Bio-Medical Materials and Engineering 26 (2015) S1845–S1853
DOI 10.3233/BME-151487
IOS Press

A novel framework for analyzing somatic copy number aberrations and tumor subclones for paired heterogeneous tumor samples

Hong Xia, Ao Li*, Zhenhua Yu, Xiaocheng Liu and Huanqing Feng
School of Information Science and Technology, University of Science and Technology of China, Huangshan Road No. 443, 230027, Hefei, China

Abstract. Application of the Next generation sequencing (NGS) technology has demonstrated that most tumor samples exhibit intra-tumor heterogeneity. Here we proposed SAPPH (Somatic Aberrations Prediction for Paired Heterogeneous tumor samples), as a new method for estimating tumor somatic copy number aberrations as well as inferring tumor subclone proportions from heterogeneous tumor sequencing data. This method is based on CBS and local proportion clustering strategy. When SAPPH is applied on simulated tumor samples, the agreement between the results analyzed by SAPPH and the sequencing signals suggests that SAPPH can find the solution to best fit the signal distributions. We benchmark the performance of SAPPH and show that it outperforms existing method in estimating tumor copy number aberrations.

Keywords: Intra-tumor heterogeneity, NGS, copy number aberration, tumor subclone, proportion

1. Introduction

Tumor subclonality, also known as intra-tumor heterogeneity, is a typical feature of human cancers that has attracted much attention over recent years [1, 2]. Somatic copy number aberrations (CNA) that appear during tumor involution result in different genomic profiles between cancer cells, which give rise to tumor subclones [3]. By studying CNA in different subclones of a large collection of tumors can lead to the discovery of novel oncogenes and tumor suppressor genes. This will in turn facilitate designing effective treatment strategies [4].

The advent of next-generation sequencing (NGS) provides a great opportunity for accurate identification of copy number aberrations and composition of tumor samples [5-7]. The NGS technology provides coverage of each genomic site by a number of short reads. Window-averaged number of reads along the genome can be used to obtain copy number, which is quantified by the Log

* Address for correspondence: Ao Li, School of Information Science and Technology, University of Science and Technology of China, Huangshan Road No. 443, 230027, Hefei, China. Tel.: +86 13865983702; Fax: +86-551-636-01800; E-mail: aoli@ustc.edu.cn.

R ratio (LRR). LLR is defined as the logarithmic ratio between the read count of tumor sample and that of the paired normal sample. B allele Frequency (BAF) for each known annotated single nucleotide polymorphism (SNP) is calculated as the fraction of B-allele read counts (B-count) to the total read counts (T-count). BAF represents the probability of mapping the B-allele [8, 9].

Determination of genomic aberrations and subclonality for intra-tumor heterogeneous NGS data is very challenging due to the following reasons. Firstly, due to the large size and high signal noise in the sequencing data make the computational analysis complicated. Secondly, copy number aberrations in the tumor samples are always diluted in sequencing signals due to the contamination of tumor sample with the normal cell. Therefore the admixed normal DNA attenuates measured signals representing genomic aberrations in tumor DNA, which results in the decrease of both signal-to-noise ratios in sequencing data and the performance of aberration detection [10-12]. Thirdly, due to the fixed total number of reads in genome sequencing, large copy number aberrations in the part of the genome will cause the observed number of mapped reads in normal genome regions to deviate from the expected value [11]. Therefore, finding the sequencing data baseline for normal genomic state is critical for the estimation of copy numbers for other genome regions. Finally, when tumor sample is known to be homogeneous, the tumor cell proportion of the sample can be deduced from the observed average copy number and BAF pattern for the genome patterns. Whereas if genomic CNAs present in only a part of the tumor cells, the average copy number and BAF may not provide sufficient information to determine proportions of different tumor subclones.

This paper focuses on modelling and analysis of the composition and genomic aberrations of heterogeneous tumor sample using NGS data. We proposed a method named SAPPH, which utilizes paired tumor-normal sequencing data. Based on an enhanced CBS method, we divided the genome intervals into two parts: high confidence genome intervals and the rest. Next, we formulated an explicit probabilistic model to estimate the mixing ratio for each CNA in high confidence genome intervals separately. In the following step, candidate tumor cell proportions were calculated by clustering the approximate local ratios. Searching the candidate proportions and computing corresponding CNAs to find a combination that best fit LRR and BAF signal distributions could realize subsequent analysis of the remaining intervals. At last, Bayesian information criterion (BIC) was used to guide model selection for final determination of the number of tumor subclones. We applied SAPPH to the simulated datasets and compared it to the results of THetA [11] to demonstrate its ability in estimating tumor subclone copy number aberrations.

2. Methods

2.1. Data generation and preprocessing

Three different tumor genomes and a normal genome were constructed to simulate intra-tumor heterogeneous samples by making different combinations between tumor and normal genomes at known proportions. We constructed each genome by randomly dividing the reference genome into a number of regions, and then assigned each region with a known aberrational state characterized by total copy number c and B-allelic number b. Simulated sequencing data were generated from a real normal sample (HCC1143_BL) that was downloaded from https://cghub.ucsc.edu/datasets/benchmark_download.html by the following steps: 1) For each region of the simulated genome, we repeatedly sampled $c \cdot n/2$ reads from the corresponding region of HCC1143_BL genome, where n is the number of reads mapped to HCC1143_BL; 2) the nucleotide

sequences of the sampled reads were modified to match BAF values of heterozygous SNPs encompassing the genome region; 3) repeating step 1 and step 2 until the reads of the whole genome were produced; and 4) the simulated reads after processing were merged and sorted to generate BAM file using Samtools software [13].

We simulated a total of 23 tumor samples, and then classified them into three simulated datasets, each containing tumor samples with one, two or three kinds of subclones, respectively. The tumor subclone with the highest cellular proportion in the sample is typically called the dominant clone (subclone 1), and those with lower cellular proportions are called subclone 2 subclone 3, and so on. These subclones may be nested within the dominant clone. It should be noted that each locus in the genome can experienced only one kind of copy number aberration in different tumor subclones. For each sample, LRR signal was computed as the read count ratio between the tumor genome and was matched to the normal genome. We excluded homozygous SNPs (with BAF > 0.95) for analysis as they were not directly informative [6]. Since BAF signal is centered on 0.5, we used the mirrored profile to better model the statistical distribution.

2.2. Segmentation

The aberrations in copy number are a result of genomic events that cause discrete gains and losses in contiguous segments of the genome. We adopted the enhanced circular binary segmentation (CBS) that has been implemented in an R package "DNAcopy" to divide chromosomes into intervals of constant copy number [14].

In any tumor, the possible range and distribution of BAF signal is useful to deduce tumor subclone proportions as well as the CNA type. Here we segmented the mirrored BAF and then synthesized the results with the LRR segmentation outcome. Therefore, the tumor genome was partitioned into J consecutive segments or intervals, and each possesses a constant copy number and B-allele number, which is reflected in mean LRR (m_lrr_j) and mean BAF (m_baf_j). Each interval contains a number of SNP sites, and we use (i, j) to denote the $i-th$ SNP site in interval j with $i \in [1,...,N_j]$, where N_j is the total number of SNP sites in interval j.

2.3. Baseline identification

During the sequencing, large amount of copy number aberrations will cause the observed number of mapped reads in an interval to deviate from expected value, even when the interval itself does not experience a CNA. Therefore, in normalized NGS data, absolute copy numbers depend not only on LRR values, but also on the baseline of normal copy number for entire chromosomes, which is reflected in LRR baseline shift.

The alignment of the tumor genome with the paired normal genome, intervals with the most similar BAF signal distributions between tumor genome and normal genome can only be either the normal genotype (AB), or balanced copy number gain (AABB, AAABBB, and so on). These mapping reads to a genomic site can be treated as a Bernoulli trial, given that T-counts override a SNP position, and the B-counts at the corresponding SNP site are modeled by a binomial distribution [15]. We used the binomial distribution with mapping probability 0.5 to calculate the probability of intervals belonging to a copy neutral state. For SNP site (i, j) in tumor and normal samples, let $bc_{i,j}^{T}$ and $bc_{i,j}^{N}$ denote the B-count, $tc_{i,j}^{T}$ and $tc_{i,j}^{N}$ denote the T-count, respectively. The probability of interval j being in

copy neutral state for tumor and normal samples (P_j^T and P_j^N) can be formulated as:

$$P_j^T = \sum_i \binom{tc_{i,j}^{\ T}}{bc_{i,j}^{\ T}} \cdot 0.5^{bc_{i,j}^{\ T}} (1-0.5)^{tc_{i,j}^{\ T} - bc_{i,j}^{\ T}} = \sum_i \binom{tc_{i,j}^{\ T}}{bc_{i,j}^{\ T}} \cdot 0.5^{tc_{i,j}^{\ T}} \tag{1}$$

$$P_j^N = \sum_i \binom{tc_{i,j}^{\ N}}{bc_{i,j}^{\ N}} \cdot 0.5^{tc_{i,j}^{\ N}} \tag{2}$$

Interval j is copy neutral if

$$abs(P_j^T - P_j^N) < \sigma_b / 2 \tag{3}$$

where σ_b is the average variance of tumor BAF in all intervals, and $\sigma_b / 2$ is the threshold for the difference between tumor and normal genotyping signals.

An interval can be ascribed to the copy neutral state when it satisfies Eq. (3). Among the copy neutral intervals, the ones that have the minimum mean LRR value can be identified as normal genotype with copy number 2, and the corresponding mean LRR is considered the baseline. This simple technique takes full advantage of the paired normal samples and is very effective in ascertaining LRR baseline shifts.

2.4. Selection and analysis of high confidence segments

The proportions of somatic aberrations for genome segments always cluster around a small number of distinct proportion 'modes'. This suggests that somatic aberrations of similar proportions may reside in the same subclone of tumor cells [15]. We addressed this issue by selecting high confidence intervals (HCI) for analysis that met one of the following criteria: (a) intervals with copy number deletion; and (b) intervals with mean mirrored BAF $> 0.5 + \sigma_b$. After filtering, we estimated a local mixing ratio for each HCI, called partial proportion. This is the ratio between the normal cell with the normal state and a single tumor subclone carrying the local CNA. This is based on the assumption that each locus experiences only one kind of copy number aberration. For genomic interval j, the average copy number y_j and average B-allele number z_j can be calculated as:

$$m\_lrr_j = \log_{10}(y_j / 2) + o \tag{4}$$

$$y_j = 10^{(m\_lrr_j - o)} \cdot 2 \tag{5}$$

$$z_j = y_j \cdot m\_baf_j \tag{6}$$

where o is the LRR baseline shift. Let w_j be the partial tumor proportion in interval j, the tumor copy number c_j and B-allele number b_j are restricted by following equations.

$$y_j = c_j w_j + 2(1-w_j) \tag{7}$$

$$z_j = b_j w_j + (1 - w_j) \tag{8}$$

Since there are limited combinations of c_j and b_j ($b_j <= c_j$), we developed a family of curved lines to model (y_j, z_j) with the change of w_j [13]. Each line corresponds to a unique combination of (c_j, b_j) and was named a canonical line. For each interval, the task is to scan all the canonical lines to find the one contains point (y_j, z_j). Due to noise and segmentation bias, (y_j, z_j) may not locate precisely on a canonical line. In this situation we find the line that contains the point closest to (y_j, z_j). Consequently, partial tumor proportions for all HCIs are calculated. We also aggregated partial tumor proportions to G centers corresponding to candidate tumor subclone proportions.

2.5. Whole genome analysis and model selection

With candidate tumor subclone proportions $R = \{r_1, r_2, ..., r_G\}$, ranging by the number of SNP probes, the complexity to analyze all genome intervals can be greatly reduced. For each interval, we could find one optimal candidate proportion with the corresponding aberrational states. This maximizies the likelihood of LRR and BAF for this interval. Supposing the optimal local tumor proportions, copy numbers and B-allele numbers for all intervals are denoted as $W = \{w_1, ..., w_j\}$, $C = \{c_1, ..., c_j\}$, $B = \{b_1, ..., b_j\}$, total likelihood of the whole genome can be calculated as follows:

$$L(W, C, B \mid \text{LRR}) = \sum_j \sum_i f(lrr_{i,j}^T \mid w_j, c_j, b_j) = \sum_j \sum_i \frac{1}{\sqrt{2\pi}\sigma_l} \exp(-\frac{(lrr_{i,j}^T - c_j w_j - 2(1-w_j))}{2\sigma_l^2}) \tag{9}$$

$$\begin{aligned} L(W, C, B \mid \text{BAF}) &= \sum_j \sum_i f(tc_{i,j}{}^T, bc_{i,j}{}^T \mid w_j, c_j, b_j) \\ &= \sum_j \sum_i \binom{tc_{i,j}{}^T}{bc_{i,j}{}^T} \cdot \left(\frac{b_j w_j + (1 - w_j)}{c_j w_j + 2(1-w_j)}\right)^{bc_{i,j}{}^T} \left(1 - \frac{b_j w_j + (1 - w_j)}{c_j w_j + 2(1-w_j)}\right)^{tc_{i,j}{}^T - bc_{i,j}{}^T} \end{aligned} \tag{10}$$

where $lrr_{i,j}^T$ represents LRR signal of SNP site (i, j) in tumor samples. σ_l is the average variance of tumor LRR for all intervals. Note that the likelihood is greatest when we use the full set of candidate proportions as the genome intervals can be fitted to the maximum extent with enough tumor subclones. However, this complex model might outstrip the real situation. Therefore, we adopted the Bayesian information criterion (BIC) to select a model with a balance between higher likelihood and fewer tumor subclones.

3. Results

3.1. Prediction of tumor subclone copy number aberrations

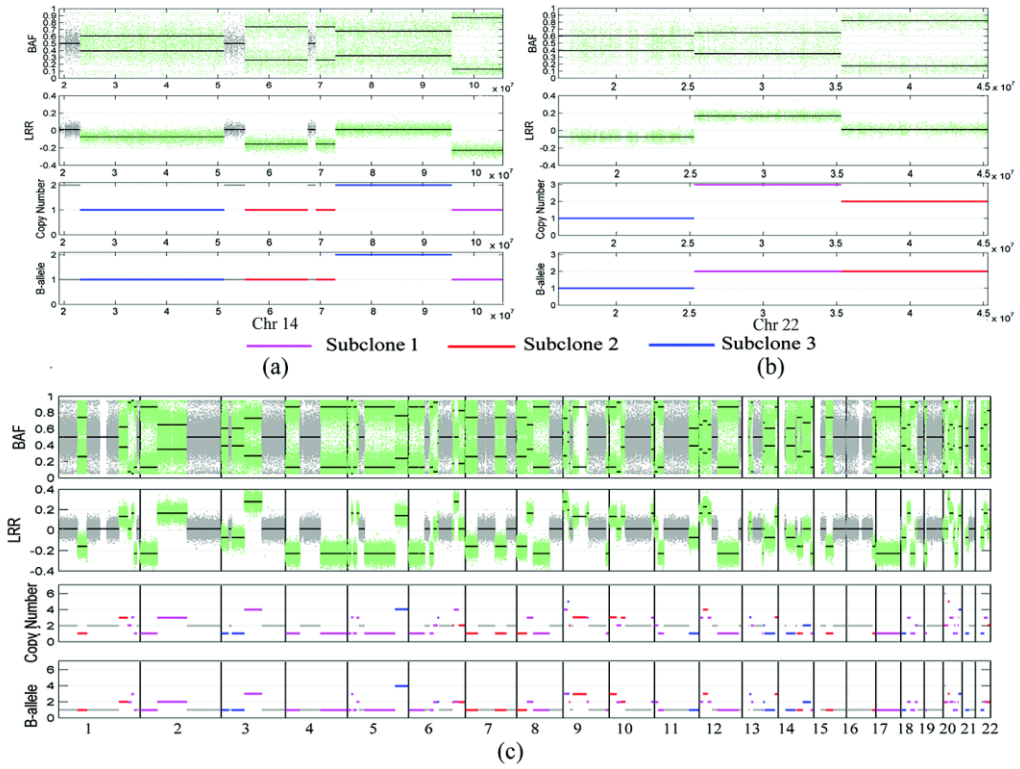

Fig. 1. Tumor subclone copy number aberrations predicted by SAPPH. For each subplot, top two panels are LRR and BAF signals, and bottom two panels are estimated copy number and B-allele number. Chromosome intervals marked by gray denote normal genomic state, and intervals with signals marked by green indicate somatic aberrations. Results show there are three tumor subclones, marked by purple, red and blue, respectively. (a) chromosome 14; (b) chromosome 22; and (c) whole genome.

SAPPH provides visualization of tumor subclone copy number aberrations for each genome segments. Figure 1 shows the signal distribution and the estimated results of chromosome 14 and 22 in a simulated sample with 20%, 30% and 35% of three kinds of tumor subclones, respectively. This sample was contaminated with 15% of the normal cells. The top two panels represent LRR and BAF signals along the chromosome and the black lines indicate average signal values for each interval. The bottom two panels are copy number and number of B-allele estimated by SAPPH. In Figure 1(a) where LRR/BAF signals for one copy deletion that is marked in red show stronger contractions to the diploid track (average LRR equals to 0, and average BAF equals to 0.5) than that marked in purple; whereas that marked in blue show even stronger contractions. Therefore, we can infer the existence of three subclones with different cellular proportions. Tumor subclone 1 can be treated as ancestor tumor cell (or dominant clone) due to the fact that both subclone 2 and 3 shares its somatic aberrations. Figure 1(b) shows that tumor subclone 2 experiences a copy-neutral LOH (loss of heterozygosity) on the part of Chr 22, while tumor subclone 3 has one copy deletion on another region of the same chromosome. The interval with local proportion of 85% marked by magenta indicates that both dominant clone and other tumor cells have one copy amplification on this region. The overall genomic profile of this sample is illustrated in Figure 1(c). SAPPH successfully identified all copy number aberrations for each tumor subclone including amplification and LOH.

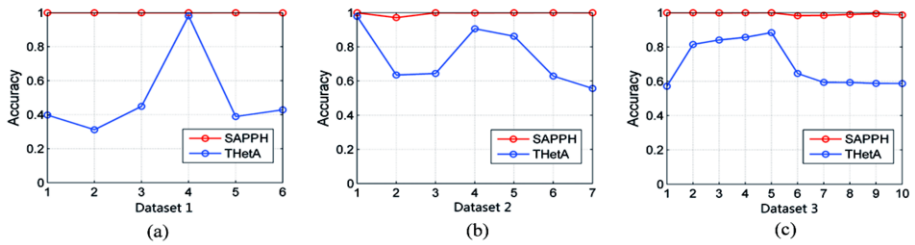

Fig. 2. Accuracy of SAPPH and THetA for three datasets. (a) Dataset 1, each sample contains only one kind of tumor subclone cell; (b) Dataset 2, each sample contains two kinds of tumor subclone cell; (c) Dataset 3, each sample contains three kinds of tumor subclone cell.

3.2. Performance evaluation on detecting somatic aberrations and subclone proportions

In order to quantitatively evaluate the performance of the proposed method, we calculated the accuracy (defined as the proportion of all correctly identified SNP copy numbers) of the simulated samples, and compared it with that of THetA. As shown in Figure 2, the accuracy of SAPPH for all the three datasets is greater than 0.95 and the performance of THetA is relatively low. This is especially true in Dataset 1 when there is only one kind of tumor subclone cells, SAPPH can correctly estimate copy numbers for almost all SNPs, whereas the average accuracy of THetA is lower than 0.5. In contrast to the consistent good performance of detection of tumor copy numbers by SAPPH, THetA shows unstable accuracy with the change of tumor subclone proportions. As shown in Table 1, we also calculated the precision, recall and F-score of the two methods in estimating genomic amplifications. These results demonstrate that SAPPH is more robust than THetA in detecting genomic amplification for heterogeneous tumor samples. To further illustrate the ability of SAPPH in detecting tumor subclone proportions, we calculated the bias of estimated proportion with respect to the ground truth for the three datasets. We did not compare with THetA in this part as it wrongly estimates tumor subclone numbers for part of samples. Figure 3 shows the bias of estimation with the change of mixed tumor subclone proportions. The overall biases are below 0.5% even for samples with tumor subclone cells as low as 10%. We conclude that the estimated tumor subclone proportions by SAPPH are highly consistent with actual values.

Table 1

Measurements of the two methods in estimating genomic amplifications for Dataset 3

Sample ID	SAPPH			THetA		
	Precision	Recall	F-score	Precision	Recall	F-score
t1_15t2_20t3_50	1.000	0.767	0.868	0.998	0.119	0.213
t1_15t2_25t3_45	1.000	0.767	0.868	0.659	0.268	0.381
t1_20t2_20t3_45	0.999	0.768	0.869	0.426	0.131	0.200
t1_20t2_30t3_35	0.999	0.768	0.869	0.607	0.244	0.348
t1_25t2_25t3_35	1.000	0.769	0.870	0.720	0.127	0.217
t1_35t2_25t3_25	0.950	0.674	0.788	0.651	0.275	0.387
t1_45t2_20t3_20	0.871	0.770	0.817	0.597	0.242	0.344
t1_45t2_30t3_10	0.999	0.705	0.826	1.000	0.242	0.390
t1_55t2_10t3_20	1.000	0.771	0.871	0.546	0.243	0.336
t1_55t2_20t3_10	0.999	0.675	0.806	0.816	0.243	0.374

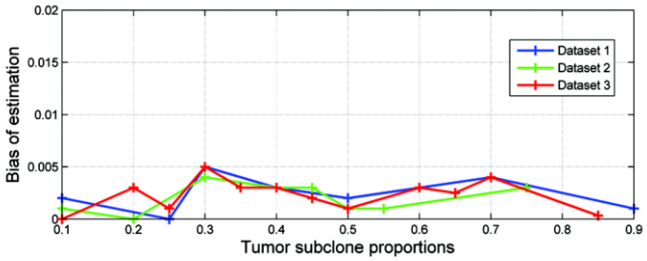

Fig. 3. Bias of estimation with the changing of tumor subclone proportions for three datasets.

4. Discussion and conclusion

In this study, we developed a novel computational method to predict somatic copy number aberrations and tumor subclone proportions from heterogeneous tumor samples. In order to decrease the computation burden we utilized SAPPH. This method avoids using general optimization approach to solve this deconvolution problem. SAPPH first estimates the mixing proportion for each interval separately, and then clusters local proportions to get the number of tumor subclones. While a separate study may bring bias estimation on some intervals, it can be refined in the final model selection step. SAPPH is highly efficient in simultaneous estimation of tumor subclone proportions as well as identification of tumor subclonal aberrations. Moreover, its computation time scales linearly with the number of somatic aberration events and is not affected by increasing the number of tumor subclones.

Acknowledgment

This work was supported by National Natural Science Foundation of China (Grant No. 31100955 and No. 61471331).

References

[1] S. Nik-Zainal, P. Van Loo, D.C. Wedge, L.B. Alexandrov, et al., The life history of 21 breast cancers, Cell **149** (2012), 994–1007.
[2] I.M. Lönnstedt, F. Caramia, J. Li, D. Fumagalli, et al., Deciphering clonality in aneuploid tumors using SNP array and sequencing data, Genome Biology **15** (2014), 470-470.
[3] P.J. Stephens, D.J. McBride, M.L. Lin, I. Varela, et al., Complex landscapes of somatic rearrangement in human breast cancer genomes, Nature **462** (2009), 1005-1010.
[4] M. Simoons, E. Topol, R. Califf and F. van de Werf, Oncogenes and cancer, New England Journal of Medicine **358** (2008), 502-511.
[5] M. Meyerson, S. Gabriel and G. Getz, Advances in understanding cancer genomes through second-generation sequencing, Nature Reviews of Genetics **11** (2010), 685-696.
[6] T. Popova, V. Boeva, E. Manié, et al., Analysis of somatic alterations in cancer genome: From SNP arrays to next generation sequencing, genomics I humans, Animals and Plants (2013), hal-01108425.
[7] R. Nielsen, JS. Paul, A. Albrechtsen and YS. Song, Genotype and SNP calling from next-generation sequencing data, Nature Reviews of Genetics **12** (2011), 443-451.
[8] Z. Yu, Y. Liu, Y. Shen, M. Wang and A. Li, CLImAT: Accurate detection of copy number alteration and loss of heterozygosity in impure and aneuploid tumor samples using whole-genome sequencing data, Bioinformatics **30** (2014), 30, 2576–2583.

[9] V. Boeva, A. Zinovyev, K. Bleakley, J.P. Vert, et al., Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization, Bioinformatics **27** (2011), 268-269.
[10] M. Mayrhofer, S. DiLorenzo and A Isaksson, Patchwork: Allele-specific copy number analysis of whole-genome sequenced tumor tissue, Genome Biology **14** (2013), R24.
[11] L. Oesper, A. Mahmoody and B.J. Raphael, THetA: Inferring intra-tumor heterogeneity from high-throughput DNA sequencing data, Genome Biology **14** (2013), R80.
[12] B. Li and J.Z. Li, A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data, Genome Biology **15** (2014), 473.
[13] H. Li, B. Handsaker, A. Wysoker, T. Fennell, et al., The sequence alignment/map format and SAMtools, Bioinformatics **25** (2009), 2078-2079.
[14] E.S. Venkatraman and A.B. Olshen, A faster circular binary segmentation algorithm for the analysis of array CGH data, Bioinformatics **23** (2007), 657-663.
[15] S.P. Shah, A. Roth, R. Goya, A. Oloumi, G. Ha, et al., The clonal and mutational evolution spectrum of primary triple-negative breast cancers, Nature **486** (2012), 395–399.