CLImAT: accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data

Zhenhua Yu[1], Yuanning Liu[1], Yi Shen[1], Minghui Wang[1,2] and Ao Li[1,2,*]

[1]School of Information Science and Technology and [2]Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH230027, China

Associate Editor: Michael Brudno

ABSTRACT

Motivation: Whole-genome sequencing of tumor samples has been demonstrated as an efficient approach for comprehensive analysis of genomic aberrations in cancer genome. Critical issues such as tumor impurity and aneuploidy, GC-content and mappability bias have been reported to complicate identification of copy number alteration and loss of heterozygosity in complex tumor samples. Therefore, efficient computational methods are required to address these issues.

Results: We introduce CLImAT (CNA and LOH Assessment in Impure and Aneuploid Tumors), a bioinformatics tool for identification of genomic aberrations from tumor samples using whole-genome sequencing data. Without requiring a matched normal sample, CLImAT takes integrated analysis of read depth and allelic frequency and provides extensive data processing procedures including GC-content and mappability correction of read depth and quantile normalization of B-allele frequency. CLImAT accurately identifies copy number alteration and loss of heterozygosity even for highly impure tumor samples with aneuploidy. We evaluate CLImAT on both simulated and real DNA sequencing data to demonstrate its ability to infer tumor impurity and ploidy and identify genomic aberrations in complex tumor samples.

Availability and implementation: The CLImAT software package can be freely downloaded at http://bioinformatics.ustc.edu.cn/CLImAT/.

Contact: aoli@ustc.edu.cn

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

1 INTRODUCTION

Various aberrations such as amplification, deletion and translocation of segmental regions are common features of cancer genomes and play an important role in tumorigenesis and progression (Albertson *et al.*, 2003; Stratton *et al.*, 2009). It is reported that dysfunction of oncogene and tumor suppressor gene is related to frequent genomic aberrations (Bignell *et al.*, 2010; Stephens *et al.*, 2009; Stratton *et al.*, 2009). Genomic aberrations in specific regions have been used as an indicator of aggressiveness of cancer and clinical outcome (Carén *et al.*, 2010; Suzuki *et al.*, 2000). Genome-wide copy number alteration

(CNA) and loss of heterozygosity (LOH) are two essential features of cancer genomes, and accurate detection of these abnormalities is a crucial step to assess genomic aberrations and cancer-related genes. Experimental technologies are now available for high-throughput profiling of genome-wide aberrations in tumor samples, such as array comparative genomic hybridization (Park, 2008), single nucleotide polymorphism (SNP) genotyping array (Li *et al.*, 2011; Peiffer *et al.*, 2006) and more recently, whole-genome sequencing (WGS) technology for massively parallel sequencing of DNA (Mardis, 2008; Metzker, 2009; Morozova and Marra, 2008; Schuster, 2007). By allowing for comprehensive analysis of genomic aberrations in cancer genomes, WGS has been demonstrated as an efficient platform for studies of human cancers (Metzker, 2009).

Although several computational approaches have been proposed for assessing genomic aberrations from tumor sequencing data (Boeva *et al.*, 2011, 2012; Carter *et al.*, 2012; Gusnanto *et al.*, 2012; Ha *et al.*, 2012; Mayrhofer *et al.*, 2013; Sathirapongsasuti *et al.*, 2011; Xi *et al.*, 2011), most of these methods do not effectively address the critical issues encountered in interpreting complex tumor samples. For example, tumor samples are often infiltrated with normal stroma, resulting in inevitable contamination of normal DNA and dilution of somatic aberration signals (Boeva *et al.*, 2011, 2012; Gusnanto *et al.*, 2012; Ha *et al.*, 2012; Mayrhofer *et al.*, 2013). Impurity of tumor sample can significantly alter WGS data; and therefore, complicates genomic aberration detection, especially when normal cells dominate in tumor samples. Recent studies, such as FREEC (Boeva *et al.*, 2011, 2012) and APOLLOH (Ha *et al.*, 2012), have been proposed to address this issue. FREEC constructs copy number and B-allele frequency (BAF) profiles to detect CNA and allelic content in cancer genomes, with optional correction for tumor impurity. APOLLOH is designed for LOH detection using tumor-normal paired samples, and the issue of tumor impurity is addressed by a two-component mixture model for allelic read counts.

In addition to tumor impurity, tumor aneuploidy is another critical issue in genomic aberration detection, which is caused by various numerical and structural chromosomal abnormalities frequently observed in cancer genome (Carter *et al.*, 2012). Although APOLLOH introduces a delicate statistical model to eliminate the effect of tumor impurity, it does not take account of tumor aneuploidy in modeling and analyzing tumor WGS data. To handle aneuploid tumor samples, FREEC provides

---

*To whom correspondence should be addressed.

an option for users to input tumor ploidy. Currently, automatic correction for tumor aneuploidy using WGS data still remains a challenging task. Theoretically, it is often difficult to determine the actual ploidy of cancer cells by sequencing technology (Gusnanto *et al.*, 2012). In some particular cases, somatic aberration signals could present similar characteristics among genomes of different ploidy (Gusnanto *et al.*, 2012; Oesper *et al.*, 2013), which makes it hard to accurately estimate the tumor ploidy. It should be pointed out that, complicated interpretation of WGS data are even more challenging in tumor samples confounded by both tumor impurity and aneuploidy, as they usually cannot be solved separately (Oesper *et al.*, 2013).

So far, only a few algorithms have been proposed for analyzing WGS data of impure tumor samples with aneuploidy (Carter *et al.*, 2012; Gusnanto *et al.*, 2012; Mayrhofer *et al.*, 2013; Oesper *et al.*, 2013). For example, CNAnorm (Gusnanto *et al.*, 2012) uses a mixture normal distribution for ratios of tumor-normal read counts to correct tumor impurity and aneuploidy. However, CNAnorm assumes that the most common component in the normal mixture is diploid, which may not hold for aneuploid tumor samples. Moreover, it cannot detect LOH in cancer genomes. Another approach, ABSOLUTE (Carter *et al.*, 2012), is originally introduced to detect CNA from SNP array data by inferring tumor impurity and ploidy. Although it can be adapted to analyze DNA sequencing data, a previous study shows that the underlying statistical models used by ABSOLUTE do not comprehensively describe the characteristics of DNA sequencing data and therefore may sometimes gravely misestimate the tumor impurity and ploidy (Oesper *et al.*, 2013). Recently, Markus *et al.* introduced a novel method called Patchwork (Mayrhofer *et al.*, 2013) for allele-specific copy number analysis of sequenced tumor tissue in consideration of tumor impurity and tumor aneuploidy, which requires intermediate arguments determined by users. In addition, it is noteworthy that another method called THetA was proposed recently to analyze tumor sequencing data (Oesper *et al.*, 2013). THetA mainly focuses on the inference of cancer subclones in heterogeneous tumor samples and cannot detect LOH in cancer genomes, as it only utilizes read count data. Therefore, it is essential to develop an efficient approach for analysis of tumor sequencing data by comprehensively addressing the challenge of tumor impurity and aneuploidy.

In this study, we present a novel method called CLImAT (CNA and LOH Assessment in Impure and Aneuploid Tumors) to detect genomic aberrations with automatic correction for both tumor impurity and aneuploidy. Without requiring a matched normal sample, CLImAT fully explores both read depth (RD) and allele frequency derived from tumor WGS data, and provides extensive data processing procedures including elimination of sequencing/mapping bias and quantile normalization (QN) of allele frequency data. By adopting an integrated Hidden Markov Model (HMM) that quantitatively delineates tumor impurity and ploidy, CLImAT provides accurate identification of various kinds of genomic aberrations even for highly impure tumor samples with aneuploidy. We apply CLImAT to both simulated and real tumor data, and the results demonstrate the superior performance of CLImAT in analysis of genomic aberrations using tumor WGS data.

2 METHODS

2.1 Simulated data by sampling reads from tumor-normal mixture

To assess the performance of CLImAT for complex tumor samples, we generate simulated tumor samples with different impurity and ploidy. Similar to the procedure proposed previously (Duan *et al.*, 2013), virtual tumor-normal mixture experiment is performed on chromosome 20 of human reference genome (NCBI build 36, hg18) by sampling reads from a control genome and a test genome with tumor impurity ranging from 0 to 0.9 with 0.1 increments (Supplementary Methods). The test genome is constructed by dividing the reference genome into 20 non-overlapping and equally sized segments, which are randomly assigned with particular kinds of genomic aberrations (Supplementary Figure S1). Sampled reads from both control and test genome are mapped to the reference using Bowtie (Langmead *et al.*, 2009) with default parameters. BAM files and pileups are generated by SAMtools (Li *et al.*, 2009). For each combination of predetermined tumor impurity and ploidy (diploidy, triploidy and tetraploidy), three BAM files are generated at $10\times$, $30\times$ and $60\times$ sampled coverage, respectively. The average copy number (ACN) is 2.48, 3.19 and 4.00 for diploid, triploid and tetraploid tumor samples, respectively. By this way, we generate totally 90 simulated tumor samples for comprehensive evaluation of prediction performance. Detailed information about construction of test genomes and read sampling process is provided in Supplementary Methods.

2.2 Real sequencing data of tumor samples

WGS data from three unpaired primary triple negative breast cancer (TNBC) samples described in a previous study (Shah *et al.*, 2012) are adopted in this study. Each sample was sequenced at $\sim30\times$ coverage on the Life/ABI SOLID sequencing platform. Reads were mapped to the reference genome hg18 using BioScope. The data was downloaded from European Genome-Phenome Archive (EGA) with accession number EGAS00001000132.

2.3 Pipeline of CLImAT

The pipeline of CLImAT is depicted in Supplementary Figure S2. RD used in this study is retrieved from the BAM file using SAMtools (Li *et al.*, 2009) and is further processed to correct GC and mappability bias. BAF signals of all known SNPs in dbSNP database (Sherry *et al.*, 2001) are normalized to eliminate allelic bias. Both RD and BAF signals are modeled by an integrated HMM for identifying genomic aberrations, including CNA and LOH, and estimating tumor impurity and ploidy.

2.4 Deriving RD and BAF from tumor WGS data

In this study, RD is obtained by counting the reads with starting position within a 1000-bp window centered at each SNP. For BAF, we count the reads that override the SNP and the reads with non-reference base at the corresponding SNP as B allelic read count. Thus, BAF of the SNP is calculated as the proportion of B allelic reads. Consisted with the procedure adopted in previous study (Ha *et al.*, 2012), data filtering is taken to further eliminate positions that have either low depth (<10 reads for 30/60× coverage and <5 reads for 10× coverage) or high depth (>250 reads).

2.5 Signal correction and normalization

GC-content and mappability may heavily affect RD signals and bring bias to CNA detection. Therefore, as the first step we perform a correction procedure to remove the bias in RD signals. For each window used in RD calculation, GC-content is measured by calculating the G + C percentage, and the mappability score is defined as the average of mappability values. The mappability file used in this study

was obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDCC/wgEncodeMapability/. Following the procedure used in (Yoon *et al.*, 2009), we scale GC-content and mappability score to integer values between 0 and 100, and perform correction of RD signals using the following equation:

$$rdc_i = rd_i \cdot \frac{m}{m_X} \qquad (1)$$

where rdc_i is the corrected RD of the ith window, rd_i is the original RD of the ith window, m is the overall median RD of all the windows and m_X is the median RD of the windows that have the same GC-content and mappability values as the ith window.

It has been reported that loss of reads (LOR) issue happens in the alignment step of sequencing data processing (Kim *et al.*, 2013). Indeed, most aligners, such as BioScope and BWA (Li and Durbin, 2009), have the preference for aligning reads to reference allele over alternative allele. Reads sequenced from alternative chromosome are inclined to be discarded because of mismatches between reference sequence and read sequence, leading to asymmetrical distribution of allelic frequencies. Therefore, it is necessary to normalize the BAF data for better estimation of LOH and other related parameters, including tumor impurity and ploidy. We adopt an efficient QN (Bolstad *et al.*, 2003) procedure to address this issue (Supplementary Methods).

2.6 Integrated HMM

We propose an integrated HMM that takes RD and BAF data as input. Supplementary Table S1 shows the hidden states defined in the HMM with detailed description of each HMM state regarding copy number, tumor genotype mutated from normal cell genotype and zygosity status. Tumor and normal genotype pairs are used to give a detailed view of the intrinsic relationship between genotypes of tumor and normal cells admixed in tumor samples. For example, (AAAB, AB) is the case that tumor genotype 'AAAB' is derived from normal cell genotype 'AB'.

2.6.1 Emission probabilities Aligning a read to a genomic position can be treated as a Bernoulli trial (Ha *et al.*, 2012). Thus, given the number of reads that override a SNP position, the number of reads that have non-reference base at corresponding SNP position is modeled by a binomial distribution. Suppose B allelic read count and total read count of the ith SNP are b_i and N_i, respectively, the observation probability for hidden state c can be formulated as:

$$p(b_i \mid w_s, N_i, c) = \sum_{k=1}^{g_c} p_0(k) \binom{N_i}{b_i} \left(\frac{z_{ck}}{y_c}\right)^{b_i} \left(1 - \frac{z_{ck}}{y_c}\right)^{(N_i - b_i)} \qquad (2)$$

where g_c is the number of tumor genotypes included in state c. The ACN y_c and average B allele copy number z_{ck} for state c are defined as:

$$y_c = n_s \cdot w_s + n_c \cdot (1 - w_s) \qquad (3)$$

$$z_{ck} = n_s \mu_s w_s + n_c \mu_{ck}(1 - w_s) \qquad (4)$$

where n_s is the normal copy number and is fixed to 2 in this study, n_c is the tumor copy number in state c and w_s is the level of tumor impurity. u_s denotes expected BAF value of normal cells and is fixed to 0.5, and u_{ck} represents the expected BAF value of the kth tumor genotype in state c.

Taking into account the over-dispersed distribution of RD values (Anders and Huber, 2010), we use a negative binomial distribution to model RD signals. Suppose that RD of the ith SNP is d_i, the observation probability for hidden state c can be formulated as:

$$p(d_i \mid w_s, o, \lambda, p_c, c) = \frac{\Gamma\left(d_i + \frac{\lambda_c(1-p_c)}{p_c}\right)}{\Gamma(d_i+1)\Gamma\left(\frac{\lambda_c(1-p_c)}{p_c}\right)} (1 - p_c)^{\frac{\lambda_c(1-p_c)}{p_c}} p_c^{d_i} \qquad (5)$$

where Γ is the gamma function and p_c is a parameter of negative binomial

distribution defined as the probability of success. The average read count λ_c for state c is defined as:

$$\lambda_c = \frac{y_c}{2}\lambda + o \qquad (6)$$

where λ is the mean value of copy neutral read count and varies with respect to tumor ploidy change. o accounts for background RD noise resulted from sequencing error and wrongly mapped reads.

2.6.2 EM algorithm for parameter estimation We employ expectation maximization (EM) algorithm for HMM training and parameter estimation. In the expectation step, the expectation of the partial log-likelihood of BAF is formulated as:

$$
E(LL_b) = \sum_{i=1}^{N}\sum_{c=1}^{C}\gamma_i(c)\sum_{k=1}^{g_c}\left(\log\left(\binom{N_i}{b_i}\right)\right.
$$
$$
\left. + b_i\log\left(\frac{z_{ck}}{y_c}\right) + (N_i - b_i)\log\left(1 - \frac{z_{ck}}{y_c}\right)\right) \qquad (7)
$$

where $\gamma_i(c)$ represents the posterior probability that the ith SNP is in state c and is calculated by the forward-backward algorithm (Rabiner, 1989). Similarly, the expectation of the partial log-likelihood function of RD can be formulated as:

$$
E(LL_d) = \sum_{i=1}^{N}\sum_{c=1}^{C}\gamma_i(c)\log\left(p(d_i \mid w_s, o, \lambda, p_c, c)\right)
$$
$$
= \sum_{i=1}^{N}\sum_{c=1}^{C}\gamma_i(c)\left(\begin{array}{l}\log\left(\Gamma\left(d_i + \frac{\lambda_c(1-p_c)}{p_c}\right)\right) + d_i\log(p_c)\\[2mm] + \frac{\lambda_c(1-p_c)}{p_c}\log(1-p_c)\\[2mm] -\log(\Gamma(d_i+1)) - \log\left(\Gamma\left(\frac{\lambda_c(1-p_c)}{p_c}\right)\right)\end{array}\right) \qquad (8)
$$

In the maximization step of the EM algorithm, we use Newton algorithm to update the parameters in emission probabilities. For example, during iteration n we update the parameter w_s by using the following formula:

$$w_{s,n+1} = w_{s,n} - \frac{\frac{\partial E(LL_b)}{\partial w} + \frac{\partial E(LL_d)}{\partial w}}{\frac{\partial^2 E(LL_b)}{\partial^2 w} + \frac{\partial^2 E(LL_d)}{\partial^2 w}} \qquad (9)$$

All the parameters are iteratively updated until the EM algorithm converges. Copy number and tumor genotype for each SNP are determined by the hidden state with the largest conditional probability. In addition, post-processing is performed for copy number annotation of highly amplified regions (copy number >7) according to the mean RD values of all SNPs within these regions (Supplementary Methods). To evaluate the reliability of CLImAT results, we also calculate a reliability score for each region to measure how well the data fit to the model (Supplementary Methods).

3 RESULTS

3.1 Correction and normalization of RD and BAF signals

We assess the performance of GC-content and mappability correction and plot the distribution of RD with respect to GC-content and mappability score for 1–3 copies (Supplementary Figure S3). Before correction, RD signals demonstrate a unimodal distribution with respect to GC-content and are positively correlated with mappability scores. After correction both

GC-content and mappability bias is significantly eliminated. Further investigation suggests the order of GC-content and mappability correction performed to tumor WGS data affects the final results and simultaneous correction for both GC-content and mappability bias shows better performance (Supplementary Figure S4).

It is observed that owing to LOR issue BAF plots of tumor samples display asymmetrical bands positioned around 0.5 (Supplementary Figure S5A). The altered distribution of BAF signals seriously hampers accurate identification of genomic aberrations in tumor samples. After applying the QN procedure, BAF signals are largely corrected with symmetrical bands positioned around 0.5 (Supplementary Figure S5B).

3.2 Appling CLImAT to simulated data

We apply CLImAT to simulated tumor data, and the results are shown in Supplementary Figure S6. The RD and BAF signals vary dramatically with increased tumor impurity for both diploid and triploid genomes. Especially, with 90% normal cells admixed in the tumor sample, both RD and BAF signals for aberrant regions are dramatically attenuated. CLImAT correctly detects all aberrant regions and provides CNA and LOH prediction with reasonable performance.

3.2.1 Estimation of tumor impurity and ploidy We examine tumor impurity estimated by CLImAT and ABSOLUTE (Carter *et al.*, 2012) on simulated data, and the results of tumor samples at 60× coverage are shown in Figure 1A. CLImAT accurately estimates tumor impurity from 0 to 0.9 with significant correlation with the ground truth (correlation coefficient = 0.999, $P = 6.24 \times 10^{-21}$ for diploid samples, correlation coefficient = 0.999, $P = 2.75 \times 10^{-12}$ for triploid samples and correlation coefficient = 0.999, $P = 1.42 \times 10^{-11}$ for tetraploid samples), indicating CLImAT can precisely recover the

proportion of cancer cells in tumor samples. In contrast, the performance of ABSOLUTE is not optimal and sometimes the results obviously deviate from the ground truth. Similar results are observed for simulated samples at 30× coverage (Supplementary Figure S7). To assess the performance of tumor ploidy estimation, we calculate the ACNs for simulated samples from the results of ABSOLUTE and CLImAT. As shown in Figure 1B, CLImAT exhibits prominent advantage over ABSOLUTE in estimating tumor ploidy. For example, CLImAT correctly identify all diploid samples at 30× coverage as diploidy, whereas ABSOLUTE tends to assign them as hyperploidy. Taken together, these results suggest that CLImAT can efficiently estimate both tumor impurity and tumor ploidy from complicated tumor samples.

3.2.2 LOH and CNA detection We adopt the performance evaluation procedure proposed in APOLLOH (Ha *et al.*, 2012), in which all the calls of the informative (heterozygous) positions are used as the golden standard to compare the abilities of different computational methods in detecting genomic aberrations. Accordingly, the CNA/LOH calls of heterozygous positions pre-determined in unpaired simulated data are treated as the ground truth. We use the standard way for performance evaluation by separately comparing the results of the computational methods investigated in this study to the ground truth in terms of sensitivity and specificity (more details of performance evaluation are provided in Supplementary Methods). The LOH detection results of three computational methods, FREEC (Boeva *et al.*, 2012), SNVMix (Goya *et al.*, 2010) and CLImAT, are shown in Figure 2. For diploid tumor samples (Fig. 2A), FREEC shows high specificity in all tests and the sensitivity is generally good at medium tumor impurity levels. Compared with the other methods, CLImAT demonstrates strong robustness to tumor impurity and maintains high sensitivity (>0.99) across all tumor samples with impurity level <0.9.
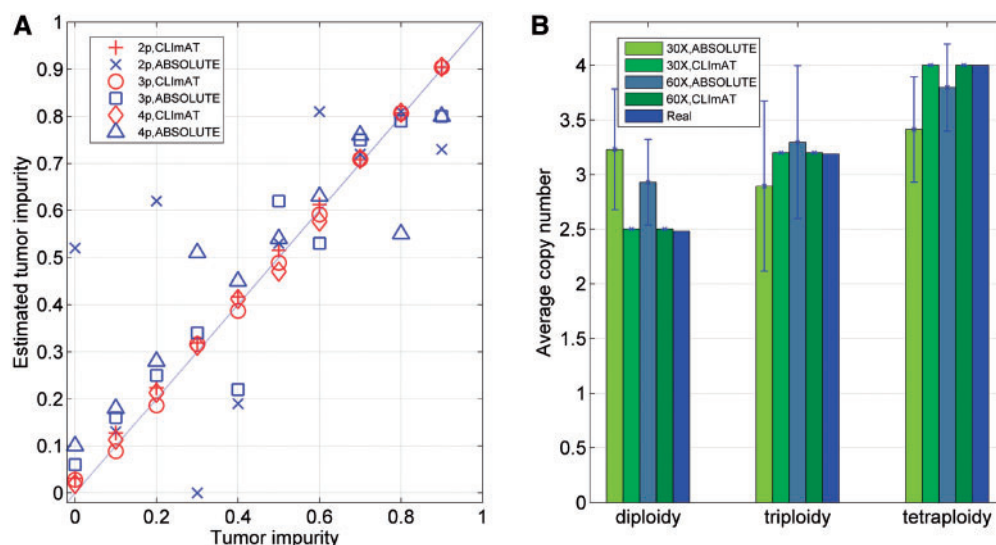

Fig. 1. Estimated tumor impurity and ACN of simulated samples. (**A**) Tumor impurity estimated by ABSOLUTE and CLImAT for samples at 60× coverage. 2p: diploid samples, 3p: triploid samples, 4p: tetraploid samples. (**B**) ACNs estimated by ABSOLUTE and CLImAT for simulated samples. Each bar shows the mean and standard deviation of estimated ACNs obtained from 10 samples with tumor impurity ranging from 0 to 0.9
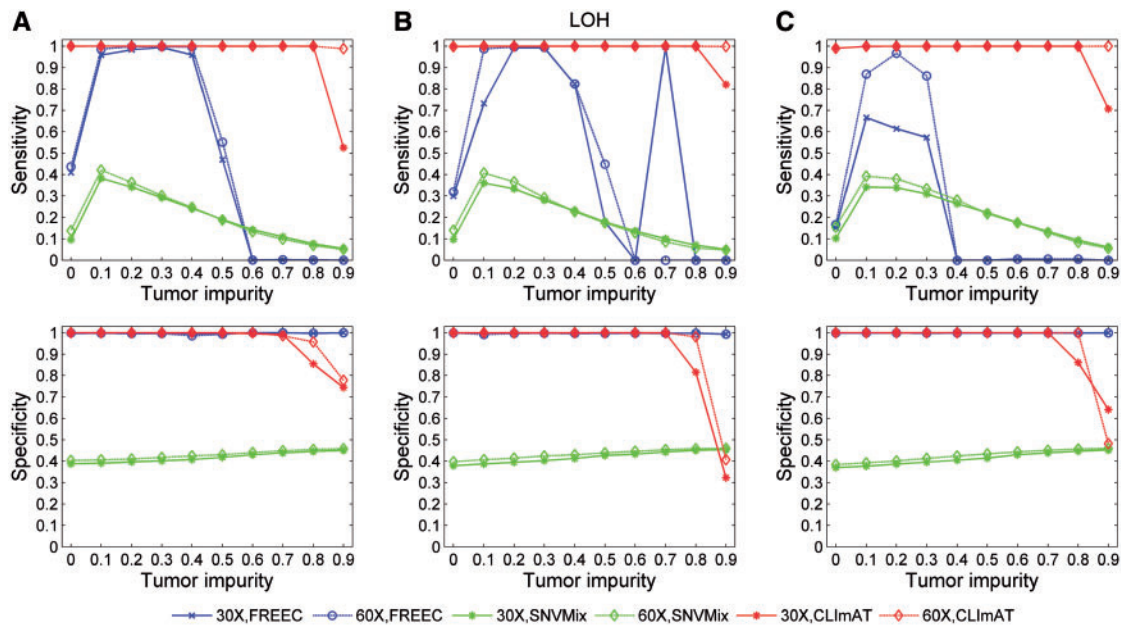
Fig. 2. LOH detection performance of FREEC, SNVMix and CLImAT on unpaired simulated data. (**A**) Results for diploid samples. (**B**) Results for triploid samples. (**C**) Results for tetraploid samples

It also keeps consistent high specificity with respect to different tumor impurity levels (<0.8) and sampled coverage. Similar results are observed for triploid and tetraploid tumor samples (Fig. 2B and C).

Next, CNA detection performance is evaluated for FREEC and CLImAT, and the results suggest that FREEC has good performance for diploid tumor samples when tumor impurity is <0.5 (Supplementary Figure S8A). At larger tumor impurity levels, the sensitivity decreases while the specificity remains high. With similar specificity across all impurity levels, CLImAT is able to retain high sensitivity (>0.99) when the tumor impurity is <0.9. For triploid and tetraploid tumor samples (Supplementary Figure S8B and C), CLImAT also performs well in identifying CNA regions. At the same time, we investigate the performance of Patchwork, and results of simulated tumor samples are shown in Table S2. We find that in general both Patchwork and CLImAT can provide accurate aberration detection with similar performance, if the intermediate arguments of Patchwork are correctly determined by the user. Furthermore, we test CLImAT on low-coverage sequencing data, and the results for simulated data with 10× coverage suggest that CLImAT may also be applied to low coverage tumor WGS data when tumor impurity level is not high (Supplementary Figure S9).

In addition to aberration detection for tumor samples, we examine the reliability score (Supplementary Methods) that is used to measure how well the data fits to the model. For simulated tumor data with two cancer subclones (Supplementary Figure S10), the reliability score for the heterogeneous region is significantly lower than those of other homogeneous regions, suggesting it can help the user to evaluate the fitness of the model and provide better interpretation of the results.

3.3 Applying CLImAT to TNBC samples

Three TNBC samples sequenced at ~30× coverage are adopted to examine the performance of CLImAT, which are also assayed by Affymetrix SNP6.0 array for comparison. By using ASCAT (Van Loo *et al.*, 2010), the results generated from SNP arrays are used as the ground truth. We first evaluate ACN and impurity of these tumor samples using different methods, and the results are shown in Table 1. From the results of ASCAT, sample 1 is identified as aneuploid tumor, whereas samples 2 and 3 are identified as hyperploid tumors. Tumor sample 1 demonstrates genome-wide deletions with ACN of 1.67, whereas tumor samples 2 and 3 include dramatic amplifications along the whole cancer genome, with ACN of 3.02 and 4.16, respectively. CLImAT provides consistent estimation of ACN for the three tumor samples. Also, the tumor impurity levels estimated by CLImAT are in good concordance with the ground truth. These results suggest CLImAT has the potential for automatically identifying and correcting for tumor impurity and aneuploidy in complicated tumor samples.

Next, we examine LOH detection performance of FREEC, CLImAT and SNVMix (Fig. 3). The same performance evaluation procedure for simulated data analysis is adopted here, and the CNA/LOH calls of heterozygous positions recognized by ASCAT are treated as the ground truth (Supplementary Methods). For all three tumor samples, CLImAT compares favorably to SNVMix and FREEC. It achieves superior sensitivity of 0.98, 0.97 and 0.94 for samples 1, 2 and 3, respectively, with specificity better than or comparable with those of the other methods. We also examine the performance of CNA detection, and the results in Supplementary Table S3 show CLImAT has high consistency with ASCAT. Furthermore, Figure 4 illustrates the WGS and SNP array data for chromosome 8, 13 and 14 of

Table 1. ACN and tumor impurity estimated by FREEC, ASCAT and CLImAT for primary TNBC samples

Methods	ACN			Impurity		
	Sample 1	Sample 2	Sample 3	Sample 1	Sample 2	Sample 3
ASCAT	1.67	3.02	4.16	0.26	0.44	0.38
CLImAT	1.87	3.15	4.13	0.19	0.43	0.32
FREEC	1.92	3.77	4.92	0.20	0.22	0.29
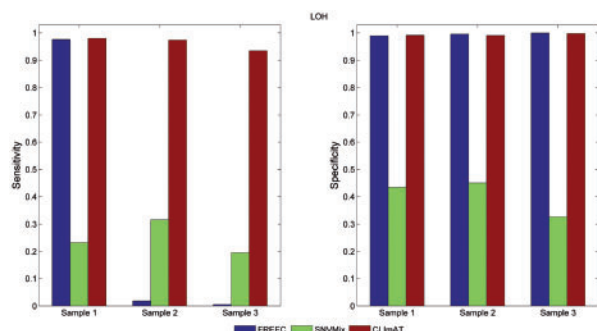
Fig. 3. LOH detection performance for primary TNBC samples. LOH detected by ASCAT from Affymetrix SNP6.0 arrays is used as ground truth

tumor sample 1, in which both BAF and LRR/RD signals generated from different platforms show similar patterns on aberrant regions. Both CLImAT and ASCAT identify consecutive LOH regions spanning chromosomes 8, 13 and 14, board hemizygous deletions on 8p(11.23–22), 8q(11.21–22.1), 13q(21.2–31.3), 14p(11.1–12) and 14q(21.3–23.1, 23.2–23.3 and 32.13–32.33), and board amplifications on chromosome 13q(12.11–13.3 and 32.1–34). In addition, benefited from high resolution of WGS platform, CLImAT provide more precise detection of small focal aberrations than ASCAT. For example, on 8p23 ASCAT only detects one homozygous deletion whereas CLImAT identify two additional homozygous deletion regions on 8p23.1, which harbors a potential tumor suppressor gene PinX1 related to telomerase activity and chromosome stability (Zhou *et al.*, 2011).

4 DISCUSSION AND CONCLUSION

Featured with finer resolution than previous genomic technologies, WGS allows more comprehensive analysis of tumor aberrations. In this study, we introduce an efficient computational approach for this purpose, which presents remarkable advantages over existing methods for interpretation of complicated tumor samples without prior knowledge of tumor impurity and ploidy. One advantage of CLImAT is the correction and normalization procedure for improving data quality of unpaired tumor samples. For example, BAF is normalized in CLImAT for elimination of LOR bias, which is indispensible for further

statistical modeling analysis of WGS data. GC-content and mappability correction of RD is also a crucial step for detecting aberrations in unpaired tumor samples.

Another advantage of CLImAT lies in the fact that it takes integrated analysis of RD and BAF using a novel HMM to provide accurate detection of genome-wide aberrations in tumor samples. The emission probabilities of HMM used in CLImAT give comprehensive description of the statistical behavior of sequencing data generated from tumor samples. Unlike previous approaches using Poisson distributions, more flexible negative binomial distribution is adopted to model over-dispersed RD signals. Moreover, the relevant parameters including tumor impurity and ploidy are automatically estimated by EM algorithm. These approaches ensure the performance of CLImAT for complex tumor samples.

Despite of the advantages mentioned above, CLImAT also has limitations in modeling and analysis of tumor sequencing data. First, CLImAT cannot be applied to exome-sequencing data, as it is originally designed to deal with unpaired WGS data. Second, although >2.6 million SNPs are investigated in CLImAT and only 1.5% adjacent SNPs have relatively large distance (>5 kb), the resolution of CLImAT may still be limited by genomic breakpoints that lie between SNPs. To further improve the resolution of CLImAT, we provide an option to estimate copy number for the regions between distant SNPs (>1 kb) by calculating the corresponding RD signals (Supplementary Methods). Third, CLImAT does not account for the issue of tumor heterogeneity (Mayrhofer *et al.*, 2013; Oesper *et al.*, 2013). The basic assumption adopted in CLImAT is that there is a single copy number for all tumor cells, which will not hold if multiple subclones exist in a tumor sample. Recently, Oesper *et al.* investigated tumor heterogeneity using DNA sequencing data and showed that multiple tumor subclones may often exist in tumor samples (Oesper *et al.*, 2013), suggesting that tumor heterogeneity is another key factor in interpreting tumor sequencing data. In heterogeneous tumor samples, the somatic aberrant signals derived from tumor sequencing data can be complicated, which makes it hard to deconvolute subclonal aberrations. Therefore, more advanced methods are required to assess tumor heterogeneity in tumor sequencing data.

In conclusion, we present CLImAT, an efficient and powerful bioinformatics tool, for detection of genomic aberrations using tumor WGS data. We expect it will be helpful for comprehensive interpretation of cancer genome and show its potential usefulness in clinical diagnosis and treatment for cancers.
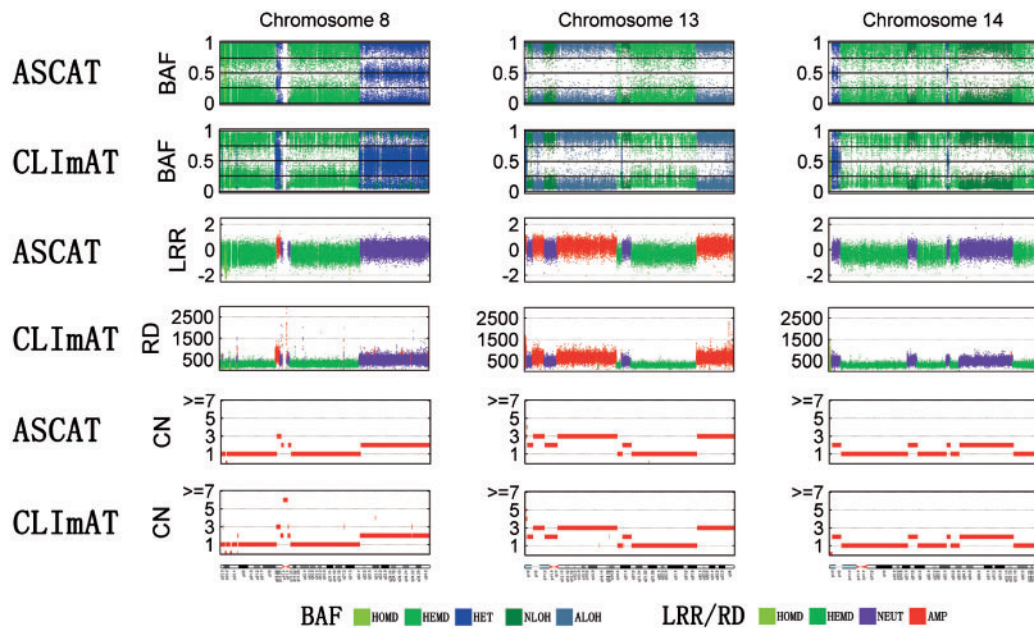
Fig. 4. Result comparison of CLImAT and ASCAT for TNBC sample 1. BAF is presented by five different aberration states: homozygous deletion (HOMD), hemizygous deletion (HEMD), heterozygous (HET), copy neutral LOH (NLOH) and amplified LOH (ALOH). LRR/RD is presented by homozygous deletion (HOMD), hemizygous deletion (HEMD), neutral (NEUT) and amplification (AMP)

ACKNOWLEDGEMENTS

This manuscript was prepared using a limited access dataset obtained from British Columbia Cancer Agency Branch (BCCA) and does not necessarily reflect the options or views from BCCA. We thank the editor and reviewers for their helpful comments and suggestions to improve this manuscript.

Conflict of Interest: none declared.

REFERENCES

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

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Bignell,G.R. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature*, **463**, 893–898.

Boeva,V. *et al.* (2011) Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics*, **27**, 268–269.

Boeva,V. *et al.* (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, **28**, 423–425.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Carén,H. *et al.* (2010) High-risk neuroblastoma tumors with 11q-deletion display a poor prognostic, chromosome instability phenotype with later onset. *Proc. Natl Acad. Sci. USA*, **107**, 4323–4328.

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

Duan,J. *et al.* (2013) Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One*, **8**, e59128.

Goya,R. *et al.* (2010) SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*, **26**, 730–736.

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.

Ha,G. *et al.* (2012) Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.*, **22**, 1995–2007.

Kim,S. *et al.* (2013) Virmid: accurate detection of somatic mutations with sample impurity inference. *Genome Biol.*, **14**, R90.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,A. *et al.* (2011) GPHMM: an integrated hidden Markov model for identification of copy number alteration and loss of heterozygosity in complex tumor samples using whole genome SNP arrays. *Nucleic Acids Res.*, **39**, 4928–4941.

Mardis,E.R. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, **9**, 387–402.

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

Metzker,M.L. (2009) Sequencing technologies—the next generation. *Nat. Rev. Genet.*, **11**, 31–46.

Morozova,O. and Marra,M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255–264.

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

Park,P.J. (2008) Experimental design and data analysis for array comparative genomic hybridization. *Cancer Invest.*, **26**, 923–928.

Peiffer,D.A. *et al.* (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.*, **16**, 1136–1148.

Rabiner,L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

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

Schuster,S.C. (2007) Next-generation sequencing transforms today's biology. *Nature*, **200**, 8.

Shah,S.P. *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**, 395–399.

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

Stephens,P.J. *et al.* (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Stratton,M.R. *et al.* (2009) The cancer genome. *Nature*, **458**, 719–724.

Suzuki,S. *et al.* (2000) An approach to analysis of large-scale correlations between genome changes and clinical endpoints in ovarian cancer. *Cancer Res.*, **60**, 5382–5385.

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.

Yoon,S. *et al.* (2009) Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.*, **19**, 1586–1592.

Zhou,X.Z. *et al.* (2011) The telomerase inhibitor PinX1 is a major haploinsufficient tumor suppressor essential for chromosome stability in mice. *J. Clin. Invest.*, **121**, 1266.