Turkish Journal of Engineering

Turkish Journal of Engineering (TUJE)
Vol. 3, Issue 3, pp. 120-126, July 2019
ISSN 2587-1366, Turkey
DOI: 10.31127/tuje.472328
Research Article

DETERMINATION OF MALIGNANT MELANOMA BY ANALYSIS OF VARIATION VALUES

Ahmet Kürşat Esim [1], Hilal Kaya [2] and Veysel Alcan *[3]

[1] Ankara Yıldırım Beyazıt University, Faculty of Engineering and Natural Science, Department of Computer Engineering, Ankara, Turkey
ORCID ID 0000 – 0002 – 3458– 2411
kursatesim@gmail.com

[2] Ankara Yıldırım Beyazıt University, Faculty of Engineering and Natural Science, Department of Computer Engineering, Ankara, Turkey
ORCID ID 0000 – 0003 – 4787 – 105X
hilalkaya@ybu.edu.tr

[3] Tarsus University, Faculty of Technology, Department of Software Engineering, Mersin, Turkey
ORCID ID 0000 – 0002 – 7786– 8591
alcan@tarsus.edu.tr

*Corresponding Author

Received: 13/11/2018 Accepted: 18/12/2018

ABSTRACT
Melanoma is a serious disease associated with mutation-based cancer cells. Genetic structure and hereditary condition play important role to understand the underlying reasons of the diseases caused by Deoxiribole Nucleic Acid (DNA). In order to identify mutation carriers and to analyze disease, researchers tend to find various gene determinations methods. Nowadays, Next Generation Sequencing (NGS) is emerging as a valuable and powerful platform to detect gene-based diseases by entiring human genome. In this study, we aimed to propose a bioinformatics application workflow to distinguish between insertions/deletions and somatic/germline mutations, by using NGS methods. We carried this study out on a data set containing 100 human genomes data (20 training, 80 testing) for the detection of Malignant Melanoma. We found that the results of diagnosis performance were 92.50% accuracy, 94.03% precision, 96.92% sensitivity and 95.45% F1 score. These results show the potential for proposed application based on NGS to improve Melanoma detection.

Keywords: *Next Generation Sequencing, Burrows-Wheeler Transform, Variant Calling, Malignant Melanoma*

1. INTRODUCTION

Melanoma is generally the most serious form of skin cancer. It tends to trigger genetic defects causing skin cells to proliferate and form malignant tumors when unrepaired Deoxyriba Nucleic Acid (DNA) damage occurs. These tumors are caused by pigment-producing melanocytes in the basal layer of the epidermis. When cancer begins to change and starts to grow, healthy cells form a mass called tumor and are out of control (Fig1). In general, about 8% of those diagnosed with new Melanoma have first-degree relatives with Melanoma. In a much smaller percentage of 1% to 2%, there are two or more close relatives having Melanoma.

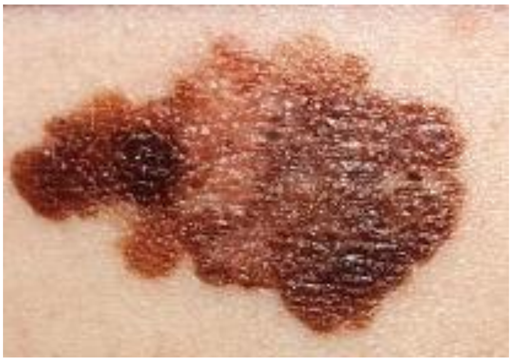

Fig. 1. A general pattern of Malignant Melanoma

Familial Melanoma is a genetic or hereditary condition which means that the risk of Melanoma can pass from one generation to another in a family. To date, familial Melanoma has been primarily associated with two genes called cyclin-dependent kinase inhibitor 2A (CDKN2A) and cyclin-dependent kinase 4 (CDK4). A mutation (change) in one of these genes gives the person an increased risk of Melanoma and plays an important role in transferring the disease to future generations when it affects the genetic sequence. Variations such as dysplastic nevus (mole), sun exposure, freckled skin and red hair occur as risk factors in genes are associated with Melanoma tendency with known genetic mutations (Audibert *et al*., 2018).

In order to identify mutation carriers in the melanocortin 1 receptor (MC1R) and CDKN2A genes, the fastest, most comprehensive and delicate method is so important for detecting this type of disease. Thus, researchers tend to find various gene determinations and to conduct any studies on mutation-based cancer cells associated with the genetic structure (Rihtman *et al*., 2016). Nowadays, Next Generation Sequencing (NGS) is emerging as a valuable and powerful platform to detect gene-based diseases by entiring human genome. The simultaneous processing of each part of a DNA molecule separated into millions of pieces taken from a single sample is called NGS. The detection of the gene architecture behind the diseases can be analyzed by using gene mapping methods (Vogelstein *et al*., 2013).

In this study, it was aimed to develop a bioinformatics application workflow to introduce both single nucleotide variants (SNV) and small diagnoses in order to obtain the molecular diagnosis of Melanoma by using NGS methods to distinguish between insertions/deletions and somatic/germline mutations.

2. METHODS

2.1. Experimental Data

In this study, an application workflow was created to prepare the data to be used to detect the tumor gene sequence. An Sh file (a scripted file for the Unix Bourne-Again type-SHell (Fileinfo (sh file), 2018)) is prepared using the Biotechnology National Information Center (NCBI) ftp servers. By the help of this Sh file, it is possible to download Melanoma and healthy reference samples from anywhere and anytime.

Sequence Read Archives (SRA) is the file type required to convert files in the toolkit. fastqz is used as a compression tool for FASTQ files. SRA and compressed files refer to softlinked, a symbolic path that represents the abstract position of another file (Cyberciti, 2018). A FASTQ record contains a sequence of quality scores for each nucleotide. Tab seperated value (TSV) files that store a data table in which data columns are separated by tabs are created in this way (Fileinfo (Tsv File), 2018). By this file, it is checked whether the sample examined has malignant tumor or not. After preparing the sample, the healthy reference sample is downloaded and converted to the required file types such as FASTA, fastfai, and dict. All required files are downloaded from Illumina's FTP servers, and then files are prepared to assist in targeting and locating the appropriate unit.

SRA tools dict, amp and files are created and the indexing process is also made from the FASTA file. Then, by Burrows-Wheeler Alignment (BWA) tools, the rules were created, the required files were taken. The sa file was created with Fastfai and Burrows-Wheeler Transformation (BWT). The pac file was created using the BWT and sa file. Indexing the genome to avoid any problems on NGS data is an important task. In order to speed up genomic searching, indexing operations on genome size, length and number can be done.

2.2. Next Generation Sequencing (NGS)

The sequencing logic behind NGS technology is similar to the Sanger Sequence. DNA polymerase strand accelerates the incorporation of genome-labeled nucleotides into the DNA template during DNA replication. During each cycle, the nucleotides are identified by fluorophore tags. Instead of sequencing a single piece of DNA, NGS executes this process in parallel between millions of DNA fragments and offers many advantages over Sanger Sequencing. Because NGS carries out multiple gene evaluation procedures in a single test, it eliminates the need for multiple tests to identify the causal mutation. This approach, while saving time, provides a more economical solution, reducing the risk of overlooking valuable clinical specimens. In addition, it provides about 5% high sensitivity in the detection of DNA mutations when compared with the existing methods (Kearse *et al*., 2012).

NGS also allows zooming to the next target regions with the help of Ribo Nucleic Acid (RNA) sorting to discover the original RNA variants, allowing all genes to be sequenced quickly. Thus, gene expression analysis allows analysis of epigenetic factors, such as DNA methylation and DNA-protein interactions, for the study of tumor types with digital messenger RNAs (mRNAs), gene-related cancer samples and somatic derivatives (Griffiths-Jones *et al*., 2006).

All NGS platforms require a library obtained by amplification or ligation with special adapter arrays. These adapter arrays provide a universal array of sequences for hybridization and sequencing pieces to the sorting cores, allowing the use of the library. By this sample preparation process, necessary data amounts for NGS were created and used. Each library fragment is raised to a solid surface by a covalently bonded DNA linker that hybridizes library adapters. This amplification process generates DNA clusters, each of which originates from a single library segment, and each cluster moves with an individual sequencing reaction. The order of each cluster is read optically from the repeated nucleotide loops. The raw data is obtained at the end of the sequencing run. This raw data is a collection of DNA sequences generated in each cluster. It may be necessary to perform further analysis to obtain meaningful results from these data (Zhang *et al*., 2011).

The use of molecular profiling exceeds the limitations of conventional solid tumor and surrounding tissue classification methods based on the morphology of tumor cells. Molecular profiling is critical to identify and characterize unique somatic mutations occured in cancer cells. Tumor profiling using NGS focuses on a preselected subset of genes / gene panel. These panels contain genes known to be involved in cancer and allow simultaneous evaluation of all potentially causing genes. Tumor profiling using NGS follows a simple workflow that can be easily scaled to hundreds of samples, allowing clinical laboratories to process and respond to samples. The targeted gene sequence analyzes more than one gene in a single assay. By optimizing the use of limited tissue samples by reducing the need for sequential testing, it provides accurate identification of rare variants in heterogeneous tumor samples (Kearse *et al*., 2012).

This study was started by downloading the mini version of the open source package management system and the environmental management system of Conda (Gao *et al*., 2017). Snakemake workflow management system was used to create transaction files. Snakemake workflows that are essentially the rules defined by extended scripts created with Python. These rules describe how to create output files using input files (Leipzig, 2016).

2.3. Proposed Application Pipeline

In this study, a multilayered architecture was used for the proposed application workflow (Fig. 2). This architecture contains multiple nested layers. Each rule has its own task identity, and many process steps work according to these task identities and threads. Faster operations were achieved by using Cloud systems. Working performance was three times higher than PCs.

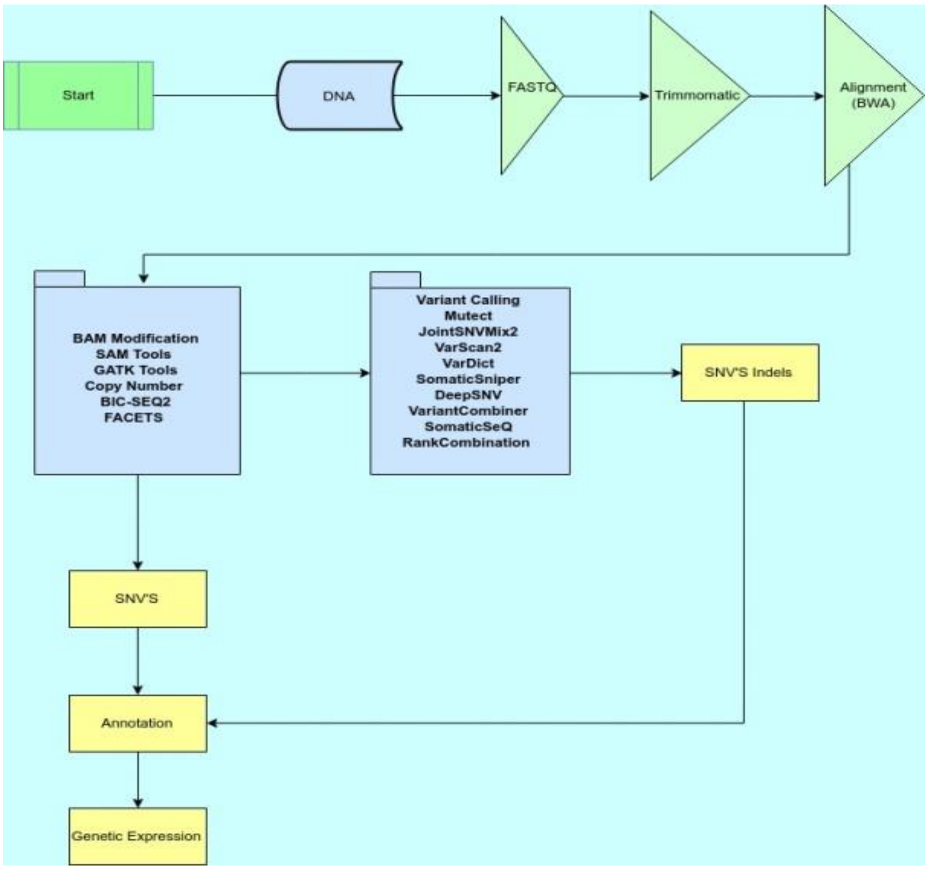

Fig.2. Proposed pipeline

BWA is a software package used to map different sequences corresponding to a large reference genome, such as the human genome, with the features such as long reading support and split alignment etc. (Michele *et al*., 2004). In this study, a BWA file was created and the indexing process was repeated.

To improve the overall quality of the skin data required for this study, cleaning and correcting the data were important pre-processing steps. Trimmomatic produces one or two DNA reading files by reading DNA from the left end and one from the right end, because a pair of reading files have been created using paired end mode (Fileinfo (Tsv File), 2018).Trimmomatic reference file called Truseq2pe was used to transmit fa.Fastq1 and Fastq2 files to reverse the fastq file. Clip trim rule was used to crop these two files and the reading file was created. For this purpose, there were two cut and soft connected fastq files. Additionally, there were two trimmed fastq files, which were fixed-linked files. After this procedure, the genomic analysis toolkit (GATK) was used for BAM modification. GATK is a programming framework designed and structured to facilitate the development of efficient and robust analysis tools for next-generation DNA sequencers using the MapReduce functional programming philosophy. GATK provides a small but rich set of data access patterns covering most of the needs of the analysis tool (McKenna *et al*., 2010).

Picard, a set of command-line tools, such as SAM / BAM / CRAM and VCF, were used to sort and process data and formats and process them with high efficiency (Ogasawara *et al*., 2016). SAM tools provide various utilities to sort, merge, index, create alignments in a preposition format, and handle alignments in Sam format (Li *et al*., 2009).

Using tracers at the end of the reading, which are below a certain threshold, create strips of cut quality adopted under the cropping threshold.

BWA-MEM is also used to deal with long reading and split alignment and is generally recommended for high-quality queries because it is faster and more accurate (Sipos *et al*., 2012).

Polymerase chain reaction (PCR) duplicates that are sets of pairs that have an unaligned start and an unaligned end and are suspected of non-independent measurements of a sequence. They are sampled from the same DNA template and violate the assumptions of the variant call. Acting as a real variant may lead to false positive variant calls, and errors in the sample are propagated to all pairs (Ebbert *et al*., 2016). Creation of SNPs and indentations are likely to be missed. Using many alignments, it may be needed to re-align the original readings to see that it is absolutely impossible to find it with a single read. This realignment-secondary alignment is part of the work of GATK haplotypecaller (Moore *et al*., 2010).

Trimmomatic-paired end-mode reading pairs can be protected, and library preparation to use additional information in better-matched readings can also be performed (Bolger *et al*., 2014).

A .txt file has been prepared to create quality data and quality mapping on the next process steps. In this file; genomic results containing BAM statistics include various statistics on genes with the total base amount found in external results and external window file types, how many of them are matched, how many of them have been processed separately, how many pairs are found in the DNA, and the quality of this data. Thus, a BAM file that has been fixed, sorted, merged and cleared from pairs has been sent to the Variant Calling phase.

2.3.1. Variant Calling

A variant call is a result of a nucleotide difference in a given position in a genome or transcriptome, usually accompanied by an estimate of the variant frequency and a measure of confidence. In this study, DNA-Seq variant calling methodology developed by Cornell Biotechnology Institute was used (Bukowski *et al*., 2017).

The presence of SNVs from NGS results that account for about 90% of all genetic variations between humans and the variant call is defined and a base is replaced by another base. SNV is only a variation in a single nucleotide without any frequency limitation. An SNP (single nucleotide polymorphism) is a single nucleotide change occured in more than 1% of a population of a single base pair difference from a reference. SNV is a special mutation while SNP is a shared mutation between a community. The variant calling is used to perform SNP genotyping, which identifies rare SNPs in a population, as well as to determine somatic SNV for each individual.

Variant calling algorithms are based on quality scores assigned to individual key calls in each widely read sequence. These scores are estimations per base of the error propagated by the sorting machine but are produced scores subject to systematic technical errors from various sources. Basic quality score recalibration (BQSR) is a process in which machine learning is applied to empirically model these errors and adjust the quality scores. This increases the accuracy of variant calls and helps to obtain more accurate basic attributes (Hsu *et al*., 2017).

Single nucleotide polymorphisms due to point mutations and single nucleotide variations; structural variations may also occur due to deletion, replication, insertion, inversion and translocation. Copy number variants (CNVs) are submicroscopic structural variations that occur due to deletion, replication and replicator transposition. Inversion is a segment of DNA in which the orientation is reversed relative to the rest of the chromosome. Translocation is a local change in the position of the chromosome segment in a genome without any change in total DNA content. In addition, Segmental uniparental dysomy as uniparental disomy describes a case where a pair of homologous chromosomes or portions of a chromosome is derived from a single parent in a diploid individual (Bulmer, 1971).

There are two SNV classes in the literature which are constitutional and germline mutations. These mutations are inherited from the parents and are found in every cell and somatic mutations that occur during the life of an individual. When looking for a rare disease, germline variations of SNV are used. For diseases not included in the literature, the contribution of somatic mutations according to disease status is examined by comparing the tumor with the normal samples.

Realign InDels (insertion/deletion) method describes the possible indices in all readings and the results given against these targets. This step is required to avoid false readings because some of the indices are misaligned during the initial alignment (Sun *et al*., 2017).

BAM Recalibration method is used to recalibrate the basic quality scores of the sequence-synthesis readings in an aligned BAM file. The quality scores of each reading in the BAM files after recalibration are more accurate because the reported quality score reduces the likelihood of misinterpreting the reference genome (McCormick *et al*., 2015).

VarScan; Illumina, SOLiD, Life/Pgm, Roche/454 etc. are platform-independent mutation search tools for targeted exhaust and whole genome reordering data. VarScan uses an intuitive / statistical approach to search for variants that meet the reading depth, basic quality, variant allele frequency, and desired thresholds (Koboldt*et al*., 2009). VarScan calls using a statistical test of somatic variants (SNPs and indices) based on an intuitive method and the number of aligned reads that support each allele. In somatic mode, VarScan reads the pileup files (linked file type) from the normal file and the tumor simultaneously (Koboldt *et al*., 2013).

BCF tools manipulate variant calls in Variant Call Format (VCF) and binary equivalent Bam-Calibration File (BCF). All commands work clearly with both uncompressed and Bgzf compressed VCFs and BCFs (Walker *et al*., 2014).

MuTect method was developed at Broad Institute for reliable and accurate identification of somatic point mutations in the NGS data of genomes (Larson *et al*., 2011). Somatic Sniper aims to identify single nucleotide positions that are different between tumor and normal BAM files. It takes a normal BAM file and compares these two files to determine the differences (Kroigard *et al*., 2016).

Strelka method for somatic SNV helps identifying small indices by sequencing the data of paired tumors and normal samples. In addition, this method utilizes a new Bayesian approach, which represents the continuous allele frequencies for both tumor and normal samples, while utilizing the normal expected genotype structure (Saunders *et al*., 2012). Strelka2 serves a rapid and accurate variant search tool for the analysis of germline variation in small cohorts and somatic variation in tumor / normal samples. Thus, data types can be found faster and safer (Koboldt *et al*., 2013).

2.3.2. Single Nucleotide Polymorphism Calling (SNP Calling)

NGS data, that commonly produced by genetic and genomic studies, attaches great importance to correct calling of SNP and genotypes. In NGS methods, a whole genome of the human or random parts of the targeted regions are randomly added to its short readings that are sequenced and then aligned or assembled to a reference genome. 'SNP calling', which aligns the DNA fragments of one or more persons to the reference genome, is used to identify variable fields. SNP calling is intended to determine which positions are polymorphisms or at least one of the bases is different from the reference sequence (Nielsen *et al*., 2011). In this study, "Hands-on Tutorial on SNP Calling" source from Plant Genome and Systems Biology Group/Pgsb was used (Haberer *et al*., 2018).

2.3.3. Genome Annotation

Genome annotation on nucleotide sequence carries out an important process to encode the gene positions and functions of a genome and to identify non-coding regions. Analysis of the DNA sequence by genome annotation software tools allows the discovery and mapping of genes, exons-introns, regulatory elements, repeats, and mutations (Peter *et al*., 2009). Human intervention is still necessary in order to produce better and meaningful results because this is an NP-hard problem. Annovar is the tool that users use to functionally identify various genes, explain genetic variants, and update information. Online Genome Scanner, hosted by the University of California, Santa Cruz, incorporates human genomes including mouse, worm, fly, yeast data as well as human genome reference names / versions of hg18-hg19-hg38. A list of variants with chromosomes starting position, end position, reference nucleotide and observed nucleotides can be accessed (Yang *et al*., 2015).

Gene-based explanation identifies whether SNPs are the most common genetic diversity among humans, while the region-based explanation describes variants in specific genomic regions. A filter-based annotation defines variants documented in specific databases. In addition, from the exhaust data, identifying the candidate gene list for Mendelian diseases, it helps to collect the nucleotide sequence at any user-specific genomic positions.

2.4. Evaluation of Classification Performance

In order to measure the classification performance of proposed application, confusion matrix was used. A confusion matrix contains information about actual and predicted results obtained by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

Table 1. Confusion Matrix

		Predicted Class	
		Yes	**No**
Actual Class	**Yes**	TP	FN
	No	FP	TN

Accuracy, precision, recall and F1 score evaluation criteria can be calculated with the formulas Eq. (1), Eq. (2), Eq. (3) and Eq. (4) respectively, taking all the correct classified samples into account.

$$\mathbf{Accuracy} = \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FN+FP}} \quad (1)$$

$$\mathbf{Precision} = \frac{\mathbf{TP}}{\mathbf{TP+FP}} \quad (2)$$

$$\mathbf{Recall} = \frac{\mathbf{TP}}{\mathbf{TP+FN}} \quad (3)$$

$$\mathbf{F1\ Score} = \mathbf{2} * \frac{\mathbf{Precision*Recall}}{\mathbf{Precision+Recall}} \quad (4)$$

Where, TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

3. RESULTS AND DISCUSSION

We designed a pipeline with open-source technologies to investige a practical and up-to-date study in this field. In this study, 100 human genome data was used. 20% of the data was used for system training; 80% of the data was used as the test data. Table 2 shows the distribution of these data by gender, average age in patient or healthy groups.

Table 2. Human genome data

Test Data			
Group	**Male**	**Female**	**Age (Mean)**
Patient	40	25	30
Healthy	5	10	25
Total	45	35	27.5

Evaluation of classification data with accuracy, precision, recall and F1 score criteria is shown in Table 3.

Table 3. The results of classification performance of the proposed application

Classification Performence			
Accuracy	Precision	Recall	F1 Score
92.50%	94.03%	96.92%	95.45%

The results show that the proposed application for the determination of Melanoma has very well classification performance. We used DNA mapping and gene fragment detection for determination of Malignant Melanoma like previous studies (Laila *et al*, 2016; Andrea *et al*., 2018). However, we have also individually determined the disease in the target patients by creating a personal set of rules. We found higher accuracy rates (92,5%) than previous studies (Laila *et al.*, found 70% accuracy rate and Andrea et al found 66% accuracy rate) when compared in classification performance.

To identify the tumor gene sequence, an application workflow was created using NGS technologies and a reference sample. Since this system is running on the cloud systems simultaneously, it provides a faster environment and larger DNA mapping capacity when compared to PC calculations. Thus the performance of the study was greatly increased. In previous studies, only mutated genes were detected but the results of the studies were not evaluated. Within the scope of this study, a workflow which is thought to be shared as open source and which is able to classify the gene mutations in all types of cancers in the human genome, is proposed. The evaluation of the developed workflow with various criteria and the comparison results with similar studies on the same disease in the literature provide promising and evident results.

There are some limitations of this study. Our data set has limited availability. Larger and more diverse testing data sets from patients with various genetic backgrounds could help on confirming our results.

3. CONCLUSION

The results of our study demonstrate that NGS is capable of a highly accurate diagnostic classification of Melanoma. We aimed to support this research area by the strength of the computer science and make the processes faster and more efficiently, and in the future to transform it into an expert system and bring a new breath to the field of medicine. In the future, it is expected that the modust computer-aided methods will entirely support clinical decisions and clinical practices through including all molecular genetic factors in diagnosis and treatment process of Melanoma, by analyzing DNA maps.

REFERENCES

Andrea, F., Franz, J.H., Tobias, S., *et al*. (2018). "Next-generation-sequencing of advanced Melanoma: Which genetic alterations have an impact on systemic therapy response?" *J Clin Oncol*, Vol.36, No.15, suppl. e21557-e21557.

Audibert, C., Stuntz M., and Glass, D. (2018). "Treatment Sequencing in Advanced BRAF-Mutant Melanoma Patients: CurrentPractice in the United States", *Journal of Pharmacy Technology*, Vol.34, No.1, pp. 17–23.

Bolger, A.M., Lohse M., Usadel B. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data", *Bioinformatics*, Vol. 30, No.15, pp. 2114-2120.

Bukowski, R, Guo, X, Lu Y, et al. (2017) “Construction of the third-generation Zea mays haplotype map.” *Gigascience*. Vol. 7, No. 4, pp. 1-12.

Bulmer, M. G. (1971). "The Effect of Selection on Genetic Variability," *The American Naturalist*, Vol.105, No. 943, pp. 201-211.

Ebbert, M.T., Wadsworth, M.E., Staley, L.A., et al. (2016) "Evaluating the necessity of PCR duplicate removal from next-generation sequencing dat aand a comparison of approaches", *BMC Bioinformatics*. Vol.17, No. 7, pp. 239.

Gao, J., Wan C., Zhang H., et al.(2017) "Anaconda: AN automated pipeline for somatic COpyNumber variation Detection and Annotation from tumor exome sequencing data*", BMC Bioinformatics*, Vol.18, No.1, pp:436.

Griffiths-Jones, S., Grocock, R. J., Dongen, S., Bateman, A., Enright, A. J. (2006). "miRBase: microRNA sequences, targets and gene nomenclature", *Nucleic Acids Research*, Vol.34, No.1, pp.140–144.

Haberer, G., Spannagl, M., "Hands-on Tutorial on SNP Calling" Plant Genome and Systems Biology Group/PGSB (Access Date: 01.11.2018)

Hsu, Y.C., Hsiao, Y.T., Kao, T.Y., Chang, J.G., Shieh, G.S. (2017). *Detection of Somatic Mutations in Exome Sequencing of Tumor-Only Samples*", Scientific Reports 7, 15959.

Kearse, M., Moir, R., Wilson, A., et al. (2012). "Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data." *Bioinformatics*, Vol.28, No.12, pp. 1647–1649.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). "The next-generation sequencing revolution and its impact on genomics." *Cell*, Vol.155, No.1, pp. 27-38.

Koboldt, D.C., Chen, K, Wylie, et al. (2009). "VarScan: variant detection in massively parallel sequencing of individual and pooled samples." *Bioinformatics* Vol. 25, No. 17, pp. 2283-2285.

Kroigard, A.B.,Thomassen M., Lænkholm A-V., Kruse T.A., Larsen M.J. (2016) "Evaluation of Nine Somatic Variant Callers for Detection of Somatic Mutations in Exomeand Targeted Deep Sequencing Data." *PLoS One.* Vol.11, No.3, e0151664.

Laila, N., Kelsey, M., Jordan, R., (2016). "Cytology Sample Based Next-Generation Sequencing for Metastatic Melanoma: A Feasible and UsefulTool", *Journal of the American Society of Cytopathology*, Vol.5, No.5, p. 67.

Larson, D.E., Harris C.C., Chen K., et al. (2011). "Somatic Sniper: identification of somatic point mutations in whole genome sequencing data." *Bioinformatics*. Vol.28, No.3, pp. 311-7.

Leipzig, J. (2016). "A review of bioinformatic pipeline frameworks", *Brief Bioinform*. Vol.18, No. 3, pp. 530-536.

Li, H., Handsaker, B., Wysoker, A., et al. (2009). "The Sequence Alignment/Map format and SAM tools", *Bioinformatics*. Vol. 25, No. 16, pp. 2078-2079.

McCormick, R.F., Truong, S.K., Mullet, J.E. (2015). "RIG: Recal ibration and inter relation of genomic sequence data with the GATK", *G3 (Bethesda).* Vol.5, No. 4, pp. 655-665.

McKenna, A., Hanna, M., Banks, E., et al. (2010). "The Genome Analysis Toolkit: A Map Reduce framework for analyzing next-generation DNA sequencing data." *Genome Research*. Vol.20, No. 9, pp. 1297-1303.

Moore, J.H., Asselbergs, F.W., Williams, S.M. (2010) "Bioinformatics challenges for genome-wide association studies." *Bioinformatics*, Vol.26, No.4, pp. 445–455.

Nielsen, R., Paul, J. S., Albrechtsen, A., Song, Y. S. (2011). "Geno type and SNP calling from next-generation sequencing data" *Nat Rev Genet*. Vol.12, No. 6, pp. 443-451.

Ogasawara, T., Cheng, Y., Tzeng, T-H.K. (2016) "Sam2bam: High-Performance Framework for NGS Data Preprocessing Tools." *PLoS One*. Vol. 11, no.11,. e0167100.

Peter, J. A., Cock, T.A., Jeffrey, T., et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." *Bioinformatics*, Vol.25, No. 11, pp. 1422–1423.

Rihtman, B., Meaden, S., Clokie, M.R., Koskella, B., Millard, A.D. (2016). "Assessing Illumina technology for the high-through put sequencing of bacteriophage genomes." *PeerJ*. Vol. 4, e2055.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L,J,, Cheetham, R.K. (2012). "Strelka: accurate somatic small variant calling from sequenced tumor-normal sample pairs." *Bioinformatics*, Vol.28, No. 14, pp. 1811-1817.

Sipos, B., Massingham, T., Stütz, A.M., Goldman N. (2012) "An Improved Protocol for Sequencing of Repetitive Genomic Regions and Structural Variations Using Muta genes is and Next Generation Sequencing." *PLoS One* Vol. 7, No. 8, e43359.

Sun, Z., Bhagwate, A., Prodduturi, N., Yang, P., Kocher, J. P.A. (2017) "Indel detection from RNA-seq data: tool evaluation and strategies for accurate detection of actionable mutations", *Briefings in Bioinformatics*, Vol. 18, No. 6, pp. 973–983.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S, Diaz LA, Kinzler KW. (2013). "Cancer genome landscapes." *Science*. Vol. 339, No. 6127, pp. 1546-1558.

Walker, B.J., Abeel, T., Shea, T., et al. (2014) "Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement." *PLoSOne.* Vol. 9, No. 11, e112963.

Yang, H., Wang, K. (2015). "Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR", *NatProtoc*. Vol.10, No.10, pp. 1556-66.

Zhang J., Chiodini, R., Badr, A., Zhang, G. (2011). "The impact of next-generation sequencing on genomics." *J Genet Genomics.* Vol. 38, No. 3, pp.95-109.