WHOLE EXOME SEQUENCING PIPELINE EVALUATION AND MUTATION DETECTION IN ESOPHAGEAL CANCER PATIENTS

Tran Thi Bich Ngoc[1]; Ho Viet Hoanh[2]; Vu Phuong Nhung[1]; Nguyen Hai Ha[1] Nguyen Van Ba[2]; Nguyen Dang Ton[1]; Tran Viet Tien[2]

SUMMARY

Background: Esophageal cancer is the eighth most common cancer in global scale with over 400,000 new cases arising during the year. Generally, the early diagnosis of this cancer remains limited, resulting to approximately 15% five year survival rate. Next generation sequencing technologies have revolutionized cancer genomics by providing a holistic approach for detecting somatic mutations. Hereby, we describe a genomic analysis of 30 esophageal cancer patients using whole exome sequencing. Subjects and methods: 10 sequencing datasets were analyzed through 3 different pipelines. Fastq2vcf modified to use MuTect2 proved to be the most optimal pipeline for esophageal cancer WES data analysis over SeqMule and IMPACT. The selected pipeline was used to analyze the remaining 20 datasets. Results and conclusion: Among 30 patient samples, variants found by Fastq2vcf restricted mostly in chr17 followed by chr9 and were very rare in chr21. Most variants found were SNVs (1,034/1,200 variants) and present in all samples; out of which 841 were non-synonymous. 4 types of damaging mutations causing changes to protein sequences and gene functions were found in exome regions as well as splicing regions. This study provides a comparison of software pipelines to identify potential mutations by analyzing whole exome sequencing data from cancer patients, which can lead to early detection and prevention of cancer. This information may be useful to other research related to cancer diagnosis using molecular biology and bioinformatics.

** Keywords: Esophageal cancer; Whole exome sequencing; Fastq2vcf; MuTect2.*

INTRODUCTION

In Vietnam, esophageal squamous cell carcinoma (ESCC) has been the most prevalent type of esophageal cancer and ranked sixth among leading causes of death by cancer [1]. Cancers occur when the molecules controlling normal cell growth (genes and proteins) are altered. In general, esophageal cancer is aggressive with poor prognosis and death rate tends to increase over time. The death rate per 100,000 increased 69% from 3 in 1990 to 5.1 in 2013, at an annual rate of 3%. Vietnam has the highest death rate from esophageal cancer in Southeast Asia, which ranked 12th in Asian region. The main risk factors include tobacco smoking, alcohol consumption, and poor nutrition.

1. Institute of Genome Research, Vietnam Academy of Science and Technology

2. 103 Military Hospital

Corresponding author: Nguyen Dang Ton (dtnguyen@igr.ac.vn)

Date received: 20/10/2018

Date accepted: 07/12/2018

Currently, next generation sequencing (NGS) is a popular strategy for genotyping, enabling more precise mutation detection than traditional methods due to its high resolution and high throughput. While whole genome sequencing provide general genetic information about variants, whole exome sequencing (WES) reduces the cost by targeting coding regions. WES sequencing of tumor samples and matched normal controls can quickly identify protein-altering mutations across a large number of patients, which may reveal causes of tumor. WES data is therefore increasingly used for somatic mutation detection in cancer genomics, with a large number of somatic alterations have been identified by WES in various tumor types. Accurate detection of somatic mutations in WES data remains one of the major challenges in cancer genomics due to various sources of errors, including artifacts occurring during polymerase chain reaction (PCR) amplification or targeted capture, machine errors and incorrect local read alignments. Tumor heterogeneity and normal tissue contamination generate additional difficulties for identifying tumor-specific somatic mutations. In recent years, several methods have been developed to improve the accuracy of somatic mutation calling. Despite the differences in methodology, all program identify tumor specific variants compare the tumor variant data of paired adjacent tissue and germline variant data in the same patient with the variants in dbSNP [2]. Until now, the Illumina platform is commonly used for WES in cancer studies. The two main steps in analyzing data include mapping raw reads into reference sequences and variant calling (SNP and indel). In this paper, we conducted a *Comparison three common analysis methods to choose a best pipeline for ESCC mutation detection.*

SUBJECTS AND METHODS

1. Sample preparation.

Samples were collected from 103 Military Hospital, Hanoi, Vietnam. Genomic DNA was extracted from the FFPE tissue samples of 30 patients (one sample from normal tissue and one sample from tumor tissue for each patient) using QIAamp DNA FFPE Tissue Kit (QIAGEN) following manufacture procedure. Concentration of total DNA was then determined by Qubit dsDNA BR Assay kit (ThermoFisher Scientific).

2. Library preparation and whole exome sequencing.

100 nano-gram of total DNA in 50 µL was normalized and fragmented using Covaris system (M220). Fragmented DNA was then cleaned up, repaired ends and library size selection. The remaining procedures including: Adenylate 3' ends, adapter ligation, DNA fragments enrichment, probe hybridization, hybridized probes capture and amplification of enriched library were performed following manufacture procedure of TruSeq Exome Kit (Illumina) and TruSeq DNA Library Prep for Enrichment (Illumina). Enriched library was quantified using Qubit dsDNA HS assay Kit (Thermo Fisher Scientific). DNA fragments distribution was checked on an 2100 Bioanalyzers using High sensitivity DNA chip (Agilent Techonologies) with expected size range

from 200 bp to 400 bp. Paired-end sequencing was carried on the Nextseq 500 platform (Illumina), at the Institute of Genome Research, VAST, following the manufacturer's instructions.

3. Data preprocessing and mapping.

Data is preprocessed to remove low quality bases using Trimmomatic. There are many software available for mapping. Most use Burrow Wheeler transform internally. Common mapping software include BWA, Bowtie, Novoalign, etc; of which many support multi-threading to increase performance, especially for large dataset, such as WES data. Bowtie2 is a fast and efficient mapping tool which can produce good mapping for large genome such as that of human. BWA, developed by Sanger Institute, is another common mapping software. It includes three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. BWA was designed for Illumina short reads while BWA-SW and BWA-MEM can handle reads from 70 bp to 1 Mbp long.

In our study, BWA was used to align short reads to the UCSC Human Reference Genome hg19 using default arguments. The produced SAM files were then converted to a sorted BAM format using SAMtools. Picard was used to mark duplicate reads, which can cause false positives. We also followed the best practices of GATK software for realignment and recalibration.

4. Variant calling.

Many options exist for variant calling with different targets: Germline variants, somatic mutations, copy number variants and structural variants.

Software such as GATK, SAMtools, Varscan are often used for detecting single nucleotide variants. In this study, the aim is to find somatic mutations in exome regions of esophageal cancer patients. Pipelines usually combine different software and methods. IMPACT only uses SAMtools while SeqMule uses both SAMtools, Varscan and Freebayes. FASTQ2VCF combines HaplotypeCaller and UnifiedGenotyper. As these two are not recommended for calling somatic variants, they are replaced by MuTect2 in our pipeline [3]. The set of variants found varies with software and input parameters. The intersection of results from three pipelines represent the final variant set. We conducted analysis on esophageal cancer dataset with all three pipelines above.

5. Downstream analysis.

Depending on the type of variants, related genes and information from databases, annotation tools will predict the potential effect and function of each variant. This helps researchers filter out potential variants for further investigation. Common annotation software such as ANNOVAR, Snpeff, etc has different methods and usage. Choice of annotation tool should depend on the research target and previous studies.

In our esophageal cancer study, ANNOVAR is used due to its ability to connect with several databases, i.e. ANNOVAR can remove SNVs from published databases such as 1000 genomes, dbSNP, cosmic, exac03, dbnsfp30a...

RESULTS AND DISCUSSION

1. Pipeline evaluation.

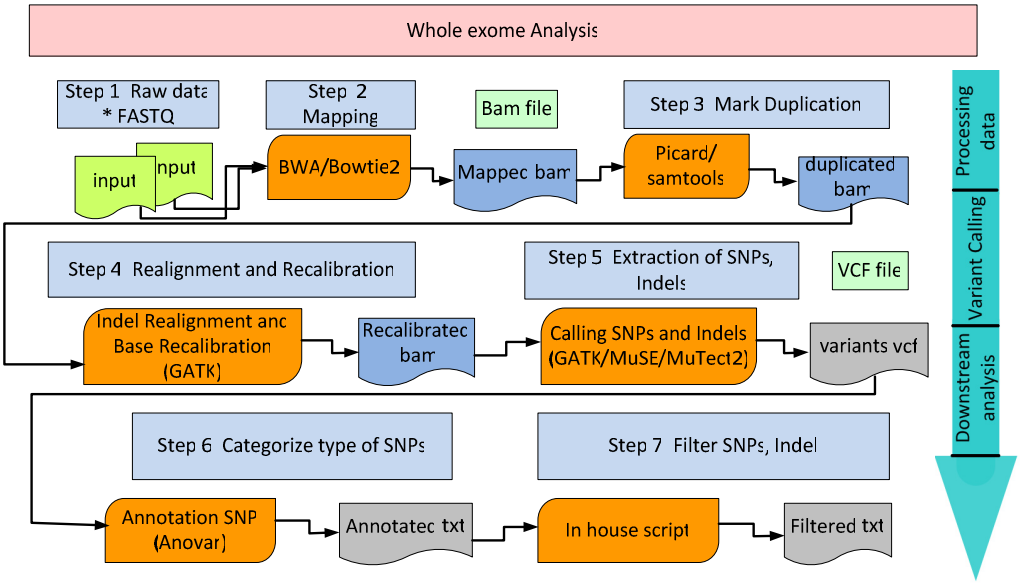

Figure 1: A common WES data analysis pipeline.

Three common WES data analysis pipeline considered in this study are SeqMule, Fastq2vcf and IMPACT. Each uses different software but follow the same steps.

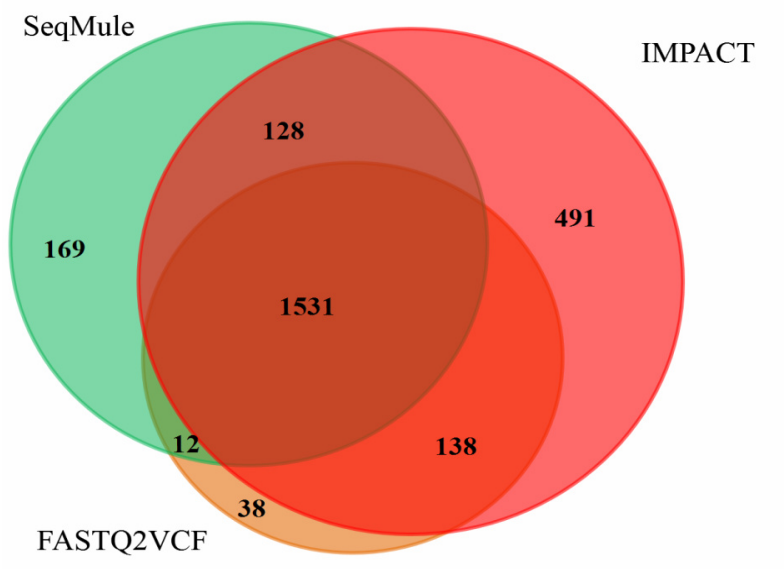

Figure 2: Variant calling results on 10 esophageal cancer datasets using 3 different pipelines.

Tumor and normal tissues pair of 10 esophageal cancer patients were analyzed with 3 pipelines. SeqMule detected 1,840 somatic mutations while IMPACT and Fastq2vcf detected 2,288 and 1,719 mutations, respectively. The intersection sets between pipelines are shown in figure 2. The number of variants found in only one pipeline were 169 (SeqMule), 491 (IMPACT) and 38 (Fastq2vcf). In the produced results, Fastq2vcf detected more than 90% the number of somatic variants called by the other 2 pipelines, higher than IMPACT (66.91%) and SeqMule (83.21%). Most somatic variants from Fastq2vcf were on genes with potential to cause esophagel cancer. Fastq2vcf also took less time to run than the other two. Hence, Fastq2vcf was used to detect variants for the remaining 20 patient samples.

Three different pipelines with several variant callers (SAMtools, FreeBayes, Varscan2 and Mutect2) were benchmarked on WES esophageal cancer data. MuTect2 produced the most accurate result, similar to research by Deng et al [1]. Fastq2vcf modified to use Mutect 2 required less time to run than the other two pipelines. We find this pipeline approriate for analyzing WES data from esophageal cancer samples. It may also be an adequate tool for other cancers as well.

2. **Prediction results.**

Whole exome data of all 30 sample pairs were shown in table 1. In exome regions, both SNVs and indels were found.

Table 1: SNV and indel numbers found on exomes of 30 patients.

Sample ID	Number of	
	SNVs	Indels
No.01	141	22
No.02	132	21
No.03	157	18
No.04	212	34
No.05	165	19
No.06	113	13
No.07	101	15
No.08	310	30
No.09	126	16
No.10	93	3
No.11	230	18
No.12	226	23
No.13	265	21
No.14	220	10
No.15	286	27

Sample ID	Number of	
	SNVs	Indels
No.16	280	26
No.17	236	14
No.18	237	24
No.19	180	13
No.20	174	16
No.21	192	22
No.22	198	22
No.23	140	12
No.24	175	19
No.25	158	13
No.26	242	15
No.27	170	20
No.28	214	23
No.29	178	30
No.30	196	16

Most variants found were SNVs (1034/1200 variants) and present in all samples; out of which 841 were non-synonymous. Variants were mainly detected on the following genes: NOTCH1 (48/841 variants/22 samples), TP53 (28/841 variants/15 samples), FAT1 (23/841 variants/15 samples), NOTCH2 (14/841 variants/10 samples), APC (11/841 variants/ 9 samples), CSMD1 (11/841 variants/8 samples), AKAP13 (10/841 variants/8 samples), FAT4 (10/841 variants/8 samples), KMT2C (10/841 variants/8 samples), AKAP9 (10/841 variants/7 samples), EP300 (10/841 variants/7 samples), ATM (8/841 variants/7 samples), PLEC (7/841 variants/7 samples), PTPN14 (7/841 variants/7 samples). Variants were rarer on genes KMT2D, FBN2, COL6A3, PALLD, SETD2, ZFHX3 (approximately 10/841 variants/6 samples).

Table 2: Annotation results in ESCC patients.

Location	Mutation types			Number of gene
Exonic	Indel	Frameshift	Deletion	43
			Insertion	16
		Nonframeshift	Deletion	20
			Insertion	10
	SNV	Non-synonymous		841
		Synonymous		193
	Stopgain			**62**
	Stoploss			**1**
	Unknown			14
Downstream	Indel			4
	SNV			25
Intergenic	Indel			176
	SNV			1,560
Intronic	Indel			212
	SNV			2,073
ncRNA_exonic	Indel			2
	SNV			72
ncRNA_intronic	Indel			17
	SNV			223
Splicing	Indel			4
	SNV			47
Upstream	Indel			9
	SNV			55
UTR3	Indel			52
	SNV			499
UTR5	Indel			7
	SNV			85

89 indels were found on 24/30 samples comprising mostly of deletions (63/89). 12 indels were found on NOTCH1 gene in 9 samples while 5 indels were found on ASXL1 gene in 4 samples. IDH2 and ATXN2 gene contained 6 and 4 indels, respectively, but only in 1 - 2 samples. 62 stopgain mutations were found in 25 samples. Only 1 stoploss mutation was present on TP53 gene in a single sample.

Splicing and downstream regions contained relatively few mutations with 51 SNPs in splicing regions (47 SNPs in 32 different genes in different samples and rarely in the same gene (1 - 2 samples)) and 29 SNPs in downstream regions (25 SNPs in different genes with only one sample has variants on the same gene).

More than 1,200 mutations were found in exon, in which chr17 had a high frequency of variants among all 30 patients, followed by chr9 (105 variants with the highest number of variants on NOTCH1 gene. No variants were found in exonic region of chr21 (*fig. 3*).

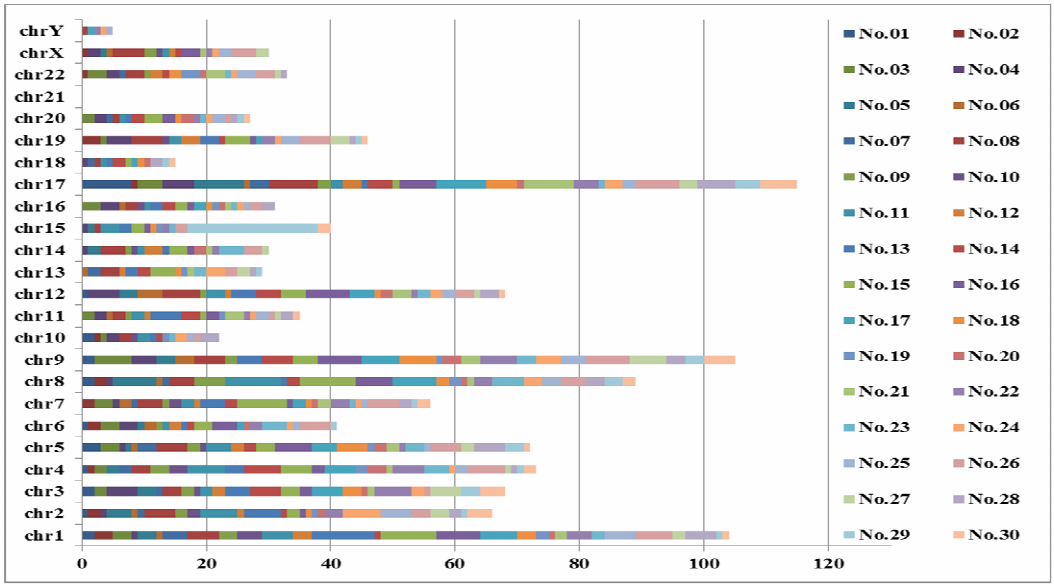

Figure 3: The number of SNVs and indels by chromosome.

Although only 30 patients were subjected for whole exome sequencing, the genes that identified in this study was previously reported by Deng et al [1]. According to their research, several genes were found that associated with esophageal cancer in 158 patients (consist of Chinese, Vietnamese and Caucasian), in which the high mutation rate was found in *CSMD3*, *TP53*, *EP300* and *NFE2L2*. Additionally, other genes discovered in current study was also in agreement with studies performed by various groups [4, 5, 6, 7, 8]. *TP53* is the most well studied tumor suppressor gene in human cancer, which was confirmed by NGS that is the most frequently mutated gene in ESCC. This gene encodes for

tp53 protein acting as tumor suppressor by regulating cell division, keeping cells from proliferating too fast or in uncontrolled way. Thereby, mutation in this gene can lead to impaired tp53 protein that is unable to control cell dividing as well as trigger apoptosis in mutated DNA containing cells. As a result, the accumulation of such cells may lead to tumor growth. The other gene that was reported commonly mutated in ESCC is *NOTCH1* with mutation rate was found at 8 - 33% [4]. *NOTCH1* encodes for Notch1 protein-a member of the Notch family receptors. Notch signaling plays an important role in cell fate determination (specialization of cells into a certain cell types in the body), cell growth and proliferation as well as differentiation and apoptosis. The Notch pathway also had been considered as both oncogene and tumor suppressor. Inactivating mutations of *NOTCH1* were identified in 21% ESCC, suggesting a role as tumor suppressor in squamous cell carcinomas [9]. Additionally, mutations of *NOTCH2* and *NOTCH3* were also detected in ESCC [7]. In addition to above well-known tumor associated genes, *EP300*-a histone modification gene was also detected in study subjects. This gene encodes for p300 protein (histone acetyltransferase), which regulate gene transcription via chromatin remodeling and plays a vital role in cell proliferation and differentiation. Besides, *KMT2C* and *KMT2D* encode for histone methyltransferase and is involved in transcription coactivation. Both *EP300* and *KMT2C* were earlier reported as histone modifier genes that frequently altered in ESCC [7, 10]. *FAT1* is an ortholog of the Drosophilla *fat* gen, this gene encodes for FAT1 protein that may act as receptor for the Hippo pathway signaling. This gene predominantly expressed in fetal epithelia and probably is important for developmental process and cell communication.

CONCLUSION

This study newly describes a comprehensive genetic screening of esophageal cancer in Vietnam, which provides mutational view and the signaling pathways likely involved in this deadly cancer. These findings are valuable for further functional examination in order to clarify the function and consequence of variants detected in study subjects.

ACKNOWLEDGEMENTS

This research was supported by program “Research on applying and developing advanced technology to support protecting and caring of public health” (Grant no. KC.10.18/16-20) and by the Institute of Genome Research, Vietnam Academy of Science and Technology (Grant No.30/QD-NCHG).

REFERENCES

1. Deng J, Chen H, Zhou D, Zhang J, Chen Y, Liu Q, Ai D, Zhu H, Chu L, Ren W. Comparative genomic analysis of esophageal squamous cell carcinoma between Asian and Caucasian patient populations. Nature Communications. 2017, 8, p.1533.

2. Liu Z.K, Shang Y.K, Chen Z.N, Bian H. A three-caller pipeline for variant analysis of cancer whole-exome sequencing data. Molecular Medicine Reports. 2017, 15, pp.2489-2494.

3. Xu H, DiCarlo J, Satya R.V, Peng Q, Wang Y. Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics. 2014, 15, p.244.

4. Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, Kong P, Wang J, Li Y, Chen W, Song B. Genomic analyses reveal mutational signatures and frequently altered genes in esophageal squamous cell carcinoma. The American Journal of Human Genetics. 2015, 96, pp.597-611.

5. Network C.G.A.R. Integrated genomic characterization of oesophageal carcinoma. Nature. 2017, 541, p.169.

6. Cheng C, Zhou Y, Li H, Xiong T, Li S, Bi Y, Kong P, Wang F, Cui H, Li Y. Whole-genome sequencing reveals diverse models of structural variations in esophageal squamous cell carcinoma. The American Journal of Human Genetics. 2016, 98, pp.256-274.

7. Gao Y.B, Chen Z.L, Li J.G, Hu X.D, Shi X.J, Sun Z.M, Zhang F, Zhao Z.R, Li Z.T, Liu Z.Y. Genetic landscape of esophageal squamous cell carcinoma. Nature Genetics. 2014, 46, p.1097.

8. Lin D.C, Hao J.J, Nagata Y, Xu L, Shang L, Meng X, Sato Y, Okuno Y, Varela A.M, Ding L.W. Genomic and molecular characterization of esophageal squamous cell carcinoma. Nature Genetics. 2014, 46, p.467.

9. Agrawal N, Jiao Y, Bettegowda C, Hutfless S.M, Wang Y, David S, Cheng Y, Twaddell W.S, Latt N.L, Shin E.J et al. Comparative genomic analysis of esophageal adenocarcinoma and squamous cell carcinoma. Cancer Discov. 2012, 2, pp.899-905.

10. Song Y, Li L, Ou Y, Gao Z, Li E, Li X, Zhang W, Wang J, Xu L, Zhou Y. Identification of genomic alterations in oesophageal squamous cell cancer. Nature. 2014, 509, p.91.