A Novel HMM for Analyzing Chromosomal Aberrations in Heterogeneous Tumor Samples

Hong Xia, Yuanning Liu, Minghui Wang, Huanqing Feng, Ao Li*
School of Information Science and Technology
University of Science and Technology of China
Hefei, China
sommer@mail.ustc.edu.cn ,
*Corresponding author: aoli@ustc.edu.cn

***Abstract*—Comprehensive detection and identification of copy number and LOH of chromosomal aberration is required to provide an accurate therapy of human cancer. As a cost-saving and high-throughput tool, SNP arrays facilitate analysis of chromosomal aberration throughout the whole genome. The performance of previous approaches has been limited to several critical issues such as normal cell contamination, aneuploidy and tumor heterogeneity. For these reasons we present a Hidden Markov Model (HMM) based approach called TH-HMM (Tumor Heterogeneity HMM), for simultaneous detection of copy number and LOH in heterogeneous tumor samples using data from Illumina SNP arrays. Through adopting an efficient EM algorithm, our method can correctly detect chromosomal aberration events in tumor subclones. Evaluation on simulated data series indicated that TH-HMM could accurately estimate both normal cell and subclone proportions, and finally recovery the aberration profiles for each clones.**

Keywords—chromosomal aberration; genotype; copy number; LOH; SNP arrray; heterogeneous tumor sample

I. INTRODUCTION

Cancer genome consists of various kinds of chromosomal aberrations, such as copy number gain or loss, loss of heterozygosity (LOH). They are related with oncogenes, tumor suppressor and gene expression during the process of tumor cell revolution [1]. Three main de facto standard tools for whole genome detection of chromosomal aberrations in tumor are aCGH (array Comparative Genome Hybridization) [2], SNP arrays (Single nucleotide polymorphism genotyping microarrays) [3] and NGS (Next Generation Sequencing) [4]. Among these, SNP arrays facilitate analysis of DNA copy number alterations (CNAs) and LOH throughout the whole genome by providing a cost-saving and high-throughput platform [5]. There are two measurements employed in SNP arrays called Log R Ratio (LRR) and B Allele Frequency (BAF). LRR is the log-normalized intensity ratio and represents for copy number intensity. BAF is a measure of normalized allelic intensity ratio of two alleles (A and B). These allele-specific measurements are widely adopted in the Illumina platform and hold the promise to tackle aberration events in tumor samples. It is notable that raw data from another SNP arrays based platform Affymetrix can also be transferred into the LRR and BAF signals through transformation and normalization [6].

Interpretation of chromosomal aberrations is complicated by several factors in real tumor samples, such as normal cell contamination, aneuploidy, and tumor heterogeneity [7]. Clinical tumor samples are usually contaminated with genetically normal cells, thereby complicating genomic analysis since the observed signal will be a combined value from both tumor and normal cells [5]. Aneuploidy occurs during cell division when chromosomes do not separate properly between the two cells resulting extra or missing chromosomes in the cell, leading to the LRR baseline shift [7]. Tumor heterogeneity can be explained by subsequent copy number alterations taking on several times during tumor progression due to the associated innate genetic and epigenetic instability of cancer cells [8]. As a consequence, multiple subclones arise and may have unique biologic characteristics including alterations in oncogenes and tumor suppressor genes. Among all these factors, tumor heterogeneity implies the biologic reasons for the natural history of neoplasms, and leads to drug resistance resulted in a diminution cure rate in clinic [9]. However, most of exiting methods are unable to address aforementioned critical issues synthetically because the issues interact with each other and cannot be settled separately.

Various approaches have been proposed to handle genomic aberrations using SNP arrays, such as GAP [5], ASCAT [6] and GPHMM [6]. However, they are not specifically designed to detect tumor heterogeneity thus will result in failure in accurately indicating potential multiple cancer cells [6]. As the "state-of the art" approach, OncoSNP is designed to model the contributions from aneuploidy, normal cell contamination and tumor heterogeneity at the same time [10]. However, a major drawback of OncoSNP is that it cannot accurately estimate normal cell proportion. Instead it is done by grid searching from 0 to 1 with step size of 0.1 for the optimal value, which will lead to poor performance if normal cell proportion cannot be accurately determined by this way. Another weakness of OncoSNP is that it processes consecutive SNPs independently and assigns

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

different kinds of subclones for them, which is computational intensive and prone to overfitting problem [10].

Motivated by the need to systematically investigate the effects of tumor heterogeneity in cancer samples as well as aneuploidy and normal cell contamination, in this paper we introduce a novel method, named Tumor Heterogeneity HMM (TH-HMM), which utilizes a parameterized Hidden Markov Model (HMM) and an efficient EM algorithm [11]. TH-HMM provides a comprehensive statistical framework to describe SNP arrays signals from heterogeneity tumor samples. In order to validate our approach, we generated a series of simulated heterogeneous tumor samples. After comprehensive analysis of the estimated results and compared to GPHMM and OncoSNP, we demonstrate TH-HMM is efficient in addressing heterogeneous tumor samples.

II. METHODS

A. Hidden states of TH-HMM

All copy number aberration events are described using a pair of tumor and normal genotypes [9]. For example, let A and B be the two alleles of each SNP locus, then the one copy amplification of either allele can be denoted by four genotype pairs associated with possible normal genotypes: (AAA, AA), (BBB, BB), (AAB, AB), (ABB, AB). Totally 21 states are defined in TH-HMM with copy number up to 7, as shown in TABLE I.

B. HMM framework and EM algorithm

TH-HMM adopts a reasonable assumption that tumor sample can be confounded by two kinds of cancer subclones with different aberration patterns, thus leading to three combinational cases for specific regions: chromosome segment with aberration only in subclone1 (t = 1); chromosome segment with aberration only in subclone2 (t = 2) and chromosomal segment with the same aberration in both tumor subclones (t = 3). We use a linear model to compute the weighted average copy number $y_{c,t}$ and the weighted BAF value $z_{c,k,t}$ as the following equations:

$$p_1 = w_1, p_2 = w_2, p_3 = w_1 + w_2 \quad (1)$$

$$y_{c,t} = p_t n_c + (1 - p_t) n_s \quad (2)$$

$$z_{c,k,t} = p_t n_c u_{c,k} + (1 - p_t) n_s u_{s,k} \quad (3)$$

where p_t is the proportion of tumor cells with the corresponding chromosomal aberration in case t. w_1 and w_2 are proportions of subclone1 and subclone2. For example, in case 3, tumor aberration appears on both tumor clones, thus the proportion of aberrant tumor cells equals to the sum of w_1 and w_2. n_s and n_c denote normal copy number and tumor copy number. $u_{s,k}$ and $u_{c,k}$ are theoretical mean BAF of normal and tumor cells.

We assume that LRR and BAF signals are normally distributed with standard deviation of σ_l and σ_b, respectively [12]. Thus, the emission probability density function of LRR and BAF in i^{th} SNP under t^{th} case can be formulated according to empirical formulas proposed in [13]:

TABLE I. HIDDEN STATES OF TH-HMM MODEL

State	CN	Description	(Tumor genotype, normal genotype)
0	N/A	Fluctuation effect	(N/A, AA), (N/A, BB), (N/A, AB)
1	0	Deletion of two copies	(N/A, AA), (N/A, BB), (N/A, AB)
2	1	Deletion of one copy	(A,AA), (B,BB), (A,AB), (B,AB)
3	2	Normal	(AA,AA), (BB,BB), (AB,AB)
4	2	Copy neutral with LOH	(AA,AA), (AA,AB), (BB,BB), (BB,AB)
5	3	Three copies with duplication of one allele	(AAA,AA), (BBB,BB), (AAB,AB), (ABB,AB)
6	3	Three copies with LOH	(AAA,AA), (AAA,AB), (BBB,BB), (BBB,AB)
7	4	Four copies with duplication of one allele	(AAAA,AA), (BBBB,BB), (AAAB,AB), (ABBB,AB)
8	4	Four copies with duplication of both alleles	(AAAA,AA), (BBBB,BB), (AABB,AB)
9	4	Four copies with LOH	(AAAA,AA), (BBBB,BB), (AAAA,AB), (BBBB,AB)
10	5	Five copies with duplication of one allele	(AAAAA,AA), (BBBBB,BB), (AAAAB,AB), (ABBBB,AB)
11	5	Five copies with duplication of both alleles	(AAAAA,AA), (BBBBB,BB), (AAABB,AB), (AABBB,AB)
12	5	Five copies with LOH	(AAAAA,AA), (BBBBB,BB), (AAAAA,AB), (BBBBB,AB)
13	6	Six copies with balanced duplication of both alleles	(AAAAAA,AA), (BBBBBB,BB), (AAABBB,AB), (AAABBB,AB)
14	6	Six copies with duplication of both alleles	(AAAAAA,AA), (BBBBBB,BB), (AAAABB,AB), (AABBBB,AB)
15	6	Six copies with duplication of one alleles	(AAAAAA,AA), (BBBBBB,BB), (AAAAAB,AB), (ABBBBB,AB)
16	6	Six copies with LOH	(AAAAAA,AA), (BBBBBB,BB), (AAAAAA,AB), (BBBBBB,AB)
17	7	Seven copies with duplication of both alleles	(AAAAAA,AA), (BBBBBB,BB), (AAAABBB,AB), (AAABBBB,AB)
18	7	Seven copies with duplication of both alleles	(AAAAAAA,AA), (BBBBBBB,BB), (AAAAABB,AB), (AABBBBB,AB)
19	7	Seven copies with duplication of both alleles	(AAAAAAA,AA), (BBBBBBB,BB), (AAAAAAB,AB), (ABBBBBB,AB)
20	7	Seven copies with LOH	(AAAAAAA,AA), (BBBBBBB,BB), (AAAAAAA,AB), (BBBBBBB,AB)

$$f(l_i|w_1, w_2, o, h, \sigma_l, c, t) = \frac{1}{\sigma_l} \Phi\left(\frac{l_i - (2log10(\frac{y_{c,t}}{2}) + o + hg_i)}{\sigma_l}\right) \quad (4)$$

$$f(l_b|w_1, w_2, \sigma_b, c, t) = \frac{1}{\sigma_b} \sum_{k=1}^{g_c} p_0(k)\, \Phi\left(\frac{b_i - \frac{z_{c,k,t}}{y_{c,t}})}{\sigma_b}\right) \quad (5)$$

where o represents LRR baseline shift, and h denotes the coefficient of GC content (g_i).

EM algorithm is employed to estimate optimal parameters (o, h, σ_l, σ_b, w_1 and w_2) by finding maximum overall likelihood in our statistical framework. In the E step, for LRR we first calculate the partial log likelihood expectation using formula (6).

$$E(LL_I) = \sum_{i=1}^{N} \sum_{c=1}^{C} \gamma_{i,t}{}^{(n)}(c)\, log\big(f(l_i|w_1, w_2, o, h, \sigma_l, c, t)\big)$$
$$= \sum_{i=1}^{N} \sum_{c=1}^{C} \gamma_{i,t}{}^{(n)}(c)\, (log\tfrac{1}{\sqrt{2\pi}} - log(\sigma_l) - \tfrac{(l_i - (2log10(\frac{y_{c,t}}{2}) + o + hg_i))^2}{2(\sigma_l)^2}) \quad (6)$$

where $\gamma_{i,t}{}^{(n)}(c)$ represents the posterior probability of the i^{th} SNP in state c in case t [11]. C denotes the total number of hidden states. We process BAF signals in the similar way as LRR signals.

$$E(LL_b) = \sum_{i=1}^{N} \sum_{c=1}^{C} \gamma_{i,t}{}^{(n)}(c)\, log\big(f(b_i|w_1, w_2, \sigma_b, c, t)\big)$$

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

$$= \sum_{i=1}^{N}\sum_{c=1}^{C}\gamma_{i,t}{}^{(n)}(c)\sum_{k=1}^{g_c}p_0(k)\left(log\frac{1}{\sqrt{2\pi}} - log(\sigma_b) - \frac{(b_i - \frac{z_{c,k,t}}{y_{c,t}})^2}{2(\sigma_b)^2}\right) \quad (7)$$

In the M step, transition probabilities between hidden states are updated from the Baum Welch algorithm [11]. The updating formulas for o, h, σ_l, σ_b are obtained by taking the partial derivative of equation (6) and (7) to 0 and then solving the equations. For example, o can be estimated as follows:

$$\frac{\partial(E(LL_l))}{\partial o} = \sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{t=1}^{M}\gamma_{i,c,j}\frac{-1}{2(\sigma_l)^2}\cdot 2\left(2log10\left(\frac{y_{c,t}}{2}\right) + o + hg_i - l_i\right) = 0 \quad (8)$$

$$o^{(n+1)} = \frac{\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{t=1}^{M}\gamma_{i,c,j}\left(l_i - 2log10\left(\frac{y_{c,t}}{2}\right) - hg_i\right)}{\sum_{i=1}^{N}\sum_{c=1}^{C}\sum_{t=1}^{M}\gamma_{i,c,j}} \quad (9)$$

However, there is no close-form formula in updating w_1 and w_2 in EM algorithm. Alternatively, Newton-Raphson method is employed, which can efficiently increase the expectation of the partial log-likelihood in each M-step numerically and therefore the overall likelihood [14]. After EM algorithm converges, parameters in the last iteration will be output as optimal values and final copy numbers and LOH status for SNPs are determined by the corresponding hidden states (c) associated with the largest conditional probability.

C. Generation of Simulated data

For the purpose to verify the performance of TH-HMM, we generate simulated samples with pre-defined aberrations and proportions of the cellular components (subclone1, subclone2 and normal cell) to imitate real heterogeneous tumor samples contaminated with normal cells. The simulated dataset is originated from the diploid HapMap sample NA06991 hybridized on an Illumina BeadChip [15]. We first define various chromosomal aberrations for three cases and compute the mean value of LRR and BAF of SNPs according to its aberration types by empirical formula [13]. By sampling from the associated normal distributions of LRR and BAF signals using equation (4) (5), simulated tumor SNP data are generated to examine the ability of TH-HMM for accurately recovering the corresponding subclone proportion and chromosomal aberrations. . For generality, we create a simulated dataset of 16 samples by considering various percentages ranging from 0 to 85% of the three proportions (normal cell proportion, w_1 and w_2) and use the names of these samples to show the actual proportions of subclones. For example, Simu-SNP-50-45 means that sub-clone1 and sub-clone2 take 50% and 45% in the simulation mixture respectively.

With the aim to mimic real tumor SNP array data, baseline shift and GC bias are also considered in data simulation. Chromosomal segments with intra-tumor heterogeneity are generated according to the proportion and aberration states for each subclone and the amount of normal cell contamination per sample. To determine genotypes in simulated tumor data, we make an assumption that tumor genotypes are generated from normal genotypes. For example, when the normal genotype is homozygous (AA), the tumor genotype can only be homozygous (A, AAA). On the other hand, when the normal genotype is heterozygous (AB), the tumor genotype can be either heterozygous (AAB, ABB) or homozygous (AAA, BBB) [9].

III. RESULTS

A. Evaluation of normal proportions

To demonstrate the validity of our approach, we applied TH-HMM to the simulated dataset of heterogeneous tumor samples. We also examined the performance of OncoSNP and GPHMM for comparison. It is of note that many chromosomal aberrations cannot be correctly detected if normal cell proportion is wrongly estimated in tumor samples. For example, tumor cells (AAAAB) mixed with 80% normal cells will be predicted as (AAB) when normal cell proportion is detected as 50%, as in this case the BAF signals for these two situations are very similar. Therefore, we first assessed the estimated normal cell proportion of simulated dataset using TH-HMM, OncoSNP and GPHMM with respect to actual proportions, as shown in Fig. 1. The estimated normal cell proportions of GPHMM have an average error of 15%. It performs well when there is only one tumor subclone (sample Simu-SNP-00-25), but incapable in predicting normal cell proportion in heterogeneity samples. The deviations of OncoSNP vary from -27%~40%. When the normal proportion is 0.1, oncoSNP cangive correct results (*Simu-SNP-50-40*, *Simu-SNP-55-35*, *Simu-SNP-65-25*, *Simu-SNP-75-15*), but in the rest cases it fails to estimate the right normal proportions. Overall, TH-HMM outperforms GPHMM and OncoSNP with precise estimation of normal cell estimations throughout the whole dataset.

B. Performance on unraveling the proportion of two subclones

When referring to each subclone, OncoSNP is helpless in determining its exact proportion in tumor sample, and cannot profile clone-specific aberrations. On the other hand, TH-HMM provides proportions of two sub-clones with high accuracy for simulated data (TABLE II). We considered

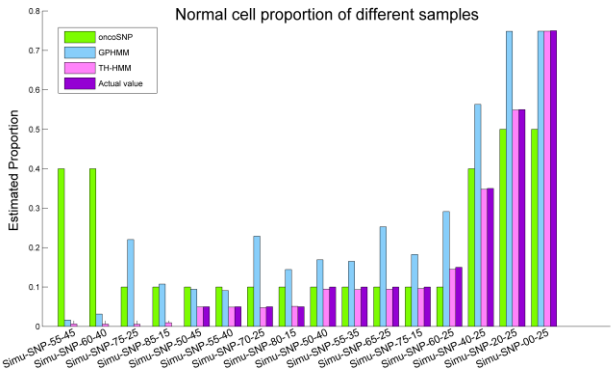

Fig. 1. Comparision of estimated normal cell proportion in 16 diluted samples of three methods. The vertical axis is the estimated proportion and its correspondence actual value. The gap around 0 between TH-HMM (red) and the expected value (purple) of the entire dataset demonstrates TH-HMM results highly coincide with the actual values. GPHMM always show higher normal cell contamination in all the samples (blue bar). The sharp fluctuation of OncoSNP (green) reveals disparate performance of OncoSNP in diverse samples, and the difference even reach to 40% in two samples.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

TABLE II. RESULTS OF THE PROPORTIONS OF TWO TUMOR SUB-CLONES.

SampleID	w_1	w_2	SampleID	w_1	w_2
Simu_SNP_00_25	0.000	0.251	Simu_SNP_60_25	0.603	0.253
Simu_SNP_20_25	0.201	0.249	Simu_SNP_60_40	0.590	0.389
Simu_SNP_40_25	0.401	0.252	Simu_SNP_65_25	0.652	0.253
Simu_SNP_50_40	0.502	0.401	Simu_SNP_70_25	0.698	0.248
Simu_SNP_50_45	0.498	0.448	Simu_SNP_75_15	0.750	0.152
Simu_SNP_55_35	0.553	0.351	Simu_SNP_75_25	0.739	0.240
Simu_SNP_55_40	0.548	0.397	Simu_SNP_80_15	0.793	0.148
Simu_SNP_55_45	0.540	0.439	Simu_SNP_85_15	0.833	0.140

samples with mixed proportions of different cellular components under three distinctive situations: 1) slight difference in the proportions between two subclones; 2) significant difference in the proportions between two subclones; and 3) pure heterogeneity tumor samples without normal cell contamination. The results show that 12 out of the 16 samples have an error less than 1%, and the rest with a maximum error about 1.7%, which indicates that TH-HMM performs very well under three distinctive situations mentioned above.

C. *Performance on detection subclone aberration*

We further investigated the utility of TH-HMM in detecting various kinds of chromosomal aberrations. Fig. 2 shows the results of chromosome 1p and 3p in a simulation sample with 45% and 35% of two subclones, respectively. Consecutive SNPs with the same copy number and LOH states are plotted as a segment. The top panel presents LRR and BAF signals of SNPs and the red lines indicate average signal values of different chromosomal segments. In Fig. 2a two subclone have the same aberration states on chromosome 1p (case 1). Under this situation, TH-HMM identifies all the aberration including one copy deletion, amplification and LOH successfully. In contrast, OncoSNP can detect the copy aberration regions and LOH events, but incapable in estimating the right copy number and neglects deletion event.

In Fig. 2b, tumor subclone2 has distinct aberrations and takes 25% on chromosome 3p, while subclone1 is normal on this region, which imitates heterogeneous tumor cells with different abnormalities of the two subclones (case3). In this case TH-HMM still correctly identifies all the copy number and LOH for all the regions of subclone2 (blue), whereas the result figure of OncoSNP does not conform to the actual solution. Although the inconsistency of normal cell proportions of consecutive chromosome regions reveals heterogeneity level in the sample [10], OncoSNP cannot discriminate heterogeneity tumor subclones. The CN plot shows that two of the LOH parts are missed and segment copy numbers are wrongly interpreted. Taken together, we come to the conclusion that the capability of TH-HMM in detection tumor subclone aberrations is much better than OncoSNP.

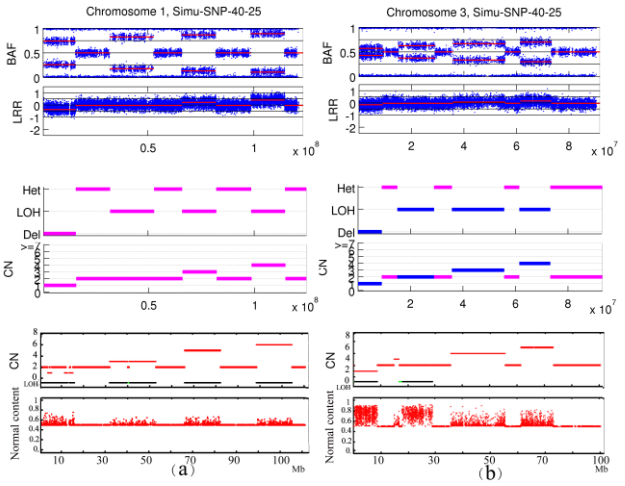

Fig. 2. Examples of heterogeneity tumor sample SNP signals (top) and the estimated aberrations of TH-HMM (middle) and OncoSNP (bottom). a) Two subclone have the same aberration states on chromosome 1 (case 1), which are correctely identifed by TH-HMM (illustrated by magenta). OncoSNP recognises the LOH events (black) while gives the wrong tumor copy numbers at aberration segments. b) When only one of the subclones has aberrations on chromosome3 (case 3), TH-HMM dissected them successfully (blue). OncoSNP misses the LOH detection in segment 3and 4, and even wrongly interprets segment copy numbers.

Finally, we calculated the accuracy (defined as the proportion of all correctly identified SNPs), and recall as fraction of the aberrations that are successfully retrieved. The average accuracy of the whole dataset is 0.953, demonstrating high consistency of our method. Meanwhile, 13 out of the 16 samples have recalls of both clones close to 1. Even for samples with extremely low subclone proportion (*Simu-SNP-75-15*, *Simu-SNP-80-15*, *Simu-SNP-85-15*), TH-HMM still achieves high performance with recalls of around 90%. These results demonstrate the excellent performance of TH- HMM in analyzing heterogeneity tumor samples.

D. *Performance on real data*

We investigated the CRL-2324 breast cell line to further evaluate the efficiency of TH-HMM. The results reveal that CRL-2324 breast cell line is complicated by the numerous genomic rearrangements associated with intratumoral heterogeneity on chromosome 4. In Fig. 3a, two magenta regions of common copy neutral LOH and one copy gain for both subclones may indicate aberrations from the cancer ancestor cells. Individual tumor subclones (red and blue) harbor private genetic aberrations (five copies amplification and three copies amplification regions for subclone1 and a small five copies amplification region for tumor subclone2) in addition to the founder mutations, and that these subclonal copy number aberrations could help scientists to study the impact of tumor heterogeneity on resistance to therapy. Moreover, for the purpose to evaluate the consistency of TH-HMM we observed the copy number and LOH profiles of CRL-2324 dilution series with 79% tumor content (Fig. 3b). The overall view of Fig. 3b is very similar with Fig. 3a, which

978-1-4799-1389-3/13/$31.00 ©2013 IEEE

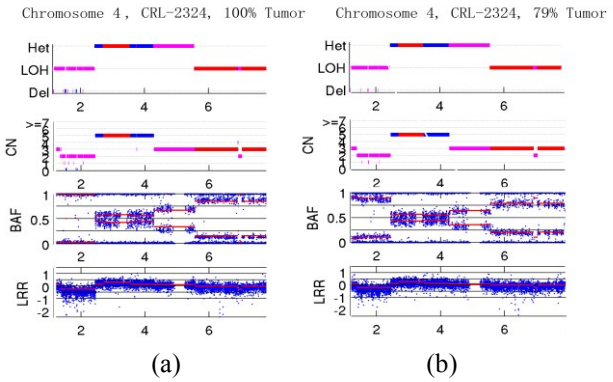

Fig. 3. Plots of aberration regions on chromosome 17 and the results of TN-HMM for CRL-2324 dilution series data. (a) Cell line; (b) Tumor sample contaminated with 21% normal cells. Red lines indicate subclone1 aberrations and blue lines indicate subclone 2 aberrations. Common copy neutral LOH and one copy gain of both tumor clones are indicated by magenta. The overall results show high consistency of TH-HMM in detecting real tumor sampkes.

proves high consistency of TH-HMM in detecting normal cell contaminated heterogeneous tumor samples.

IV. CONCLUSION

Studying subclonal heterogeneity on analysis of tumor biopsies is a crucial factor to explain tumor cell evolutionary and pathology. In this study we proposed a novel method to identify deletion, amplification, and LOH events in heterogeneity tumor samples. The experimental results demonstrate the robust performance of our algorithm. Fundamental to the success for our method is the integrated statistical TH-HMM framework taking the combination effects of normal cell contamination, aueploidy and tumor heterogeneity jointly. The ability of our proposed method to estimate the proportions of multiple tumor subclones and specify detailed information of genomic aberrations for them will greatly facilitate cancer research.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (31100955, 61101061), Fundamental Research Funds for the Central Universities (WK2100230007).

REFERENCES

[1] Stratton MR, Campbell PJ, Futreal PA: The cancer genome. Nature 2009, 458:719-724.

[2] Park PJ: Experimental design and data analysis for array comparative genomic hybridization. Cancer Invest 2008, 26:923-928.

[3] Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, et al: High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 2006, 16:1136-1148.

[4] Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A: Statistical challenges associated with detecting copy number variations with next-generation sequencing. Bioinformatics 2012, 28:2711-2718.

[5] Sun, W., Wright, F.A., Tang, Z., Nordgard, S.H., Van Loo, P., Yu, T., Kristensen, V.N. and Perou, C.M. Integrated study of copy number states and genotype calls using high-density SNP arrays. Nucleic Acids Res, 2009, 37, 5365-5377.

[6] Li A, Liu Z, Lezon-Geyda K, Sarkar S, Lannin D, Schulz V, Krop I, Winer E, Harris L, Tuck D : 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 Research, 2011, 39:4928-4941.

[7] Rasmussen M, Sundstrom M, Goransson Kultima H, Botling J, Micke P, Birgisson H, Glimelius B, Isaksson A: Allele-specific copy number analysis of tumor samples with aneuploidy and tumor heterogeneity. Genome Biol 2011, 12:R108.

[8] Parisi F, Ariyan S, Narayan D, Bacchiocchi A, Hoyt K, Cheng E, Xu F, Li P, Halaban R, Kluger Y. Detecting copy number status and uncovering subclonal markers in heterogeneous tumor biopsies. BMC Genomics 2011, 12:230.

[9] Diaz-Cano SJ: Tumor heterogeneity: mechanisms and bases for a reliable application of molecular marker design. Int J Mol Sci. 2012;13(2):1951-2011.

[10] Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G, Harris A,Ragoussis J, Sieber O, Holmes CC: A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data. Genome Biol 2010, 11:R92.

[11] Rabiner,L.R: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, 1989, 77, 257–286.

[12] Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, Bucan M: PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007, 17:1665-1674.

[13] Nancarrow, D.J., Handoko, H.Y., Stark, M.S., Whiteman, D.C. and Hayward, N.K: SiDCoN: a tool to aid scoring of DNA copy number changes in SNP chip data. PLoS One. 2007, 2, e1093.

[14] Dempster, A.P.; Laird, N.M.; Rubin, D.B.: Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 1977, 39 (1): 1–38

[15] Staaf J, Lindgren D, Vallon-Christersson J, Isaksson A, Göransson H, Juliusson G, Rosenquist R, Höglund M, Borg A, Ringnér M: Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. Genome Biol. 2008;9(9):R136.

978-1-4799-1389-3/13/$31.00 ©2013 IEEE