, 2007; Newton and Lee, 2000; Newton *et al.*, 1994, 1998; Taylor *et al.*, 2008). Yet, these techniques do not explicitly attempt to estimate the order of occurrence.

Many methods do explicitly estimate a common temporal ordering among samples based on the co-occurrence across patients. Fearon and Vogelstein (1990) first proposed a temporal ordering of mutations based on the mutations in colorectal tumors from different stages. Since then, a great deal of methodological work has formalized this work. For example, oncogenetic tree models (Desper *et al.*, 2000) cast this notion in a probabilistic setting, which later work extended and generalized (Beerenwinkel *et al.*, 2005a, b, 2006; Gerstung *et al.*, 2009; Hjelm *et al.*, 2006; Liu *et al.*, 2009; Newton, 2002; Rahnenführer *et al.*, 2005; Simon *et al.*, 2000). Bilke (2005) modeled the critical elements of tumor progression in neuroblastoma by analyzing the sets of shared mutations between the stages of a tumor and finding the most likely model of progression between stages of aberrations. Other approaches rely on stochastic models of cellular

*To whom correspondence should be addressed.

growth, such as the algorithm RESIC (Attolini *et al.*, 2010), which models the overall accumulation of abnormalities in a population of tumors based on a probabilistic model of cell division and fitness of mutations. This is not an exhaustive review, as many other approaches to this problem exist, but a common feature is estimating a common temporal ordering by comparing across many samples.

The simulation study of Sprouffske *et al.* (2011) compares estimates of tumor progression that use multiple samples from a single tumor/patient with those obtained from cross-sectional samples from many independent tumors and finds that the cross-sectional estimates can be quite misleading due to the heterogeneity of paths to tumorigenesis seen in tumors. In Durinck *et al.* (2011), we introduced a novel approach for temporal ordering of genomic abnormalities that instead focused on assessing *from a single sample of a single individual* the internal ordering of chromosomal abnormalities. In that work, we dealt with the narrow case of ordering regions of copy-neutral loss-of-heterozygosity (CNLOH), when one copy of the chromosome is deleted and replaced by the other copy. Greenman *et al.* (2012) gave a generalization to general chromosomal amplifications, and in Nik-Zainal *et al.* (2012) applied the technique to 21 breast cancer samples.

The model proposed by Durinck *et al.* (2011) and Greenman *et al.* (2012) is general in principle and offers a remarkable ability to analyze the history of a single tumor. However, there has been little examination of the performance of the estimates of the temporal orderings. We show that with higher levels of amplification, most events result in non-identifiable models, meaning that most regions with high levels of amplifications cannot be timed in this way. Furthermore, differences in estimation procedures can have important effects on the quality of the estimate, particularly for events that occur early in tumorigenesis, and therefore are of particular biological interest. The method of Greenman *et al.* (2012) uses a partial maximum-likelihood estimation (MLE) technique, which can perform badly for early (and late) events. We introduce a full MLE procedure that accounts for sequencing variation, which we show performs better for estimating early events, particularly for samples with moderate read coverage. We also introduce a Bayesian estimation procedure, which can, in some situations, stabilize estimates for early events when there are low numbers of mutations in a region.

The implementation of these methods on ovarian tumors, which contain a large number of amplifications, allows for the examination of the general pattern of amplification in ovarian tumors and suggests that there might be two distinct patterns of amplification present: steady accumulation of amplifications over time versus whole-genome amplification.

Traditional copy-number analysis examines regions of the normal genome as to whether they are amplified in the tumor, and as such is largely our focus; indeed, this is the only alternative in the case of exon sequencing. If limited to this approach, only regions with one of three types of allelic copy number are viable candidates for the temporal analysis. However, an amplification and insertion of one region of the genome into another creates a different genome with connections between regions that do not exist in the normal genome. In the case of whole-genome sequencing, reads that span such breakpoints will be sequenced and algorithms have been proposed to use these breakpoints to estimate the relationships between these separate regions (Greenman *et al.*, 2012). The additional information regarding the connections between regions, when available, has the potential to make specific estimates of timing more feasible.

2 METHODS

We consider regions that have chromosomal copy-number changes, in other words a region in the genome that has been amplified or deleted a known number of times resulting in S copies of the region. What we observe as a single region with abnormal copy number is generally the culmination of a series of K events resulting in the final observed copy number. This results in $K+1$ stages in the life of the tumor where the region's copy number is stable, and the goal is to estimate the proportion of the lifetime of the tumor spent in each stage. As we make clear later, we are generally only able to consider regions that have a history of only amplifications, but for now we will keep the terminology general.

At each stage, individual point mutations could have been introduced into one of the existing copies. The amount of time for which the tumor rested in a particular stage will determine the probability of a mutation accumulating in that region during that time, as will the mutation rate of the tumor at that time. More precisely, let the vector $\pi = (\pi_0, \ldots, \pi_K)$ consist of the probabilities that a mutation originated in the corresponding stage of the tumor progression. Comparing vectors π calculated for different regions allows for precise comparison of the temporal order of aberrations in different regions. Of particular interest might be π_0, the proportion of time before any chromosomal change.

2.1 Model

We now describe a basic probablistic model for linking the vector π to the observed set of N mutations, as proposed by Durinck *et al.* (2011) for CNLOH events and generalized by Greenman *et al.* (2012).

Assume there are N total mutations in the region. Let P_i be the allele frequency for a mutated location i, defined as the proportion of the sequenced copies that are mutated in that location. We can only consider locations that have been mutated *and* have $P_i > 0$; locations that have been mutated in the past but have $P_i = 0$ at the time of observing the tumor cannot be distinguished from locations that were never mutated. We do not assume that we know which of the S copies hold the mutation or from which of the original copies (maternal or paternal) the mutation is descended. The set of possible values for of P_i are given by $\{1/S, \ldots, S/S\}$ for a pure tumor sample. For generality, we will denote the set of them as $\{a_1, \ldots, a_S\}$ to handle possible contamination in our sample (see Supplementary Appendix 3). Depending on the types of copy-number changes that have occurred, only a subset of the set $\{a_1, \ldots, a_S\}$ may actually be possible. For example, if only histories with amplifications events are considered, then S/S is not possible.

Then P_i is a multinomial random variable defined by a probability vector $q = (q(a_1), \ldots, q(a_S))^T$ that gives the probability of a randomly acquired mutation having allele frequency for each a_j, $q(a_j) = P(P_i = a_j | P_i > 0)$. The values of the vector q depends on the random process of mutagenesis over the life of the tumor. Specifically, P_i is completely determined by two random events: (i) the stage in which mutation i occurred and (ii) which copy in existence during that stage was mutated. We can formulate a probability model that links q with the parameter of interest, π, by making the following assumptions:

(1) Each location was mutated once in the history of the tumor

(2) If a mutation occurred in stage k, it is equally likely to be on any of the copies in existence during stage k

(3) The probability of a mutation occurring in a stage k is assumed proportional to π_k and the number of copies of the region in existence during the stage k. As we are concerned only with locations

i with $P_i > 0$, we can use Bayes rule and assumption 2, previously mentioned, to obtain

$$P(\text{mutation at } i \text{ originated in stage } k | P_i > 0) = S_k \pi_k / c_\pi$$

where S_k is the number of copies at stage k that survive to the end point at which the tumor is observed, and c_π is a normalizing constant equal to $\sum_{k=0}^{K} S_k \pi_k$

Let $\mathcal{C}_{kj}$ be the set of copies in existence in stage k that lead to an allele frequency a_j in the observed tumor. All mutations originating from stage k on the same copy s will have the same allele frequency. We notate that allele frequency as $p(k,s)$; $p(k,s)$ will be the proportion of the existing S copies that trace their descent from the parental copy s that existed in stage k. Conditioning we can write that

$$\begin{aligned} q(a_j) &= \sum_{k=0}^{K} \sum_{s \in \mathcal{C}_{kj}} P(\text{mutation on copy } s | \text{mutated in } k\text{th stage}, P_i > 0) \\ &\quad \times P(\text{mutated in } k\text{th stage} | P_i > 0) \\ &= \sum_{k=0}^{K} \sum_{s \in \mathcal{C}_{kj}} \frac{1}{S_k} \frac{S_k \pi_k}{c_\pi} = \frac{1}{c_\pi} \sum_{k=0}^{K} A_{jk} \pi_k, \text{ where} \\ A_{jk} &= \{\#\text{copies in stage } k \text{ that result in allele frequency } a_j\} \end{aligned}$$

We can collect the elements A_{jk} into an $S \times (K+1)$ matrix A and write the model simply as $q = A\pi / c_\pi$. As noted earlier, only a subset of the set $\{1/S, \ldots, S/S\}$ may actually be possible. When this is the case, the corresponding row of A will be all zeros, and can be removed.

2.2 Identifiability

To formulate the model, we must assume that we can determine the number of copies in stage k that could have resulted in allele frequency a_j to construct the matrix A. This requires precise knowledge of the history of amplifications. If we have three copies of the maternal copy (M) and two of the paternal copy (P), these could have been acquired in a variety of ways. For example, the M copy could be duplicated, then the P copy and then one of the existing copies of M. This will result in a different matrix A than if the P copy were duplicated, and then the M copy was duplicated two times (see Supplementary Table S2).

If there is only one event ($K = 1$), then the only possible event histories are a deletion, a gain, or a CNLOH event where one of the copies deletes and replaces the other copy simultaneously (a deletion would not result in an identifiable model, see later in text). These events can generally be distinguished from each other based on estimating the total copy number S as well as the allelic copy number (the number of copies of the maternal and paternal alleles). Similarly, for case $S = 4$, if we assume K-2—i.e. two events where at each event a single copy of the region was amplified—the allelic copy number is sufficient to distinguish the two possible history matrices A corresponding to (1,3) and (2,2) allelic copy numbers. When there is >2 events in the history of the region ($S > 4$), even if they are simple amplifications, then there are multiple event histories that can lead to the same allelic copy number but different histories A, and thus different resulting probabilities, q, of observed allele frequencies. If only allelic copy number is available, as with exome sequencing, this means that only these five cases where K equals 1 or 2 can we know the matrix A.

With whole-genome sequencing, algorithms have been proposed to use reads spanning the breakpoints to reconstruct the event history A (Greenman *et al.*, 2012), though not all amplified regions will have a unique construction. With exome sequencing, however, the event histories of large amplification regions will not be distinguishable.

Even if the event history matrix A is completely known, the question remains as to whether π is identifiable, i.e. does complete knowledge of the probability distribution of the data, q, allow for reconstruction of the parameter π? It is clear that π is only identifiable if the matrix A has rank $K+1$. For this reason, we can only estimate π in cases where there have been no deletions (Greenman *et al.*, 2012), so only regions with a history of pure amplification can be considered. An exception is the setting of CNLOH, where the assumption is that one of the copies deletes and replaces the other copy simultaneously so the time for the "stage" corresponding to a deletion is zero.

We show that in the case of sequential amplification (where each event is the addition of only one copy of the region) the sequence of events for which this will be true is limited: for a total number of copies equal to S, there is always exactly one history such that A is invertible and it is a history where all of the gains are on a single line of descent. This necessitates that the minor copy must have copy number 1, but this is not a sufficient condition if $S > 4$. When $S > 4$, even if the minor copy number is equal to 1, there can still be multiple histories associated with it and only one of these histories will be identifiable (see Fig. 1). This implies that of the five cases where the A matrix can be identified based solely on allelic copy number, only three of them are identifiable: CNLOH (2,0), single gain (1,2), and unbalanced two gains (1,3). For the two tumor types we analyzed—ovarian and skin—this was around 40% of the amplified regions, see Supplementary Table S3.

In the sequential amplification setting, the only identifiable matrix A has a simple form and its inverse has the same simple form,

$$\begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & 0 \\ \vdots & & \iddots & & \vdots \\ 0 & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{pmatrix} + k e_1 x^T, \tag{1}$$

where e_1 is the unit vector. For A, $k = 1$ and $x = (1, 2, \cdots, S-1)^T$, and for A^{-1}, $k = 1/S$ and $x = (S-1, \cdots, 2, 1)^T$. See Supplementary Appendix 2 for the proof.

Not all amplifications are the result of a single copy gain at each event. For example, if two copies are adjacent to one another, a further duplication event can replicate both copies simultaneously. We can consider that at each event a random set of the existing copies are chosen to be duplicated. In this setting, we have explored the possible histories via simulation (see Supplementary Appendix 6 for details). Generally, the histories that result in identifiable models are a small proportion of all models simulated, though the histories generated by our simulations may not be representative of likely biological scenarios. There is no obvious condition that appears to guarantee identifiability, but the simulations continue to suggest that identifiability is loosely a property of having a concentration of duplications along a small number of lineages.

2.3 Modeling sequencing variability

With sequencing data, we will not observe the true allele frequencies P_i, but rather the m_i sequenced fragments that overlap the location i, of

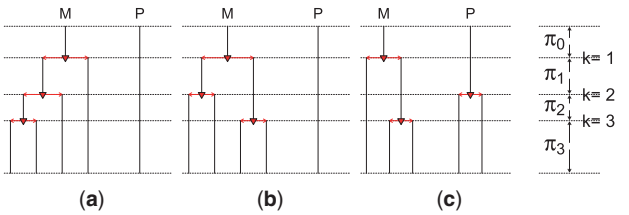

Fig. 1. Example of three histories of amplification that can lead to copy number $S = 5$: starting at the top with normal copy-number state of one allele from the maternal (M) and one from the paternal (P), at each time point k there is an amplification of an existing copy resulting in five copies at the bottom, which represents the point at which the tumor sample was removed. Only (**a**) is identifiable because all amplification occurs on one lineage

which X_i of them are mutated. Greenman *et al.* (2012) estimate P_i by first classifying each mutated location as one of the possible $\{a_j\}$ based on which a_j results in the largest likelihood, and then by finding the MLE of the model given in Section 2.1 using the estimated $\hat{P}_i$ as the true P_i. This will ignore variability in the estimates of P_i, which can have an important impact if m_i is moderate. For example, normal contamination will result in a set of alleles that encompass a smaller range, making it harder to infer the true P_i from the X_i without greater sequencing depth.

We propose directly taking into account the variability in the X_i by modeling the distribution of X_i. These adjustments allow us to reliably include mutations with lower values of m_i, increasing the total number of mutations and thus the power. This distribution allows us to account for sequencing error, as well as the fact that we only consider locations where $X_i > 0$. However, in practice this will also make little difference in most settings unless the coverage is very low.

We model X_i as Binomial($m_i, \tilde{P}_i$), where $\tilde{P}_i$ is the expected allele frequency after accounting for normal contamination and sequencing error (see Supplementary Appendix 3). In general the sequencing error will be fairly small (1–2%), and will only affect small values of P_i (<0.2), for example, if there is a large amount of normal contamination or large copy number. A larger concern are sub-clonal populations, where some of the mutations, or even the entire region, are not variant in all tumor cells. Mild levels of sub-clonal populations that do not contain the region will not necessarily affect the results badly if there is a large read depth and the number of events K is small, but in more difficult cases can severely bias the results, see Supplementary Appendix 3.

2.4 Estimating π from tumors

We can then estimate $\hat{\pi}$ using maximum-likelihood techniques; unlike Greenman *et al.* (2012), we expand our likelihood to include the sequencing variability and sequencing error described earlier in text. For large amount of sequencing depth, there is likely to be little difference in the two methods, but for lower levels of sequencing, explicitly accounting for the sequencing variability brings improved stability.

We assume in what follows that A defines an identifiable model. As $q \propto A\pi$, and both q and π must sum to 1, q lies in a constrained set Ω. If A is square, we can write this as $\Omega = \{q : A^{-1}q \pm 0, 1^T q = 1.\}$. Then by the invariance property of the MLE it is sufficient to find the MLE of q, with the constraint that $\hat{q} \in \Omega$ and then solve for $\hat{\pi}$. We use an EM algorithm to estimate $\hat{q}$ from the data $X_1, \ldots, X_N$, where the M-step involves a constrained maximization (see Supplementary Appendix 4 for details).

The most important factor in the ability to estimate the timing of events is the true value of the π vector. The allele frequencies of mutated locations follow a multinomial distribution with the number of categories equal to the number of alleles. When one of the alleles has a low probability of occurrence, as given by the parameter q, then the estimates of π become more unstable. Mutations acquired in every mutational stage contribute to the allele frequency $1/S$ (or its corresponding allele frequency after adjusting for normal contamination and sequencing error). The corresponding element of q, notated as q_1, absorbs much of the probability; indeed, when the history is sequential amplification, it is easy to show that q_1 is guaranteed to be largest element of the vector q, meaning that the most likely allele frequency must always be $1/S$. As probabilities in q are far from balanced regardless of the value of π, then when π has small values the probabilities in q are even smaller; this will lead to instability in the estimates. Furthermore, as the total copy number of a region grows, the number of possible alleles does as well, making the estimation problem even more difficult. To observe all the possible alleles with a high probability requires a large number of mutations N as the copy number grows or a value of π that is small (see Supplementary Fig. S1).

Some of the most important events to time accurately are those that are early (with small π_0), and therefore to counteract this instability, we introduce a Bayesian model for estimating π. Specifically, we assume that π follows a Dirichlet distribution that puts uniform probability on the K-simplex where π lies. This is commonly done in the case of simple multinomial estimation, where a Dirichlet prior is equivalent to adding pseudo-counts to the data to stabilize the estimates. The Dirichlet distribution is not a conjugate prior for our distribution, and therefore we sample from the posterior distribution of π using sampling importance resampling to calculate the posterior mean and credible intervals (see Supplementary Appendix 5 for details).

3 RESULTS

3.1 Simulation data

We simulated mutation data for different histories using the model described earlier in text: calculating the q vector of multinomial probabilities for a given π and A, generating P_i from a $Multinomial(N, q)$, and then generating mutation counts using a $Binomial(m, \tilde{P}_i)$. We varied the parameter π, the read depth, numbers of mutations N and the normal contamination to evaluate the performance of our estimation procedures; no sequencing error was comprehensively simulated because the effect was so small. In all situations, estimation of π_0 when $\pi_0 < 0.1$ is highly variable with few mutations, and furthermore the MLE is a biased estimate, underestimating π_0, until N becomes large. Even in the simplest example of $K = 1$ (CNLOH or single gain), if $\pi_0 = 0.01$ values of N as high as 200 or 300 are needed to remove this bias (Supplementary Fig. S2). For $\pi_0 = 0.01$, the MLE estimate $\hat{\pi}_0$ equals zero for almost all simulations, reflecting the low probability of observing an allele identified with the earliest stage.

For larger values of π_0, the estimates are unbiased starting around $N = 50$, with continually greater precision for larger N (Supplementary Fig. S4). The read depth has much less effect on the estimation, particularly if the read depth is >30; even for read depth as low as 10, the loss of precision due to low read depth is not nearly as striking as that due to reducing the number of mutations (assuming that the mutations are correctly identified as mutated, which is problematic with only 10× coverage). This implies that including more mutations with lower read depth will increase N and lead to greater precision in the estimate of π despite the lower precision of each individual location. We also note that the 95% bootstrap confidence intervals are slightly biased, with coverage probability somewhat <95% even with large N (Supplementary Fig. S6).

Therefore, in evaluating our proposed procedures, we concentrate on two different contexts, $\pi_0 \leq 0.1$ and $\pi_0 > 0.1$, and assume that we have at least 50 mutations in a region. We focus on the estimation of π_0 as being of the greatest biological interest.

Full Maximum Likelihood We expect that the difference between the partial MLE method of Greenman *et al.* (2012) and our full MLE method will be the largest when the question of classifying mutated locations to a particular allele frequency has the greatest uncertainty: lower read coverage and/or higher levels of normal contamination. Simulation results show that with no normal contamination, the partial MLE method can be biased even in the relatively simple case of the single-gain case with read coverage as high as 30× (Fig. 2). By 75× coverage the two methods are indistinguishable for low numbers of events, but for larger K, the partial MLE still remains biased even with 75× coverage, see Supplementary Figure S8. In particular, the

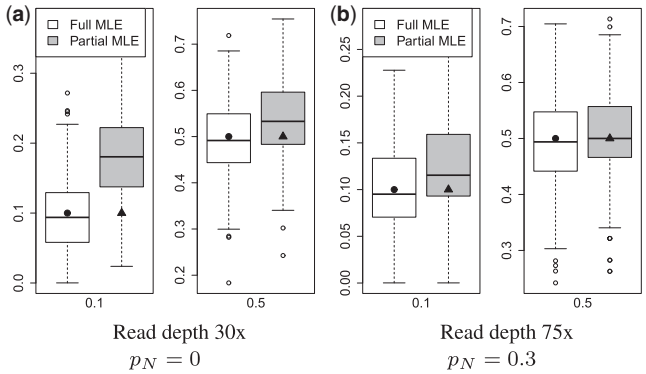

Fig. 2. Boxplots of $\hat{\pi}_0$ based on simulated data for values of $\pi_0 = 0.1, 0.5$ in the single-gain case. (**a**) Read depth of 30× and no normal contamination. (**b**) Read depth of 75× and 30% normal contamination. The number of mutations, N, was fixed at 125

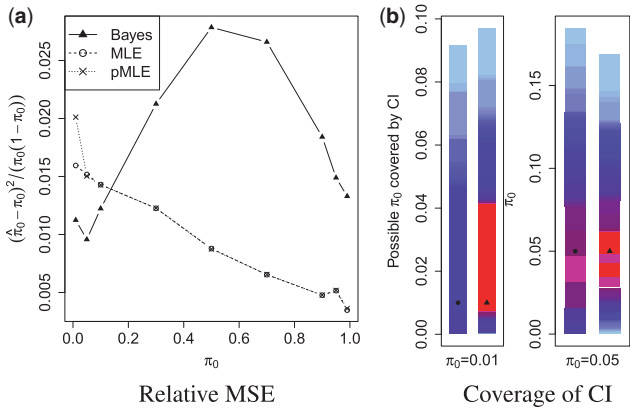

Fig. 3. Comparison of Bayesian and MLE estimates for CNLOH, see Supplementary Figures S10 and S11 for the single-gain case. (**a**) Plots of relative mean squared error on simulated data of CNLOH for three different estimates. Relative MSE is the MSE scaled by the value of $\pi_0(1-\pi_0)$ to reflect the size of the MSE relative to the size of π_0. (**b**) Comparison of CI coverage for MLE and Bayesian on simulated data. For each possible π_0, CI coverage was calculated as the percentage of CIs from simulated data that covered π_0; a color scale indicates the CI coverage, with red indicated $\geq 95\%$ coverage and magenta indicated $90-95\%$ coverage. The solid points indicate the true value (circle for MLE and triangle for Bayesian). Shown are the results for when the true π_0 is small (0.01, 0.05) and the Bayesian estimates are not extremely biased; see Supplementary Figure S12 for all values of π_0

partial MLE method overestimates π_0 for small π_0, and conversely for large π_0. Even where the full MLE tends to be biased and underestimates π_0, the partial MLE goes the other direction and overestimates π_0 by a larger margin (and conversely for large π_0), resulting in worse average error, Figure 3. When normal contamination increases, the partial MLE does worse, so that even for 75× coverage and K = 1, estimation of moderately low values of π_0 (e.g. $\pi_0 = 0.1$) is noticeably still biased (Figs 2 and 3). For large K, where the allele frequencies are closer together and harder to distinguish, the problems are magnified across a wide spectrum of π_0 and larger coverage is required before the bias disappears, see Supplementary Figures S6.

We note that the difference between the partial and full estimation methods also depends on the complexity of the problem. For CNLOH, where there is a direct identification between allele frequency and the stage in which the mutation occurred, there is less difference in the methods—only when normal contamination reaches ~0.3 is there a difference if the read depth is 30×. This is likely due to the fact that with CNLOH there is no constraint on the space in which the vector q lies. With more complex models (i.e. larger K), small variations in the estimation of q result in larger perturbations of the vector π (see Supplementary Fig. S9).

Bayesian Estimation From a frequentist perspective, Bayesian estimates will be on average biased, but can offer less variability and thus less overall error. Simulations definitely reflect this bias, with Bayesian estimates on average underestimating π_0 across the board for CNLOH regions and generally for single-gains shrinking the estimates toward $\pi_0 = 0.5$.

The Bayesian estimates for gains generally are similar to that of the MLE, with the overall error similar for most values of π_0. For extreme π_0 ($\pi_0 = 0.01$ or 0.99), the Bayesian estimates have worse overall error than the MLE, so that they do not improve those estimates as was hoped (Supplementary Fig. S10). However, the Bayesian interval estimates have a better coverage probability, particularly in extreme values of π_0, than the bootstrap CI. For CNLOH, however, the Bayesian estimates have a different behavior than the MLE. For small π_0 (≤ 0.1), the Bayesian estimates have a better error rate as well as better coverage probability; indeed for extreme π_0, the MLE bootstrap confidence intervals are bad, often giving extremely small or zero-width intervals. The Bayesian estimates have a much worse error for $\pi_0 > 0.1$ because they are extremely biased downward (Fig. 3).

3.2 Cancer data

We use the exome sequencing data from six of the eight tumors that we previously analyzed in Durinck *et al.* (2011). In that work, we analyzed only CNLOH events and found that the CNLOH event on chromosome 17 covering the tumor suppressor gene TP53 to be an early event. Here we analyze both CNLOH and single copy gains, and compare the performance of estimates of $\hat{\pi}$ (no events with S=4 were observed). We note that for these tumors, there is a high mutation rate, and many CNLOH and single-gain abnormalities, with few higher level amplifications. We also evaluated the timing of regions for five ovarian tumors with WGS available through the TCGA project (Cancer Genome Atlas Research Network, 2011). The ovarian tumors have many more rearrangements than the skin cancers and a much lower mutation rate. See Supplementary Appendix 1 and Supplementary Table S3 for more details regarding the datasets.

We first observe that the estimate of π_0 for the two skin tumors containing a CNLOH event on chromosome 17 continues to imply 17p CNLOH is an early event (Fig. 4a), even after the addition of the single-gain regions, with estimates of π_0 on the order of 0.05 [additional CNLOH events were found by Durinck *et al.* (2011) in samples that we did not analyze, see Supplementary Appendix 1]. CNLOH events involving the region containing TP53 are present in four of the five ovarian samples, and TP53 mutations are found in all four of these regions. Three of these TP53 mutations clearly occurred before the CNLOH event (i.e. homozygous) with the remaining mutation

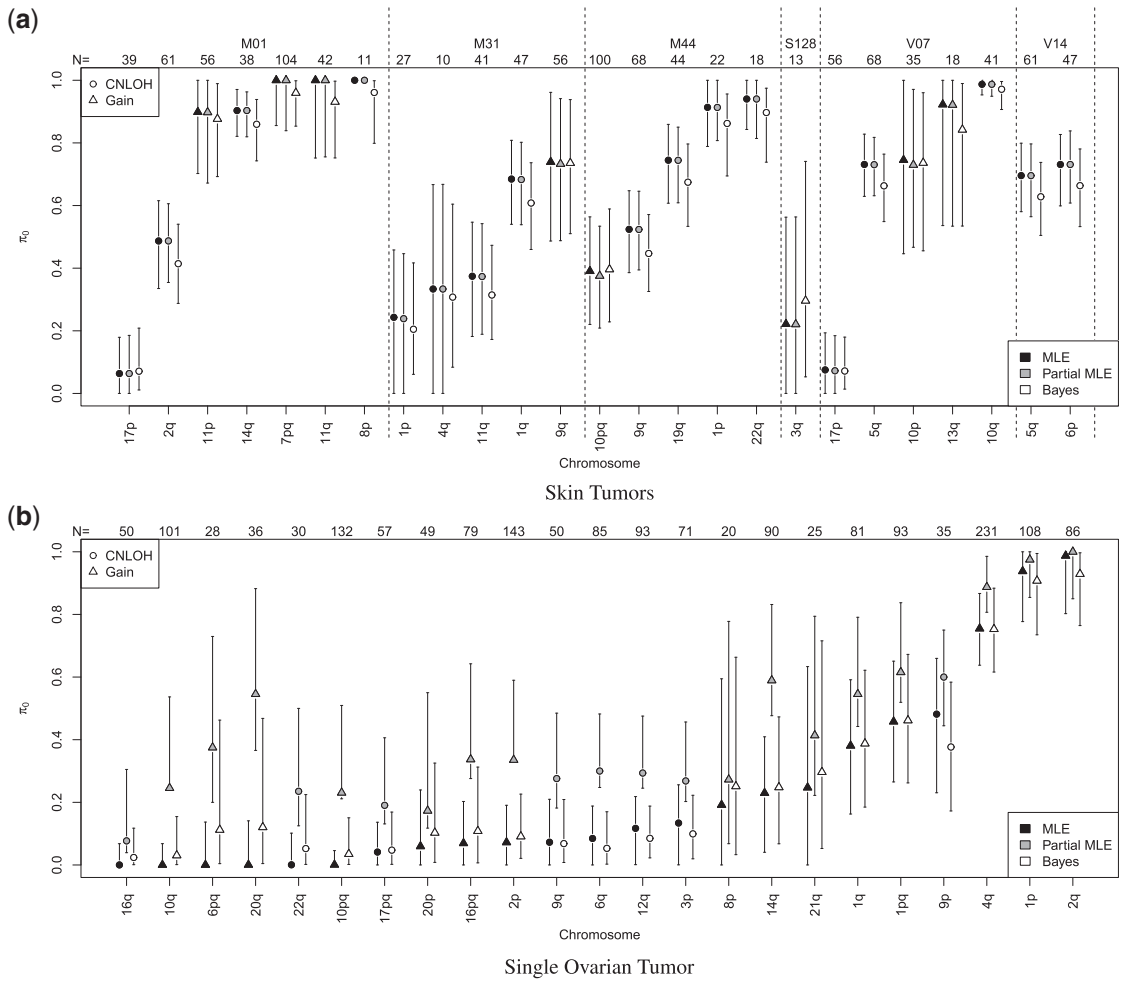

Fig. 4. Estimates of π_0 with bootstrap confidence intervals for the (**a**) skin tumors and (**b**) a single ovarian tumor (13–1411). The full MLE (black symbol), partial MLE (gray symbol) and Bayesian estimates (white symbol) with their corresponding confidence intervals are shown side by side for each region. Regions of CNLOH are marked with a circle, and those from a single gain are marked with a triangle. Regions from different samples are separated by dashed lines. Each region is labeled by the chromosome and the arm that contain the region. Above each region, the number of mutations (N) identified in the region is indicated. The CNLOH of 17p is shown in red

ambiguous (see Supplementary Table 4 and Supplementary Fig. S15). However, only one sample (13–1411) has an estimate of π_0 for the region containing TP53 that is as early in the life of the tumor as seen in the skin cancers.

The large number of aberrations present in ovarian tumors allows us to observe the general trajectory of chromosomal amplifications (Supplemental Fig. S15). The five tumors present different general profiles. Two tumors (13-0890 and 13-1411) have events that are clearly separated through time and span the range of (relative) time, whereas events from other three tumors are estimated to have occurred over a small range of time, suggesting a short duration of rapid copy-number change. This suggests the possibility of two different biological mechanisms in use, with some tumors starting with a whole-genome duplication whereas other tumors steadily accumulate copy-number changes.

Comparison with the partial MLE We have seen in simulations that the partial MLE implemented in Greenman *et al.* (2012) tends to overestimate π_0, particularly for small values of π_0. For the skin data, the estimates of π_0 for the two early events on chromosome 17p given by the partial MLE method are much larger (Fig. 4a). For sample M01, the 95% confidence intervals based on the partial MLE for the CNLOH of chromosome 17p overlaps that of chromosome 2, making it ambiguous whether 17p is the first event. The early events in the ovarian tumors show an even larger difference between the two estimates. For these early events of interest, the difference in estimation can be important and accounting for sequencing variability identifies early events more conclusively.

Aside from the early events, we see that the estimates for the full and partial MLE methods are generally equivalent for the skin tumors. For the ovarian tumors, however, the differences are more striking even for events that have only moderate estimates of π_0, see Figure 4b. This is due to the fact that the ovarian samples are sequenced at $\sim 35\times$ coverage compared with $100\times$ coverage for the skin tumors (Supplementary Table S3); furthermore, gains are more heavily represented for the ovarian tumors, and the gains show much greater differences between the estimates.

Bayesian Estimation For regions where the MLEs are approaching the boundary of the parameter space (estimates that are essentially 0 or 1) the Bayesian estimates, as expected, shrink the estimates away from the boundary and toward the prior mean. Particularly for CNLOH regions, the bootstrap confidence intervals for the MLE estimates often do not sufficiently capture the variation in the estimates. One example is the CNLOH event in 8pq in the skin tumor M01, where only 11 mutations are observed, but all of them are heterozygous. Bootstrap confidence intervals give great confidence to the parameter estimate of $\pi_0 = 1$ even though $N = 11$; the Bayesian analysis, as hoped, modulates these estimates, both decreasing the estimate in accordance with the prior distribution and giving greater levels of uncertainty corresponding to the small sample size.

The Bayesian estimates, as seen in simulation, also generally give lower estimates of all π_0 for CNLOH and estimates closer to 0.5 for extreme values of π_0 for single gains, which is also reflected in both cancer datasets. However, the difference in estimates is not large relative to the confidence intervals of the estimates, and is probably offset by the advantage of increased accuracy of the confidence intervals, particularly for early events.

4 CONCLUSION

Precise timing of chromosomal abnormalities provides a wonderfully detailed glimpse of the etiology of a single tumor. However, we have demonstrated that there are limitations to this technique. In particular, we have shown that for high-level amplifications, most of the possible combinations of events that result in large amounts of amplification will not retain enough information in the allele frequencies to be able to estimate the ordering. Only regions where the amplification follows one single lineage can be timed using this model. This may result in a biased impression of the etiology, as this type of amplification may be predominant for the promotion of certain types of abnormalities and may miss many other types of oncogenes.

As we note in the introduction, our focus has been the traditional one of copy-number analysis, where each region in the normal genome is analyzed separately as to its behavior in the tumor. With exome sequencing, this traditional viewpoint is still the only one available. With whole-genome sequencing, as we noted, one can analyze the relationships between the regions and order the events using the information from other regions. In this case, a single region A can share an event with another region B if the amplification brought the two into proximity to each other through an insertion. Then there are additional constraints on estimating π_A and π_B jointly, as that event must occur at the same moment for both. This implies that with reasonably deep whole-genome sequencing such that these relationships are reliably determined, there will be a larger percentage of histories that are identifiable.

In addition, early events, which are of particular biological significance, are sensitive to estimation procedures and large numbers of mutations are necessary to be able have stable estimates of the time of occurrence. Of even greater difficulty is the ordering of a collection of early events. Even with whole-genome sequencing, some regions will not have the hundred or more mutations that our simulations show are necessary to distinguish early events, particularly in tumors with low mutation rates.

However, we have also shown that differences in estimation techniques can help provide better estimates and confidence intervals for temporal estimates. We have introduced a full MLE to handle sequencing variability due to lower read coverage, as well as a Bayesian estimation technique. We have shown the full MLE can provide improvement with read depths as large as 30×, and even up to 75× or higher if there is normal contamination or early events. The Bayesian estimates have a varying performance for different values of the parameter space, but can provide increased stability, particularly in their estimates of confidence intervals for the estimates.

Ultimately, the ability to successfully estimate π also relies on intrinsic properties of the cancer. In the skin tumors, only half of the samples had CNLOH over the tumor suppressor gene TP53 (not all of which were examined in this work); in the other samples, both copies of TP53 were also inactivated but through multiple mutations, not a chromosomal abnormality. Other important regions may be too small in a particular sample to have sufficient mutations—the regions we ordered were large, sometimes entire chromosomal arms. Some tumors, such as the ovarian, have low mutation rates so that even with whole-genome sequencing many regions will have few mutations or not enough to confidently distinguish between events. While 30–60% of the abnormal regions could theoretically be timed in our sample, the percentage that had enough mutations was generally 20–30%. Therefore, timing of the chromosomal abnormalities of a single sample remains extremely fragmentary, and an insight into tumor etiology will still ultimately be gained by comparing the temporal ordering of many tumors.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviews for their helpful suggestions. The results published here are in whole or part based on data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions that constitute the TCGA research network can be found at http://cancergenome.nih.gov/.

Funding: National Institute of Health TCGA grant (U24 CA143799); the Anna Fuller fund; the Dermatology Foundation; the American Skin Association; and National Science Foundation grants (DMS-0636667, DMS-1026441).

Conflict of Interest: none declared.

REFERENCES

Attolini,C. *et al.* (2010) A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl Acad. Sci. USA*, **107**, 17604–17609.

Beerenwinkel,N. *et al.* (2005a) Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.*, **12**, 584–598.

Beerenwinkel,N. *et al.* (2005b) Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, **21**, 2106–2107.

Beerenwinkel,N. *et al.* (2006) Evolution on distributive lattices. *J. Theor. Biol.*, **242**, 409–420.

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl Acad. Sci. USA*, **104**, 20007–20012.

Bilke,S. (2005) Inferring a tumor progression model for neuroblastoma from genomic data. *J. Clin. Oncol.*, **23**, 7322–7331.

Brodeur,G. *et al.* (1982) Statistical analysis of cytogenetic abnormalities in human cancer cells. *Cancer Genet. Cytogenet.*, **7**, 137–152.

Campbell,P.J. *et al.* (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. *Proc. Natl Acad. Sci. USA*, **105**, 13081–13086.

Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.

Cancer Genome Atlas Research Network. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Desper,R. *et al.* (2000) Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.*, **7**, 789–803.

Durinck,S. *et al.* (2011) Temporal dissection of tumorigenesis in primary cancers. *Cancer Discov.*, **1**, 137–143.

Fearon,E. and Vogelstein,B. (1990) A genetic model for colorectal tumorigenesis. *Cell*, **61**, 759–767.

Frumkin,D. *et al.* (2008) Cell lineage analysis of a mouse tumor. *Cancer Res.*, **68**, 5924–5931.

Gerlinger,M. *et al.* (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Eng. J. Med.*, **366**, 883–892.

Gerstung,M. *et al.* (2009) Quantifying cancer progression with conjunctive Bayesian networks. *Bioinformatics*, **25**, 2809–2815.

Greenman,C.D. *et al.* (2012) Estimation of rearrangement phylogeny for cancer genomes. *Genome Res.*, **22**, 346–361.

Hjelm,M. *et al.* (2006) New probabilistic network models and algorithms for oncogenesis. *J. Comput. Biol.*, **13**, 853–865.

Huang,H. *et al.* (2007) Bayesian analysis of frequency of allelic loss data. *J. Am. Stat. Assoc.*, **102**, 1245–1253.

Liu,J. *et al.* (2009) Inferring progression models for CGH data. *Bioinformatics*, **25**, 2208–2215.

Navin,N.E. and Hicks,J. (2010) Tracing the tumor lineage. *Mol. Oncol.*, **4**, 267–283.

Newton,M.A. (2002) Discovering combinations of genomic aberrations associated with cancer. *J. Am. Stat. Assoc.*, **97**, 931–942.

Newton,M.A. and Lee,Y. (2000) Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics*, **56**, 1088–1097.

Newton,M.A. *et al.* (1994) Assessing the significance of chromosome-loss data: where are suppressor genes for bladder cancer? *Stat. Med.*, **13**, 839–858.

Newton,M. *et al.* (1998) On the statistical analysis of allelic-loss data. *Stat. Med.*, **17**, 1425–1445.

Nik-Zainal,S. *et al.* (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.

Nishizaki,T. *et al.* (1997) Genetic alterations in primary breast cancers and their metastases: direct comparison using modified comparative genomic hybridization. *Genes Chromosomes Cancer*, **19**, 267–272.

Rahnenführer,J. *et al.* (2005) Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, **21**, 2438–2446.

Sasatomi,E. *et al.* (2002) Comparison of accumulated allele loss between primary tumor and lymph node metastasis in stage II non-small cell lung carcinoma: implications for the timing of lymph node metastasis and prognostic value. *Cancer Res.*, **62**, 2681–2689.

Siegmund,K.D. *et al.* (2009) Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. *Proc. Natl Acad. Sci. USA*, **106**, 4828–4833.

Simon,R. *et al.* (2000) Chromosome abnormalities in ovarian adenocarcinoma: III. Using breakpoint data to infer and test mathematical models for oncogenesis. *GGenes Chromosomes Cancer*, **28**, 106–120.

Sprouffske,K. *et al.* (2011) Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prev. Res.*, **4**, 1135–1144.

Taylor,B.S. *et al.* (2008) Functional copy-number alterations in cancer. *PLoS One*, **3**, e3179.