BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 24 2014, pages 3532–3540
doi:10.1093/bioinformatics/btu651

Sequence analysis

Advance Access publication October 8, 2014

Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data

Layla Oesper[1,†,*], Gryte Satas[1,†] and Benjamin J. Raphael[1,2,*]

[1]Department of Computer Science and [2]Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Most tumor samples are a heterogeneous mixture of cells, including admixture by normal (non-cancerous) cells and subpopulations of cancerous cells with different complements of somatic aberrations. This intra-tumor heterogeneity complicates the analysis of somatic aberrations in DNA sequencing data from tumor samples.

Results: We describe an algorithm called THetA2 that infers the composition of a tumor sample—including not only tumor purity but also the number and content of tumor subpopulations—directly from both whole-genome (WGS) and whole-exome (WXS) high-throughput DNA sequencing data. This algorithm builds on our earlier Tumor Heterogeneity Analysis (THetA) algorithm in several important directions. These include improved ability to analyze highly rearranged genomes using a variety of data types: both WGS sequencing (including low $\sim 7\times$ coverage) and WXS sequencing. We apply our improved THetA2 algorithm to WGS (including low-pass) and WXS sequence data from 18 samples from The Cancer Genome Atlas (TCGA). We find that the improved algorithm is substantially faster and identifies numerous tumor samples containing subclonal populations in the TCGA data, including in one highly rearranged sample for which other tumor purity estimation algorithms were unable to estimate tumor purity.

Availability and implementation: An implementation of THetA2 is available at http://compbio.cs.brown.edu/software

Contact: layla@cs.brown.edu or braphael@brown.edu

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

Received on April 6, 2014; revised on August 8, 2014; accepted on September 29, 2014

1 INTRODUCTION

Several recent studies indicate that most tumor samples are a heterogeneous mixture of cells, including admixture by normal (non-cancerous) cells and subpopulations of cancerous cells with different complements of somatic aberrations (Gerlinger *et al.*, 2012; Nik-Zainal *et al.*, 2012). Characterizing this *intra-tumor heterogeneity* is essential for several reasons. First, an estimate of tumor *purity*, the fraction of cancerous cells in a tumor, is necessary for accurate identification of somatic aberrations of all types in the sample. Most cancer genome sequencing studies use a re-sequencing approach to detect somatic aberrations. Reads from a tumor sample (and usually a matched normal sample) are aligned to the human reference genome. Differences in the sequence of aligned reads, the number of aligned reads or the configuration of aligned reads (e.g. split reads or discordant pairs) are used to infer the presence of single nucleotide or other small variants, copy number aberrations or structural aberrations, respectively (Ding *et al.*, 2010; Meyerson *et al.*, 2010). However, the presence of intra-tumor heterogeneity can dilute the signals required to identify somatic aberrations.

Second, estimates of the *composition* of a tumor sample—including not only the tumor purity, but also the number and fractions of subpopulations of tumor cells—provide useful for understanding tumor progression and determining possible treatment strategies (Greaves and Maley, 2012; Mullighan *et al.*, 2008). In particular, *clonal* somatic aberrations that exist in all tumor cells are likely early mutational events and their identification sheds light on the early stages of cancer. Conversely, *subclonal* somatic aberrations might reveal properties shared by a subset of tumor cells, such as drug resistance or ability to metastasize. Identification of such aberrations and subpopulations of tumor cells might inform treatment strategies, and/or help predict metastasis/relapse.

In the past few years, several methods to infer tumor purity and/or tumor composition have been developed. These methods generally fall into two categories: (i) methods that use somatic single-nucleotide variants (SNVs) and (ii) methods that use somatic copy number aberrations. SNV-based methods such as EXPANDS (Andor *et al.*, 2014), PyClone (Roth *et al.*, 2014) and many others (Jiao *et al.*, 2014; Larson and Fridley, 2013) use clustering of variant allele frequencies to determine tumor populations and frequencies. While these types of methods are able to derive multiple tumor subpopulations, they often require estimates of copy number for each region containing SNVs. Deriving such estimates for highly rearranged aneuploid tumors is as difficult as the estimation of intra-tumor heterogeneity itself. Moreover, these approaches require high-coverage sequencing to overcome the high variance in read counts at individual SNVs. For example, both PyClone (Roth *et al.*, 2014) and PhyloSub (Jiao *et al.*, 2014) explicitly require deeply sequenced data. Thus, less expensive low-coverage sequence data as generated in TCGA (Cancer Genome Atlas Network, 2012) is not amenable to these approaches.

Copy number-based methods such as ABSOLUTE (Carter *et al.*, 2012) and CNAnorm (Gusnanto *et al.*, 2012) use observed shifts in read depth due to copy number aberrations to predict

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

tumor purity, but do not explicitly consider multiple tumor subpopulations, and therefore may return purity estimates that only reflect a single subpopulation of tumor cells in a sample. In Oesper *et al.* (2013), we introduced the Tumor Heterogeneity Analysis (THetA) algorithm to infer the composition of a tumor sample—including both the percentage of normal admixture and the fraction and content of one *or more* tumor subpopulations that differ by copy number aberrations.

In this article, we present THetA2, which extends the THetA algorithm in several important directions. First, we substantially improve the computation for the case of multiple distinct tumor subpopulations in a sample. Second, we extend THetA to infer tumor composition for highly rearranged genomes using a two-step procedure where initial estimates are made using high-confidence regions of the genome, and then are extended to the entire genome. Third, we devise a probabilistic model of B-allele frequencies (BAFs), which can be used to solve the identifiability issue when read depth alone is consistent with multiple possible tumor compositions. Finally, we extend THetA to analyze whole-exome (WXS) sequencing data. We apply our new algorithm to both whole-genome (WGS) (including low-pass) and WXS sequence data from 18 samples from The Cancer Genome Atlas (TCGA). We find that the improved algorithm is substantially faster and able to analyze highly rearranged genomes—identifying numerous tumors with subclonal tumor populations in the TCGA data. Where available, we compare our purity estimates to published values for ABSOLUTE (Carter *et al.*, 2012). While the purity estimates are largely in agreement for higher purity samples, we find cases where ABSOLUTE fails or underestimates purity, but THetA2 identifies multiple tumor subpopulations. These improvements greatly expand the range of sequencing data and tumors for which we can infer tumor composition.

2 METHODS

2.1 Notation and problem formulation

We assume that the reference genome is partitioned into a sequence $\mathbf{I} = (I_1, \ldots, I_m)$ of non-overlapping intervals, according to changes in the density, or depth, of reads aligning to each position in the reference (Xi *et al.*, 2011). Given $\mathbf{I}$, we define a corresponding *read depth vector* $\mathbf{r} = (r_1, \ldots, r_m) \in \mathbb{N}^m$ where r_j is the number of reads with a (unique) alignment within I_j. A cancer genome is defined by an *interval count vector* $\mathbf{c} \in \mathbb{N}^m$, where c_j is the integer number of copies of interval I_j in the cancer genome.

A tumor sample $\mathcal{T}$ is a mixture of cells that contain different collections of somatic mutations, and in particular somatic copy number aberrations. Each subpopulation has a distinct interval count vector representing the genome of the subpopulation. Following the model introduced in Oesper *et al.* (2013), we represent $\mathcal{T}$ by: (i) an *interval count matrix* $\mathbf{C} = [c_{j,k}] \in \mathbb{N}^{m \times n}$ where $c_{j,k}$ is the number of copies of interval I_j in the k^{th} distinct subpopulation; and (ii) a *genome mixing vector* $\mu \in \Delta_{n-1} = \{(\mu_1, \ldots, \mu_n)^T | \sum_{j=1}^{n} \mu_j = 1, \text{ and } \mu_j \geq 0 \text{ for all } j\}$ where μ_k is the percentage of cells in $\mathcal{T}$ that belong to the k^{th} distinct subpopulation.

Let the interval count matrix $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_n)$, where $\mathbf{c}_j$ is the j^{th} column of $\mathbf{C}$. We assume that $\mathbf{C}$ satisfies three constraints. (i) The first column $\mathbf{c}_1 = 2^m$ so that the first component of the tumor sample is the normal genome. (ii) The number n of subpopulations is less than the number m of intervals. (iii) The copy numbers of the intervals are bounded below by 0 and above by a maximum copy number $k \geq 2$. Thus, $\mathbf{C} \in \{0, \ldots, k\}^{m \times n}$. We define $\mathcal{C}_{m,n,k}$ to be the set of all such $\mathbf{C}$, and define $\Omega_{m,n,k} = \{(\mathbf{C}, \mu) | \mathbf{C} \in \mathcal{C}_{m,n,k}, \mu \in \Delta_{n-1}\}$ to be the domain of pairs $(\mathbf{C}, \mu)$ satisfying all constraints.

We model the observed read depth vector $\mathbf{r}$ using a multinomial probability distribution with parameter $\mathbf{p} = (p_1, \ldots, p_m)$, where p_j is the probability that a randomly chosen read will align to interval I_j. A pair $(\mathbf{C}, \mu)$ defines a value for the multinomial parameter $\mathbf{p} = \widehat{\mathbf{C}\mu} = \frac{\mathbf{C}\mu}{|\mathbf{C}\mu|_1}$. Thus, the negative log likelihood $L(\mathbf{C}, \mu | \mathbf{r}) = -\log(\text{Mult}(\mathbf{r}; \widehat{\mathbf{C}\mu}))$ is the negative log of the multinomial probability of observing counts $\mathbf{r}$ in the intervals given the probability of a read aligning to an interval is defined by $\widehat{\mathbf{C}\mu}$. The goal is to find the interval count matrix $\mathbf{C}^*$ and genome mixing vector μ^* that minimize the negative log likelihood:

$$(\mathbf{C}^*, \mu^*) = \text{argmin}_{(\mathbf{C},\mu) \in \Omega_{m,n,k}} L(\mathbf{C}, \mu; \mathbf{r}) \quad (1)$$

2.2 Interval count matrix enumeration

In this section, we derive an improved procedure to solve the optimization problem (1). In Oesper *et al.* (2013), we showed that the function $L(\mathbf{C}, \mu; \mathbf{r})$ is a convex function of μ. Thus, for a fixed interval count matrix $\mathbf{C}$, the optimal value of μ can be computed efficiently. In the important special case of a mixture of normal cells and a single tumor population ($n = 2$), we reduce the domain of the interval count matrix $\mathbf{C}$ to a set whose size is polynomial in m and guaranteed to contain the optimal $\mathbf{C}^*$. This set is easy to enumerate, and we obtain an efficient algorithm. However, when a tumor sample contains multiple tumor subpopulations ($n > 2$), the algorithm in Oesper *et al.* (2013) enumerates all $\mathbf{C} \in \mathcal{C}_{m,n,k}$ and checks whether each such $\mathbf{C}$ satisfies a particular ordering constraint that is a necessary, but not sufficient, condition for the optimal $\mathbf{C}^*$.

In this section, we derive an algorithm that explicitly enumerates only those matrices $\mathbf{C}$ that satisfy a more restrictive necessary ordering constraint for a mixture of *any number* of tumor genomes. All proofs are contained in the Supplementary Material.

2.2.1 Compatible order We say that vectors $\mathbf{v} = (v_1, \ldots, v_m)$ and $\mathbf{w} = (w_1, \ldots, w_m) \in \mathbb{R}^m$ have *compatible order* provided all $1 \leq i, j \leq m$, $v_i \leq v_j$ if and only if $w_i \leq w_j$. In Oesper *et al.* (2013) we proved that if $(\mathbf{C}^*, \mu^*)$ is optimal [i.e. satisfies Equation (1)] then $\widehat{\mathbf{C}^*\mu^*}$ and $\mathbf{r}$ have compatible order. We define $\mathcal{S}_{m,n,k}$ to be the set of matrices $\mathbf{C} \in \mathcal{C}_{m,n,k}$ that satisfy this ordering constraint: i.e. $\mathcal{S}_{m,n,k} = \{\mathbf{C} | \mathbf{C} \in \mathcal{C}_{m,n,k} \text{ and } \exists \mu \in \Delta_{n-1} \text{ such that } \widehat{\mathbf{C}\mu} \text{ is in compatible order with } \mathbf{r}\}$. Thus, to find the optimal solution $(\mathbf{C}^*, \mu^*)$, it is sufficient to examine matrices $\mathbf{C} \in \mathcal{S}_{m,n,k}$.

Without loss of generality, we assume that the read depth vector $\mathbf{r} = (r_1, \ldots, r_m)$ satisfies $r_1 \leq r_2 \leq \ldots \leq r_m$. Thus, the set $\mathcal{S}_{m,n,k} = \{\mathbf{C} \in \mathcal{C}_{m,n,k} | (\mathbf{C}\mu)_1 \leq (\mathbf{C}\mu)_2 \leq \ldots \leq (\mathbf{C}\mu)_m \text{ for some } \mu \in \Delta_{n-1}\}$.

For a matrix $\mathbf{C} \in \mathcal{C}_{m,n,k}$, the set of μ that result in a compatible ordering can be calculated using the function $\Phi(\mathbf{C})$ as follows:

$$\Phi(\mathbf{C}) = \bigcap_{j=1}^{m-1} \{\mu | \mu \in \Delta_{n-1} \text{ such that } (\mathbf{C}\mu)_j \leq (\mathbf{C}\mu)_{j+1}\}. \quad (2)$$

Thus, a matrix $\mathbf{C} \in \mathcal{S}_{m,n,k}$ if and only if $\Phi(\mathbf{C})$ is not empty. Corollary 2.1 follows directly from Equation (2).

COROLLARY 2.1. *Suppose* $\mathbf{C} \in \mathcal{C}_{m,n,k}$. *If there exists an* $i \in \{1, \ldots, m-1\}$ *such that for all* $t \in \{2, \ldots, n\}$, $c_{i,t} \geq c_{i+1,t}$ *and there exists a* $t \in \{2, \ldots, n\}$ *such that* $c_{i,t} > c_{i+1,t}$, *then* $\Phi(\mathbf{C}) = \varnothing$.

2.2.2 Using a graph to enumerate $\mathcal{S}_{m,n,k}$ We now present an algorithm to enumerate $\mathcal{S}_{m,n,k}$ for $n \geq 2$. Consider a complete (including self loops) directed graph $G_{n,k}$, with a vertex for each possible row in a matrix in $\mathcal{C}_{m,n,k}$. Paths on $G_{n,k}$ of length $m-1$ correspond to matrices in $\mathcal{C}_{m,n,k}$ (See Fig. 1 and Supplementary Fig. S1).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

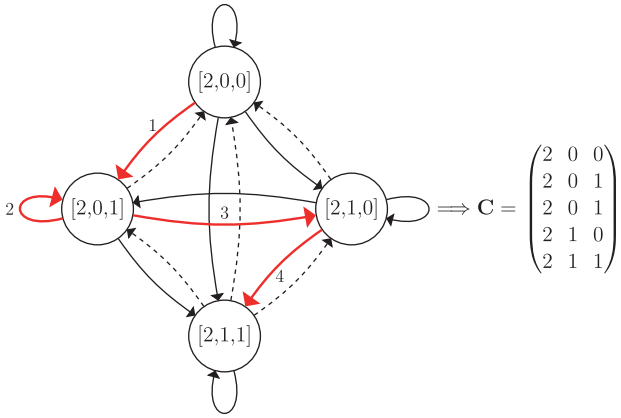

Fig. 1. The graph $G_{3,1}$ is used to enumerate the matrices $\mathcal{S}_{m,3,1}$ as a subset of the paths of length $m-1$. The dashed edges can be removed by applying Corollary 2.1. The highlighted path corresponds to the matrix on the right

To enumerate the subset of paths on $G_{n,k}$ that correspond to matrices in $\mathcal{S}_{m,n,k}$, we use a depth-first search. While building paths, we calculate the set Φ for the matrix implied by the current path, and only proceed down branches that do not result in the empty set (see Supplementary Material for details). As a result, we are guaranteed to enumerate only the matrices in $\mathcal{S}_{m,n,k}$.

Corollary 2.1 allows us to reduce the graph $G_{n,k}$ by showing that there are certain edges that will never appear in paths that correspond to matrices in $\mathcal{S}_{m,n,k}$, and thus can be removed from the graph before matrix enumeration. In the case where $n = 3$, the calculation of Φ is reduced to a problem in a single variable, $\frac{\mu_2}{\mu_3}$.

2.3 A two-step procedure for genome-wide inference of copy numbers

In Oesper *et al.* (2013), we inferred tumor composition using a relatively coarse interval partition **I** of the reference genome, considering only large copy number aberrations. As a result, the published approach could not readily be applied to highly rearranged genomes that are segmented into many intervals. Moreover, manual selection of a subset of intervals was typically required when analyzing samples containing multiple tumor populations. Even with the improved enumeration procedure described in the previous section, when more than one tumor subpopulation is considered, the number of matrices **C** that need to be enumerated is exponential in the number m of intervals. Moreover, the number of matrices **C** is also exponential in the maximum copy number state k considered, making analysis of genomes with extensively amplified regions more difficult.

In this section, we present a two-step procedure for interval selection that overcomes the limitations stated above, and allows us to infer the composition of highly rearranged genome that are highly fragmented and/or contains amplified segments with more than k copies. Our two-step procedure consists of the following steps: (i) Select a set of high-confidence intervals and determine the most likely **C** and μ for those intervals. (ii) Use the estimates of **C** and μ to determine copy numbers for all other intervals in **I** not used in the first step, thus allowing for analysis of both highly amplified regions and fragmented genomes.

2.3.1 Interval selection We automate the selection of a subset of high-confidence intervals used to determine the optimal $(\mathbf{C}^*, \mu^*)$ for those intervals. Further details are included in the Supplementary Material. Briefly, we partition **I** into two sets of intervals: (i) $\mathbf{I}_H$—high-confidence intervals; (ii) $\mathbf{I}_L$—lower-confidence intervals. $\mathbf{I}_H$ is selected to contain up to a fixed integer d longest intervals from **I** such that each interval selected is longer than a predetermined minimum length and is not obviously amplified beyond the specified max copy number k. $\mathbf{I}_L$ contains all remaining intervals from **I**. Additionally, $\mathbf{I}_H$ must represent $>10\%$ of the total length of all provided intervals, otherwise the sample is determined not to be a good candidate for analysis using THetA2. Once $\mathbf{I}_H$ and $\mathbf{I}_L$ have been selected, we use the improved THetA2 algorithm described in the previous section to calculate $\mathbf{C}^*_H$ and μ^*_H for just the intervals in $\mathbf{I}_H$.

2.3.2 Determining additional copy numbers: single row Given $(\mathbf{C}^*_H, \mu^*_H)$ predicted for high-confidence intervals $\mathbf{I}_H$, we infer copy numbers for the remaining intervals $\mathbf{I}_L$. We start with the simplifying assumption that $|\mathbf{I}_L| = 1$. We prove the following theorem.

THEOREM 2.1. *Let* $\mathbf{C} = [c_{i,j}]$ *be an interval count matrix.* $L(\mathbf{C}, \mu|\mathbf{r})$ *is a convex function of* $c_{i,j}$.

We use Theorem 2.1 to find the optimal real-valued solution for the $c_{i,j}$'s corresponding to the single interval $\mathbf{I} \in \mathbf{I}_L$, given $\mathbf{C}^*_H$ and μ^*_H. We then check the surrounding integer values to find the integral solution, which, by convexity, is guaranteed to find the optimal integer solution.

2.3.3 Determining additional copy numbers: multiple rows In the previous section, we showed how to find the optimal copy number for a single additional interval in $\mathbf{I}_L$ given optimal values $\mathbf{C}^*_H$, and μ^*_H for a set of high-confidence intervals $\mathbf{I}_H$. To estimate copy numbers when $I \in \mathbf{I}_L$ contains more than one interval, we estimate the optimal copy numbers for each interval in $I \in \mathbf{I}_L$ when appended to $\mathbf{C}_H$ individually as described in the previous section, and then jointly append all inferred copy numbers to $\mathbf{C}^*_H$ to obtain a new matrix $\mathbf{C}_{H\cup L}$. We then return the solution $(\mathbf{C}_{H\cup L}, \mu^*_H)$. We note that this approach provides no guarantee for finding the optimal copy numbers across *all* $I \in \mathbf{I}_L$ given $\mathbf{C}^*_H$, and μ^*_H. However, in practice, we find that the solutions returned by our procedure are generally similar to this optimum (Supplementary Table S1).

2.4 Model selection

As in Oesper *et al.* (2013), we use the Bayesian information criterion (BIC) to select from different sized models (i.e. different numbers n of tumor populations) and their corresponding maximum likelihood solutions. We use the standard BIC of $-2log(L) + alog(b)$ where L is the likelihood of a solution, $a = (m+1)(n-1)$ is the number of free parameters in the model and b is the number of data points (the total number of tumor and normal reads). In contrast, Oesper *et al.* (2013) used a modified BIC that more strongly penalized solutions with more tumor populations. Such a modification is not necessary here, as our improved algorithm considers copy number data across the entire genome, rather than only a small number of intervals, reducing the possibility of overfitting. Thus, we are able to more robustly identify samples with multiple subpopulations of tumor cells.

2.5 Probabilistic model of BAFs

THetA2 may return multiple equally like pairs $(\mathbf{C}, \mu)$ when using read depth alone. We derive a probabilistic model of BAFs—the fraction of reads containing the minor allele—that may be used to distinguish between multiple pairs $(\mathbf{C}, \mu)$. Let $\mathbf{v} = (v_1, v_2, \ldots, v_q)$ be the observed BAFs for q heterozygous germ line SNPs in the normal sample and $\mathbf{w} = (w_1, w_2, \ldots, w_q)$ be the corresponding BAFs from the tumor genome. We model **w** as being drawn from Gaussian distributions whose parameters depend

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

on $\mathbf{v}$, $\mathbf{C}$ and μ. We then select the $(\mathbf{C}, \mu)$, which maximizes the likelihood of the observed BAFs in the tumor sample:

$$L(\mathbf{C}, \mu|\mathbf{v}, \mathbf{w}) = P(\mathbf{w}|\delta, \sigma^2) = \prod_{i=1}^{q} \mathcal{N}(w_i|sgn(0.5 - w_i)\delta_i, \sigma_i^2) \quad (3)$$

Here σ_i^2 is the observed variance for all heterozygous SNPs in $\mathbf{v}$ that lie within interval I_j, and δ_j is the expected BAF deviation away from 0.5 given $\mathbf{C}$ and μ. See the Supplementary Material for further details.

2.6 Application to WXS data

Finally, we extend THetA2 for WXS data, where only the coding regions of the genome have been targeted for sequencing. From WXS data we need to infer the following two values: (i) a set of non-overlapping intervals $\mathbf{I} = (I_1, \ldots, I_m)$ in the reference genome; and (ii) a corresponding read depth vector $\mathbf{r} = (r_1, \ldots, r_m)$.

To infer the interval partition $\mathbf{I}$, we rely on recently developed algorithms such as ExomeCNV (Sathirapongsasuti *et al.*, 2011) and EXCAVATOR (Magi *et al.*, 2013) for segmentation and detection of copy number aberrations from WXS data. The segmentation returned by one of these algorithms may contain gaps rather than being a complete partition of the reference genome, but still provides a set of non-overlapping intervals that may be used as input to THetA2. We note that some methods use normalization procedures for GC content, mappability and even exon length and this information is therefore implicitly incorporated into the input provided to THetA2.

We compute the read depth vector $\mathbf{r} = (r_1, \ldots, r_m)$ for WXS data as follows. Given a set $\mathbf{I}$ of non-overlapping intervals in the reference genome, a set $\mathbf{E}$ of exons in the reference genome and a read length ℓ, we set $r_j = \frac{x_j}{\ell}$ where x_j is the total number of sequenced nucleotides that have a unique alignment to some exon $e \in \mathbf{E}$ within interval I_j. Thus, r_j is approximate count of the number of reads aligning to some exon located in interval I_j.

3 RESULTS

We ran THetA2 on simulated data, WGS (including low-pass data 5–7× coverage) and WXS data from 18 breast carcinoma, ovarian carcinoma, glioblastoma multiforme, kidney renal clear cell and lung squamous cell carcinoma samples from TCGA (Supplementary Table S2). Where available, we compare our estimates of tumor purity to the estimates reported by the ABSOLUTE algorithm (Carter *et al.*, 2012) that estimates purity from SNP array data.

The rest of this section is organized as follows. First, we discuss results on simulated data. Second, we demonstrate THetA2's performance on WXS data, including comparison of results for samples for which both WGS and WXS data were available. Next, we present in-depth analysis of several WGS samples to demonstrate the efficacy of THetA2 on highly rearranged genomes, using both low-pass and moderate coverage sequence data. Finally, we apply our probabilistic model of BAFs to one sample and disambiguate between two equally likely solutions.

3.1 Simulated data

We tested THetA2 on simulated data to demonstrate the improvements in THetA2 over the original THetA as well as ABSOLUTE. We created simulated mixtures using real sequencing data from an AML tumor sample and matched normal sample (TCGA-AB-2965) from The Cancer Genome Atlas Research Network (2013). This sample was chosen because of its high purity (~95% pure) and lack of copy number aberrations as predicted by array data, providing high confidence that our simulated mixture and implanted copy number aberrations are not confounded by impurity and aberrations in the real data. Simulated mixtures are created by implanting random amplifications and deletions (see Supplementary Material) to create different tumor populations, and then creating a mixture representing different tumor compositions.

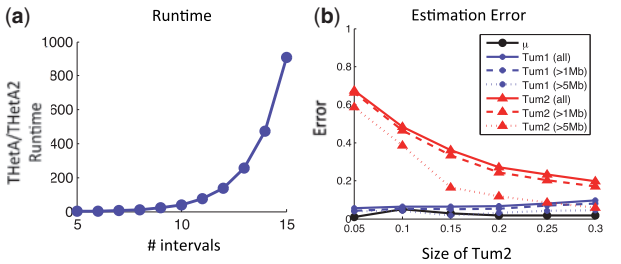

Fig. 2. Runtime comparison and estimation error for THetA2 on simulated data containing a mixture of normal cells and two tumor subpopulations. (**a**) The ratio of runtimes for the old and new enumeration procedures as a function of the number of intervals used in the first step of the algorithm. (**b**) Estimation error for both μ and C for each tumor population (Tum1 and Tum2) as the proportion of Tum2 increases and the proportion of Tum1 is fixed at 0.5. Error for μ is the Euclidean distance from the true value and error for each tumor population is the fraction of the genome for which the copy number is incorrectly inferred for the all copy number estimates, and when only considering intervals that are longer than 1 Mb and 5 Mb

3.1.1 Mixtures with three subpopulations We find that THetA2 computes the optimal solution orders of magnitude faster than the original THetA (Fig. 2a). Using 30× simulated data, THetA2 demonstrates consistent accuracy at estimating μ (error < 0.05) and copy numbers in the larger tumor population (error < 0.1). In addition, the accuracy in estimating copy numbers improves for the smaller tumor population as its proportion increases (Fig. 2b). Furthermore, THetA2 has increased performance at estimating copy numbers for both populations when considering only longer intervals. For example, when the smaller subpopulation comprises 0.3 of cells in the sample and we consider only intervals longer than 5 Mb, the error rate for both populations drops <0.06. We see similar trends using 7× simulated data, but the lower coverage results in slightly worse copy numbers estimates (Supplementary Fig. S2).

We also directly compare THetA2 with the original THetA on this simulated data. The two-step method enables THetA2 to infer copy numbers for 100% of the genome compared with only 6–11% of the genome with the original THetA (Supplementary Fig. S3). The expanded fraction of the genome analyzed also translates into a substantial increase in the fraction of genome with correct copy number estimates. In our simulations, THetA2 correctly infers copy numbers for the larger and smaller tumor subpopulation in 83–87% and 28–72% more of the genome, respectively, than THetA (Supplementary Fig. S4).

Using the simulated mixture in Figure 3a, we demonstrate that the improved enumeration procedure in combination with the

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

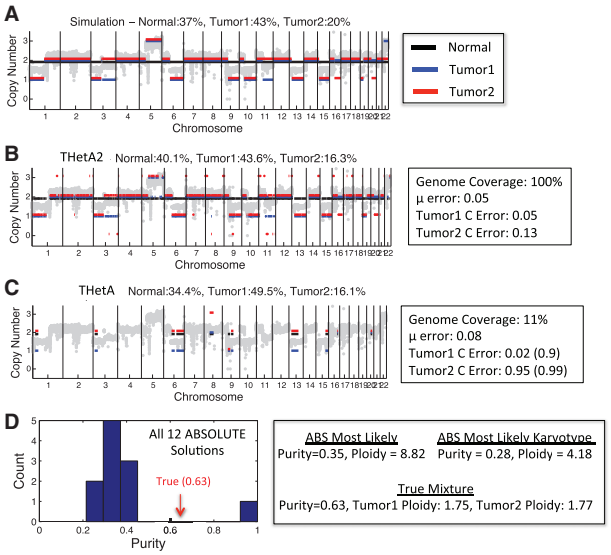

Fig. 3. Comparison of THetA2, THetA and ABSOLUTE on a simulated mixture of three subpopulations. (**a**) True simulated mixture including read depth ratios (dots) within 50 kb bins and the true copy numbers for a mixture of normal cells and two tumor subpopulations. (**b**) Tumor composition inferred by THetA2 using default parameters. Genome coverage is the fraction of the genome for which copy number estimates are made. μ error is the Euclidean distance from the true μ and **C** error is the fraction of the genome with the incorrect copy number estimate. (**c**) Similar to (b) but shows composition inferred by the original THetA and also shows **C** error across both predicted regions and the complete genome. (**d**) (left) Histogram of all 12 purity estimates output by ABSOLUTE. (right) The purity and ploidy reported in the most likely and most likely using only Karyotype solutions output by ABSOLUTE

two-step method can lead to improved estimates of both μ and **C**. On this mixture, THetA2 is able to reconstruct both tumor populations with accuracy >0.87 (Fig. 3b) across the entire genome. However, because THetA is only able to consider a small fraction of the genome, when applied to this mixture, it has increased error at estimating μ and completely misestimates the smaller tumor subpopulation with error of 0.95 across the regions for which copy number estimates were made and error of 0.99 across the whole genome (Fig. 3c). We also applied ABSOLUTE (Carter *et al.*, 2012) to this mixture, run with default parameters, using the same partition of the genome output by BIC-seq (Xi *et al.*, 2011). ABSOLUTE returns a collection of 12 different solutions, each with a different purity and likelihood (Fig. 3d). The most likely solutions returned by ABSOLUTE underestimate purity by at least 0.28 and estimated a tetraploid solution, whereas the true sample has mean ploidy 1.75 and 1.77 in the two tumor populations. Further details are located in the Supplementary Material.

3.1.2 Mixtures with four subpopulations To demonstrate the extensibility of the model to greater numbers of subpopulations, we create a simulated 30× coverage mixture containing four distinct subpopulations. Because of the increased runtime when considering larger numbers of subpopulations, we use an alternative segmentation procedure to reduce the total number of intervals (see Supplementary Material for details). We find that on this simulation, THetA2 was able to estimate μ with 0.05 error, comparable with the accuracy achieved for smaller numbers of subpopulations, and was able to correctly infer copy number for 99.6% of the intervals considered, with the tradeoff of only considering 87.6% of the total genome (Supplementary Fig. S5). Further, we demonstrate how the output of THetA2 changes when the number of subpopulations (n) is fixed below the true number of subpopulations. In particular, we show that in this case THetA2 still provides useful information about the true mixture (Supplementary Fig. S6).

3.2 Extension to WXS data

To demonstrate THetA2's effectiveness on WXS data, we ran THetA2 on Illumina WXS data for the subset of 16 of the 18 tumor samples from TCGA for which WXS data were available (Supplementary Table S2). For each sample, we used both ExomeCNV (Sathirapongsasuti *et al.*, 2011) and EXCAVATOR (Magi *et al.*, 2013) with default parameters to determine an interval partition **I** (see Supplementary Fig. S7 for the complete WXS workflow). If we assume that the tumor sample is a mixture of normal cells and a single tumor population, then the purity estimates obtained by THetA2 on the ExomeCNV and EXCAVATOR interval segmentations were similar for most samples (Supplementary Fig. S8). The two exceptions were two tumor samples where we find subclonal copy number aberrations (for one example see Supplementary Fig. S9). We found that the presence of subclonal aberrations can result in estimates of purity that are artificially low. For example, a segmentation may not accurately distinguish all the present subclonal aberrations. Thus, in the results below, we use the THetA2 solution with higher purity estimate from the ExomeCNV and the EXCAVATOR segmentations. Further details are in the Supplementary Material.

3.2.1 Comparison of THetA2 with ABSOLUTE On most samples, THetA2 purity estimates are within 0.08 of the estimates reported by the ABSOLUTE algorithm (Carter et al., 2012) (Fig. 4a). One example is the glioblastoma sample TCGA-06-0214, for which we estimate purity of 0.67 compared with 0.66 reported by ABSOLUTE. However, although the purity estimates are similar, THetA2 is additionally able to identify two subpopulations of tumor cells, in 46.4 and 20.1% of cells in sample (Fig. 4b) and determine which copy number aberrations are part of each subpopulation.

There are two samples where THetA2 purity estimates are not in agreement with those reported for the ABSOLUTE algorithm (Carter *et al.*, 2012) (Fig. 4a). The first is the ovarian carcinoma sample TCGA-29-1768 where we infer multiple tumor subpopulations and report a purity of 0.87 compared with 0.55 reported by Carter *et al.* (2012). Notably, one of the tumor subpopulations returned by THetA2 is in 54% cells. A possible explanation is that ABSOLUTE reported the purity for the major tumor subpopulation. The second is glioblastoma sample TCGA-06-0188 which we infer to contain two tumor subpopulations consisting of 43.1 and 20.3% cells. In comparison, ABSOLUTE reports that the sample is highly non-clonal and is unable to estimate purity. Our purity estimate of 0.7 is in the range of 0.6–0.8 reported by TCGA histopathology reports. We perform

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

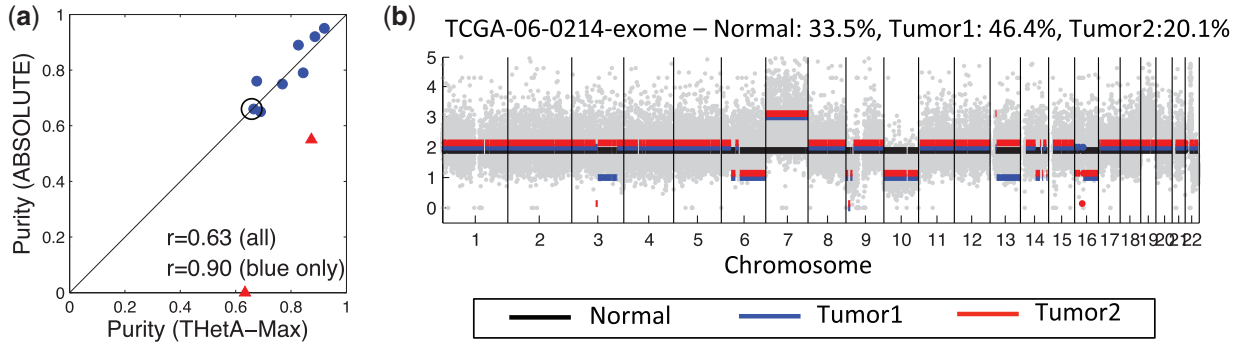

Fig. 4. THetA2 results on WXS data. (**a**) Comparison of purity estimates by THetA2 and ABSOLUTE (as reported in Carter *et al*., 2012). With exception of two outlier samples (red triangles; TCGA-29-1768 and TCGA-06-0188), both approaches predict similar estimates on high purity samples: $r = 0.9$ from Pearson correlation coefficient. Circled sample is TCGA-06-0214, for which both methods agree on sample purity. (**b**) Tumor composition inferred by THetA2 on glioblastoma multiforme sample TCGA-06-0214. Read depth ratios (dots) within 50 kb bins and the copy numbers (for all intervals >2 Mb) inferred by THetA2 for a mixture of normal cells and two tumor subpopulations. We detect rearrangements common to glioblastoma multiforme (Sturm *et al*., 2014) such as amplification of chromosomes 7, and loss of chromosomes 6q, 9p, 10, 13q and 14q

further analysis of this sample and find supporting evidence for our estimated tumor composition (Supplementary Fig. S10). These results demonstrate that consideration of multiple tumor populations may be important for determining tumor purity, especially for samples with large subclonal populations.

3.2.2 Consistency across sequencing platforms To further validate the results of THetA2 on WXS data, we compared results for the 7 of the 18 TCGA samples for which both WGS (including low-pass with 5–7× coverage) and WXS sequence data were available. For WGS samples, we partition the reference genome using the BIC-seq algorithm (Xi *et al*., 2011) run with default parameters (see Supplementary Fig. S7 for WGS workflow). We found that purity estimates for WXS data to be within 0.04 of purity estimates for WGS data for 4 of the 7 samples (Table 1).

We also compare the copy number aberrations predicted for the different subpopulations between the WXS and WGS data using a similarity measure described in Table 1 caption. We find that four of the genomes have ≥ 0.89 similarity under our measure (Table 1) for the major subpopulation. Notably, we find that THetA2 infers three subpopulations for sample TCGA-06-0214 on both WXS and WGS data—selecting the $n = 3$ solution over both $n = 2$ and $n = 4$ for WGS data (see Supplement) and has similarity 0.92 between the data types for the minor subpopulation. We also found similar copy number similarity results using a less stringent measure that only considers copy number state rather than exact copy number value (Supplementary Table S3). These results demonstrate the consistency of THetA2—including the inference of multiple tumor subpopulations—across different types of sequencing data.

3.3 Analysis of highly rearranged and heterogeneous genomes

One of the main advantages of THetA2 is the ability to analyze highly rearranged genomes containing many copy number aberrations in one or more tumor subpopulations. We analyze in further detail several highly rearranged genomes that THetA2 predicted to contain subclonal populations from WGS data.

3.3.1 Low-pass breast cancer samples TCGA-A2-A0EU and TCGA-AO-A0JL We used THetA2 to analyze two breast cancer genomes, TCGA-A2-A0EU and TCGA-AO-A0JL, that were sequenced with low-pass (5-7X) WGS sequencing. These are the most rearranged of the breast cancer genomes that we analyzed—containing many intervals in BIC-seq segmentation (493 and 675 intervals respectively) and more predicted copy number aberrations. We attempted to run ABSOLUTE (Carter *et al*., 2012) on these genomes using the BIC-seq segmentation. However, despite trying a range of values for the parameters, we obtained purity <0.3 for both samples. For comparison, we cite the results reported by Yadav and De (2014) on these samples, using ABSOLUTE and a different segmentation.

In both samples, THetA2 identifies multiple subclonal populations. We infer that breast cancer sample TCGA-A2-A0EU contains normal admixture with two distinct tumor subpopulations, one with 42.7% cells and another with 34.6% cells (Supplementary Fig. S11a). We note that our estimate of tumor purity (0.77) is below the reported histopathology purity of 0.90 for this sample, but closer than the ABSOLUTE estimate of 0.49. We infer that breast cancer sample TCGA-AO-A0JL contains normal admixture with two distinct tumor subpopulations, one with 57.0% cells and another with 30.5% cells (Supplementary Fig. S11b). Despite being the most rearranged of the breast cancer genomes analyzed, our estimated tumor purity of 0.88 is near the reported histopathology value of 0.80. In comparison, ABSOLUTE inferred purity of 0.50 for this sample. We are also able to identify a number of clonal and subclonal chromosome arm level events for both genomes (see Supplemental Material), as well as many other small events, thus demonstrating that THetA2 can analyze highly rearranged genomes with low-coverage WGS sequencing data.

3.3.2 Lung squamous cell sample TCGA-56-1622 We ran THetA2 on a highly rearranged lung squamous sample TCGA-56-1622, containing 2847 intervals in the segmentation.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 1. Comparison of THetA2 results on WGS and whole-exome data

Sample	Path.	ABS	WGS Purity (# populations)	WXS Purity (# populations)	Overlap	CNA Sim
TCGA-06-0185	0.95	0.89	0.87 (3)	0.83 (2*)	0.97	0.91
TCGA-06-0188	**0.6–0.8**	**NA**	**0.70 (3)**	**0.63 (3)**	0.96	0.79, 0.62
TCGA-06-0214[a]	0.25–0.8	0.66	0.67 (3)	0.67 (3)	0.96	0.97, 0.92
TCGA-56-1622	0.9	–	0.68 (3)	0.78 (3)	0.96	0.89, 0.57
TCGA-A2-A0EU	0.9	–	0.77 (3)	0.90 (3)	0.91	0.61, 0.22
TCGA-AO-A0JJ	0.8	–	0.52 (3)	0.52 (2)	0.85	0.67
TCGA-BH-A0W5	0.7	–	0.51 (2*)	0.54 (2*)	0.98	0.97

Notes. Path. are purity estimates reported in TCGA histopathology reports. ABS are ABSOLUTE purity estimates reported by Carter *et al.* (2012) (samples marked with '–' do not have published purity estimates from ABSOLUTE). WGS Purity, WXS Purity and # populations are values predicted by THetA2. Overlap is $\frac{\mathbf{I}^*}{|\mathbf{I}_{WGS}| \cup |\mathbf{I}_{WXS}|}$ where $\mathbf{I}_{WGS}$ and $\mathbf{I}_{WXS}$ are the interval partitions for the WGS and whole-exome data, respectively, and $\mathbf{I}^*$ is the set of intervals longer than 100 kb contained in both $\mathbf{I}_{WGS}$ and $\mathbf{I}_{WXS}$. CNA Sim is the fraction of $\mathbf{I}^*$ where the copy number estimates are the same between the two data types. *Indicates that the sample did not pass the criteria to be considered for multiple tumor populations (see Supplemental Material). [a]For sample TCGA-06-0214, WGS data were aligned to hg18 and WXS data aligned to hg19. See Supplementary Table S3 for purity estimates across all genomes analyzed and results using an additional similarity metric. Bolding indicates the sample for which THetA2 can estimate purity, but ABSOLUTE reports as highly non-clonal and is unable to estimate purity.

We note that this genome is so fragmented that ABSOLUTE (Carter *et al.*, 2012) does not attempt to estimate tumor purity when run with default parameters. Moreover, this sample has so many copy number changes that SNV-based algorithms (Andor *et al.*, 2014; Roth *et al.*, 2014) would have extreme difficulty in defining regions of normal copy number to analyze. THetA2 infers that sample contains normal admixture with two distinct tumor subpopulations, one with 50.1% cells and another with 18.1% cells (Fig. 5a). Using the new two-step procedure, THetA2 also identifies many smaller copy number aberrations (Supplementary Fig. S12) and we find that the read depth predicted using our reconstruction closely matches the observed read depth (Fig. 5b).

We examine this sample in further detail using B-allele frequency (BAF) information not used by THetA2. We constructed a virtual SNP array defining the BAF at a known germ line SNP to be the fraction of reads containing the minor allele as described in Oesper *et al.* (2013). In diploid regions of the genome that have not undergone any copy number changes, we expect that the BAFs for germ line heterozygous SNPs to be near a value of 0.5, as approximately half of the reads should contain the B-allele. In a pure tumor sample a deletion of a segment on a single chromosome will lead to a loss of heterozygosity (LOH) and BAFs at 0 or 1 in a symmetric *double banded pattern* centered around 0.5. As the sample become less pure (i.e. more admixture by normal cells), the double banded pattern will shift closer to 0.5.

In many of the regions where THetA2 predicted a clonal deletion (i.e. in all subpopulations), such as chromosomes 3, 5q and 18 (Fig. 5b), we observe that the BAFs cluster near 0 and 1, as expected for a deletion occurring in a majority of cells in the sample. Similarly, we find that the shifts in BAF are consistent with THetA2's predictions of subclonal deletions in 50.0 and 18.1% of cells (Fig. 5b). On chromosome 1p, we observe a discrepancy between THetA2's predictions and BAF. THetA2 predicts that 1p is a clonal deletion; however, the BAFs are clustered tightly around 0.5, indicating an equal number of both parental copies of this region in the tumor sample. One explanation is that 1p is homozygously deleted in one of the tumor subpopulations, rather than a heterozygous deletion in both subpopulations, which would keep the balance of the parental copies of 1p in the tumor sample.

3.4 Using BAFs

For glioblastoma sample TCGA-06-0145, THetA2 outputs two possible $(\mathbf{C}, \mu)$ pairs using only read depth – one largely haploid and one largely diploid. We apply our probabilistic model of BAFs and find that the diploid reconstruction, which includes rearrangements characteristic to glioblastoma such as amplification of chr7 and deletion of chr10 (Sturm *et al.*, 2014), is determined to be the more likely tumor composition (Supplementary Fig. S13).

4 DISCUSSION

We introduced an algorithm to infer tumor composition – of highly rearranged genomes from WGS (high or low coverage) or WXS DNA sequencing data. These are implemented as improvements to our THetA algorithm. The THetA2 algorithm is able to analyze highly rearranged, aneuploid samples that are beyond the scope of existing algorithms that infer tumor heterogeneity. A recently published comparison of algorithms for inferring tumor purity (Yadav and De, 2014) showed that our original THetA algorithm (Oesper *et al.*, 2013) performed well, but sometimes underestimated tumor purity when run to only consider normal cells and one tumor subpopulation. We argue that this purity underestimation is likely a result of not directly considering all tumor subpopulations in the sample. In every sample that we analyzed with the new algorithm, tumor purity was higher when considering multiple tumor subpopulations.

Although the improved THetA2 presented here is useful on a wide range of sequencing data from different tumors, some limitations remain. First, THetA2 is unable to distinguish tumor subpopulations that are not differentiated by copy number aberrations. As copy number aberrations are ubiquitous in most solid tumors (Albertson *et al.*, 2003), we expect that THetA2 will be applicable to many genomes. However, for some diploid tumors, SNV analysis is preferable. Incorporation of additional

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

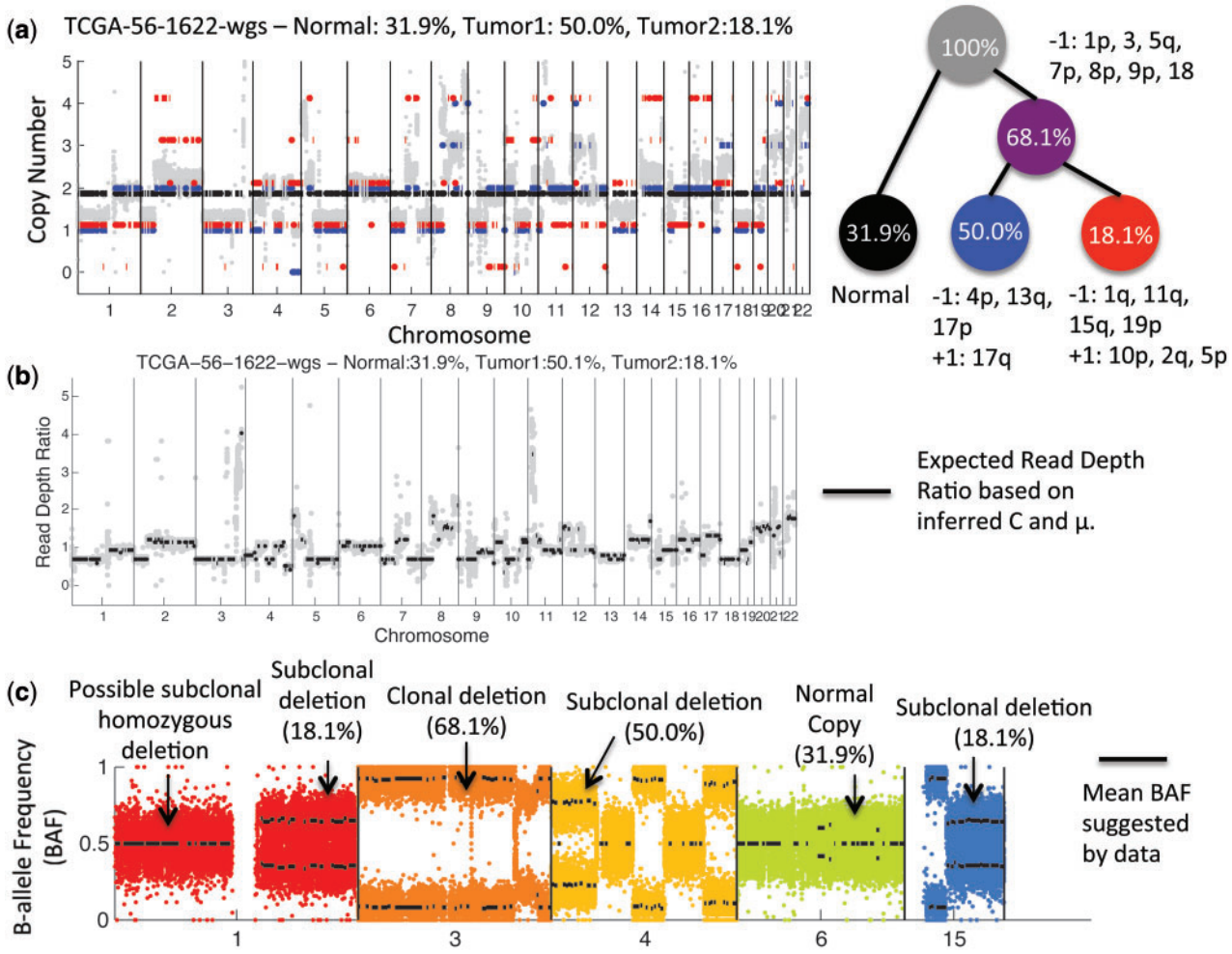

Fig. 5. Analysis of squamous cell lung cancer sample TCGA-56-1622. (**a**) (Left) Read depth ratios (gray dots) within 50 kb bins and the inferred copy number aberrations calculated by THetA2 when the tumor is considered to be a mixture of three subpopulations: normal cells (black) and two tumor subpopulations (blue and red). (Right) A reconstruction of the tumor mixture along with ancestral clonal population (purple) with the inferred aberrations and estimated fraction of cells in each population (see Supplementary Material). (**b**) Expected read depth ratios (see Supplementary Material) for intervals longer than 2 Mb based on inferred **C** and μ (black) overlaid on observed read depth ratios (gray dots). **(c)** Virtual SNP array showing BAFs at germ line SNPs on indicated chromosomes and the mean BAF in each segment (see Supplementary Material)

information, such as BAFs for somatic and germ line SNPs, into the model (Andor *et al.*, 2014; Roth *et al.*, 2014) may also increase the scope of samples for which THetA2 is applicable.

Second, while the improvements presented here greatly decrease the computational burden of the algorithm when considering multiple tumor populations, the algorithm remains exponential in the size of the interval partition of the reference genome—making it impractical to infer tumor composition with more than a handful of subpopulations in many cases. Identification of further mathematical restrictions to the domain of interval count matrices, or use of sampling techniques in place of complete enumeration are future avenues of investigation, which may prove useful in this respect. Additionally, when considering multiple tumor subpopulations, the quality of the results is limited by features of the data including the presence of copy number aberrations that distinguish subpopulations as well as the number of sequence reads available to identify these aberrations. The later is a function of sequencing coverage, aberration length and proportion of cells that have the aberration.

While the limited number of tumor subpopulations that THetA2 analyzes may not be sufficient to fully analyze tumor progression, THetA2's ability to recover subpopulations with relatively low-coverage sequencing data can provide some insight into tumor subpopulations in cases where methods that rely on high-coverage data (Jiao *et al.*, 2014; Roth *et al.*, 2014) cannot. Combining THetA2s output with other methods that do explicitly consider the phylogenetic history of a tumor such as Jiao *et al.* (2014) or Hajirasouliha *et al.* (2014) may prove a useful avenue of exploration.

The two-step procedure introduced here allows us to infer subclonal copy number aberrations at much smaller scales. However, some care is required to avoid overfitting the data,

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

particularly for small, subclonal copy number aberrations where GC bias or other sequencing artifacts may lead to incorrect inferences. Incorporating more sophisticated segmentation procedures that account for such effects and appropriately scale read counts (Benjamini and Speed, 2012) are useful directions for future research.

Finally, this work focuses on the important first step of quantifying intra-tumor heterogeneity from a single mixed tumor sample. Downstream analysis including the clinical and functional impact of the inferred tumor composition is an important area for future work.

5 CONCLUSION

We present a new algorithm, THetA2, to infer the composition of a tumor sample—including both the percentage of normal admixture and the fraction and content of one *or more* of tumor subpopulations that differ by copy number aberrations. The new algorithm builds on our THetA algorithm (Oesper *et al*., 2013), and includes several improvements that allow us to analyze highly rearranged genomes from WGS (high and low coverage) or WXS sequencing data. In addition, the new algorithm is orders of magnitude faster and allows us to use BAFs to distinguish between different reconstructions.

ACKNOWLEDGEMENTS

The results published here are in whole or part based on data generated by TCGA research network established by the National Cancer Institute and the National Human Genome Research Institute.

Funding: This work is supported by a National Science Foundation (NSF) graduate research fellowship DGE0228243 (to L.O.); National Science Foundation (NSF) career award CCF-1053753 (to B.J.R.); and grant RO1HG005690 from the National Institutes of Health to B.J.R. B.J.R. is also supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund, an Alfred P Sloan Research Fellowship.

Conflict of interest: none declared.

REFERENCES

Albertson,D.G. *et al.* (2003) Chromosome aberrations in solid tumors. *Nat. Genet.*, **34**, 369–376.

Andor,N. *et al.* (2014) Expands: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics*, **30**, 50–60.

Benjamini,Y. and Speed,T.P. (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.*, **40**, e72.

Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Cancer Genome Atlas Research Network. (2013) Genomic and epigenomic landscapes of adult *de novo* acute myeloid leukemia. *N. Engl. J. Med.*, **368**, 2059–2074.

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

Ding,L. *et al.* (2010) Analysis of next-generation genomic data in cancer: accomplishments and challenges. *Hum. Mol. Genet.*, **19**, R188–R196.

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

Greaves,M. and Maley,C.C. (2012) Clonal evolution in cancer. *Nature*, **481**, 306–313.

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.

Hajirasouliha,I. *et al.* (2014) A combinatorial approach for analyzing intra-tumor heterogeneity from high-throughput sequencing data. *Bioinformatics*, **30**, i78–i86.

Jiao,W. *et al.* (2014) Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics*, **15**, 35.

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

Magi,A. *et al.* (2013) Excavator: detecting copy number variants from whole-exome sequencing data. *Genome Biol.*, **14**, R120.

Meyerson,M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.*, **11**, 685–696.

Mullighan,C.G. *et al.* (2008) Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, **322**, 1377–1380.

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

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

Roth,A. *et al.* (2014) Pyclone: statistical inference of clonal population structure in cancer. *Nat. Methods*, **11**, 396–398.

Sathirapongsasuti,J.F. *et al.* (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: exomecnv. *Bioinformatics*, **27**, 2648–2654.

Sturm,D. *et al.* (2014) Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge. *Nat. Rev. Cancer*, **14**, 92–107.

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.

Yadav,V.K. and De,S. (2014) An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. In: *Brief. Bioinform*, in press. [Epub ahead of print, doi:10.1093/bib/bbu002, February 20, 2014].

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016