Integrated modeling of GC-content, mappability, tumor impurity and aneuploidy for accurate detection of genomic aberrations

ZHENHUA YU¹, AO LI², LIANG ZOU³⁴, XUEHONG SUN¹, MINGHUI WANG², AND PENG ZHANG¹

¹School of Information Engineering, Ningxia University, Yinchuan 750021, China
²School of Information Science and Technology, Centers for Biomedical Engineering, University of Science and Technology of China, Hefei 230027, China
³National Engineering Research Center for Agro-Ecological Big Data Analysis & Application, Anhui University, Hefei 230601, China
⁴Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC V6T 1Z4, Canada

Corresponding author: L. Zou (liangzou@ece.ubc.ca)

This work is supported by the Natural Science Foundation of Ningxia Province (Grant No. 2018AAC03020) and National Natural Science Foundation of China (Grant Nos. 61571414, 61471331 and 61561039)

ABSTRACT
Next-generation sequencing has been widely used in cancer-focused studies for comprehensive landscape of tumor genomes. Detection of genomic aberrations is one of the focal points in this area. Analysis of tumor sequencing data is usually complicated by several critical issues such as GC-content bias, mappability bias, tumor impurity and aneuploidy. Efficient computational methods are still in great demand for comprehensively addressing these issues. We introduce GPHMM-SEQ, a novel algorithm for inferring tumor impurity and ploidy, as well as detecting copy number alterations and loss of heterozygosity from paired tumor-normal samples. Read depth signals derived from sequencing data are analyzed using a novel hidden Markov model that employs integrated representation of GC-content bias, mappability bias, tumor impurity and aneuploidy. The evaluation on simulated and real tumor sequencing data demonstrates GPHMM-SEQ has the superior performance compared to existing methods.

INDEX TERMS
Hidden Markov model, next-generation sequencing, aberration detection, tumor impurity and aneuploidy

I. INTRODUCTION
Cancer genomes are characterized by various somatic alterations such as segmental deletions and amplifications of chromosomal regions. These alterations are acquired by tumor cells during the procedure of tumourigenesis and progression and play an important role in tumor persistence and growth [1]. Frequently occurring genomic alterations are reported to result in the dysfunction of cancer-associated genes such as oncogenes and tumor suppressor genes [2, 3]. Among these alterations, somatic copy number alterations (CNAs) and loss of heterozygosity (LOH) have emerged as two main categories of cancer genetic structural variations. Identifying these abnormalities is of long-standing interest in cancer-focused studies with the purpose of finding pivotal molecular events that drive tumor progression [4–10].

The advent of next-generation sequencing (NGS) [11] now allows an unprecedented view on the comprehensive landscape of cancer genomes. With the nucleotide resolution, NGS is widely used for high-throughput profiling of cancer genomes and detecting genomic aberrations such as CNAs and LOH [1]. However, there are several critical issues that complicate the analysis of tumor sequencing data. First, difference in GC-content and mapping ability of reads of different genomic loci results in the non-uniform read depth [12, 13]. The biased distribution of reads will complicate accurate detection of CNAs and LOH. Second, tumor samples are often impure due to normal cell contamination [10]. Unfortunately, the tumor impurity representing the proportion of normal cells is usually unknown. Normal cell contamination makes it difficult to distinguish among different types of aberrations, especially when the impurity of tumor sample is severe. Third, various numerical aberrations such as CNAs and other structural abnormalities result in the aneuploidy of genomes [14, 15],...
and the underlying tumor ploidy is unknown when dealing with tumor sequencing data. Finally, clonal evolution may result in emergence of distinct cell populations, and tumor samples are heterogeneous in this case [5, 16, 17]. Analysis of heterogeneous samples will be much challenging due to divergent aberrations represented in different tumor subclones at the same genomic locus. Critically, the aforementioned issues often simultaneously arise in the same tumor sample and are strongly intertwined [16], therefore cannot be effectively addressed when dealing with each of the issues separately. For instance, different combination of tumor impurity and ploidy could provide similar explanations of the somatic aberrant signals [18], which results in interpretation ambiguity of the sequencing data. Therefore, computational methods for analyzing tumor sequencing data should account for these issues. Statistical methods such as hidden Markov models [19, 20] for integrated representation of GC-content bias, mappability bias, tumor impurity, and aneuploidy will lead to improved performance for aberration detection.

So far, several approaches are proposed for correction for GC-content and mappability biases when analyzing read depth data. CONTRA [21] simply uses log-ratios at base resolution to remove the effect of GC-content. Nevertheless, it has been reported that the read depth ratio between tumor and normal pair cannot completely remove GC-content or mappability bias [22]. APOLLOH [23] filters highly mappable regions as the first step, then uses loess curve fit between GC-content and read depth for tumor and normal samples separately, and corrects read depth by dividing the fitted value. FREEC [24] employs a quadratic or cubic polynomial fit to approximate the dependency of read depth on GC-content when no control sample is available, then scales the read depth by the fitted value. It also provides option for users to correct the mappability bias by using a mappability track. EXCAVATOR [25] and CLImAT [12] extend the median normalization approach initially introduced in EWT [26] for minimizing the effect of GC-content and mappability. In addition, OncoSNP-SEQ [5] adopts a simple linear regression model for GC-content and mappability correction. However the existing studies have demonstrated linear regression is unable to adequately describe the relationship between read depth and GC-content [24]. A recent study called FACETS [27] uses loess regression of log-ratio over GC-content, and subtracts the fitted value from the log-ratio. No approaches for mitigating mappability bias are employed in FACETS’ pipeline. All of the aforementioned methods implement the GC-content and mappability correction procedures as the preprocessing steps before aberration calling.

For accurate detection of CNAs and LOH, many methods have been designed in recent years for addressing the issues of tumor impurity and aneuploidy. For example, ABSOLUTE [15] uses a two-component mixture model to quantitatively represent the average copy number of a given genomic locus. The two-component mixture model is widely used in tumor NGS data analysis methods such as APOLLOH [23], CLImAT [12], FACETS [27], and CloneCNA [17], to quantize the effect of tumor impurity. Some methods [24] address the tumor aneuploidy issue by giving option for users to specify the most abundant copy number of the tumor genome, however such information is usually not available for human tumor samples. Therefore, computational methods that automatically correct the tumor aneuploidy are necessary to complement the arsenal of the analysis tools. ABSOLUTE measures the average ploidy of tumor-normal mixed genomes by parameterizing tumor impurity and aneuploidy in inferring relative copy number. CLImAT estimates the mean read depth of copy neutral regions by introducing ploidy-related parameters in the emission model of the hidden Markov model. In addition, FACETS solves the aneuploidy of tumor genome by inferring the location shift of copy neutral loci from segmentation results. Similarly, CloneCNA infers the baseline shift of the log-ratios that varies with respect to tumor ploidy. These methods have greatly improved the accuracy of aberration detections.

Despite the advantages of the aforementioned methods, their performance may be degraded if GC-content and mappability biases are not effectively removed in the data preprocessing steps. To our knowledge, there are few methods that directly incorporate the GC-content and mappability effect into the aberration calling framework, meanwhile take into account the issues of tumor impurity and aneuploidy. Therefore, it is essential to develop algorithms for aberration detection by integrated representation of GC-content bias, mappability bias, tumor impurity, and aneuploidy in one framework.

In this study, we design GPHMM-SEQ, a novel algorithm for detecting somatic CNAs and LOH using sequencing data of complex tumor samples by addressing the aforementioned issues. GPHMM-SEQ is developed by borrowing the statistical strength of GPHMM [10] algorithm which is based on SNP-array data. By fully investigating read depth data derived from paired tumor-normal samples, GPHMM-SEQ is able to accurately infer tumor impurity and ploidy, and detect CNA and LOH segments. A novel hidden Markov model (HMM) that employs integrated representation of GC-content bias, mappability bias, tumor impurity, and aneuploidy is implemented for analyzing read depth signals. Joint analysis of the log-ratio and B allele frequency (BAF) data with integration of different sources of variation enables GPHMM-SEQ to accurately identify genomic aberrations even for highly contaminated and aneuploid tumor samples. Evaluation of the ability of GPHMM-SEQ on multiple sequencing datasets demonstrates our method has better performance than existing methods.

II. MATERIALS AND METHODS
A. SEQUENCING DATA

To examine the performance of GPHMM-SEQ in detecting CNAs and LOH, we simulate 27 tumor samples by combining different values of tumor impurity and ploidy. Three tumor genomes (near-diploidy, near-triploidy and near-tetraploidy) and a normal genome are constructed and mixed at known proportions. We construct each genome by following the approach described in DeAnnCNV [28]. For the simulation of sequencing experiment, the whole-genome sequencing (WGS) data of the chromosome 1 of a real sample HCC1143_BL is used to produce the sequencing reads of each simulated genome. The data is downloaded by following the instructions at https://gdc.cancer.gov/resources-tcga-users/tcga-mutation-calling-benchmark-4-files. By mixing each of the tumor genomes with the normal genome, 9 mixtures at ~30X coverage are generated by sampling reads from the BAM files at different proportions. By this way, total 27 mixed samples are generated.

Additionally, the real whole-exome sequencing (WES) data of 9 paired breast cancer samples [29] is used in this study. These samples are sequenced to ~30X coverage on the Illumina Genome Analyzer IIx sequencing platform, and available from European Genome-Phenome Archive (EGA) (No. EGAS00001000132). We further download the WES data of 9 paired primary invasive breast cancer samples from the Sequence Read Archive (SRA) under Accession number PRJNA273304, and the Run Ids of these samples are listed in Table I. The reads are aligned to the hg19 human reference genome.

B. THE STATISTICAL MODEL IN GPHMM-SEQ

The workflow of GPHMM-SEQ is shown in Fig. 1. As the first step we use SAMtools [30] to identify germline heterozygous SNP positions $L$ from the normal genome. For each SNP position in $L$, read depth and B allelic read depth of the tumor genome, read depth of the normal genome, and GC-content and mappability score are extracted using an in-house tool. To measure the mappability score of a given SNP position, the mean mappability value of a fixed-size (1000bp) window centered at the SNP is calculated by using “wgEncodeCrgMapabilityAlign36mer” track from UCSC (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/encodeDC C/wgEncodeMapability/ for reference hg18, and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDC C/wgEncodeMapability/ for reference hg19). GC-content is calculated as the percentage of ‘G’ and ‘C’ nucleotides within the window. The read depth is then sample-normalized to make it comparable between tumor and normal pair. The log-ratio (LCR) between the normalized tumor and normal read depths, and BAF measured by the proportion of B allelic reads, are prepared for further analysis.

FIGURE 1. Workflow of GPHMM-SEQ.

For simultaneous detection of CNAs and LOH, we introduce an integrated HMM that takes joint analysis of LCR and BAF data by fully representing GC-content and mappability biases, tumor impurity and aneuploidy. The hidden states of the HMM are defined in Table II. Each state is depicted with detailed information including copy number, tumor genotypes and the type of chromosomal abnormality.
where \( \phi(x) \) is the density of standard normal distribution, \( w_c \) denotes the tumor impurity, \( \sigma_i \) is the standard deviation of LCR signal, and \( h=(h_0, h_1, h_2, h_3, h_4) \) denote the coefficients associated with GC-content and mappability. For hidden state \( c \), the mean value \( \mu_{ic}^h \) of the LCR signal of the \( i \)th SNP is defined as:

\[
\mu_{ic}^h = \log\left(\frac{y_c}{2}\right) + h_0 + h_1 g_i + h_2 g_i^2 + h_3 g_i^3 + h_4 m_i
\]

(2)

where \( g_i \) and \( m_i \) denote the values of GC-content and mappability respectively, and \( y_c \) is the mean copy number and defined as:

\[
y_c = 2w_c + n_c(1-w_c)
\]

(3)

where \( n_c \) represents the tumor copy number. Note that in (2) we incorporate the effects of GC-content, mappability, tumor impurity, and aneuploidy on the LCR signals. We employ a cubic polynomial fit for GC-content based on the results reported in FREEC [24] which demonstrates cubic polynomial fit is effective enough to model the relationship between read depth and GC-content, and a linear fit for mappability supported by our previous study [12].

Similarly, we assume the BAF signal is distributed according to a mixture of normal distributions, and the probability density function for hidden state \( c \) is defined as:

\[
f(b_i \mid c, w_c, \sigma_b) = \frac{1}{\sum_{k=1}^{K_c} p_{ck}} \phi\left(\frac{b_i - \mu_{ic}^b}{\sigma_b}\right)
\]

(4)

where \( K_c \) denotes the number of tumor genotypes associated with state \( c \). We use \( \sigma_b \) to denote the standard deviation of BAF signal. \( p_{ck} \) represents the prior probability of the \( k \)th genotype. \( \mu_{ic}^b \) is the mean value of BAF signal and formulated as:

\[
\mu_{ic}^b = \frac{w_c + b_{ck}(1 - w_c)}{y_c}
\]

(5)

where \( b_{ck} \) is the B allele copy number. Note that from (5) we take into account the effect of tumor impurity on the BAF signal.

We use expectation maximization (EM) [31] algorithm to estimate the model parameters:

\[
\theta = (\pi, A, w_c, h, \sigma_i, \sigma_b)
\]

(6)

where \( \pi \) is the initial state distribution, and \( A \) is the matrix of transition probability. In the E-step of the EM algorithm during the \( n \)th iteration, for LCR and BAF signals the partial log-likelihood are defined as follows:

\[
E(\mathbb{L}_t) = \sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \log(f(l_i \mid c, w_c, \sigma_i, h))
\]

(7)

\[
E(\mathbb{L}_b) = \sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \log(f(b_i \mid c, w_c, \sigma_b))
\]

(8)

where \( \gamma^{(n)} \) is the posterior probability and inferred by standard forward-backward algorithm [32]. \( T \) is the number of SNP positions, and \( C \) denotes the number of the HMM hidden states. In the maximization step, we aim to maximize the object function

\[
F(w, \sigma, h, \sigma_b) = E(\mathbb{L}_t) + E(\mathbb{L}_b)
\]

(9)

By setting the first partial derivative of \( F \) for parameter \( h_0 \) to 0, we derive the formula to update \( h_0 \) for the next iteration:

\[
h_0^{(n+1)} = \frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \left( l_i - \log\left(\frac{y_c^{(n)}}{2}\right) - h_0^{(n)} g_i \right) - \sum_{i=1}^{C} \gamma_{ic}^{(n)} m_i}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} g_i}
\]

(10)

Similarly, we update \( h_1, h_2, h_3, h_4, \sigma_i \) and \( \sigma_b \) using following formulas:

\[
h_1^{(n+1)} = \frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \left( l_i - \log\left(\frac{y_c^{(n)}}{2}\right) - h_0^{(n)} g_i \right) - h_0^{(n+1)} g_i}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} g_i}
\]

(11)

\[
h_2^{(n+1)} = \frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \left( l_i - \log\left(\frac{y_c^{(n)}}{2}\right) - h_0^{(n)} g_i - h_1^{(n+1)} g_i \right) - h_0^{(n+1)} g_i}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} g_i}
\]

(12)

\[
h_3^{(n+1)} = \frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \left( l_i - \log\left(\frac{y_c^{(n)}}{2}\right) - h_0^{(n)} g_i - h_1^{(n+1)} g_i - h_2^{(n+1)} g_i \right) - h_0^{(n+1)} g_i}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} g_i}
\]

(13)

\[
h_4^{(n+1)} = \frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \left( l_i - \log\left(\frac{y_c^{(n)}}{2}\right) - h_0^{(n+1)} g_i - h_1^{(n+1)} g_i - h_2^{(n+1)} g_i \right) - h_0^{(n+1)} g_i}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} g_i}
\]

(14)

\[
\sigma_i^{(n+1)} = \sqrt{\frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \left( l_i - \log\left(\frac{y_c^{(n)}}{2}\right) - h_0^{(n+1)} g_i - h_1^{(n+1)} g_i - h_2^{(n+1)} g_i \right)^2}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} g_i}}
\]

(15)

\[
\sigma_b^{(n+1)} = \frac{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)} \sum_{k=1}^{K_c} p_{ck} (b_{ck} - (w_c + b_{ck}(1 - w_c)) \gamma_{ic}^{(n)} \right)^2}{\sum_{i=1}^{C} \sum_{c=1}^{C} \gamma_{ic}^{(n)}}
\]

(16)

Due to the closed form of formula for updating \( w_c \), if not available, we adopt numerical methods such as Newton–Raphson method [33] to update \( w_c \):

\[
w_c^{(n+1)} = w_c^{(n)} - \frac{\partial F}{\partial w_c} \left/ \frac{\partial^2 F}{\partial^2 w_c} \right.
\]

(17)

We use the approach described in [32] to estimate the \( \pi \) and \( A \). At the beginning of the EM algorithm, we assume a
uniform distribution of the initial state with the following formula:

\[ \pi^{(0)} = \{\pi_i\}, \pi_i = \frac{1}{C} \]  

(18)

The initial values of the state transition matrix are defined as follows:

\[ A^{(0)} = \{a_{ij}\}, a_{ij} = \begin{cases} 1 - p_t & i = j \\ \frac{p_t}{(C - 1)} & i \neq j \end{cases} \]  

(19)

where \( p_t \) is the total transition probability from a state to all other different states, and set to \( 10^{-6} \) in this study. The parameters \( \sigma_l \) and \( \sigma_b \) are initialized to sample standard deviation of the LCR and BAF signals, respectively. Moreover, the parameters \( h_1, h_2, h_3 \), and \( h_4 \) are set to near-zero values (0.001) initially.

We iteratively update the model parameters until the algorithm converges. To find the global optimal parameters, we adopt a grid search of the initial values of the parameters \( w_s \) and \( h_0 \). The aberration type for each SNP can be inferred by the Viterbi algorithm [34] or from the state with the maximum posterior probability. All SNPs are then segmented according to the aberration types.

Finally, a reliability score is measured for each segment to evaluate how well the model explains the data by following the approach described in CLImAT.

C. CALIBRATION OF INFERRED COPY NUMBER

Same to the model architecture of GPHMM algorithm, the hidden states of the HMM employed in GPHMM-SEQ are defined by a finite set of aberrations. Due to the complexity of tumor genomes, there may be extremely amplified genomic regions which cannot be accurately explained by the limited hidden states of the HMM. Therefore, we propose a post-processing scheme to calibrate the copy number states of highly amplified regions using the following formula:

\[ cn = R\left(\frac{2^{m+1} - 2w_i}{1 - w_i}\right) \]  

(20)

where \( cn \) is the total copy number, \( R \) denotes the function taking the nearest integer, and \( m \) is the mean value of LCR signals. After the total copy number is determined, the optimal tumor genotype is inferred from the candidate hidden state that maximizes the likelihood of the BAF signals.

D. IMPLEMENTATION OF GPHMM-SEQ

We implement the algorithm by using Matlab and C, and the source code is available at http://bioinformatics.ustc.edu.cn/gphmmseq.

III. RESULTS

A. APPLY GPHMM-SEQ TO SIMULATED DATA

We first apply GPHMM-SEQ to simulated tumor data, and examine its ability of inferring tumor impurity and ploidy, as well as detecting CNAs and LOH.

To examine the accuracy of tumor impurity predictions, we make a comparison between the results of FACETS and GPHMM-SEQ. Fig. 2(a) shows that GPHMM-SEQ accurately estimates the tumor impurity with high significant positive correlation (correlation coefficient = 0.99, \( p = 5.60 \times 10^{-29} \) and mean absolute error = 0.02) with the underlying tumor impurity. In comparison, the results of FACETS sometimes obviously deviate from the expected tumor impurity. For example, for tetraploid sample with tumor impurity of 0.5, FACETS predicts a tumor impurity of 0.1. We further compare the tumor ploidy inferred by FACETS and GPHMM-SEQ with the ground truth ploidy, and the results are shown in Fig. 2(b). The mean and standard deviation of estimated tumor ploidy are calculated, and the results suggest GPHMM-SEQ has significant advantage in inferring tumor ploidy, especially on the tetraploidy samples. For example, GPHMM-SEQ correctly identifies all tetraploid samples as near-tetraploidy with standard deviation of 0.05, however FACETS tends to predict them as near-triploidy with standard deviation of 0.66.

FIGURE 2. (a) Estimated tumor impurity and (b) average copy number for simulated samples.
We proceed to evaluate the abilities of FACETS and GPHMM-SEQ in detecting CNAs and LOH. All germline SNPs predefined in simulation procedure are used for evaluation. A summary of the GPHMM-SEQ detection results on diploid samples is shown in Fig. 3. When compared with the ground truth represented in the bottom of the figure, CNA and LOH segments are correctly identified with consistent high performance when tumor impurity is less than 0.8. For the sample with 80% normal cells, two amplified regions are predicted as normal due to the significantly attenuated LCR and BAF signals in these regions. We first assess the performance of FACETS and GPHMM-SEQ in calling integer copy numbers. As shown in Fig. 4(a), FACETS has distinct recognition abilities across different tumor impurities and copy numbers. It correctly identifies copy numbers of 1, 6 and 7 when the tumor impurity is less than 0.4, but exhibits low performance for larger copy numbers (>2) at severe tumor impurities. Compared with FACETS, GPHMM-SEQ demonstrates strong robustness to copy number and tumor impurity, and gets high sensitivity for different copy numbers even in samples with tumor impurity of 0.8 as shown in Fig. 4(b). We further investigate confusion matrix for the concordance of the copy number calls and the underlying truth using all simulated mixtures. As shown in Fig. 4(c) and (d), GPHMM-SEQ has consistent better performance in quantifying the CNAs.

![Figure 4](image)

**FIGURE 4.** Sensitivity of FACETS (a) and GPHMM-SEQ (b) in detecting CNAs. Confusion matrix of CNA calls from FACETS (c) and GPHMM-SEQ (d).

For evaluation of LOH detection, we measure the respective sensitivity for different aberration types including HOMD, DLOH, NHET, NLOH, AHET, and ALOH under different tumor impurities. Fig. 5(a) shows that FACETS maintains high sensitivity for all aberration types at lower tumor impurities (<0.7) except for NHET in pure samples and ALOH at tumor impurity of 0.6. In comparison, GPHMM-SEQ accurately identifies all aberration types at tumor impurities less than 0.7, and also maintains comparable sensitivity at high tumor impurities as demonstrated in Fig. 5(b). Moreover, the confusion matrix is also investigated and represented in Fig. 5(c) and (d), demonstrating GPHMM-SEQ has higher power than FACETS to distinguish different aberration types.

![Figure 5](image)

**FIGURE 5.** Sensitivity of FACETS (a) and GPHMM-SEQ (b) in detecting aberration types. Confusion matrix of aberration type calls from FACETS (c) and GPHMM-SEQ (d).

### B. APPLY GPHMM-SEQ TO REAL DATA

We first assess the quality of model fitting on the EGA dataset. Table III shows the estimated parameters for all samples, we can find significant evidence that log-ratio cannot completely remove GC-content and mappability bias, which is consistent with the result reported by previous study [22]. The relationship between the real LCR ($l_i$) and estimated mean LCR ($\hat{l}_i$) for SNPs in 2-copy regions is then investigated and fitted using a linear regression, and we get significant p-values ($<10^{-15}$) for all investigated samples. The results show LCR signals are distributed with different mean values with respect to GC-content and mappability, and our model successfully discovers the subtle difference as shown in Fig. 6. These results in different samples demonstrate integrated representation of GC-content, mappability, tumor impurity, and aneuploidy in our model effectively characterizes the distribution of LCR signals.

The tumor impurity and ploidy inferred by different methods on EGA samples are shown in Table IV. Both FACETS and GPHMM-SEQ predict most of the samples to be highly contaminated, and the maximum tumor impurity of 0.69 and 0.61 respectively for sample SA031. In addition, all samples are identified as aneuploid by FACETS and GPHMM-SEQ. The inferred tumor ploidy changes from 1.64 for SA029 to 3.62 for SA018 from GPHMM-SEQ predictions. The different combinations of tumor impurity and ploidy result in significantly different distributions of the LCR signals as shown in Fig. 6. We then evaluate the accuracy of tumor impurity and ploidy estimations by comparing the results of FACETS and GPHMM-SEQ with the ground truth values. Here we apply ASCAT [35] software to Affymetrix SNP6.0 array data of same samples, and use the predictions as gold standard for evaluation. For tumor impurity inferring, GPHMM-SEQ gets more accurate results (MAE = 0.03) when compared to FACETS (MAE = 0.05). High consistency of the ploidy predictions (MAE = 0.11) between GPHMM-SEQ and ASCAT is observed for all samples, while FACETS overestimates the ploidy of some
near-diploid samples. These results indicate the good concordance between sequencing and microarray data is achieved by appropriate representation of tumor impurity and ploidy in our model.

### TABLE III

<table>
<thead>
<tr>
<th>Sample</th>
<th>$h_0$</th>
<th>$h_1$</th>
<th>$h_2$</th>
<th>$h_3$</th>
<th>$\sigma_l$</th>
<th>$\sigma_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA018</td>
<td>-0.50</td>
<td>-0.40</td>
<td>-0.16</td>
<td>0.28</td>
<td>0.15</td>
<td>0.41</td>
</tr>
<tr>
<td>SA029</td>
<td>-0.06</td>
<td>0.61</td>
<td>0.04</td>
<td>-1.12</td>
<td>-0.04</td>
<td>0.51</td>
</tr>
<tr>
<td>SA030</td>
<td>-0.33</td>
<td>-0.11</td>
<td>-0.23</td>
<td>-0.36</td>
<td>0.06</td>
<td>0.37</td>
</tr>
<tr>
<td>SA031</td>
<td>0.35</td>
<td>-0.46</td>
<td>-0.85</td>
<td>-1.19</td>
<td>0.14</td>
<td>0.42</td>
</tr>
<tr>
<td>SA051</td>
<td>-0.47</td>
<td>0.33</td>
<td>0.54</td>
<td>0.68</td>
<td>-0.03</td>
<td>0.38</td>
</tr>
<tr>
<td>SA052</td>
<td>0.07</td>
<td>-0.10</td>
<td>-0.20</td>
<td>-0.36</td>
<td>0.02</td>
<td>0.36</td>
</tr>
<tr>
<td>SA065</td>
<td>-0.14</td>
<td>-0.35</td>
<td>-0.48</td>
<td>-0.47</td>
<td>0.12</td>
<td>0.46</td>
</tr>
<tr>
<td>SA069</td>
<td>0.09</td>
<td>-0.60</td>
<td>-0.45</td>
<td>0.05</td>
<td>0.15</td>
<td>0.38</td>
</tr>
<tr>
<td>SA071</td>
<td>-0.11</td>
<td>0.31</td>
<td>0.25</td>
<td>0.06</td>
<td>0.01</td>
<td>0.36</td>
</tr>
</tbody>
</table>

### FIGURE 6. Model fitting analysis on real samples.

### TABLE IV

<table>
<thead>
<tr>
<th>Sample</th>
<th>Tumor impurity FACETS</th>
<th>Tumor impurity GPHMM-SEQ</th>
<th>Tumor impurity ASCAT</th>
<th>Tumor ploidy FACETS</th>
<th>Tumor ploidy GPHMM-SEQ</th>
<th>Tumor ploidy ASCAT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA018</td>
<td>0.43</td>
<td>0.37</td>
<td>0.37</td>
<td>3.59</td>
<td>3.62</td>
<td>3.62</td>
</tr>
<tr>
<td>SA029</td>
<td>0.56</td>
<td>0.51</td>
<td>0.52</td>
<td>1.74</td>
<td>1.64</td>
<td>1.50</td>
</tr>
<tr>
<td>SA030</td>
<td>0.48</td>
<td>0.54</td>
<td>0.53</td>
<td>3.23</td>
<td>3.43</td>
<td>3.40</td>
</tr>
<tr>
<td>SA031</td>
<td>0.69</td>
<td>0.61</td>
<td>0.59</td>
<td>3.50</td>
<td>2.04</td>
<td>1.85</td>
</tr>
<tr>
<td>SA051</td>
<td>0.49</td>
<td>0.59</td>
<td>0.56</td>
<td>2.48</td>
<td>2.75</td>
<td>2.56</td>
</tr>
<tr>
<td>SA052</td>
<td>0.49</td>
<td>0.60</td>
<td>0.49</td>
<td>2.86</td>
<td>1.95</td>
<td>1.74</td>
</tr>
<tr>
<td>SA065</td>
<td>0.29</td>
<td>0.31</td>
<td>0.33</td>
<td>2.69</td>
<td>2.62</td>
<td>2.54</td>
</tr>
<tr>
<td>SA069</td>
<td>0.62</td>
<td>0.55</td>
<td>0.58</td>
<td>3.12</td>
<td>2.41</td>
<td>2.30</td>
</tr>
<tr>
<td>SA071</td>
<td>0.35</td>
<td>0.30</td>
<td>0.28</td>
<td>1.89</td>
<td>1.88</td>
<td>1.87</td>
</tr>
</tbody>
</table>

| MAE*   | 0.05                   | 0.03                     | N/A                 | 0.48                | 0.11                   | N/A               |

*MAE: mean absolute error.

We further investigate the CNA and LOH segments on real samples. Fig. 7 and Fig. 8 show the result of GPHMM-SEQ on samples SA018 and SA029, respectively. GPHMM-SEQ predicts SA018 as near-tetraploidy, amplification of almost whole chromosome of 1, 2, 3, 6, 7, 10, 11, 15, 17, 18, 19 and 21, consecutive amplified regions on chromosomes 8q, 9p, 13q, 20q and 22p, are identified for the sample. SA029 is predicted as highly contaminated and near-diploid sample, and represented by broad hemizygous deletions in chromosomes 2, 4, 10, 12, 13, 15 and 17.

As the ground truth CNA and LOH events are not available for EGA samples, we use the results of SNPs inferred by ASCAT as the baseline for evaluation. The confusion matrix for the concordance of the copy number calls and the underlying truth is measured and shown in Fig. 9(a) and (b). FACETS tends to predicts real copy numbers to their near values and gets mean sensitivity of 0.53. GPHMM-SEQ shows better performance in quantifying the CNAs and achieves a mean sensitivity of 0.76. We also calculate the confusion matrix for aberration type predictions for both methods as shown in Fig. 9(c) and (d). The similar results as on the simulated data are observed, and GPHMM-SEQ performs well in distinguishing different aberration types. These results demonstrate GPHMM-SEQ’ ability of accurately identifying CNAs and LOH in complex tumor samples.
To further examine the performance of our proposed method, we apply GPHMM-SEQ and FACETS to the SRA dataset. The prediction results of tumor impurity and ploidy are shown in Table. V. For samples SRA1–SRA8, two algorithms get the similar inferences for both tumor impurity and ploidy. For instance, SRA1 is identified as triploid and highly contaminated sample, and the inferred proportions of normal cells are 71% and 64% respectively. Most of the samples are predicted to be near-triploidy, and the tumor ploidy ranges from 2.36 to 4.08 from the results of GPHMM-SEQ. In addition, sample SRA9 is identified as near-diploidy by both methods, but the estimated tumor impurities are significantly different with respective values of 0.02 and 0.42.

The computational complexity of GPHMM-SEQ mainly lies in calculating the posterior probability and updating model parameters. The computational complexity of GPHMM-SEQ is $O(TN)$, here $T$ is the number of investigated SNPs and $N$ denotes the number of iterations executed in EM algorithm. The average time used to process single sample on EGA and SRA datasets are 7.6 and 9.5 minutes respectively when running on a standard desktop PC with 3.6GHz CPU and 8GB RAM, and the memory usage is less than 400 MBytes.

We then analyze the runtime performance of GPHMM-SEQ on real samples, and the results are shown in Table. VI. The computational complexity of GPHMM-SEQ mainly lies in calculating the posterior probability and updating model parameters. The computational complexity of GPHMM-SEQ is $O(TN)$, here $T$ is the number of investigated SNPs and $N$ denotes the number of iterations executed in EM algorithm. The average time used to process single sample on EGA and SRA datasets are 7.6 and 9.5 minutes respectively when running on a standard desktop PC with 3.6GHz CPU and 8GB RAM, and the memory usage is less than 400 MBytes.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Tumor impurity</th>
<th>Tumor ploidy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FACETS</td>
<td>GPHMM-SEQ</td>
</tr>
<tr>
<td>SRA1</td>
<td>0.71</td>
<td>0.64</td>
</tr>
<tr>
<td>SRA2</td>
<td>0.31</td>
<td>0.29</td>
</tr>
<tr>
<td>SRA3</td>
<td>0.48</td>
<td>0.51</td>
</tr>
<tr>
<td>SRA4</td>
<td>0.48</td>
<td>0.46</td>
</tr>
<tr>
<td>SRA5</td>
<td>0.69</td>
<td>0.69</td>
</tr>
<tr>
<td>SRA6</td>
<td>0.63</td>
<td>0.67</td>
</tr>
<tr>
<td>SRA7</td>
<td>0.44</td>
<td>0.51</td>
</tr>
<tr>
<td>SRA8</td>
<td>0.56</td>
<td>0.54</td>
</tr>
<tr>
<td>SRA9</td>
<td>0.02</td>
<td>0.42</td>
</tr>
</tbody>
</table>

We then analyze the runtime performance of GPHMM-SEQ on real samples, and the results are shown in Table. VI. The computational complexity of GPHMM-SEQ mainly lies in calculating the posterior probability and updating model parameters. The computational complexity of GPHMM-SEQ is $O(TN)$, here $T$ is the number of investigated SNPs and $N$ denotes the number of iterations executed in EM algorithm. The average time used to process single sample on EGA and SRA datasets are 7.6 and 9.5 minutes respectively when running on a standard desktop PC with 3.6GHz CPU and 8GB RAM, and the memory usage is less than 400 MBytes.

To further examine the performance of our proposed method, we apply GPHMM-SEQ and FACETS to the SRA dataset. The prediction results of tumor impurity and ploidy are shown in Table. V. For samples SRA1–SRA8, two algorithms get the similar inferences for both tumor impurity and ploidy. For instance, SRA1 is identified as triploid and highly contaminated sample, and the inferred proportions of normal cells are 71% and 64% respectively. Most of the samples are predicted to be near-triploidy, and the tumor ploidy ranges from 2.36 to 4.08 from the results of GPHMM-SEQ. In addition, sample SRA9 is identified as near-diploidy by both methods, but the estimated tumor impurities are significantly different with respective values of 0.02 and 0.42. The copy number, BAF and log ratio data of GPHMM-SEQ and FACETS on the sample is compared, and the result is shown in Fig. 10. Both methods identify the amplified heterozygous regions in chromosomes 1q and 8q, and GPHMM-SEQ additionally detects the hemizygous deletions in chromosome 8p whereas FACETS erroneously predicts it as neutral heterozygosity. The results indicate that FACETS fails to explain some of the aberrations due to inaccurate estimation of the tumor impurity.
TABLE VI
The time usage of GPHMM-SEQ on real samples

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sample</th>
<th>SNP number</th>
<th>EM iterations*</th>
<th>Time(min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGA</td>
<td>SA018</td>
<td>28715</td>
<td>40</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>SA029</td>
<td>26568</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA030</td>
<td>29172</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA031</td>
<td>28146</td>
<td>79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA051</td>
<td>32979</td>
<td>89</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA052</td>
<td>24252</td>
<td>44</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA065</td>
<td>31176</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA069</td>
<td>26178</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA071</td>
<td>27380</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>SRA</td>
<td>SA1</td>
<td>73912</td>
<td>35</td>
<td>9.7</td>
</tr>
<tr>
<td></td>
<td>SA2</td>
<td>71319</td>
<td>40</td>
<td>11.4</td>
</tr>
<tr>
<td></td>
<td>SA3</td>
<td>72060</td>
<td>40</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>SA4</td>
<td>68802</td>
<td>40</td>
<td>11.6</td>
</tr>
<tr>
<td></td>
<td>SA5</td>
<td>85693</td>
<td>18</td>
<td>8.2</td>
</tr>
<tr>
<td></td>
<td>SA6</td>
<td>60606</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SA7</td>
<td>44460</td>
<td>16</td>
<td>6.1</td>
</tr>
<tr>
<td></td>
<td>SA8</td>
<td>50817</td>
<td>38</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>SA9</td>
<td>62483</td>
<td>40</td>
<td>10.5</td>
</tr>
</tbody>
</table>

*: the maximum number of iterations is set to 40.

IV. DISCUSSION AND CONCLUSION
GPHMM-SEQ is a novel algorithm for estimation of tumor impurity and ploidy, as well as detection of CNAs and LOH using sequencing data of paired tumor-normal samples. The read depth signals, GC-content, and mappability of germline heterozygous SNP positions are fully investigated. For data analysis, an integrated HMM is employed in GPHMM-SEQ to represent copy number changes of tumor genomes. Unlike existing methods that correct GC-content and mappability biases as a preprocessing procedure, GPHMM-SEQ takes joint representation of GC-content bias, mappability bias, tumor impurity, and aneuploidy in the emission models of the HMM, and makes integrated analysis of LCR and BAF data to simultaneously detect CNAs and LOH. The emission models in GPHMM-SEQ are carefully designed for analyzing tumor sequencing data while GPHMM itself cannot be used for this purpose. We employ cubic polynomial fit for GC-content and linear fit for mappability in the emission model of the HMM, which effectively characterizes the statistical behavior of the LCR data. Although the cubic polynomial fitting strategy is previously adopted in [24] as the preprocessing step, a significant disadvantage of such approach lies in the fact that it cannot effectively eliminate GC-content bias, because read depth data is also confounded by the issues of mappability bias, tumor impurity and aneuploidy, which often simultaneously arise in the same tumor sample and are strongly intertwined [16].

In addition, we use EM algorithm to automatically estimate the global model parameters such as tumor impurity and coefficients of GC-content. These features enable GPHMM-SEQ to have significant advantages over existing methods in two aspects. First, GPHMM-SEQ provides more accurate inferring of tumor impurity and ploidy. Second, the proper modelling of miscellaneous signals enables accurate detection of CNAs and LOH. By examining the performance of GPHMM-SEQ in both emulated and real sequencing datasets which provide reasonable representations of tumor impurity and aneuploidy issues, we demonstrate GPHMM-SEQ ability in accurately analyzing tumor sequencing data.

Despite of the superior performance mentioned above, GPHMM-SEQ has several limitations in practical applications. First, the model may be under fitted if the tumor genome is extremely amplified as the maximum copy number considered in GPHMM-SEQ is 7. Second, since GPHMM-SEQ only investigates germline heterozygous SNP positions, the resolution of aberration detection may be limited by the breakpoints in non-SNP genomic loci. Third, GPHMM-SEQ assumes the tumor samples are homogeneous. This assumption will not hold for tumor samples containing multiple subclonal populations. A potential improvement of GPHMM-SEQ is to extend the emission models in the HMM by accounting for multiple tumor subclones, which may lead to multiple solutions that give the similar likelihood of the data.

In conclusion, we demonstrate that integrated representation of GC-content bias, mappability bias, tumor impurity, and aneuploidy in single model enables superior performance of detecting CNAs and LOH using paired tumor-normal sequencing data. We expect that GPHMM-SEQ will be greatly beneficial for cancer-focused studies that aim to elucidate critical events involved in tumourigenesis and progression.

ACKNOWLEDGMENT
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 BCCA.

REFERENCES


