Life Science Informatics Publications

Research Journal of Life Sciences, Bioinformatics, Pharmaceutical and Chemical Sciences

Journal Home page http://www.rjlbpcs.com/

RJLBPCS

ISSN 2454-6348

Original Research Article **DOI: 10.26479/2018.0406.17**

WHOLE EXOME SEQUENCE DATA ANALYSIS FOR DETECTION OF GENE VARIANTS OF OVARIAN CANCER ANDRELATED CLINICAL STUDY

Maheswari L Patil, Shivakumar B Madagi

Department of Bioinformatics, Akkamahadevi Women's University,

Vijayapura, Karnataka, India.

ABSTRACT: The Whole Exome Sequencing (WES) is most commonly used application in next generation sequencing (NGS) involving sequencing and only the exons of all protein coding regions in whole genome. This helps to detect disease causing variants and discover of targets in gene. The present investigation uses the human ovarian cancer NGS samples and analysis using WES analysis. The samples is analyzed using FastQC tool for quality of samples followed by alignment of quality checked sampled with reference genome hg19 of human using Bowtie2. Data was generated in SAM format and converted into BAM format using SAM tool. The generated BAM file is converted to sorted bam file and then removal of duplicates using Picard tool. Finally generation of VCF file consists of variants of genes involved in causing ovarian cancer. The results showed the generation of excel file after annotation of VCF file using SIFT annotator. The results showed the genes MLH1, MSH2, BRCA1, BRCA2, ATM, PRSS1, PTEN, TP53, ERCC2, PIK3CA and EGFR are involved in causing cancer. The clinical study of the genes was carried out that indicated the clinical study provide information about variant drug pairs based variant annotations. This will help to develop personalized medicine for ovarian cancer and find out the biomarker for ovarian cancer. Whole exome sequencing data helps to identify clinical variants to predict biomarkers to detect the diseases in an early stage of diseaseandalsotointerpretpharmacogenomiccharacteristicofdrugsusedtocurethe disease.

KEYWORDS: Ovarian cancer, Whole Exome sequencing, clinical study, Next generation Sequencing.

Corresponding Author: Mrs. Maheswari L Patil*

Department of Bioinformatics, Akkamahadevi Women's University, Vijayapura, Karnataka, India.

Email Address:navi.maheswari@gmail.com

1. INTRODUCTION

Next generation sequencing (NGS) methods have the capability of performing massively parallel sequencing of large areas of the genome with high accuracy [1]. NGS methods have provided a great impetus to the discovery of genetic aberrations and their establishment as prognostic and predictive markers of diseases [2, 3]. The genomics field of research has undergone improvements lead NGS to provide higher accuracy, larger throughput and more applications than other platforms [4,5]. NGS useful for many applications on human genomes research such as, de novo genome sequencing, whole-genome resequencing or more targeted sequencing, cataloguing the transcriptomes of cells tissues and organisms (RNA–seq), genomic variation and mutation detection, genome-wide profiling of epigenetic marks and chromatin structure using methyl– seq, DNase–seq and ChIP–seq (chromatin immunoprecipitation coupled to DNA microarray) and personal genomics [6]. Next-generation sequencing strategies allow single-nucleotide resolution and reduced sequencing time and cost [7,8] facilitating larger projects such as whole-genome sequencing (WGS) and whole-exome sequencing (WES). WES not only involves finding DNA sequence of protein-coding exons it may also include in finding DNA regions that encode RNA molecules that are not involved in protein synthesis [9]. Also it is used in the development of personalized medicine [10].Ovarian cancer is fifth most common cancer among the women's [11]. There are number of risk factors available for causing ovarian cancer [12]. Whole exome sequencing used to identify gene variants in the ovarian cancer [13] this helps in detection, diagnosis, prognosis, therapy response and targets of ovarian cancer.Mutations in DNA repair genes have shown to be associated risk in causing ovarian cancer [14]. These genes include BRCA1 and BRCA2 [15, 16], the mismatch repair genes [17, 18], RAD51C [19, 20], RAD51D [21] and BRIP1 [22]. Until the ovarian cancer reaches to advance stage it not recognized in about 70% of affected women [23]. WES studies for epithelial ovarian cancer have identified FANCM as novel susceptibility gene for high grade serous ovarian cancer [14]. Using WES technology identified some of the variants in patients of ovarian cancer that are sensitivity to platinum drugs [24].The current investigation involves the analysis of NGS samples of human ovarian cancer resulting in genes causing cancer. The clinical study of these variants was carried out to know the gene drug pairs of variant annotations.

2. MATERIALS AND METHODS

Sample collection for WES analysis was retrieved from ENA database. The samples ERP035486_1 and ERP035486_2, ERP035487_1 and ERP035487_2 and ERP035488_1 and ERP035488_2 were the human ovarian cancer and are in NGS standard format file fastq. The samples were reportedly sequenced using illumina sequencer and were paired end type. The steps involved for analysis were as followed.

Analysis of Samples

The downloaded samples of the ovarian cancer were quality checked using FastQC [25] tool resulted with HTML report representing %GC content; overrepresented sequences explain low quality bases and others.

Data Preprocessing

Involves the removal of low quality bases reported to occur during analysis of samples. There are tools available for this step that removes the overrepresented sequences in samples.

Alignment of samples

The quality checked samples are aligned with the human reference genome hg19 downloaded from UCSC genome database of size 3.5GB. This is carried out using Bowtie2 [26] generating file in SAM (Sequence alignment mapping) format. There are many tools available for alignment but the Bowtie2 tool is faster in aligning [27], hence used in current work.

Post processing alignment

The step involves the conversion of SAM format file to BAM (Binary alignment mapping) format using SAM tools [28]. This is followed by generation sorted bam file and index. Also the duplicates in sorted file are removed using Picard tools Markduplicate program [28].

Variant analysis

There are two parts in the analysis of variants; one involves the generation of the mpileup file using SAM tool secondly generation of Variant calling format (VCF) using BCF tools [29]. This VCF file was used for the annotation of the genes.

Variant annotation

SIFT 4g annotator [30] was used for the annotation of genes. Generates excel file consist of the annotation of genes.

Clinical studies

The genes reported from the above WES analysis were used for the clinical annotations. This was done using the data from the PharmaGKB. The clinical annotations describe the gene drug pairs of the variant annotations of the PharmaGKB database.

3. RESULTS AND DISCUSSION

Quality Analysis of Samples

The quality analysis of experimental samples of ovarian cancer datasets can be predicted using FastQC. The overall summary of FastQC results has basic statistical information that can predict the sequence quality and duplicate reads.

Alignment summary of Samples

The alignment summary of the human ovarian cancer is summarized as follows. This was generated on console when the alignment is carried out using Bowtie2 tool.

ERP035486_1 and ERP035486_2

53295762 reads; of these:

53295762 (100.00%) were paired; of these;

44043139 (82.64%) aligned concordantly 0 times

5218549 (9.79%) aligned concordantly exactly 1 time

4034074 (7.57%) aligned concordantly >1 times

44043139 pairs aligned concordantly 0 times; of these:

1086666 (2.47%) aligned discordantly 1 time

42956473 pairs aligned 0 times concordantly or discordantly; of these:

85912946 mates make up the pairs; of these:

62946703 (73.27%) aligned 0 times

3842384 (3.54%) aligned exactly 1 time

19923859 (23.19%) aligned >1 times

40.95% overall aligned rate.

ERP035487_1 and ERP035487_2

67958352 reads; of these:

67958352 (100.00%) were paired; of these;

56425665 (83.03%) aligned concordantly 0 times

6521120 (9.60%) aligned concordantly exactly 1 time

5011567 (7.37%) aligned concordantly >1 times

56425665 pairs aligned concordantly 0 times; of these:

1396112 (2.47%) aligned discordantly 1 time

55029553 pairs aligned 0 times concordantly or discordantly; of these:

110059186 mates make up the pairs; of these:

77185465 (70.13%) aligned 0 times

4158740 (3.78%) aligned exactly 1 time

28714901 (26.89%) aligned >1 times

43.21% overall aligned rate.

ERP035488_1 and ERP035488_2

65162502 reads; of these:

65162502 (100.00%) were paired; of these;

54303131 (83.33%) aligned concordantly 0 times

6253167 (9.60%) aligned concordantly exactly 1 time

4606204 (7.07%) aligned concordantly >1 times

54303131 pairs aligned concordantly 0 times; of these:

1248482 (2.30%) aligned discordantly 1 time

53054649 pairs aligned 0 times concordantly or discordantly; of these:

106109298 mates make up the pairs; of these:

78040281 (73.55%) aligned 0 times

4193399 (3.95%) aligned exactly 1 time

23875618 (22.50%) aligned >1 times

40.12% overall aligned rate.

Variant Analysis

The annotation of WES resulted with the excel file describing following. The results showed the different variant types of gene. The variant types of genes involved in causing ovarian cancer involves Non-synonymous, Non-coding, Frameshift Deletion, Frameshift Insertion, Synonymous,Substitution, Non-Frameshift Deletion, Non-Frameshift Insertion, Start Lost, Stop Loss And Stop Gain. The results are tabulated in Table 1.

Table 1: Indicates the variant type of the annotated results of the samples

Parameters	**Sample 1 (ERP035486_1 and ERP035486_2)**	**Sample 2 (ERP035486_1 and ERP035486_2)**	**Sample 3 (ERP035486_1 and ERP035486_2)**
FRAMESHIFT DELETION	1357	3171	3537
FRAMESHIFTINSERTION	1389	3011	3252
NONCODING	16543	35999	17321
NONFRAMESHIFT DELETION	18	58	86
NONFRAMESHIFT INSERTION	26	39	73
NONSYNONYMOUS	2393	3965	2210
START-LOST	7	21	12
STOP-GAIN	108	177	115

STOP-LOSS	18	16	4
SUBSTITUTION	446	860	1396
SYNONYMOUS	961	1425	955

The table summarizes the existence of the number variant type ie mutations in 3samples. The Non synonymous variant type shows the mutation occurs due to insertion and deletion of the single nucleotide in the sequence, hence does not translate into amino acid. In the current work chose the nonsynonymous mutations occurred in all three samples and common genes were chosen. The mutations with highest mutations were selected indicating those are responsible in causing ovarian cancer. Here present work listed several genes such as MLH1, MSH2, BRCA1, BRCA2, ATM, PRSS1, PTEN, TP53, ERCC2, PIK3CA and EGFR is mainly observed in ovarian cancer and these genes also associated with other types of cancers such as breast cancer, pancreatic cancer, gastric cancer and brain tumor. These can be used as novel biomarkers for ovarian cancer.

Clinical Study of genes

The Table 2 indicates the Gene variant and drug pair's base of variant annotations. This study of variant annotations reveals genotype based summaries and describes the impact of phenotype information of the variant. The table describes the clinical study of the gene variants where most of genes that have number mutations in causing ovarian cancer with their variants are mentioned. The study revealed the phenotypic impact of the variants and summarizing the drug molecule available indicating level of annotation of the drug molecule. The genes EGFR, PIK3CA, ERCC2, PTEN and TP53 are studied, where the table describes the variants of gene, drug molecule and the phenotype.

Table 2: Summarizes the gene variant and drug pair base of clinical study

Gene	Variant	Molecule	Phenotype
TP53	rs1042522(level 2B)	antineoplastic agents cisplatin cyclophosphamide fluorouracil paclitaxel	Breast Neoplasms Neoplasms Neutropenia Ovarian Neoplasms Stomach Neoplasms
PTEN	rs17431184(level 4)	capecitabine fluorouracil(Efficacy)	Neoplasm Metastasis
ERCC2 KLC3	rs13181 (Level 3)	cisplatin oxaliplatin platinum Platinum compounds	Colorectal Neoplasms Esophageal Neoplasms Osteosarcoma Ovarian Neoplasms Pancreatic Neoplasms
ERCC2	rs1052555 (level3)	Platinum compounds	Carcinoma, Non-Small-

			Cell Lung
ERCC2	rs1799793 (level 3)	cisplatin	Neoplasms
ERCC2 KLC3	rs13181 (Level 4)	cisplatin gemcitabine	Mesothelioma
PIK3CA	rs870995 (lavel4)	docetaxel	Carcinoma, Non-Small-Cell Lung
EGFR	rs121434568(level 1b 2a 2b)	gefitinib, erlotinib carboplatin, gefitinib paclitaxel, docetaxel gemcitabine	Carcinoma, Non-Small-Cell Lung
EGFR	rs2293347(level3)	gefitinib	Carcinoma, Non-Small-Cell Lung
EGFR	rs712829 (level 3)	gefitinib	Neoplasms
	rs11506105 (level 3)	peginterferon alfa-2a peginterferon alfa-2 bribavirin	Hepatitis C, Chronic
	rs2227983(level3	cetuximab	Head and Neck Neoplasms
EGFR	rs2293347(level 3)	fluorouracil	Stomach Neoplasms
EGFR	rs712829 (level 3)	cetuximab irinotecan panitumumab	Colorectal Neoplasms
EGFR	rs712829(level 4)	Alkylating Agents geldanamycin topoisomerase I inhibitors erlotinib	Neoplasms

The levels of annotations indicate 1b, 2a, 2b, 3 and 4. The level1b reported that here the annotation for a variant-drug combination where the predominance of evidence shows an association. The level2a reported that here annotation for a variant-drug combination indicates variants are within very important pharamacogenes defined by PharamaGKB. The level2b reported that here Annotation for a variant-drug combination is with moderate evidence of an association. The level3 reported that here annotation for a variant-drug combination evaluated in multiple studies but lacking clear evidence of an association. The level4 reported that here annotation for a variant-drug is non-significant study or in vitro, molecular or functional assay evidence only.The gene TP53

indicates the variant rs1042522 annotated at level 2 has phenotype of ovarian cancer and the drug molecules antineoplastic agents, cisplatin, cyclophosphamide, fluorouracil and paclitaxel. Khrunin Andrey et al., reported in TP53 Genotype GG is not associated with increased risk of Drug Toxicity when treated with cisplatin and cyclophosphamide in women with Ovarian Neoplasms as compared to genotypes CC + CG [31]. The gene ERCC2 indicates the variant rs13181 annotated at level3 has phenotype of ovarian cancer and the drug molecules cisplatin, Platinum compounds, platinum and oxaliplatin. In ERCC2 Khrunin Andrey et al. reported that Genotypes GG + GT are not associated with decreased risk of progression-free survival or overall survival when treated with cisplatin and cyclophosphamide in women with Ovarian Neoplasms as compared to genotype TT [31].The genes EGFR, PIK3CA and PTEN showed its variants and the drug molecule and phenotypic impact of the variants as listed in the table. The phenotypes listed are other than ovarian cancer such as Colorectal Neoplasms, Head and Neck Neoplasms, Stomach Neoplasms, Carcinoma, Non-Small-Cell Lung, Mesothelioma and many others.

4. CONCLUSION

Exome-wide analysis strongly supports and extends results from previous studies employing candidate gene approaches for discovery of ovarian cancer genes. The investigation predicted MLH1, MSH2, BRCA1, BRCA2, ATM, PRSS1, PTEN, TP53, ERCC2, PIK3CA and EGFR genes mainly involved causing in ovarian cancer. The novel biomarkers developed by new strategies such as genome sequencing will provide the best opportunity to reduce ovarian cancer mortality by increasing the detection rate of early-stage disease which can be cured by surgery with or without adjuvant chemotherapy. The clinical annotations of these genes reveal the gene variant drug pair indicating variant annotations and create genotype-based summaries describing the phenotypic impact of the variant.

CONFLICT OF INTEREST

Authors have no any conflict of interest.

REFERENCES

1. Raza K, Ahmad S. Principle, analysis, application and challenges of next-generation sequencing: a review. arXiv preprint ar-Xiv:160605254 2016.
2. Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R, et al. The somatic genomic landscape of glioblastoma. Cell, 2013; 155:2, 462-77.
3. Hoadley, K. A., Yau, C., Wolf, D. M., Cherniack, A. D., Tamborero, D., Ng, S., Leiserson, M., et al. (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell, 2014, 158:4, 929-944.
4. Git, A., Dvinge, H., Salmon-Divon, M., Osborne, M., Kutter, C., Hadfield, J., Bertone, P. Caldas, C. et. al., Systematic comparison of microarray profiling, real-time PCR, and next-

generation sequencing technologies for measuring differential microRNA expression. RNA (New York, N.Y.), 2012, 16:5, 991-1006.

5. Sîrbu, A., Kerr, G., Crane, M., & Ruskin, H. J., RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. PloS one, 2012, 7:12, e50986.
6. Pareek, C. S., Smoczynski, R., &Tretyn, A, Sequencing technologies and genome sequencing. Journal of applied genetics, 2011, 52:4, 413-35.
7. Tucker, T., Marra, M. & Friedman, J. M. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Geneics 2008; 85, 142–154.
8. Shendure, J. & Ji, H. Next-generation DNA sequencing. Nat. Biotechnology 2008; 26, 1135–1145.
9. Valencia, C. A., Husami, A., Holle, J., Johnson, J. A., Qian, Y., Mathur, A., et.al., Clinical Impact and Cost-Effectiveness of Whole Exome Sequencing as a Diagnostic Tool: A Pediatric Center's Experience. Frontiers in pediatrics, 2015 3, 67.
10. Majewski J, Schwartzentruber J, Lalonde E, Montpetit A, Jabado N. What can exome sequencing do for you? J Med Genet. 2011; 48:580–9.
11. Dong A, Lu Y, Lu B. Genomic/Epigenomic Alterations in Ovarian Carcinoma: Translational Insight into Clinical Practice. Journal of Cancer. 2016; 7(11):1441-1451.
12. Burges A, Schmalfeldt B. Ovarian Cancer: Diagnosis and Treatment. DeutschesÄrzteblatt International. 2011;108(38):635-641
13. McLemore MR, Miaskowski C, Aouizerat BE, Chen L, Dodd MJ. Epidemiologic and Genetic Factors Associated with Ovarian Cancer. Cancer nursing. 2009; 32(4):281-290.
14. Dicks, E., Song, H., Ramus, S. J., Oudenhove, E. V., Tyrer, J. P., Intermaggio et.al., Germline whole exome sequencing and large-scale replication identifies FANCM as a likely high grade serous ovarian cancer susceptibility gene, 2017, Oncotarget, 8(31), 50930-50940.
15. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science. 1994, 266:66–71.
16. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature. 1995; 378:789–92.
17. Aarnio M, Sankila R, Pukkala E, Salovaara R, Aaltonen LA, de la Chapelle A, et.al., Cancer risk in mutation carriers of DNA-mismatch-repair genes. Int J Cancer. 1999, 81:214–18.
18. Song H, Cicek MS, Dicks E, Harrington P, Ramus SJ, Cunningham JM, et.al., The contribution of deleterious germline mutations in BRCA1, BRCA2 and the mismatch repair genes to ovarian cancer in the population. Hum Mol Genet. 2014, 23:4703–09.
19. Meindl A, Hellebrand H, Wiek C, Erven V, Wappenschmidt B, Niederacher D, et.al. Germline

mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. Nat Genet. 2010, 42:410–14.

20. Loveday C, Turnbull C, Ruark E, Xicola RM, Ramsay E, Hughes D, et al., Germline RAD51C mutations confer susceptibility to ovarian cancer. Nat Genet. 2012, 44:475–76.
21. Loveday C, Turnbull C, Ramsay E, Hughes D, Ruark E, Frankum JR, Bowden G, et.al., Germline mutations in RAD51D confer susceptibility to ovarian cancer. Nat Genet. 2011, 43:879–82.
22. Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A, Sigurdsson A, Jonasdottir A, et al., Mutations in BRIP1 confer high risk of ovarian cancer. Nat Genet. 2011, 43:1104–07.
23. AssaadSemaan, Kristin Delfino, Andrew Wilber, Kathy Robinson, Laurent Brard, Shaheen et.al., Exome sequencing of ovarian cancer patients to identify variants predictive of sensitivity to chemotherapy, Journal of clinical Ocology, 2017, 34:15.
24. Burges, A. &Schmalfeldt, B, Ovarian Cancer: Diagnosis and Treatment, DeutschesÄrzteblatt International, 2011, 108(38):635–641.
25. Andrews S. FastQC. Babraham Bioinformatics; 2010. Cambridge, UK.
26. Langmead B. and S. L. Salzberg. 2012. Fast gapped- read alignment with Bowtie 2. Nature Methods, 9(4):57– 359,
27. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010; 26:873–81.
28. Ogasawara, T., Cheng, Y., &Tzeng, T. K. Sam2bam: High-Performance Framework for NGS Data Preprocessing Tools. PloS one, 2016, 11:11, e0167100.
29. www. samtools.github.io/bcftools/bcftools.html [Last accessed 14-10-2018].
30. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protocols. 2016; 11: 1-9.
31. Khrunin Andrey, Feodosia, IvanovaAlexey, Moisseev,Denis, Khokhrin, YuliyaSleptsova, Vera Gorbunova et.al., Pharmacogenomics of cisplatin-based chemotherapy in ovarian cancer patients of different ethnic origins. Pharmacogenomics. 2012; 13:2.