Hindawi
Scientific Programming
Volume 2021, Article ID 6536470, 6 pages
https://doi.org/10.1155/2021/6536470

Research Article

rmvPFBAM: Removing Primers from BAM Files Based on Amplicon-Based Next-Generation Sequencing and Cloud Computing When Analyzing Personal Genome Data

Yanjun Ma

Public Security Information Department, Liaoning Police College, Dalian, Liaoning, China

Correspondence should be addressed to Yanjun Ma; mayanjun0010@yeah.net

Received 23 September 2021; Accepted 1 November 2021; Published 16 November 2021

Academic Editor: Punit Gupta

Personal genomic data constitute one important part of personal health data. However, due to the large amount of personal genomic data obtained by the next-generation sequencing technology, special tools are needed to analyze these data. In this article, we will explore a tool analyzing cloud-based large-scale genome sequencing data. Analyzing and identifying genomic variations from amplicon-based next-generation sequencing data are necessary for the clinical diagnosis and treatment of cancer patients. When processing the amplicon-based next-generation sequencing data, one essential step is removing primer sequences from the reads to avoid detecting false-positive mutations introduced by nonspecific primer binding and primer extension reactions. At present, the removing primer tools usually discard primer sequences from the FASTQ file instead of BAM file, but this method could cause some downstream analysis problems. Only one tool (BAMClipper) removes primer sequences from BAM files, but it only modified the CIGAR value of the BAM file, and false-positive mutations falling in the primer region could still be detected based on its processed BAM file. So, we developed one cutting primer tool (rmvPFBAM) removing primer sequences from the BAM file, and the mutations detected based on the processed BAM file by rmvPFBAM are highly credible. Besides that, rmvPFBAM runs faster than other tools, such as cutPrimers and BAMClipper.

1. Introduction

Genomic variations are associated with the pathogenesis and treatment of many diseases, especially cancer. Identifying genomic variations of genetic biomarkers is important for the clinical diagnosis and treatment of cancer patients. Nowadays, there are several technologies to detect genomic variations, such as polymerase chain reaction (PCR), Sanger Sequencing, and next-generation sequencing [1]. Next-generation sequencing is the most effective way for detecting genomic variations because it can obtain hundreds of millions of bases of DNA molecules at one time.

Targeted sequencing is one commonly useful solution of next-generation sequencing focused on specific genomic regions [1]. Because targeted sequencing is cost-effective and could produce high-depth sequencing data which are able to detect low-frequency genomic variations, targeted sequencing is the most widely used approach in clinical cancer diagnosis [2]. There are two methods commonly used for targeted sequencing: capture hybridization-based sequencing and amplicon-based sequencing [3]. Amplicon-based sequencing uses multiplex PCR technology to generate thousands of amplicons for massively parallel sequencing and is one of the widely used targeted sequencing technology because of its easier operation and higher amplification efficiency [4–9]. However, during the primer amplification, false-positive mutations could be introduced in the primer region because of nonspecific primer binding and primer extension reactions [10]. So, it is necessary to remove the primers before executing the downstream analysis, such as detecting mutations.

The existing removing primer tools usually remove primers from FASTQ files, such as cutPrimers [11] and pTrimmer [12], and the only one tool removing primers from BAM (Binary sequence Alignment/Map) files is

BAMClipper [13]. There are some drawbacks of cutting primers from FASTQ files. For example, because the length of read is shorter after cutting the primer, the probability of the reads misalignment may be improved, and the shorter reads may lead to inaccurate detecting of copy number variation, even unable to use copy number variation tools such as ONCOCNV [14].

BAMClipper clips primers from BAM file by only modifying the CIGAR (Concise Idiosyncratic Gapped Alignment Report) value in BAM files instead of removing primers from reads. If downstream analysis such as detecting mutations was processed based on the cutting primer BAM file processed by BAMClipper, false-positive mutations existed in primer region can still be detected when using VarScan [15], that is, a classical tool for detecting mutations. So, we developed a tool removing primers based on the BAM file (rmvPFBAM) by creating a new BAM, and the tool is available at github (https://github.com/YanjunMasir/rmvPrimer). rmvPFBAM runs faster than the other tools, and mutations detected from rmvPFBAM's cutting primer BAM file are more accurate than those from BAMClipper's BAM.

2. Methodology

2.1. Datasets. For clinical research, the patient's surgical tissue can be collected to construct sequencing library that will be running on the sequencing machine, and the sequencing data will be produced after finishing running. The sequencing data are one input of rmvPFBAM, and the other input of rmvPFBAM is the primers used for capturing the target genomic region.

All reads and primers in the article are from the dataset of Ultradeep Targeted Sequencing of a set of Cancer Genes Project (SRP019940), presenting the molecular profile of 38 breast cancer species [16, 17]. This project used a panel of 47 genes involving 1,736 amplicons. The average length of the amplicon was 20 bp, and all reads were obtained by sequencing on matched normal and tumor tissues using Illumina MiSeq sequencer under a 150 bp long-paired reads protocol. For rmvPFBAM demonstration purpose, six patients from the SRP019940 were randomly selected to form the dataset (SRR866441, SRR866442, SRR866443, SRR866444, SRR866445, and SRR948507). The raw reads were downloaded from the European Nucleotide Archive datasets. Then, the reads were aligned using the BWA software [18]. rmvPFBAM and BAMClipper were executed based on the aligned BAM files, and cutPrimers was executed based on the FASTQ files.

2.2. Implementation. In target sequencing, each target region is covered by millions of reads (Figure 1). For pair-end reads, read 1 and read 2 contain the forward primer and reverse primer, respectively, but if the fragment is shorter than 150 bp, then read 1 contains part or whole of the reverse primer and read 2 contains part or whole of the forward primer (Figure 1). Due to the high error rate of base pairing during primer extension, the forward and reverse primer needed to be removed before processing the reads.

The rmvPFBAM workflow is implemented with python language and is available to Linux platform. The program rmvPFBAM uses BAM file and primer file as input. The primer file must contain the amplicon information, such as chromatin, amplicon's start position, amplicon's end position, front primer sequence, and reverse primer sequence. rmvPFBAM uses pysam package to process the BAM file and regex package to search primer sequences with regular expressions and multiprocessing for multithreading. Pysam is the most widely used python module that can manipulate mapped short read sequence data stored in SAM/BAM files. Because the primer files usually contain thousands of primers, rmvPFBAM splits the primers into several parts with each part containing hundreds of primers. Then, these several parts of primers are processed at the same time.

We compared our tool with already existing tools cutPrimers and BAMClipper. cutPrimers removed primers from FASTQ file instead of BAM file. BAMClipper removed primer sequences from BAM files, and it only modified the CIGAR value in the BAM file instead of modifying the BAM files, but rmvPFBAM not only modified the CIGAR value but also the BAM file. Examples of commands used for execution are available in Supplementary Material.

The flowchart (Figure 2) shows the details of processing one amplicon in the program. For one amplicon, all reads mapping to this amplicon were extracted. Then, for one primary mapping read, find the read pair of this read and remove the primer sequence from the reads only if the reads contain the primer sequence. After removing primer sequence, save the processed reads to one file and continue to process the next read mapping to this amplicon. Reads not containing the primer sequence are not allowed to save to the file.

3. Results

We performed comparative evaluation of the three programs using six samples from the SRP019940. Besides that, we also downloaded three datasets from the amplicon-based sequencing data [5] to compare rmvPFBAM, cutPrimers, and BAMClipper. The results were displayed in the Supplementary Material.

As cutPrimers remove primers from FASTQ files, only the runtime of this tool is compared with rmvPFBAM. Comparison of the functionality of rmvPFBAM, cutPrimers, and BAMClipper included the following parameters: (1) time of running, (2) no. of paired reads after cutting primer, (3) no. of target region reads after cutting primer, (4) no. of nontarget region reads after cutting primer, (5) no. of mutations detected based on the cutting primer BAM, and (6) no. of mutations based on the cutting primer BAM (falling in the target region). Results of the comparative analysis are presented in Table 1. We executed all processes on a personal computer (Intel(R)Xeon(R) CPU E5-2620 v2 2.10 GHz, 32 G RAM).

'No. of paired reads after cutting primers (%)' indicates the count of reads in the BAM file before cutting primers and

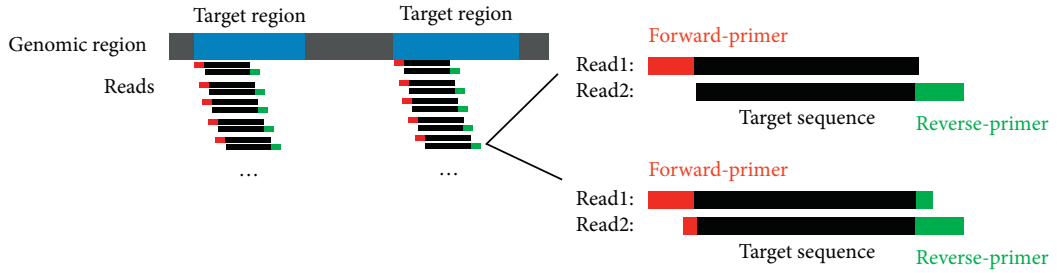


Figure 1: Scheme of the target sequencing by multiplex PCR. The blue region is the target region that we want to get the sequence of that and millions of reads are produced to cover it. For each read, the red sequence is the forward primer and the green sequence is the reverse primer. For each pair of read, read 1 contains forward primer (red) and read 2 contains reverse primer (green).

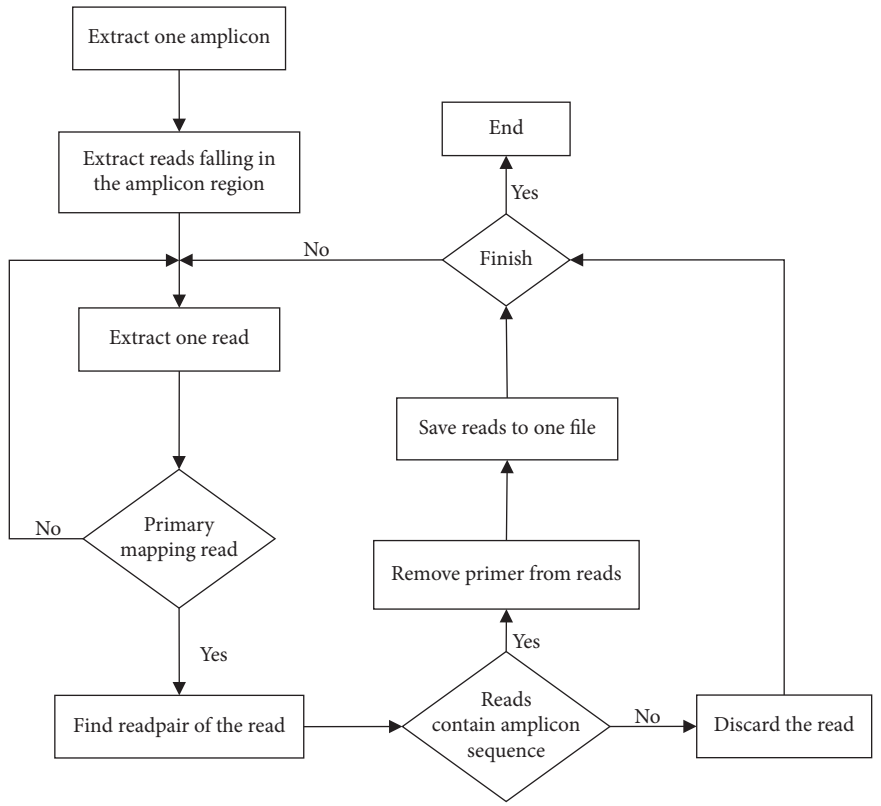


Figure 2: Workflow of rmvPFBAM.

the percent of those reads in the total reads. 'No. of target region reads after cutting primers (%)' indicates the reads falling in the target region and the percent of those reads in the total reads. 'No. of nontarget region reads after cutting primers (%)' indicates the reads falling in the nontarget region and the percent of those reads in the total reads. 'No. of mutations detected based on the cutting primer BAM' indicates the count of mutations detected by VarScan2 based on the cutting primer BAM file. 'No. of mutations based on the cutting primer BAM (fall in target region)' indicates the count of mutations falling in the target region based on the cutting primer BAM.

rmvPFBAM showed a much higher speed of processing reads than cutPrimers and BAMClipper in the six datasets (Figure 3(a)). The average running time of rmvPFBAM was 162 s, the cutPrimers was 526 s, and the BAMClipper was 1137 s. rmvPFBAM's running time was almost four or seven times that of the other two tools. Besides that, we compared the number of reads falling in the target region. Because BAMCliper only modifies the CIGAR value in the BAM file, the modified BAM by BAMClipper contains the same number of reads as the before modified BAM. However, the number of reads mapping to the target region is less different between rmvPFBAM and BAMClipper (Figure 3(b)).

Only reads mapping to the target region are useful for the downstream analysis, such as detecting mutations. So, we compared the number of mutations detected based on rmvPFBAM's and BAMClipper's BAM file. One classical tool for detecting mutations-Varscan2 was used to detect mutations from BAM files processed by BAMClipper and rmvPFBAM. About a third of mutations detected from the BAMClipper's BAM files were in the nontarget region

Table 1: Comparative data among the three tools by using six samples from amplicon-based next-generation sequencing data.

Sample (no. of paired reads)	Parameter	CutPrimers (err = 3, threads = 4)	BAMClipper (with –g)	rmvPFBAM (err = 3)
SRR866441 (2238436)	Time of running(s)	451	869	121
SRR866442 (2738246)		507	1135	159
SRR866443 (2628572)		485	1052	150
SRR866444 (3586450)		672	1467	229
SRR866445 (2760844)		513	1148	149
SRR948507 (2749834)		525	1149	164
SRR866441 (2238436)	No. of paired reads after cutting primers (%)	–	2238436(100)	1840262(82.2)
SRR866442 (2738246)			2738246(100)	2518726(92.0)
SRR866443 (2628572)			2628572(100)	2320172(88.3)
SRR866444 (3586450)			3586450(100)	3342890(93.2)
SRR866445 (2760844)			2760844(100)	2380948(86.2)
SRR948507 (2749834)			2749834(100)	2430968(88.4)
SRR866441 (2238436)	No. of target region reads after cutting primers (%)	–	1882305(84.1)	1840262(82.2)
SRR866442 (2738246)			2609012(95.3)	2518726(92.0)
SRR866443 (2628572)			2365634(90.0)	2320172(88.3)
SRR866444 (3586450)			3390685(94.5)	3342890(93.2)
SRR866445 (2760844)			2424973(87.8)	2380948(86.2)
SRR948507 (2749834)			2473225(89.9)	2430968(88.4)
SRR866441 (2238436)	No. of nontarget region reads after cutting primers (%)	–	269848(12.1)	0
SRR866442 (2738246)			103128(3.8)	0
SRR866443 (2628572)			186292(7.1)	0
SRR866444 (3586450)			161479(4.5)	0
SRR866445 (2760844)			254749(9.2)	0
SRR948507 (2749834)			221905(8.1)	0
SRR866441 (2238436)	No. of mutations detected based on the cutting primer BAM	–	981	676
SRR866442 (2738246)			527	329
SRR866443 (2628572)			584	398
SRR866444 (3586450)			544	349
SRR866445 (2760844)			847	515
SRR948507 (2749834)			793	488
SRR866441 (2238436)	No. of mutations based on the cutting primer BAM (fall in target region)	–	727	676
SRR866442 (2738246)			351	329
SRR866443 (2628572)			415	398
SRR866444 (3586450)			361	349
SRR866445 (2760844)			552	515
SRR948507 (2749834)			527	488

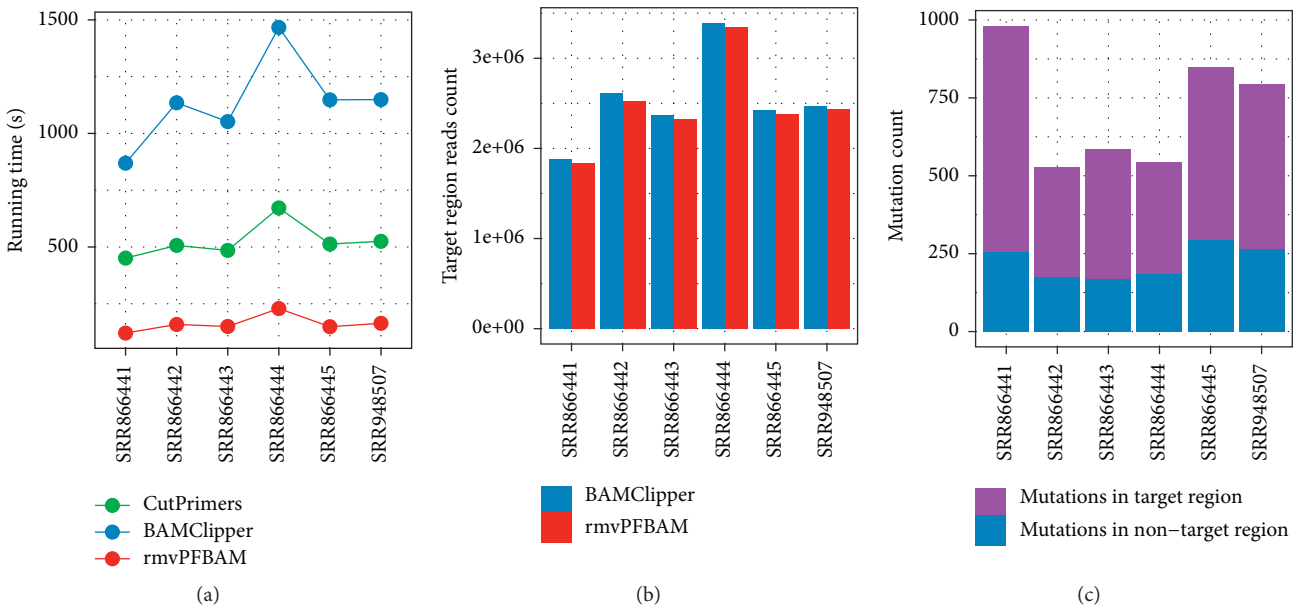
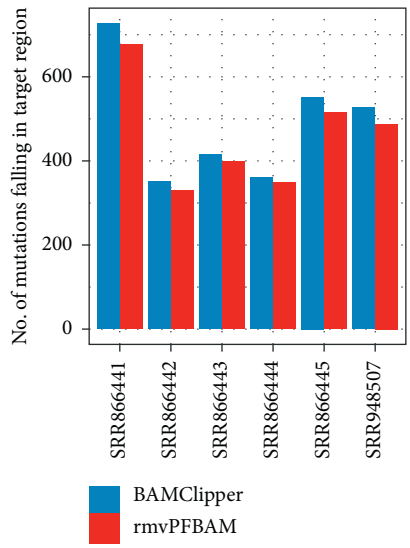


Figure 3: Continued.



(d)

Figure 3: Comparative results of the three tools: (a) the running time of the three tools; (b) the target region reads count of BMAClipper and rmvPFBAM; (c) the counts of SNVs falling in the target region and nontarget region detected by BAMClipper; (d) the number of mutations falling in the target region detected by BAMClipper and rmvPFBAM.

(Figure 3(c)), and they were mostly like to be false-positive [8]. When only comparing mutations falling in the target region, the count of mutations from the BAM file processed by rmvPFBAM was almost the same as that of from BAMClipper (Figure 3(d)). So, the mutations detected from the rmvPFBAM's cutting primer BAM file mostly fall in the target region and show little false-positive mutations compared with BAMClipper.

4. Discussion

rmvPFBAM is one tool cutting primer sequences from the BAM file of amplicon-based next-generating sequencing. The processing speed of this tool is much faster than that of the other tools. Also, owing to its creation of a new BAM file instead of only modifying the CIGAR value of the BAM file, the downstream analysis based on rmvPFBAM's created BAM file is more accurate than from BAMClipper's BAM files.

Amplicon-based next-generation sequencing is widely used in the diagnosis of clinical cancer patients. Accurate and reliable detection of mutations could improve diagnostics and find new potential targets. Although there are many tools to remove primers, most tools remove primers from FASTQ files instead of BAM files. Removing primers from FASTQ file is simple to implement, but it is slightly limited for the downstream analysis. Nowadays, BAMClipper is the only published tool removing primers from BAM files. It processes the primers for a long time and only modified the CIGAR value of the BAM file so we developed rmvPFBAM. The processing speed of rmvPFBAM is much faster than that of BAMClipper, and it modified all the items in the BAM file including CIGAR and sequence. The result of downstream analysis such as detecting mutations is more accurate than BAMClipper. However, because rmvPFBAM applies a more strict strategy to remove primers from pair-end reads, about 10 percent of reads were discarded from the original BAM file. It is valuable to get more accurate results through the loss of 10 percent of the data.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Conflicts of Interest

The author declares that there are no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

The author acknowledges the Innovation Team Support Program of Liaoning Police College and the Soft Science Research Project of the Ministry of Public Security (Grant 2020LLYJLNST050).

Supplementary Materials

The supplementary file contains the "Commands that were used for tool execution" and the "Comparative data among the three tools by using six samples from amplicon-based next-generation sequencing data." (*Supplementary Materials*)

References

[1] S. Goodwin, J. D. McPherson, and W. R. McCombie, "Coming of age: ten years of next-generation sequencing technologies," *Nature Reviews Genetics*, vol. 17, no. 6, pp. 333–351, 2016.

[2] F. Bewicke-Copley, E. Arjun Kumar, G. Palladino, K. Korfi, and J. Wang, "Applications and analysis of targeted genomic sequencing in cancer studies," *Computational and Structural Biotechnology Journal*, vol. 17, pp. 1348–1359, 2019.

[3] S. S. Hung, B. Meissner, E. A. Chavez et al., "Assessment of capture and amplicon-based approaches for the development of a targeted next-generation sequencing pipeline to personalize lymphoma management," *Journal of Molecular Diagnostics*, vol. 20, no. 2, pp. 203–214, 2018.

[4] N. Guibert, Y. Hu, N. Feeney et al., "Amplicon-based next-generation sequencing of plasma cell-free DNA for detection of driver and resistance mutations in advanced non-small cell lung cancer," *Annals of Oncology*, vol. 29, no. 4, pp. 1049–1055, 2018.

[5] Y. Liao, Z. Ma, Y. Zhang et al., "Targeted deep sequencing from multiple sources demonstrates increased NOTCH1 alterations in lung cancer patient plasma," *Cancer Medicine*, vol. 8, no. 12, pp. 5673–5686, 2019.

[6] Y. Onda, K. Takahagi, M. Shimizu, K. Inoue, and K. Mochida, "Multiplex PCR targeted amplicon sequencing (MTA-Seq): simple, flexible, and versatile SNP genotyping by highly multiplexed PCR amplicon sequencing," *Frontiers of Plant Science*, vol. 9, p. 201, 2018.

[7] B. Snetsinger, C. K. Ferrone, and M. J. Rauh, "Targeted, Amplicon-Based, Next-Generation Sequencing to Detect Age-Related Clonal Hematopoiesis," *Methods Molecular Biology*, vol. 2045, pp. 167–180, 2019.

[8] E. Strengman, F. A. S. Barendrecht-Smouter, C. d. Voijs, P. d. Vree, and I. J. Nijman, "Amplicon-based targeted next-generation sequencing of formalin-fixed, paraffin-embedded tissue," *Methods in Molecular Biology*, vol. 1908, pp. 1–17, 2019.

[9] J. Tian, Y. Geng, D. Lv et al., "Using plasma cell-free DNA to monitor the chemoradiotherapy course of cervical cancer," *International Journal of Cancer*, vol. 145, no. 9, pp. 2547–2557, 2019.

[10] C. M. McCall, S. Mosier, M. Thiess et al., "False positives in multiplex PCR-based next-generation sequencing have unique signatures," *Journal of Molecular Diagnostics*, vol. 16, no. 5, pp. 541–549, 2014.

[11] A. Kechin, U. Boyarskikh, A. Kel, and M. Filipenko, "cutPrimers: a new tool for accurate cutting of primers from reads of targeted next generation sequencing," *Journal of Computational Biology*, vol. 24, no. 11, pp. 1138–1143, 2017.

[12] X. Zhang, Y. Shao, J. Tian et al., "pTrimmer: an efficient tool to trim primers of multiplex deep sequencing data," *BMC Bioinformatics*, vol. 20, no. 1, p. 236, 2019.

[13] C. H. Au, D. N. Ho, A. Kwong, T. L. Chan, and E. S. K. Ma, "BAMClipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing," *Scientific Reports*, vol. 7, no. 1, p. 1567, 2017.

[14] N. Rieber, R. Bohnert, U. Ziehm, and G. Jansen, "Reliability of algorithmic somatic copy number alteration detection from targeted capture data," *Bioinformatics*, vol. 33, no. 18, pp. 2791–2798, 2017.

[15] D. C. Koboldt, Q. Zhang, D. E. Larson et al., "VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing," *Genome Research*, vol. 22, no. 3, pp. 568–576, 2012.

[16] O. Harismendy, R. B. Schwab, H. Alakus et al., "Evaluation of ultra-deep targeted sequencing for personalized breast cancer care," *Breast Cancer Research*, vol. 15, no. 6, p. R115, 2013.

[17] S. E. Yost, H. Alakus, H. Matsui et al., "Mutascope: sensitive detection of somatic mutations from deep amplicon sequencing," *Bioinformatics*, vol. 29, no. 15, pp. 1908-1909, 2013.

[18] H. Li and R. Durbin, "Fast and accurate long-read alignment with Burrows-Wheeler transform," *Bioinformatics*, vol. 26, no. 5, pp. 589–595, 2010.