BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 15 2014, pages 2121–2129
doi:10.1093/bioinformatics/btu174

Genome analysis

Advance Access publication April 2, 2014

Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity

Yi Li[1] and Xiaohui Xie[1,2,3,*]

[1]Department of Computer Science, [2]Institute for Genomics and Bioinformatics and [3]Center for Machine Learning and Intelligent Systems, University of California, Irvine, CA 92697, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Next-generation sequencing (NGS) has revolutionized the study of cancer genomes. However, the reads obtained from NGS of tumor samples often consist of a mixture of normal and tumor cells, which themselves can be of multiple clonal types. A prominent problem in the analysis of cancer genome sequencing data is deconvolving the mixture to identify the reads associated with tumor cells or a particular subclone of tumor cells. Solving the problem is, however, challenging because of the so-called ‘identifiability problem’, where different combinations of tumor purity and ploidy often explain the sequencing data equally well.

Results: We propose a new model to resolve the identifiability problem by integrating two types of sequencing information—somatic copy number alterations and loss of heterozygosity—within a unified probabilistic framework. We derive algorithms to solve our model, and implement them in a software package called PyLOH. We benchmark the performance of PyLOH using both simulated data and 12 breast cancer sequencing datasets and show that PyLOH outperforms existing methods in disambiguating the identifiability problem and estimating tumor purity.

Availability and implementation: The PyLOH package is written in Python and is publicly available at https://github.com/uci-cbcl/PyLOH.

Contact: xhx@ics.uci.edu

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

Received on November 7, 2013; revised on March 23, 2014; accepted on March 27, 2014

1 INTRODUCTION

The advent of next-generation sequencing (NGS) and launch of comprehensive cancer genome sequencing projects (Collins and Barker, 2007; Hudson *et al*., 2010) have yielded an unprecedented view on the complex landscape of cancer genomes, leading to the discovery of new cancer-causing genes and pathways, and novel therapeutic targets for treating cancers. Analyzing the data from cancer genome sequencing remains, however, computationally challenging because of the shear size of the sequencing data and the complexity of the tumor genomes and samples.

Cancer genomes are often characterized by wide-spread somatic copy number alterations (CNAs), where genomic segments are deleted or duplicated one or more times. Identifying somatic CNAs associated with specific tumor genomes is of long-standing interest in the study of cancer genomes and is one of the focal points of the cancer genome analysis. Many computational methods have been proposed to discover copy number changes directly from DNA microarrays (Bignell *et al*., 2004; Lindblad-Toh *et al*., 2000; Mei *et al*., 2000; Pinkel *et al*., 1998; Zhao *et al*., 2004) or sequencing data (Campbell *et al*., 2008; Chiang *et al*., 2008). However, most of these methods aim at identifying the relative copy numbers of segments of the same tumor genome. Discovering copy numbers in an absolute scale is biologically more relevant (Carter *et al*., 2012) but more challenging. This is because of the fact that the absolute copy number changes can be affected by two confounding factors: (i) tumor purity, the fraction of all cancerous cells within a heterogeneous tumor sample, and (ii) tumor ploidy, the baseline copy number of genomic segments or entire chromosomes (Carter *et al*., 2012; Oesper *et al*., 2013), both of which are unknown and themselves need to be estimated to infer absolute copy number changes. It is possible to estimate tumor purity and ploidy using experimental techniques such as quantitative image analysis (Yuan *et al*., 2012) and single-cell sequencing (Navin *et al*., 2011); however, these techniques are still too expensive or time-consuming to support large-scale studies. Hence, it is of great interest to use computational approaches to estimate tumor purity and ploidy, and consequently absolute copy number changes, directly from the NGS data.

Tumor purity and ploidy affect not only copy number changes in different segments of genomes but also the distribution of allele frequencies in these segments. In the NGS data, these two types of information can be summarized in terms of the total number of reads mapped to each segment (total read count) and the frequencies of reads matching B-alleles (B-allele frequencies) at different sites. Computational methods have been proposed to estimate tumor purity alone (Larson and Fridley, 2013; Su *et al*., 2012) or jointly with tumor ploidy (Carter *et al*., 2012; Gusnanto *et al*., 2012; Oesper *et al*., 2013) based on these two types of information extracted from NGS data.

Depending on how copy number changes and B-allele frequency information are used, the existing methods can be roughly grouped into two categories: one category of methods use B-allele frequencies (BAFs) at somatic mutation sites to estimate tumor purity, including PurityEst (Su *et al*., 2012) and PurBayes (Larson and Fridley, 2013). These methods leverage the fact that the BAFs at somatic mutation sites are expected to be ~0.5 if the tumor purity is 100%, and any addition of normal cells will lead to a reduction in the observed BAFs at these sites. The second category of methods relies on copy number changes

*To whom correspondence should be addressed.

to estimate tumor purity and/or ploidy, including CNAnorm (Gusnanto *et al.*, 2012), THetA (Oesper *et al.*, 2013) and ABSOLUTE (Carter *et al.*, 2012). It has been shown that the methods in the second category are often more accurate and robust than those in the first category because of the fact that (i) the total read counts are large in NGS data, and thus, methods relying on copy number changes are statistically more stable than methods relying on BAFs at somatic mutation sites, the number of which is often small, and (ii) the determination of somatic mutations is not perfect and the inclusion of false-positive findings can significantly bias the estimation (Koboldt *et al.*, 2012; Oesper *et al.*, 2013; Roberts *et al.*, 2013).

However, the utility of the methods relying on copy number changes to estimate tumor purity and ploidy is severely hindered by the so-called 'identifiability problem', where different combinations of tumor purity and ploidy can explain the observed data equally well (Carter *et al.*, 2012; Oesper *et al.*, 2013; Reiersøl, 1950). This is because tumor purity and ploidy are often intertwined—changes in one can be offset by compensations from the other, allowing the same copy number to be explained by multiple combinations of tumor purity and ploidy. For example, a homozygous deletion combined with 30% tumor purity can also be explained as a heterozygous deletion combined with 60% tumor purity. Resolving this ambiguity is key to accurate estimation of tumor purity and ploidy. Existing methods try to solve this identifiability problem by using heuristics, e.g. favoring solutions that have the smallest deviations from diploid (e.g. CNAnorm; Gusnanto *et al.*, 2012), seeking additional experiential data (e.g. ABSOLUTE; Carter *et al.*, 2012) or simply outputting all possible solutions (e.g. THetA; Oesper *et al.*, 2013).

Here we provide a more principled way to solve the identifiability problem by combining the information revealed from copy number changes and B-allele frequencies. Instead of using B-allele frequencies extracted from somatic mutation sites as in the previous cases, we use B-allele frequencies calculated at sites that are heterozygous with respect to the normal genomes, and most of which are common single nucleotide polymorphisms (SNPs). These heterozygous sites are much more abundant (Sachidanandam *et al.*, 2001) and easier to identify, leading to more statistically stable results. Copy number changes in the cancer genome often result in loss of heterozygosity (LOH) at these heterozygous sites, and the extent of LOH is closely related to absolute instead of relative copy number changes. We will use BAFs to gauge the extent of LOH, and provide information on the absolute copy number changes by examining the patterns of BAFs at the heterozygous sites within the same genomic segment. For example, although a homozygous deletion with 30% tumor purity results in the same copy number as a heterozygous deletion with 60% tumor purity in a tumor sample, B-allele frequencies at heterozygous sites of the tumor sample cluster at different values in the two combinations and therefore are able to distinguish these two cases. Based on this insight, we propose a full probabilistic model implemented as a software package called PyLOH to integrate the information gathered from CNAs and LOH. Estimations of tumor purity and absolute copy numbers are then formulated as an optimization problem in which we choose those values that maximally explain both total read counts and B-allele frequency information.

Our method is similar in spirit to some of the earlier methods proposed for SNP array analysis, where both the signal intensity and BAF of each SNP are used in estimating copy number changes. The combination of these two signals has been shown to improve the estimation accuracy of tumor ploidy (Greenman *et al.*, 2010), or both tumor purity and ploidy (Rasmussen *et al.*, 2011; Van Loo *et al.*, 2010; Yau *et al.*, 2010). Recently, some of these methods have been extended to sequencing data, including OncoSNP-SEQ by Yau (2013) and Patchwork by Mayrhofer *et al.* (2013). However, the OncoSNP-SEQ algorithm uses only the reads mapped to the SNP sites, although our algorithm uses all reads, and thus should be able to yield a more accurate estimation of copy number changes. Similar to our work, the Patchwork algorithm also uses all reads, but it requires manual interpretation through data visualization to determine the initial copy numbers of clusters of genomic segments, which could be useful when the tumor genome is too complex for the algorithms to resolve different solutions by themselves. Here we seek an alterative approach that is based on a generative model and requires no manual intervention. In addition, the Patchwork algorithm requires the existence of copy-neutral loss of heterozygosity within the tumor genome to run the algorithm, whereas our algorithm has no such constraints.

The outline of this article is as follows: in Section 2, we describe the full probabilistic model of PyLOH. In Section 3, we first present cluster patterns of BAFs in NGS data of paired tumor-normal samples, then introduce a visualization tool called 'BAF heat map' to characterize such patterns. Finally, we compare tumor purity estimates of PyLOH and other methods on both simulated datasets and 12 breast cancer sequencing datasets. Our results show that explicitly incorporating both CNAs and LOH information can resolve the identifiability problem and significantly improve the accuracy of tumor purity estimation. Finally, we discuss the limitations of PyLOH and propose future directions in Section 4.

2 METHODS

In this section, we present the probabilistic model of PyLOH that combines CNAs and LOH information to infer absolute copy numbers and tumor purity. We first introduce some notations, then propose a generative mixture model incorporating both total read counts and B-allele frequency information and finally introduce algorithms to solve the model.

2.1 Basic definitions and notations

Similar to previous work (Carter *et al.*, 2012; Oesper *et al.*, 2013), we assume the tumor genome has already been segmented into J segments, each of which has the same CNAs. Denote the copy number of the j-th segment of the tumor genome by C_j with $j = 1, \ldots, J$. In addition, we assume each segment has a number of heterozygous sites (single nucleotide changes) in the corresponding normal (i.e. control) genome. We use (i,j) to index the i-th heterozygous site in segment j with $i = 1, \ldots, I_j$, where I_j is the total number of heterozygous sites in segment j.

The observed data are summarized and grouped into two categories: one category is the copy number information, represented as the total number of reads mapped to each segment. Let D_j denote the number of reads mapped to segment j. The second category of observed data is the allele frequency information, represented by the total number of reads matching each of two alleles at a heterozygous site. For notational purpose, for each heterozygous site, we define the *A allele* to be the allele matching the reference genome and the *B allele* to be the corresponding unmatched one. Using the notation from (Roth *et al.*, 2012), let a_{ij} and b_{ij} denote the number of reads matching A and B alleles, respectively, at site (*i*,*j*). As most of the data we consider are from paired tumor–normal samples, we use a superscript N (from normal samples) and T (from tumor samples) to denote the sample origin of the data. For example, D_j^T and D_j^N will denote the total number of reads mapped to segment j from the tumor and normal samples, respectively.

To account for the contamination of normal cells, we assume the tumor sample yielding the sequence data consists of a mixture of normal and tumor cells. Denote the fraction of tumor cells within the tumor sample by ϕ, which will also be called tumor purity. Consequently, the average copy number of each segment within the tumor sample is

$$\bar{C}_j = \phi C_j + (1-\phi)2 \tag{1}$$

for $j = 1, \ldots, J$, assuming that the default copy number within normal cells is always 2. Our goal is to use both the total read count information and site-specific allele count information to infer both the absolute copy number $\{C_1, \ldots, C_J\}$ and the tumor purity ϕ.

2.2 Modeling CNAs

Following the Lander–Waterman theory (Lander and Waterman, 1988), the probability of a read originating from a specific segment depends on three main factors: (i) the copy number of the segment, (ii) the total genomic length of the segment and (iii) the mappability of the segment (depending on factors such as GC content, repetitive sequence and so on) (Oesper *et al.*, 2013). Borrowing the concept of interval weight factor from (Oesper *et al.*, 2013), we associate a coefficient θ_j to segment j accounting for the effect of its genomic length and mappability. We assume the expected number of reads mapped to segment j, denoted by λ_j, in the tumor sample is proportional to $\bar{C}_j\theta_j$. That is, given two segments a and b, we have

$$\frac{\lambda_a}{\lambda_b} = \frac{\bar{C}_a\theta_a}{\bar{C}_b\theta_b} \tag{2}$$

In Equation (2), the mapping coefficient θ_j's matters only in their relative values. For simplicity, we take $\theta_a/\theta_b = D_a^N/D_b^N$, the ratio of the mapped read counts between these two segments in the normal sample, as it reflects intrinsic sequence properties of these segments and therefore should be the same between the normal and tumor samples.

The above formula determines the relative value of the expected number of reads mapped to each segment. To further specify the absolute value of λ_j of segment j, we make use of the allele frequency information and curate a list of segments that contain no loss of heterozygosity. Where there is no loss of heterozygosity, the only possible copy numbers at these segments in tumor cells must be even numbers. From the list, we further remove 'outlier' segments whose copy numbers deviate from the bulk of the segments in the list based on the observed read counts at these segments. At the end, we are left with a set of segments (denoted by set S containing the indices of these segments) that both contain no loss of heterozygosity and likely share the same copy number. Details are given in Supplementary Material *Data preprocessing*.

The set of segments in S will be the baseline segments that we use to specify the expected read counts λ_js. To reduce complexity, we assume that the same even copy number c_s shared by all segments in S can only be either 2 or 4. (The other possible values are 0 for homozygous deletion, which is unlikely, as each segment in S is supported by a certain amount of reads, or values that are >4 for ploidy higher than tetraploid, which is likely to be rare.) Our algorithm will check both cases and select the one most compatible with the observed data (in terms of the likelihood function). Given the values of c_ss for each $s \in S$, the average copy number of these segments in the tumor sample, taking the contamination of normal cells into account, is then given by Equation (1).

With the average copy numbers in the baseline segments given, we then specify the expected read count for each segment $j = 1, \ldots, J$ in the tumor sample as follows:

$$\lambda_j = \frac{1}{|S|} \sum_{s \in S} \frac{\bar{C}_j\theta_j}{\bar{C}_s\theta_s} D_s^T \tag{3}$$

which is the average expected read count suggested by the baseline segments through Equation (2), where the observed read counts in segment s of the tumor sample are denoted as D_s^T. Here $|S|$ denotes the number of segments in set S.

Given the expected read count at each segment, we model the probability of observing D_j^T reads in segment j as a Poisson distribution with parameter λ_j,

$$D_j^T \mid C_j, \phi \sim \text{Poisson}(\lambda_j) \tag{4}$$

for each $j = 1, \ldots, J$, where λ_j is a parameter depending on the absolute copy numbers and is calculated based on Equation (3). More discussion about using the Poisson distribution is given in Supplementary Material.

2.3 Modeling loss of heterozygosity

To model the loss of heterozygosity at heterozygous sites (i.e. with genotype AB in the normal cells), we need to consider the genotypes of these sites in tumor cells. Let

$$\mathcal{G} = \{\emptyset, \text{A}, \text{B}, \text{AA}, \text{AB}, \text{BB}, \text{AAB}, \text{ABB}, \text{AABB}\}$$

be the set of possible genotypes that we will consider at each heterozygous site in tumor cells. By focusing on this set, we have excluded some other genotypes that are less likely to occur in tumor cells. For instance, we will not consider genotypes AAA or BBB, as any copy number change from AB to these two genotypes will involve at least one deletion and two insertions. Instead, all genotypes included in $\mathcal{G}$ can be derived from AB with a minimum of one operation on each allele. Although we formulated the set of possible genotypes here by assuming the maximum copy number of each allele is 2 in tumor cells, PyLOH allows the user to change this value. However, there is a trade-off in choosing the maximum copy number threshold. On one hand, increasing the threshold can accommodate genomes with high instability, but on the other hand, it can also significantly increase the complexity of the model and thus make it more susceptible to overfitting.

The corresponding copy number and BAFs associated with each genotype in $\mathcal{G}$ are $\{0,1,1,2,2,2,3,3,4\}$ and $\{\frac{1}{2}, \epsilon, 1-\epsilon, \epsilon, \frac{1}{2}, 1-\epsilon, \frac{1}{3}, \frac{2}{3}, \frac{1}{2}\}$, respectively, written in the same order as the genotypes in set $\mathcal{G}$. We have included a small $\epsilon \ll 1$ in the calculation of BAF to account for sequencing and/or read-mapping biases or errors. In practice, we choose $\epsilon = 0.01$, corresponding to a Phred quality score of 20 (Ewing and Green, 1998). We will use n_g and μ_g to denote the corresponding copy number and BAF, respectively, for genotype g.

As the tumor sample consists of a mixture of normal and tumor cells, the fraction of B alleles in the tumor sample is the weighted average of BAFs between normal and tumor cells, with weights depending on tumor purity ϕ and copy numbers,

$$\bar{\mu}_g = \frac{\phi n_g \mu_g + (1-\phi)2\mu_0}{\phi n_g + (1-\phi)2} \tag{5}$$

where $\mu_0 = 0.5$ is the BAF at the heterozygous sites in normal cells.

Using the notation from (Roth *et al.*, 2012), let G_{ij} be a random variable denoting the genotype of site (*i*,*j*) in tumor cells. Conditional on its genotype, we model the probability of the B allele count at each site as a binomial distribution, that is, given $d_{ij}^T = a_{ij}^T + b_{ij}^T$ reads mapped to site (*i*,*j*), the chance of observing b_{ij}^T reads matching B allele is given by the equation

$$b_{ij}^T \mid G_{ij} = g, \phi \sim \text{Binomial}(d_{ij}^T,\ \bar{\mu}_g) \tag{6}$$

with the total number of trials specified by d_{ij}^T and the chance of success at each trial specified by $\bar{\mu}_g$.

2.4 Combining CNAs and LOH information

For heterozygous sites located within the same segment, their genotypes are constrained by the underlying copy number associated with the segment. We model this constraint through a conditional probability distribution $P(G_{ij} = g|C_j = c) = Q_{gc}$ for all *i* and *j*. Here Q_{gc} is a predefined matrix specifying the chance of a site being genotype *g* conditional on the underlying copy number being *c*. In practice, we assign a small probability σ to any genotypes incompatible with the copy number *c* conditional on the heterozygosity in normal cells, and equal probabilities to other compatible genotypes.

Conditional on the underlying copy number, we can then write down the probability of observing B-allele read count at each site as follows:

$$\text{P}(b_{ij}^T|C_j = c, \phi) \ = \ \sum_{g \in \mathcal{G}} Q_{gc} \ \text{P}(b_{ij}^T|G_{ij} = g, \phi)$$

We will assume that conditional on the underlying copy number, the B-allele read counts at different sites of the same segment are independent of each other and are independent of the total read count from the segment. Let $\mathbf{b}_j^T = (b_1^T, \ldots, b_{I_j}^T)$ denote all B-allele read counts at heterozygous sites of segment *j*. Under the conditional independence assumption outlined above, the joint probability of observing D_j^T and $\mathbf{b}_j^T$ conditional on the underlying copy number $C_j = c$ and the tumor purity being ϕ is as follows:

$$\begin{aligned} \text{P}(D_j^T, \mathbf{b}_j^T|C_j = c, \phi) &= \text{P}(D_j^T|C_j = c, \phi) \\ &\times \prod_{i=1}^{I_j} \sum_{g \in \mathcal{G}} Q_{gc} \ \text{P}(b_{ij}^T|G_{ij} = g, \phi) \end{aligned} \tag{7}$$

where the probability of D_j^T conditional on C_j and ϕ is the Poisson distribution (4), and the probability of b_{ij}^T conditional on G_{ij} and ϕ is the binomial distribution (6).

2.5 Likelihood model

So far, we have specified the probability of observing total read count and site-specific B-allele read counts at each segment conditional on the underlying copy number. Next we further treat the copy number C_j at each segment as a random variable, and model its probability as a categorical distribution with support $\mathcal{C} = \{0, 1, 2, 3, 4\}$, denoting the range of considered copy numbers, and parameters $\rho_j = (\rho_{j0}, \ldots, \rho_{j4})$, where ρ_{jc} denotes the probability of having $C_j = c$ in segment *j*. In other words, we have

$$C_j \mid \rho_j \sim \text{Categorical}(\mathcal{C}, \rho_j) \tag{8}$$

for each $j = 1, \ldots, J$.

We treat $\rho = (\rho_1, \ldots, \rho_J)$ and ϕ as parameters of our model $\Theta = (\phi, \rho)$, and the goal of our model is to infer the values of these parameters based on the total read count information in each segment and site-specific allele count information at each heterozygous site of these segments. Let $\mathbf{D} = (D_1, \ldots, D_J)$ and $\mathbf{b} = (\mathbf{b_1}, \ldots, \mathbf{b_J})$. By Equation (7), and assuming the observations from different segments are conditionally independent conditional on the tumor purity ϕ, the likelihood of observing the combined read count information is then given as follows:

$$\begin{aligned} \text{P}(\mathbf{D}, \mathbf{b}|\phi, \rho) &= \prod_{j=1}^{J} \sum_{c \in \mathcal{C}} \text{P}(C_j = c|\rho_j)\text{P}(D_j^T, \mathbf{b}_j^T|C_j = c, \phi) \\ &= \prod_{j=1}^{J} \sum_{c \in \mathcal{C}} \rho_{jc} \frac{\lambda_j^{D_j^T} e^{-\lambda_j}}{D_j^T!} \left[\prod_{i=1}^{I_j} \sum_{g \in \mathcal{G}} Q_{gc} \binom{d_{ij}^T}{b_{ij}^T} \bar{\mu}_g^{b_{ij}^T} (1 - \bar{\mu}_g)^{a_{ij}^T} \right] \end{aligned} \tag{9}$$

Given the likelihood function, we can then estimate the model parameters using maximum likelihood estimation. Alternatively, we can also add a prior into the model by incorporating our prior knowledge on the copy numbers and/or tumor purity. For instance, we can use the Dirichlet distribution to incorporate the prior on the distribution of copy numbers and beta distribution to incorporate the prior on tumor purity,

$$\rho_j \sim \text{Dirichlet}(\omega), \quad \phi \sim \text{Beta}(\alpha, \beta) \tag{10}$$

where ω is a vector having the same dimension as ρ_j and gives a weight to each copy number. If the priors are specified, we can then estimate the values of the parameters by maximizing their posterior probability, i.e. using the method of maximum a posteriori (MAP) estimation. In this article, we use non-informative prior for ϕ and a Dirichlet prior configured based on the compatible genotypes of each copy number for ρ_j. We solve the MAP problem using the Expectation-Maximization (EM) framework (Dempster *et al.*, 1977). An alternative approach would be to take a Bayesian approach to calculate the posterior probabilities of the tumor purity and copy number changes. We do not take the Bayesian approach owing to computational considerations, as it would require more time-consuming inference procedures. The complete details about prior configurations and EM updates are given in Supplementary Material.

3 RESULTS

Next we demonstrate the utility of combining loss of heterozygosity with CNAs to infer tumor purity and ploidy. For this purpose, we first present a clustering pattern of B-allele frequencies derived from NGS data in paired tumor-normal samples. Then we show that this clustering pattern can be used to resolve the ambiguous combinations of tumor purity and copy number changes, using both a toy example and real data. Afterward, we apply our method PyLOH, developed to infer tumor purity and absolute copy numbers by integrating the information from total read counts and B-allele frequencies, to simulated data and compare its performance with exiting state-of-the-art methods, CNAnorm-1.4.0 (Gusnanto *et al.*, 2012), THetA-0.0.3 (Oesper *et al.*, 2013) and PurBayes-1.3 (Larson and Fridley, 2013). Finally, we test the performance of our methods and other methods on real data, consisting of 12 whole genome sequencing datasets from breast cancer samples (Banerji *et al.*, 2012).

3.1 BAFs patterns in NGS data and BAF heat map

As discussed in the introduction, the distribution of BAFs is closely related to the underlying copy number changes. In particular, copy number changes at sites that are heterozygous with respect to the normal genome may result in a deviation of the BAFs from 0.5 (loss of heterozygosity), and the extent of this deviation depends on the absolute copy number changes and tumor purity. We illustrate this idea using a heat map plot

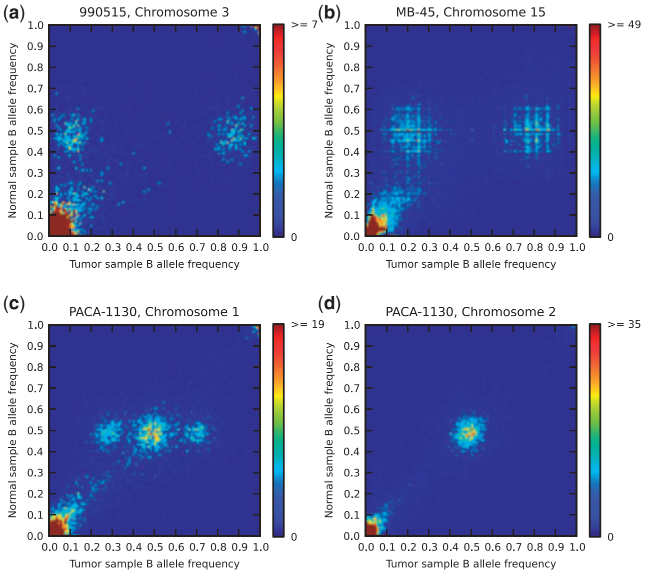
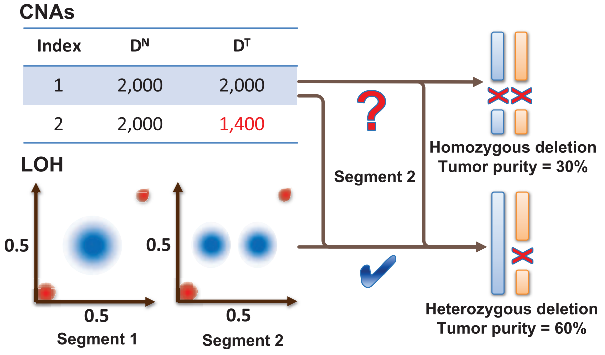


Fig. 1. Frequencies of BAF pairs in paired tumor–normal samples shown as heat maps. (**a**) Chromosome 3 of patient 990 515 (Zang *et al.*, 2012); (**b**) Chromosome 15 of patient MB-45 (Banerji *et al.*, 2012); (**c**) Chromosome 1 of patient PACA-1130 (Biankin *et al.*, 2012); (**d**) Chromosome 2 of patient PACA-1130 (Biankin *et al.*, 2012). The x-axis and the y-axis are divided into 100 bins representing the BAF resolution of 1%, thus each BAF heat map is a 100 × 100 mesh grid. The color of each grid quantifies the number of sites that has such a paired BAF across a specific genomic segment



Index	D^N	D^T
1	2,000	2,000
2	2,000	1,400

Fig. 2. A toy example illustrating the utility of BAFs patterns in resolving the identifiability problem

(Fig. 1), which shows the frequencies of BAF pairs, with one calculated from the normal sample and the other calculated from the matched tumor sample at the same site, coded in pseudo colors. Figure 1 shows the BAF heat maps of paired tumor–normal samples from three independent cancer genome NGS datasets, including both exome sequencing (Biankin *et al.*, 2012; Zang *et al.*, 2012) and whole genome sequencing (Banerji *et al.*, 2012).

The BAF heat maps demonstrate a clear cluster pattern on the distribution of BAF pairs (Fig. 1). Two clusters are shared by all four heat maps, including the bottom left cluster, containing sites with homozygous A-allele in both normal and tumor genomes, and the top right cluster, containing sites with homozygous B-allele in both normal and tumor genomes. (The small deviations of BAFs away from 0 or 1 of sites in these two clusters are likely because of sequencing and/or read-mapping errors.) Without changes in BAFs, these two clusters reveal no information with regard to the underlying copy number changes. Thus, we focus our attention on the other clusters in the heat maps, which all have BAFs centering at 0.5 in the normal samples, and thus contain mostly the heterozygous sites that underscore our method. For this reason, these clusters will be referred to as *heterozygous clusters* in the following.

The heterozygous clusters demonstrate distinct cluster patterns in different genomic segments of same/different samples. Although the BAFs of these clusters all center at 0.5 in the normal samples, the BAFs of the corresponding matched tumor samples can center at 0.5 (Fig. 1c and d) or at values away from 0.5 (Fig. 1a–c). In fact, these values provide a measure on the extent of LOH in the segments of tumor samples. For example, the tumor BAFs of the heterozygous cluster center at 0.5 in Figure 1d, suggesting no loss of heterozygosity in this segment. Without LOH, the absolute copy number in this case can only be even numbers, with diploid being the most plausible answer. (We can eliminate homozygous deletion, as there are reads mapped to this region.)

A different cluster pattern emerges in Figure 1a and b, which show two heterozygous clusters, with tumor BAFs centering at 0.1 and 0.9 in Figure 1a, and at 0.2 and 0.8 in Figure 1b, suggesting significant loss of heterozygosity in these two segments. (The appearance of two heterozygous clusters in these two cases and the symmetry of the two clusters with respect to the tumor BAF = 0.5. This is because the B-alleles are determined according to the human reference genome, which is not phased.) If the underlying copy number changes are a single-copy deletion in both cases, then the cluster with tumor BAFs centering at 0.1 (Fig. 1a) would correspond to a larger LOH, and consequently a higher tumor purity than the other case (Fig. 1b).

Figure 1c shows an interesting case with three heterozygous clusters with one center cluster showing no LOH and two symmetric clusters suggesting LOHs. This more complex cluster pattern suggests more than one type of CNAs within the segment being considered, most likely owing to the presence of both diploid and single-copy deletion changes.

Overall, these BAF heat maps provide a convenient and intuitive way to examine the overall CNAs of a chromosomal segment, and illustrate the utility of BAFs at heterozygous sites for inferring tumor purity and absolute copy numbers.

3.2 Using BAFs to solve the identifiability problem

The BAF patterns shown in Figure 1 can be used to resolve the identifiability problem, as the heterozygous clusters in each BAF heat map will center at different values with respect to different combinations of tumor purity and copy number changes. We demonstrate this idea using a toy example (Fig. 2). In this example, we have total read counts in two segments of the genome from both normal and tumor samples. The segment 2 has much smaller total read counts from the tumor sample than the normal sample. The differences can be explained by either a heterozygous deletion with 60% tumor purity or a homozygous deletion

Table 1. Three SNP sites from the exome sequencing data of patient MB-154

Index	Pos	d_N	d_T	BAF_N (%)	BAF_T (%)	dbSNP ID
1	chr6:112,147,822	141	87	49	25	rs28763978
2	chr7:131,842,835	98	29	51	52	rs156961
3	chr7:82,225,896	317	352	50	51	rs62465931

Note: BAF_N and BAF_T denote BAFs of the normal and tumor sample, respectively. d_N and d_T denote the read depth at the SNP site of the normal and tumor sample, respectively.

with 30% tumor purity in this segment. The total read counts themselves cannot distinguish these two possibilities. However, if we add in the information from the BAFs of the sites in segment 2, an observation of heterozygous clusters centering at tumor BAFs away from 0.5 would eliminate the homozygous deletion solution (Fig. 2).

We can observe similar cases in real cancer genome sequencing data as those in the above toy example. For instance, Table 1 shows the total read counts and BAFs at three SNP sites [dbSNP 130 ID listed (Sherry *et al.*, 2001)] observed in the exome sequencing of a breast cancer patient MB-154 (Banerji *et al.*, 2012). The mean coverage of the exome sequencing data was 141X for the tumor samples and 133X for the normal samples, respectively (Banerji *et al.*, 2012). The first SNP site shows an example of a heterozygous deletion as d_T is significantly lower than d_N, whereas BAF_T significantly deviates from 0.5. The second site shows an example of a homozygous deletion as d_T is significantly lower than d_N, whereas BAF_T is around 0.5. As a control, the third site shows an example without LOH or CNAs.

3.3 Results from simulated data

We have developed a probabilistic model to infer tumor purity and absolute copy numbers by integrating the LOH information described above and the information based on total read counts (see Section 2). Next we benchmark the performance of our new method on simulated data and compare it with other algorithms. By using simulated data, we know the ground truth of both tumor purity and absolute copy numbers, thereby providing us an objective way of comparing the performance of different algorithms.

We first created an artificial diploid human genome by using the human reference genome as a template and inserting SNP sites with a frequency similar to those observed in the human population (Sachidanandam *et al.*, 2001). This diploid genome will be treated as the normal genome in our follow-up simulation and analysis. The tumor genome was generated by adding somatic mutations and copy number changes to the normal genome. NGS reads were then simulated from the tumor sample consisting of a mixture of the normal and tumor genome, with the fraction of the tumor genome determined by the tumor purity. To reduce computational time, we use only data from chromosome 1 in our analysis. Details on how the genomes and reads were generated are described in Supplementary Material.

We created four tumor genomes that differ in their copy number configurations. The absolute copy numbers of each tumor genome were configured to introduce the identifiability problem, and the detailed configurations are given in Supplementary Table S1. For each tumor genome, we then simulated eight different sets of NGS reads from both normal and tumor samples by varying tumor purity. Overall, 32 sets of paired tumor-normal reads, each with 60X coverage, were generated. We applied PyLOH to each of these datasets, and compared its performance with three other methods, including PurBayes, CNAnorm and THetA. THetA and PyLOH require a segmentation of the tumor genome based on relative CNAs as an input. To avoid issues related to genome segmentation, we used ground truth segmentations in our analyses. Similarly, we used ground truth somatic mutation sites as the input for PurBayes. Details on how reads were preprocessed are given in Supplementary Material.

The tumor purities estimated by PyLOH and three other existing methods are shown in Figure 3. Because of the space limitation, we only show the results of the first three simulated datasets in Figure 3. The complete tumor purity estimates for all the simulated datasets are shown in Supplementary Table S2. The absolute copy numbers estimated by PyLOH and THetA for each simulated dataset are shown in Supplementary Tables S4–S7. A few observations emerge from the figure and tables. First, PyLOH significantly outperforms the other three methods on these datasets, providing a more accurate estimation of both tumor purity and absolute copy numbers and returning ground true values in most of the tested cases. Second, the THetA method, based only on total read counts information, is able to identify the ground truth as one of its possible solutions for tumor genomes 1 and 2 but fails to resolve the identifiability problem. Finally, the PurBayes method, based on information from somatic mutations, can return the true tumor purity in some cases, but has larger deviations than PyLOH, likely reflecting the statistical fluctuation associated with relatively small number of somatic mutation sites.

In addition to the four simulated tumor genomes discussed above, we simulated two additional tumor genomes based on copy number configurations derived from Sanger COSMIC v68 (Forbes *et al.*, 2011). The tumor purities and absolute copy numbers of the two COSMIC samples estimated by PyLOH and other methods are shown in Supplementary Tables S9–S11. PyLOH still outperformed the other methods on these two new datasets. Further details on these two datasets are described in Supplementary Material.

3.4 Results from breast cancer sequencing data

Having illustrated the utility of our method on simulated data, we proceed to test the performance of PyLOH on a real cancer genome dataset, consisting of whole genome sequencing of 12 breast cancer samples (Banerji *et al.*, 2012). As the ground truth tumor purities are unknown for this dataset, we used the tumor purities calculated by ABSOLUTE based on SNP array data and reported in the paper by Banerji *et al.* (2012) as a baseline for our comparison. Although this baseline is by no means absolutely correct, it offers clues on the performance of different algorithms, as it was derived from SNP array data instead of NGS data as in

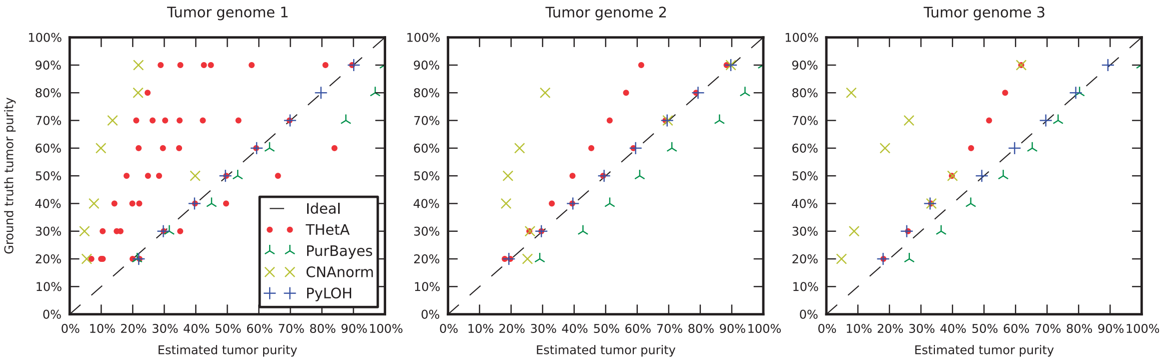


Fig. 3. The tumor purity estimates of the first three simulated datasets given by THetA, CNAnorm, PurBayes and PyLOH. The x-axis is the estimated tumor purity and the y-axis is the ground truth tumor purity

Table 2. The tumor purity estimates of the 12 breast cancer whole genome sequencing datasets given by THetA, CNAnorm, PurBayes, PyLOH and ABSOLUTE

Patient ID	THetA	CNAnorm	PurBayes	PyLOH	ABSOLUTE
MB-15	0.288	0.245	0.999	0.589	0.22
MB-45	0.526	0.291	0.999	0.566	0.25
MB-50	0.193	0.224	0.999	0.532	0.47
MB-82	0.129	0.274	0.999	0.192	0.74
MB-98	0.598	0.437	0.999	0.698	0.54
MB-106	0.409[a] 0.817[a]	0.135	0.999	0.831	0.89
MB-116	0.325	0.510	0.769	0.325	0.66
MB-123	0.358	0.353	0.999	0.377	0.65
MB-154	0.645	0.187	0.999	0.664	0.70
MB-165	0.668[a]	0.172	0.999	0.662	0.68
MB-198	0.301	0.293	0.999	0.607	0.64
MB-200	0.515	0.158	0.999	0.523	0.55
MAE[b]	0.220	0.320	0.397	0.186	n/a

[a]THetA outputted multiple solutions and here we only show the solutions with the smallest deviation from the diploid. [b]For tumor purities reported by THetA with multiple solutions, we used the average of the solutions with the smallest deviation from the diploid to calculate the MAE.

our case. We used BIC-seq-1.2.1 (Xi *et al.*, 2011) to obtain segmentation files for THetA and PyLOH, and used VarScan-2.3.5 (Koboldt *et al.*, 2012) to call somatic mutation sites for PurBayes. As THetA often outputs multiple solutions, we selected the ones with the smallest deviation from the diploid whenever this happens, as recommended by THetA (Oesper *et al.*, 2013). Further details on this dataset are described in Supplementary Material.

The tumor purities estimated by PyLOH and three existing algorithms, THetA, CNAnorm and PurBayes, for the 12 breast cancer sequencing datasets are summarized in Table 2. If the tumor purities estimated by ABSOLUTE are used as our comparison baseline, we find PyLOH to be the most accurate algorithm among the four—it yields a mean absolute error (MAE) of 0.186, as compared with a MAE of 0.22 by the second best algorithm, THetA. PurBayes, which uses somatic mutations to estimate tumor purity, produced poorest results, likely because of the inclusion of false-positive results in the somatic mutation calling procedure.

Although PyLOH returned closer solutions to ABSOLUTE (as measured by MAE) than any of the other methods, the tumor purities estimated by PyLOH and ABSOLUTE deviate in six samples: MB-15, MB-45, MB-82, MB-98, MB-116 and MB-123. To find out why such a discrepancy arises, we carefully studied each of these six cases. In two of these cases (MB-15 and MB-45), we believe that the results obtained by PyLOH are more accurate because the tumor purities and absolute copy numbers inferred by total read counts information are consistent with those inferred by BAFs information, and both support the results obtained by PyLOH. For sample MB-82, MB-116 and MB-123, the contribution of BAFs information to estimating tumor purities is not significant compared with the total read count information, likely because of a low tumor purity in these samples. As a result, the estimation of tumor purity is mainly contributed by information from total read counts, and in fact produces a similar estimation compared with THetA. For the remaining case MB-98, the estimated tumor purities given by the four algorithms are all inconsistent—one possible reason for this may be the existence of subclonal tumor populations in the tumor sample.

Aside from the accuracy comparison described above, we note that PyLOH is fast, with a running time scaling linearly with the number of segments. This is in contrast to the THetA method, the running time of which scales exponentially with the number of segments, as it explores all combinations of copy number changes across all segments (Oesper *et al.*, 2013). As a result, THetA takes a prohibitively long time to run when the number of segments is above 150 and the maximum copy number is >6, although PyLOH has no such constraints. Further details about the run time of each algorithm are given in Supplementary Material.

4 DISCUSSION

In this article, we examined the problem of estimating tumor purity and absolute copy number changes from NGS data, and, in particular, focused on solving the identifiability problem that has not been properly solved by the existing methods. We demonstrated that the distribution of B-allele frequencies at sites that are heterozygous with respect to the normal genome provides key, but under used, information to solve the identifiability problem. We further developed a full probabilistic model to integrate the copy number change and BAF information, and derived a principled way to estimate tumor purity and absolute CNAs. We benchmarked the performance of our method, PyLOH, on both simulated data and real whole genome sequencing data, showing that our method outperforms existing methods in both cases.

PyLOH requires a segmentation of the genome into segments with different CNAs as input. Many algorithms have been developed to segment genomes based on copy number changes and BAFs of SNP array data with varying levels of accuracy (Olshen *et al.*, 2004; Sun *et al.*, 2009; Van Loo *et al.*, 2010; Yau *et al.*, 2010). A few of these array-based methods have recently been translated to the sequencing domain (Mayrhofer *et al.*, 2013; Yau, 2013). A future direction of PyLOH would be to integrate these existing methodologies and combine them with the probabilistic model of PyLOH to carry out both genome segmentation and absolute copy number estimation.

Another important future direction is to use our model to study tumor heterogeneity. So far, we have focused on separating genetic changes from a mixture of normal and tumor cells. It is well known that multiple tumor clonal types may coexist in the tumor sample, each with an associated mutation landscape (Parsons, 2008). To further model intra-tumor heterogeneity on top of the current probabilistic framework, we can assume there are multiple populations of tumor cells. Thus, the model likelihood given by Equation (9) can be extended to account for subclonal tumor populations (details in Supplementary Material). We plan to further extend PyLOH in this direction to tackle the more challenging problem of deconvolving tumor heterogeneity by combining copy number change and allele frequency information.

ACKNOWLEDGEMENTS

The authors thank J Biesinger, D Quang, and J Wu for helpful discussion.

Funding: The work was partly supported by National Institute of Health (grant R01HG006870).

Conflict of Interest: none declared.

REFERENCES

Banerji,S. *et al.* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, **486**, 405–409.

Biankin,A.V. *et al.* (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.

Bignell,G.R. *et al.* (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.*, **14**, 287–295.

Campbell,P.J. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet.*, **40**, 722–729.

Carter,S.L. *et al.* (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.*, **30**, 413–421.

Chiang,D.Y. *et al.* (2008) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.

Collins,F. and Barker,A. (2007) Mapping the cancer genome. *Sci. Am. Mag.*, **296**, 50–57.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B Methodol.*, **39**, 1–38.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.

Forbes,S.A. *et al.* (2011) Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39** (**Suppl. 1**), D945–D950.

Greenman,C.D. *et al.* (2010) Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics*, **11**, 164–175.

Gusnanto,A. *et al.* (2012) Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics*, **28**, 40–47.

Hudson,T.J. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

Koboldt,D.C. *et al.* (2012) Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.*, **22**, 568–576.

Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Larson,N.B. and Fridley,B.L. (2013) Purbayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics*, **29**, 1888–1889.

Lindblad-Toh,K. *et al.* (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nat. Biotechnol.*, **18**, 1001–1005.

Mayrhofer,M. *et al.* (2013) Patchwork: allele-specific copy number analysis of whole genome sequenced tumor tissue. *Genome Biol.*, **14**, R24.

Mei,R. *et al.* (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res.*, **10**, 1126–1137.

Navin,N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90–94.

Oesper,L. *et al.* (2013) Theta: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol.*, **14**, R80.

Olshen,A.B. *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.

Parsons,B.L. (2008) Many different tumor types have polyclonal tumor origin: evidence and implications. *Mutat. Res.*, **659**, 232–247.

Pinkel,D. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, **20**, 207–211.

Rasmussen,M. *et al.* (2011) Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. *Genome Biol.*, **12**, R108–R108.

Reiersøl,O. (1950) Identifiability of a linear relation between variables which are subject to error. *Econometrica*, **18**, 375–389.

Roberts,N.D. *et al.* (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*, **29**, 2223–2230.

Roth,A. *et al.* (2012) Jointsnvmix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics*, **28**, 907–913.

Sachidanandam,R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Su,X. *et al.* (2012) Purityest: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics*, **28**, 2265–2266.

Sun,W. *et al.* (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.*, **37**, 5365–5377.

Van Loo,P. *et al.* (2010) Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA*, **107**, 16910–16915.

Xi,R. *et al.* (2011) Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proc. Natl Acad. Sci. USA*, **108**, E1128–E1136.

Yau,C. (2013) OncoSNP-SEQ: a statistical approach for the identification of somatic copy number alterations from next-generation sequencing of cancer genomes. *Bioinformatics*, **29**, 2482–2484.

Yau,C. *et al*. (2010) A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. *Genome Biol*., **11**, R92.

Yuan,Y. *et al*. (2012) Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci*. *Transl*. *Med*., **4**, 157ra143.

Zang,Z.J. *et al*. (2012) Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat*. *Genet*., **44**, 570–574.

Zhao,X. *et al*. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*., **64**, 3060–3071.