Exhaustive capture of ovarian cancer transcriptional and genomic variant integrating canonical and mapping-free protocols

Zhiqin Fu[1], Chao Lu [2], Chao Ding[1], Chengxi Zhou[1], Tingting Yu[3], Yue Yang[1,*], Liang Shi[1,*]

[1] The Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital) Institute of Basic Medicine and Cancer(IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang Province, 310022, China;

[2] Department of Gastrointestinal and Pancreatic Surgery, Zhejiang Provincial People's Hospital, Hangzhou, Zhejiang Province, 310014, China;

[3] Xihu District Hangzhou SanDu Town Community Health Service Center, Hangzhou, Zhejiang Province, 310000, China

[*]**corresponding author:** Yue Yang and Liang Shi,YueYang,yangyue@zjcc.org.cn.

Liang Shi, shiliang@zjcc.org.cn

Zhiqin Fu :19870613cc@163.com;

Chao Lu: lc1342@163.com

Chao Ding: dingchao@zjcc.org.cn

Chengxi Zhou: zhoucx@zjcc.org.cn

Tingting Yu:435517746@qq.com

Yue Yang: yangyue@zjcc.org.cn

Liang Shi: shiliang@zjcc.org.cn

Abstract

Ovarian cancer is the most frequent cause of deaths in gynecologic malignancies. Many possible mechanisms have been proposed via RNAseq and DNAseq technique recently. However, the driving factors are still obscure. The possible reasons are attributed to the incomplete human reference. This study integrated the canonical mapping-based and mapping-free protocols to extract reliable variations and novel events. We eventually obtained 450 reliable SNVs from the WES data and novel events from the RNAseq data, including 154 SNVs, 462 intron events, two repeats and six splice events. We identified six differentially expressed genes and six contigs that are significantly related to survival prognosis. The recurrent SNVs in significantly differentially expressed genes can be validated in an independent cohort of 20 Chinese ovarian cancer patients.

Keywords: Ovarian cancer, mapping-based, mapping-free, SNVs, survival prognosis

Introduction

Ovarian cancer (OC) is a highly heterogeneous cancer. There are three main tumor areas in the ovary. The surface epithelium is where most malignant tumors occur[1]. It is presented in different types of histology. Serous ovarian cancer (SOC) is the most common type and often develops in old age[2]. Endometrioid cancer tends to develop at a young age and is associated with endometriosis[3]. Mucinous carcinoma and clear cell carcinoma also appear at a young age[4].

These malignant tumors' complexity stems from multi-dimensional genetic variation affected by changes in genetic factors, including transcription levels and genome levels. With next-generation sequencing, we can use RNAseq data to analyze genes or transcripts significantly different in OC patients at the transcriptome level. At the same time, we can use WES data to obtain genes or loci with high-frequency mutations in OC patients. Many studies have detected and confirmed differentially expressed genes and mutations in OC patients[5–7]. However, these variants were detected either using one approach or using one dataset of cohorts. The consistency between multiple studies is much lower than expected[8].

The canonical sequencing analysis protocols have identified a large number of genetic variations through comparison with reference sequences. However, the current human reference sequence is still incomplete[9]. There are many gaps in the human genome, including telomere regions, centromeres, as well as a large number of repetitive regions and other low complexity regions. Due to these regions' low mappability, it is challenging to obtain confidential coverage during the sequencing process, so we also call them the “dark genome”[10]. Many cancer-related genetic variants exist in these gaps, but the canonical sequencing analysis protocols often ignore or discard these results.

Another alternative approach is mapping-free protocols. In recent years, researchers have developed many analysis methods that do not rely on reference sequences. One of the commonly used algorithms is De Bruijn graph (DBG)[11]. DBG is widely used in the de novo assembly, especially for the species without available or complete reference. DBG is the graph algorithm based on the k-mers approach, which decomposes the reads into smaller k-mers. Then a graph is constructed according to the overlap between k-mers. A variant generates a bubble structure due to there are two alleles in each branch. Therefore, variants are captured through searching for bubbles from the DBG. Generally, each library is used to construct its DBG and capture variants independently. Another mapping-free protocol dealing with large numbers of cohorts is called DEkupl[12]. This algorithm first screen kmers that are absent in the reference and then uses a differential test to further select significant kmers between two conditions. The differential test methods include T-test, DESeq2[13] and LimmaVoom[14]. Finally, these selected kmers are merged into contigs. All these contigs are supposed to harbor variants or belong to the genome gaps.

This study integrated the mapping-based protocol and mapping-free protocol to achieve a comprehensive analysis of OC patients. Both RNAseq and WES data were applied to obtain novel transcriptional events and convincing SNVs. The novel events were validated using an independent cohort of 20 Chinese OC patients.

Methods and materials

1. **data extraction**
The raw fastq files were retrieved from The Cancer Genome Atlas (TCGA)database[15]. Both of the RNAseq and WES sequencing data were involved in this study. All patients' clinical information was also obtained, including the survival time, stage, relapse, or metastasis. Patients were divided into two groups: complete remission and progressive group. Adapters sequences were trimmed using the cutadapt software[16]. Duplicated reads were removed since these reads are majority generated from the PCR process instead of natural status in cells.

2. **Canonical protocol based on reference**
The latest version of the genome and annotation files were downloaded from the Gencode[17]. Reads from RNAseq data were mapped to the human genome of the hg38 version using the STAR software with default parameters[18]. Reads from the whole exome sequencing (WES) data were mapped to the human genome using the BWA algorithm with default parameters[19]. The read counts were normalized to RPKM. DESeq2 was used to screen differentially expressed genes. GATK was applied to call variants, and all variants were stored in the Variant Call Format (VCF) format files[20].

3. **Mapping-free protocol without reference**
As the human genome is not completed and gaps are present, all variations within the 'unannotated' regions are not captured by canonical methods. Herein, we introduced a mapping-free protocol named DEkupl[12]. The reads are decomposed to kmers and all the kmers different to reference are retained. In this way, we can exhaustively capture all the variations in the "dark genome regions". Besides, the DEkupl estimates the gene expression using the Kallisto, which is also a reference-free software[21]. The DEkupl software provides both differentially expressed genes detected by limma and novel events with variations exhibited as contigs.

4. **Gene-level candidates**
As the gene expression estimators in two protocols are based on different rationales, we compared the differential genes detected by both protocols. Only the consistent differential genes were considered to be cancer-related candidates. The P-value cutoff was set to 0.05 after false discovery rate correction. Genes with log2FC over than 1 were deemed up-regulated, and genes with log2FC smaller than -1 were down-regulated.

Eventually, the differentially expressed genes (DEGs) between responsive patients and progressive patients were selected.

5. Contig-level candidates

Since we have no prior information on the genome loci in our mapping-free protocol, we listed all the contigs harboring variations. We can benefit from this way because variations within the repeat or low complexity regions can also be kept, which would be ignored or discarded by the mappers such as STAR or BWA. Only the differentially expressed contigs were retained and further mapped to the genome for annotation. The annotation process was done using GSNAP of version 2020-06-04[22]. The contigs contain multiple events, including single nucleotide variants (SNV), splice, split, lincRNA, polyA, repeat, and unmapped. The unmapped contigs may either come from the exogenous microorganisms or unannotated human genome. The consistent variations with Genome Analysis Toolkit (GATK) results were considered as convincing variations. Meanwhile, the mapping-free protocol specific products were regarded as 'novel' events.

6. Correlation analysis between convincing candidates and prognosis

The convincing candidates were composed of differentially expressed genes and recurrent SNVs detected by mapping-based and mapping-free protocols. The candidates were compared to the patients' survival prognosis using the log-rank test[23]. We aim to detect the diagnostic and prognostic candidates. The Kaplan–Meier curves[24] were drawn for the top prognosis related candidates. Besides the univariate regression, we also applied the Cox proportional-hazards model (CoxPh), which is a multivariable regression model[25].

7. Diagnostic model construction integrating the gene and contig signatures

The significant gene candidates and contig candidates were screened using a log-rank test. The further lasso regression[26] feature selection method was used to select diagnostic values signatures. The samples were randomly separated into two groups, in which 75% were training set and the rest 25% were test set. Using the selected signatures, a support vector machine (SVM) model was constructed[27]. Cross-validation[28] was used to assess the performance of the diagnostic model on the trainset. The ROC curve was shown using the test set in terms of sensitivity and specificity[29].

8. Deeply investigate the novel signatures

To deeply investigate the correlation between the novel signatures and cancer progression, we compared the expression profile of signature contigs between the early stage (stage I/II) and late-stage (stage III/IV) patients. The role of signature contigs regarding prognosis was also investigated. Survival analysis was performed to show the potential signature contigs to be used as prognostic indicators.

9. Independent validation

To validate the novel events we identified using the mapping-free protocol, we performed the WES sequencing using 20 paired clinical patients from the Cancer

Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital). All procedures performed in this study involving human participants were consistent with the ethics standards of the Declaration of Helsinki and its later amendments or comparable ethics standards. The ethics committee of Zhejiang Cancer Hospital approved this study. All tissues were collected from the biobank of Zhejiang Cancer Hospital (Reference number IRB-2019-5). All patients were diagnosed according to the International Federation of Obstetrics and Gynaecology (FIGO) classification. The patients were 56 years old in average with stage III to IV high-grade serous ovarian cancer. The white blood cells from patients were used as negative control. The WES data were obtained using novaseq 6000 platform. The sequencing depth and read length are 100 and pair-end 150 bp. Detailed information can be seen from the supplementary table S1.

Results

1. Differentially expressed genes from RNAseq data

Differentially expressed genes (DEGs) between patients responsive and progressive were extracted from the RNAseq data using the limma algorithm. The selected genes present diverse expression levels between two groups of patients. Up/down-regulated genes were extracted according to the P values and fold change values. The volcano graphs of both canonical RNAseq pipeline and DEkupl were drawn, as shown in Fig. 1A.

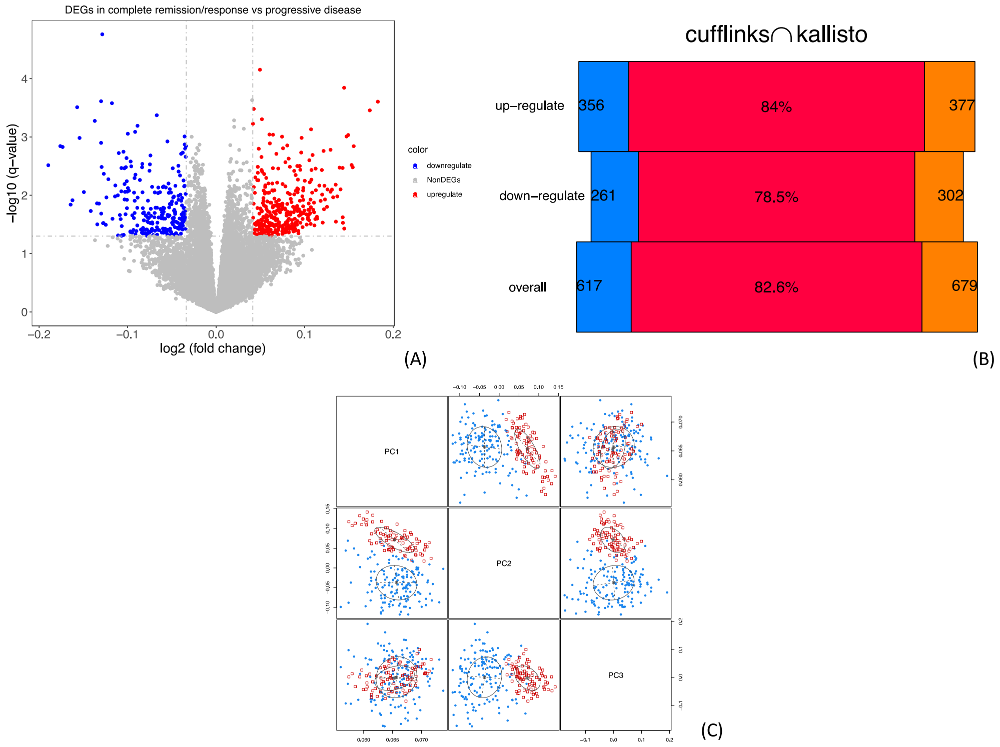


Figure 1 differentially expressed genes (A) volcano graph of genes using cufflinks. Red/blue dots represent the up/down-regulated genes. (B) The results generated by cufflinks are marked in blue and the results of Kallisto are marked in orange. The red bars in the middle indicate the Jaccard index. (C) Principle component analysis. Response/progressive patients are marked in blue and red, respectively.

Fig. 1B shows the overlapped DEGs between two protocols. The overlapped DEGs were used as stable prognostic risk genes. Eventually, we obtained 587 DEGs in total, including 226 up-regulated and 251 down-regulated genes.

To check if two groups of patients present differences on genes level, we performed the principal component analysis (PCA)[30]. As seen in Fig. 1C, the two groups of patients can be distinguished using the first three components.

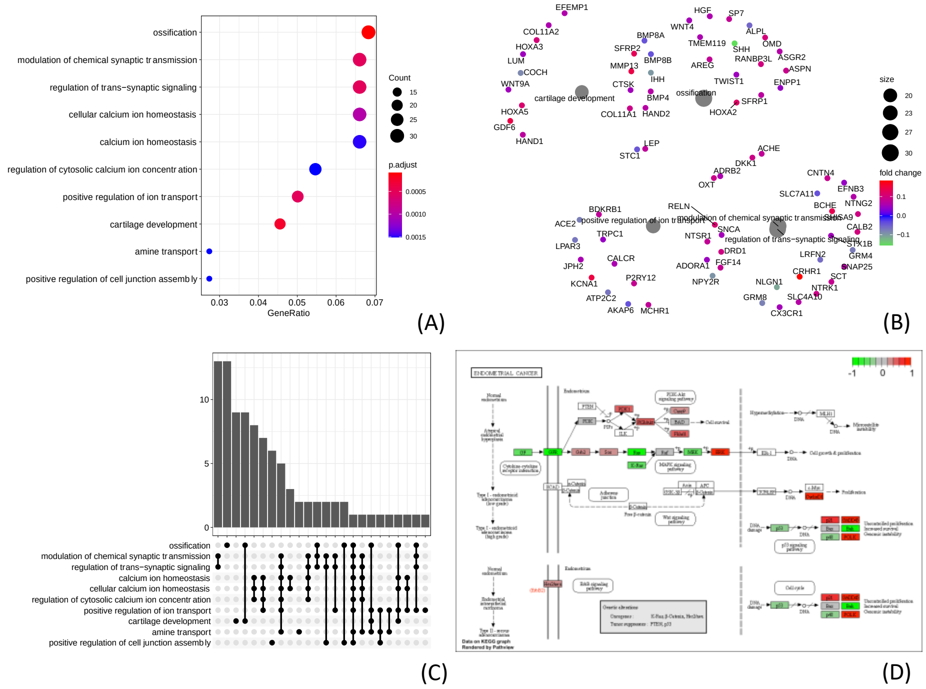


Figure 2 Function enrichment using DEGs (A) The dotplot of the top 10 enriched gene ontology biological process terms. (B) The regulatory relationships between genes and corresponding functions. (C) The upset graph is showing the number of overlapped genes among functions. (D) The most significant KEGG pathway

To further clarify the functions regulated by the DEGs, we enriched the DEGs to both of gene ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG pathway)[31] using the clusterProfiler package[32]. The top 10 GO biology process terms are shown in Fig. 2A. The significant functions include ossification and calcium ion homeostasis. The ossification has been widely proved to be a risk factor that drives multiple kinds of cancers[33]. And the calcium ion homeostasis plays a crucial role in the cancer cell differentiation[34]. Fig. 2B shows the four hub functions and corresponding genes. These functions are supposed to be activated as most of the involved genes were up-regulated in progressive patients. Fig. 2C suggests that some enriched functions share cross-talk genes. These genes achieve the communication between functions as bridges. Fig. 2D is the most significant KEGG pathway that was enriched by 26 DEGs. It implies the genetic association between ovarian cancer and endometrial cancer.

2. Recurrent SNVs from WES data

Large numbers of mutated genes do not express in the ovarian tissues. The significant number of SNVs is attributed to the gene's length, for instance, TTN[35]. Therefore, the frequently mutated genes have little impact on the disease progression if they are not expressed. Given this hypothesis, we only investigated the SNVs from the differentially expressed genes. The most recurrently mutated genes were identified using the GATK across all patients. The frequency and patients harboring each mutated gene can be seen from Fig. 3A.

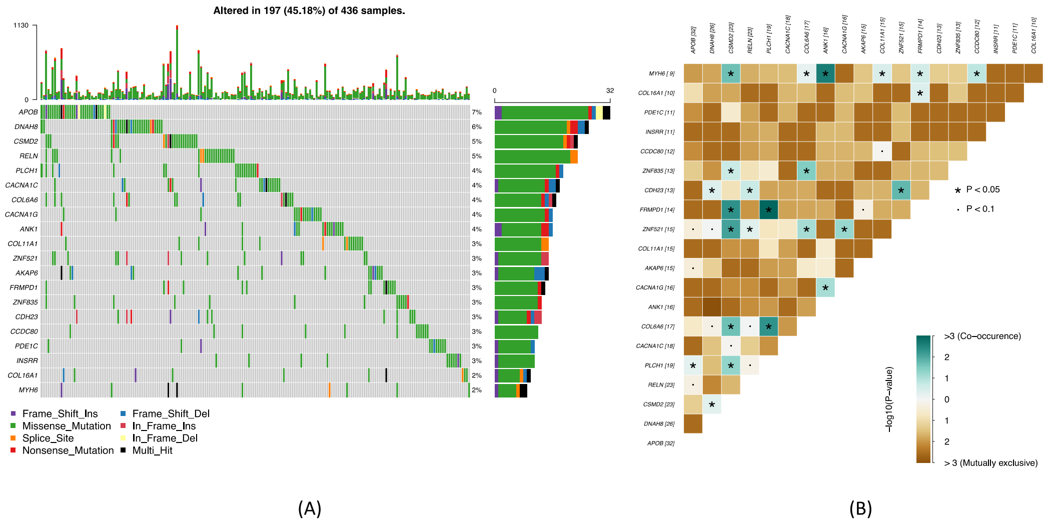


Figure 3 Top mutated genes that are also differentially expressed (A) Oncoplot of top mutated genes. Each row indicates one gene and each column represent one patient. (B) Interactions between mutated genes. The co-occurrence and mutually exclusive interactions are marked in blue and yellow, respectively.

As seen in Fig. 3A, the top 20 differentially expressed genes with the highest mutational frequency can be observed in at least 2% of patients, in which the highest frequency is 7%. Interestingly, the most well-known cancer suppressor gene, TP53[36], is absent in our top 20 gene list, which has a mutational frequency of 88% in ovarian cancer patients. The reason is TP53 is not differentially expressed in progressive ovarian cancer patients despite a high mutational frequency. In contrast, the genes in our list were recurrently mutated in ovarian cancer patients and presented significant differences as the tumor progressed (Table 1).

Table 1 Limma result of the top 20 genes

gene	logFC	P.Value	mean_response	mean_progressive	direction

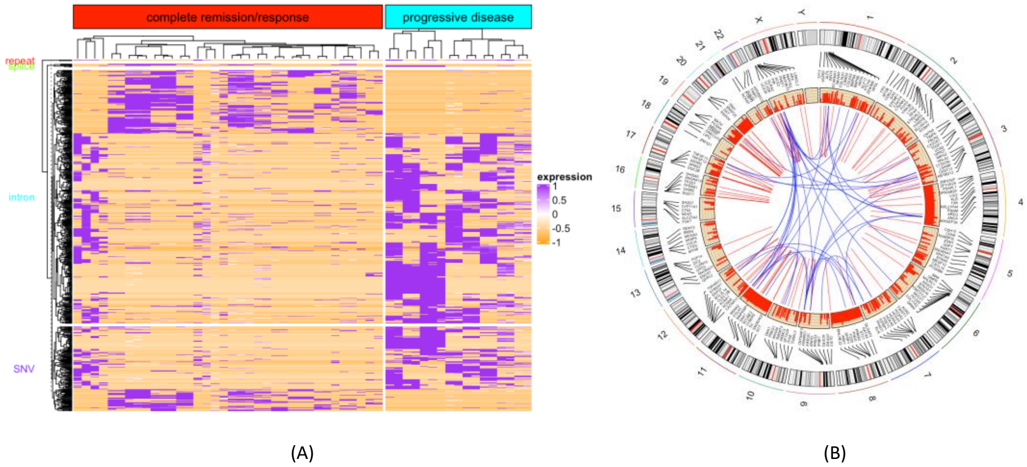
APOB	0.095	0.012	0.318	0.413	upregulate
DNAH8	-0.067	0,035	0,289	0,222	downregulate
CSMD2	0,084	0,02	0,905	0,989	upregulate
CACNA1C	0,059	0,009	1,016	1,075	upregulate
RELN	0,112	0,026	0,774	0,886	upregulate
PLCH1	-0,101	0,002	1,038	0,936	downregulate
COL11A1	0,093	0,041	1,143	1,236	upregulate
COL6A6	0,143	0,003	0,689	0,832	upregulate
CACNA1G	0,08	0,014	0,855	0,936	upregulate
ZNF835	-0,055	0,039	1,023	0,968	downregulate
CDH23	0,047	0,039	1,163	1,21	upregulate
FRMPD1	-0,063	0,017	1,044	0,981	downregulate
ZNF521	0,058	0,019	1,139	1,197	upregulate
ANK1	-0,078	0,024	0,808	0,731	downregulate
AKAP6	-0,039	0,032	1,089	1,05	downregulate
PDE1C	0,105	0,003	0,372	0,477	upregulate
CCDC80	0,051	0,004	1,365	1,417	upregulate
COL16A1	0,047	0,003	1,266	1,313	upregulate
MYH6	-0,106	0,048	0,391	0,285	downregulate
INSRR	0,112	0,003	0,319	0,432	upregulate

To further investigate the internal relationship between the most mutated genes, we applied the maftools[37] to estimate the interactions between genes. There are two types of interactions between genes that are the co-occurrence and mutually exclusive relationship. Co-occurrence interaction indicates that two genes tend to be mutated simultaneously[38]. The mutually exclusive interaction suggests that two genes are barely observed to be mutated simultaneously[39]. It's worth mentioning that mutually exclusive interaction does not mean two genes could not be mutated simultaneously. Instead, it means the cells carrying both of the two mutated genes tend to trigger the apoptosis program and turn out to die. This phenomenon is called synthetic lethality[40], which is a significant avenue of cancer therapy.

As seen in Fig. 3B, there are many mutually exclusive gene pairs whose P values are less than 0.1, marked as dots. The predominant interaction among the top 20 genes is co-occurrence. The gene pairs with co-occurrence relationships are marked with stars, suggesting significant P values less than 0.05.

3. Novel events identified by DEkupl

Since we also applied a mapping-free protocol, we were able to capture novel events that were unsolvable to the mapping-based methods. DEkupl captured all the kmers that were absent in the human reference and then merged the kmers into contigs. These contigs were therefore considered to be the context of local events. We eventually identified contigs composed of 154 SNVs, 462 intron events, 2 repeats, 6 splice and 24 unmapped events. As DEkupl masked all the kmers present in the reference, these contigs extended from the retained kmers can be considered as novel events. The detailed results can be seen from the supplementary table. All these novel events present significant differences between the two groups (Fig. 4A).



(A) (B)

Figure 4 DEkupl contigs results (A) The heatmap of novel events identified by DEkupl using RNAseq data. (B) The circus graph of SNVs identified by DEkupl using the WES data.

In Fig. 4A, most of the novel contigs were highly expressed in the progressive group. Meanwhile, we observed that some complete remission patients exhibit similar expression patterns as the progressive patients. It suggests that patients sensitive to therapy at the initial phase might carry a risk of relapse in the following phases.

4. Candidates SNVs screening

Besides the shared DEGs, we also screened the SNVs detected by both of the two protocols as convincing SNVs. The contigs generated by DEkupl were mapped to the human genome using GSNAP software. Then the genomic coordinates were compared with the GATK results. Finally, we obtained 450 convincing SNVs belonging to the DEGs (Fig. 4B). Fig. 4B shows the locations of SNVs and the correlation between the host genes. Some SNVs are found to locate at closed loci forming a local cluster. And genes from the same chromosome tend to present a positive correlation. In contrast, genes from distant genomic loci tend to have a negative correlation.

5. Prognosis related indicators

Since the SNVs whose host genes belong to the DEGs were considered potential indicators, we intended to extract prognosis related indicators from the whole DEGs list. We initially used the log-rank test to estimate each gene. The patients were divided into two groups according to the median expression value of the queried gene. Then the log-rank test P value was computed. After ranking the genes P values in ascending order, we selected the top 6 genes and drew the KM curves, as shown in Fig. 5.

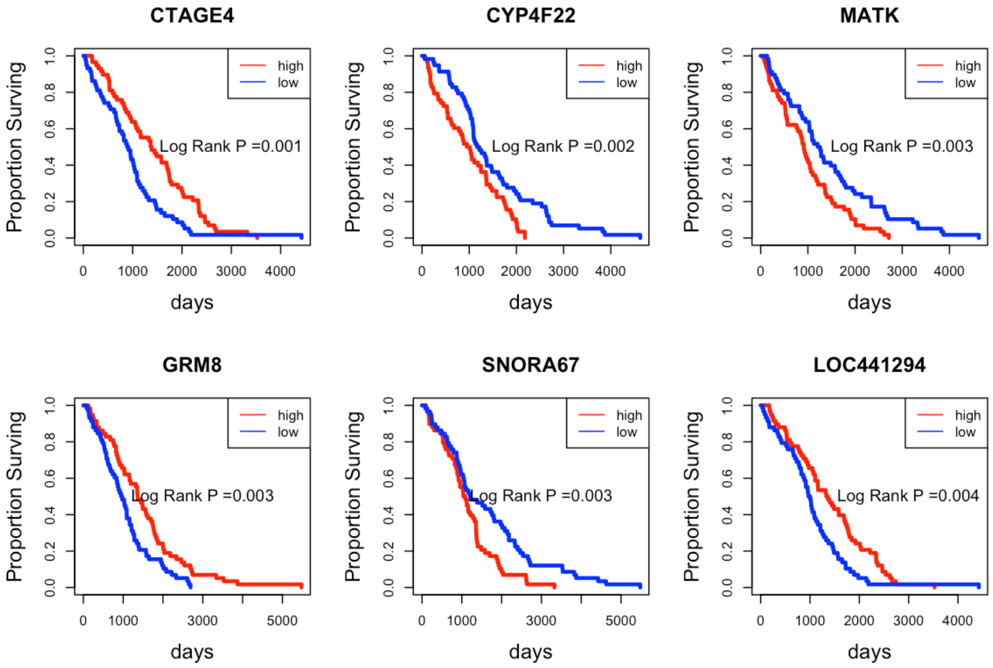


Figure 5 Kaplan–Meier curves of the top 6 prognostic indicator genes. The patients with high/low expression levels are marked in red and blue, respectively.

As seen from Fig. 5, patients with different expression levels of the six genes tend to present various surviving time. The high expression levels of gene CYP4F22, MATK and SNORA67 always imply a poor surviving prognosis. In contrast, the low expression levels of gene CTAGE4, GRM8 and LOC441294 indicate a poor surviving prognosis.

We also applied the coxPh, a multivariable regression method, to select prognostic indicators from the DEGs. The top 10 genes can be seen from Table 2.

Table 2 CoxPh results

gene	beta	HR (95% CI for HR)	wald.test	p.value
ARHGEF38	-1.2	0.31 (0.16-0.6)	12	0.00057
PRSS16	-1.2	0.31 (0.15-0.64)	10	0.0016
SNORA52	1.4	4.1 (1.7-9.8)	9.8	0.0018
SNORA84	1	2.8 (1.4-5.3)	9.1	0.0026
ST7OT4	-1.4	0.24 (0.093-0.61)	8.9	0.0029
LOC728606	-1	0.36 (0.18-0.71)	8.6	0.0033
CACNA1C	1.7	5.4 (1.7-17)	8.4	0.0038
TPO	1.1	3 (1.4-6.5)	8.3	0.0039
INSRR	0.83	2.3 (1.3-4.1)	7.6	0.0058
RNF183	-1.3	0.27 (0.1-0.69)	7.6	0.006

The top indicator lists in coxPh and log-rank test are not the same. This is because coxPh considers the internal interactions among genes, while the log-rank test treats each gene independently. However, the top 6 genes detected by the log-rank test all have higher ranks among the 587 genes. The gene SNORA67 ranks at the 15th position and the gene with the lowest rank is MATK, whose rank is 220th.

6. Diagnostic model construction

To screen the diagnostic signatures, we herein applied a machine learning feature selection algorithm called lasso regression. We combined all the DEGs and convincing SNVs as features. Lasso regression determined the best combination of features according to the Log gamma criteria.

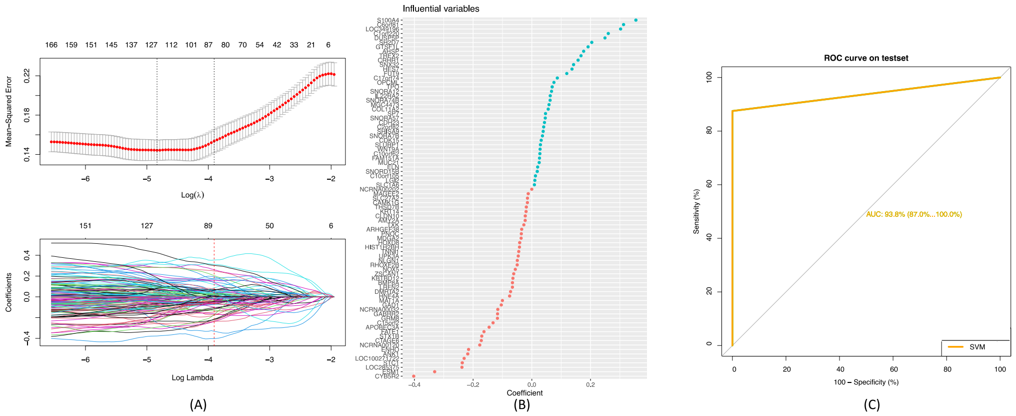


Figure 6 machine learning results (A) Feature selection process performed by lasso regression. (B) Feature importance ranking according to the coefficient. (C) ROC curve of the model on the test set.

In Fig. 6A, 82 signatures were selected by lasso regression via the cutoff of log gamma. The best log gamma cutoff was determined automatically by lasso regression. The coefficient of each signature can be seen from the Fig. 6B. Positive and negative signatures were marked in blue and red, respectively. Finally, the SVM model was trained using the 82 signatures and the AUC was 93.8% on the test set. Our results demonstrate that the signatures in our findings can distinguish progressive patients from the responsive patients.

7. Survival analysis using the contigs of novel events

Besides DEGs, we also captured some novel events from the RNAseq data. These novel events were composed of SNVs, intron, splice, repeat and unmapped contigs. It's worth mentioning that the unmapped contigs may either come from transcripts produced by rearranged genes or result by exogenous viral genomes and could, thus, be highly relevant biologically. We therefore investigated the correlation between surviving time and the novel events, including unmapped contigs.

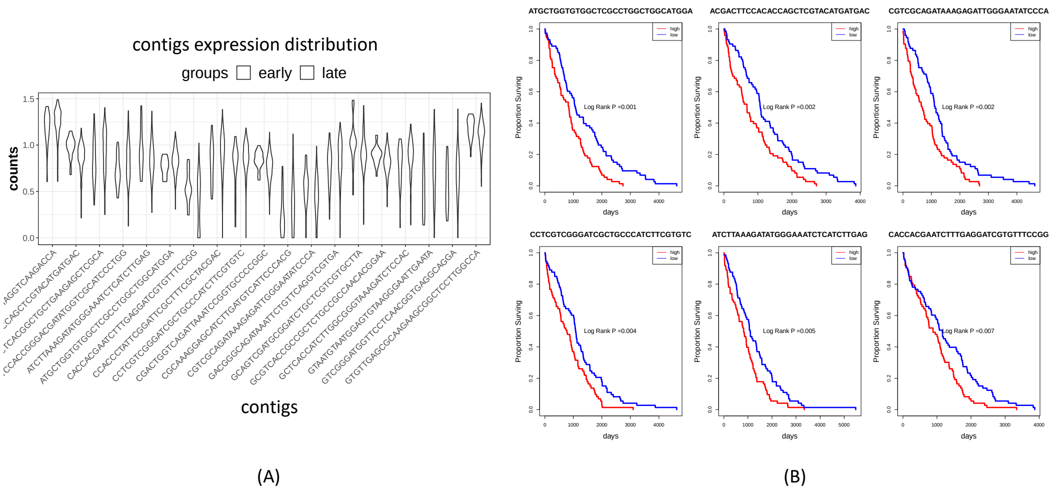


Figure 7 expression of contigs in different cancer groups (A) The violin graph shows the expression of the top 20 contigs of novel events in the early and late-stage patients. (B) Survival analysis shows the top 6 most significant contigs

We compared the expression of contigs corresponding to the novel events, as shown in Fig. 7A. The violin graph indicates the transcriptional difference between the early and late-stage patients. Survival analysis also proves the correlation between these novel events and survival prognosis. Interestingly, we showed the top 6 most significant contigs in Fig. 7B, all the patients with high expressions of these 6 contigs tend to have a worse prognosis than the others. Our findings imply that the overexpression of these unannotated events may promote cancer progression.

8. Validation using independent dataset

To validate the convincing SNVs identified by DEkupl from the TCGA patients, we recruited 20 matched ovarian cancer patients from the hospital. Twenty tumor tissues and 20 matched blood tissues were used to do the whole-exome sequencing. To select the tumor-specific SNVs, we collected all the libraries from blood tissues as a normal panel. Then the recurrent SNVs observed in at least two patients were screened as tumor-specific SNVs. We eventually obtained 3013 contigs corresponding to 1872 genes. There were 719 contigs mapping to the exons and 1355 contigs mapping to the introns. The rest contigs were mapped to the intergenic regions.

We investigated the 450 convincing SNVs identified by both mapping-based and mapping-free protocols in the TCGA cohorts from the independent dataset. We found that only 23 of the 450 SNVs were considered recurrent SNVs in the validation cohort (Fig. 8).

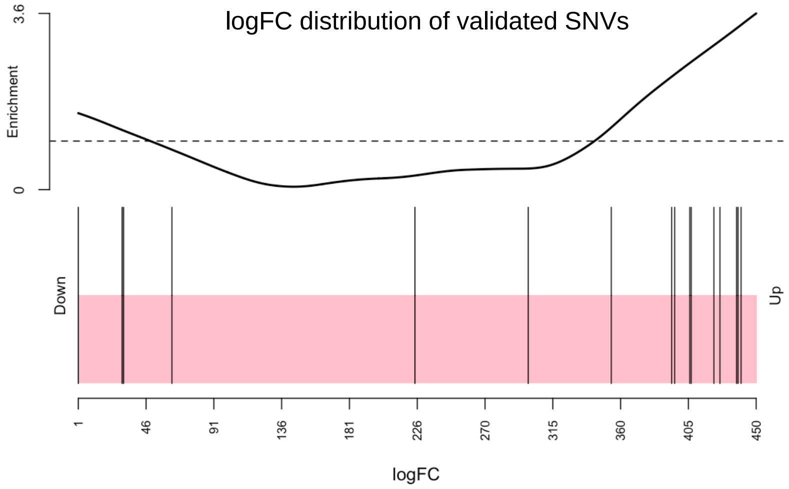


Figure 8 LogFC distribution of validated SNVs. The x-axis represents the genes logFC of all 450 validated SNVs. The vertical lines represent the 23 validated SNVs in the independent cohorts. The curve represents the enrichment score from down to up-regulated genes.

As seen from Fig. 8, even though only 23 SNVs are consistent between the TCGA cohorts and the independent cohorts, these common SNVs tend to present in significantly differentially expressed genes. The enrichment P-value calculated after 10000 times permutation was 0.04, which means the shared SNVs significantly enrich in the down/up-regulated genes.

Discussion

Recurrent variations are valuable tools in cancer diagnosis and treatment. Within this scope, sequencing technology has revealed the universality and diversity of the human transcriptome and genome. However, there are a large number of false positives in these cancer-related genes. For example, TTN, a high-frequency mutation gene detected in many cancers, but TTN is also one of the longest genes known to humans. Thus, it has a higher probability of accumulating more mutations.

On the other hand, many cancer mutations are found in some implausible genes (such as those encoding olfactory receptors and the muscle protein titin). However, many genes that carry mutations are not expressed in cancer tissues. Therefore, we combined the RNAseq and WES data to screen out genes with high-frequency mutations from genes significantly differentially expressed in ovarian cancer patients.

We first compared responsive patients to the progressive patients and screened out 226 up-regulated and 251 down-regulated genes. These DEGs were concentrated in the calcium ion homeostasis related functions. The endometrial cancer KEGG pathway was also enriched. The calcium ion homeostasis is widely proved to be a cancer promoter. The abnormal calcium ion homeostasis may result in the dysfunction of multiple biological processes, including cellular repair processes and cell proliferation[41].

Highly mutated genes were selected from the DEGs. These genes present diverse expression levels between the responsive and progressive OC patients. On the other hand, these genes harbor recurrent mutations in at least 2% of patients. Thus, we avoided the false positive genes that are either non-differentially expressed or have nothing to do with cancer progression. To further investigate these mutated genes' internal relationships, we drew a heatmap showing the co-occurrence and mutually exclusive relationship. We observed some significant co-occurrence gene pairs, like ANK1 and MYH6, FRMPD1 and PLCH1, COL6A6 and PLCH1. These genes tend to be mutated in the same patients. Meanwhile, we found the gene pairs with mutually exclusive relationships, indicating the potential synthetic lethality candidates. This kind of gene pairs are barely observed to be mutated simultaneously. Therefore, for tumor cells with one of the mutated genes in a synthetic lethality pair, the other gene can be considered as a potential therapeutic target. Tumor cells can be killed when block the other gene using inhibitors, which mimics the condition of two genes mutated together.

Besides, the canonical sequencing analysis protocol relies on comparison with reference sequences to detect mutations. This method is highly dependent on the accuracy and completeness of the reference sequence. At the same time, it cannot do anything about mutations out of the reference sequence. However, many cancer-related mutations are hidden in these "dark genome". In-depth exploration of these unmappable regions is essential to complement the current human understanding of cancer. Therefore, in this study, we also used a mapping-free method called Dekupl. On the one hand, the two mapping-based and mapping-free protocols validate each other to screen out variants with higher confidence. We have unearthed many contigs related to cancer prognosis, including unmapped contigs. These contigs cannot be mapped to human reference sequences by software such as BWA or STAR, but they present significant differences in cancer patients with diverse responses to therapies. Besides, combined with survival analysis, we found that these contigs are also significantly related to survival prognosis.

The novel events identified by DEkupl include SNVs, repeat, splice, intron and also unmapped contigs. Except for the repeat and unmapped contigs, all the other contigs can be mapped to the genome. In this way, we obtained the convincing SNVs comparing with the mapping-based approach. For the 24 unmapped contigs, they were only captured by DEkupl. Even though the source of these unmapped contigs is still not clear, these contigs were recurrently observed in multiple patients. Therefore, to some extent, these novel events complete the puzzle of cancer mechanisms and the novel events can be used as alternative indicators for diagnosis and prognosis.

Nevertheless, due to the sample size of the validation data, only 23 of the convincing SNVs were verified. Besides the sample size, another factor resulting in the low consistency is the use of two different WES measurement techniques for the TCGA cohorts and the independent validation cohorts. What's more, the independent data cohorts are Chinese patients, which is a diverse population to the TCGA cohorts. But through enrichment analysis, we found that these variants detected in large TCGA cohorts and small independent cohorts tend to have more obvious differential expressions. Therefore, these 23 SNVs are supposed to be stable OC related factors in regardless of sequencing platform and population.

In addition to clinical value, the newly discovered sources of embedded DE-kupl contigs are also crucial, especially unmappable contigs. These contigs may be derived from exogenous RNA and DNA or viral sequences embedded in the human genome.

Abbreviations

OC: Ovarian cancer

SOC: Serous ovarian cancer

DBG: De Bruijn graph

TCGA: The Cancer Genome Atlas

WES: whole exome sequencing

VCF: Variant Call Format

DEGs: differentially expressed genes

GATK: Genome Analysis Toolkit

CoxPh: Cox proportional-hazards model

SVM: support vector machine

PCA: principal component analysis

FIGO: Federation of Obstetrics and Gynaecology

GO: gene ontology

KEGG: Kyoto Encyclopedia of Genes and Genomes

SNV: single nucleotide variants

Ethics Statement

All procedures involving human participants in this study complied with the ethics standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethics standards. This study was approved by the ethics committee of the Zhejiang Cancer Hospital. Informed consent was obtained from all participants included in the study.

Author Contributions

Zhiqin Fu, Chao Lu performed the experiments. Zhiqin Fu, Chao Lu, Liang Shi, Tingting Yu analysed the data and wrote the manuscript. Chao Ding, Chengxi Zhou collected the medical records of the enrolled patients. Yue Yang and Liang Shi designed the study. All authors have read and approved the final manuscript

Funding

This study was supported by grants from the Natural Science Foundation of Zhejiang Province (Grant No. Q20H160028, LYY19H310001), Medical and Health Research Project of Zhejiang Province (2020KY059, 2020KY020, 2021KY589). The funders had no role in the study design and implementation.

Declaration of competing interest

The authors declare that they have no competing interests.

Reference

1. Kurman RJ. Origin and molecular pathogenesis of ovarian high-grade serous carcinoma, Ann Oncol. 24(10)(2013) x16–x21. doi: 10.1093/annonc/mdt463
2. Encinas G, Maistro S, Pasini FS, et al. Somatic mutations in breast and serous ovarian cancer young patients: A systematic review and meta-analysis, Revista da Associacao Medica Brasileira. 61(5)(2015)474-83. doi: 10.1590/1806-9282.61.05.474.
3. Ramalingam P. Morphologic, Immunophenotypic, and Molecular Features of Epithelial Ovarian Cancer, Oncology (Williston Park, N.Y.).30(2)(2016)166-76 .

4. Khalique S, Lord CJ, Banerjee S, et al. Translational genomics of ovarian clear cell carcinoma,Seminars in Cancer Biology. 61(2020) 121–131. doi: 10.1016/j.semcancer.2019.10.025.

5. Smith B, Agarwal P, Bhowmick NA. MicroRNA applications for prostate, ovarian and breast cancer in the era of precision medicine, Endocrine-Related Cancer.24(5)(2017)R157-R172 . doi: 10.1530/ERC-16-0525.

6. Fehrmann RSN, Li X, van der Zee AGJ,et al. Profiling Studies in Ovarian Cancer: A Review, Oncologist.12(8)(2007)960-6 . doi: 10.1634/theoncologist.12-8-960.

7. Lee J-Y, Kim HS, Suh DH, et al. Ovarian Cancer Biomarker Discovery Based on Genomic Approaches, J Cancer Prev.18(4)(2013)298-312. doi: 10.15430/jcp.2013.18.4.298.

8. Marigorta UM, Rodríguez JA, Gibson G, et al. Replicability and Prediction: Lessons and Challenges from GWAS, Trends in Genetics.34(7)(2018)504-517. doi: 10.1016/j.tig.2018.03.005.

9. Breschi A, Gingeras TR, Guigó R. Comparative transcriptomics in human and mouse, Nature Reviews Genetics. 18(7)(2017)425-440. doi: 10.1038/nrg.2017.19.

10. Oprea TI. Exploring the dark genome: implications for precision medicine, Mammalian Genome. 30(7-8)(2019)192-200. doi: 10.1007/s00335-019-09809-0.

11. Li Z, Chen Y, Mu D, et al. Comparison of the two major classes of assembly algorithms: Overlap-layout-consensus and de-bruijn-graph, Brief Funct Genomics.11(1)(2012)25-37 . doi: 10.1093/bfgp/elr035.

12. Audoux J, Philippe N, Chikhi R, et al. DE-kupl: Exhaustive capture of biological variation in RNA-seq data through k-mer decomposition, Genome Biol.18(1)(2017)243. doi: 10.1186/s13059-017-1372-2.

13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol. 15(12)(2014)550. doi: 10.1186/s13059-014-0550-8.

14. da Silveira WA, Hazard ES, Chung D, et al. Molecular profiling of RNA tumors using high-throughput RNA sequencing: From raw data to systems level analyses, Methods in Molecular Biology. 1908(2019)185-204. doi: 10.1007/978-1-4939-9004-7_13.

15. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge, Wspolczesna Onkologia. 19(1A)(2015)A68-77. doi: 10.5114/wo.2014.47136.

16. Chen C, Khaleel SS, Huang H, Wu CH. Software for pre-processing Illumina next-generation sequencing short read sequences, Source Code Biol Med. (2014)8:9. doi: 10.1186/1751-0473-9-8.

17. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes, Nucleic Acids Res. 47(D1)(2019)D766-D773 . doi: 10.1093/nar/gky955.

18. Dobin A, Davis CA, Schlesinger F, et al. STAR: Ultrafast universal RNA-seq aligner, Bioinformatics. 29(1)(2013)15-21. doi: 10.1093/bioinformatics/bts635.

19. Houtgast EJ, Sima VM, Bertels K,et al.Hardware acceleration of BWA-MEM genomic short read mapping for longer read lengths, Comput Biol Chem. (2018)75:54-64. doi: 10.1016/j.compbiolchem.2018.03.024.

20. do Valle ÍF, Giampieri E, Simonetti G, et al. Optimized pipeline of MuTect and GATK tools to improve the detection of somatic single nucleotide polymorphisms in whole-exome sequencing data, BMC Bioinformatics. 17(Suppl 12)(2016)341 . doi: 10.1186/s12859-016-1190-7.

21. Du Y, Huang Q, Arisdakessian C, et al. Evaluation of STAR and kallisto on single cell RNA-seq data alignment. G3 Genes, Genomes, Genet. 10(5)(2020)1775-1783. doi: 10.1534/g3.120.401160.

22. Wu TD, Reeder J, Lawrence M, et al. GMAP and GSNAP for genomic sequence alignment: Enhancements to speed, accuracy, and functionality, Methods in Molecular Biology. (2016)1418:283-334. doi: 10.1007/978-1-4939-3578-9_15.

23. Kleinbaum DG, Klein M. Kaplan-Meier Survival Curves and the Log-Rank Test(Third edn),. Springer.2012.

24. Rich JT, Neely JG, Paniello RC, et al. A practical guide to understanding Kaplan-Meier curves, Otolaryngol - Head Neck Surg. 143(3)(2010)331-6. doi: 10.1016/j.otohns.2010.05.007.

25. Fisher LD, Lin DY. Time-dependent covariates in the cox proportional-hazards regression model, Annual Review of Public Health. 20(1990)145-57. doi: 10.1146/annurev.publhealth.20.1.145.

26. Hans C. Bayesian lasso regression, Biometrika. (2009)1-11. http://dx.doi.org/10.1093/biomet/asp047

27. Chang CC, Lin CJ. LIBSVM: A Library for support vector machines, ACM Trans Intell Syst Technol. 2(3)(2011) 27. doi: 10.1145/1961189.1961199.

28. Astorino A, Fuduli A. The Proximal Trajectory Algorithm in SVM Cross Validation, IEEE Trans Neural Networks Learn Syst. 27(5)(2016)966-77. doi: 10.1109/TNNLS.2015.2430935.

29. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation, Caspian Journal of Internal Medicine. 4(2)(2013)627-35.

30. Kiers HAL, Van Mechelen I. Three-way component analysis: Principles and illustrative application, Psychol Methods. 6(1)(2001)84-110. doi: 10.1037/1082-989x.6.1.84.

31. Xing Z, Chu C, Chen L, et al. The use of Gene Ontology terms and KEGG pathways for analysis and prediction of oncogenes, Biochim Biophys Acta - Gen Subj. 1860(11 Pt B)(2016)2725-34 . doi: 10.1016/j.bbagen.2016.01.012.

32. Yu G, Wang LG, Han Y, et al. ClusterProfiler: An R package for comparing biological themes among gene clusters, Omi A J Integr Biol. 16(5)(2012)284-7. doi: 10.1089/omi.2011.0118.

33. Liming Zhang , Qian Lei , Hongxiang Wang, et al.Tumor-derived extracellular vesicles inhibit osteogenesis and exacerbate myeloma bone disease, Theranostics. 9(1)(2019)196-209. doi: 10.7150/thno.27550.

34. Stewart TA, Yapa KTDS, Monteith GR. Altered calcium signaling in cancer cells, Biochim Biophys Acta. 1848(10 Pt B)(2015)2502-11. doi: 10.1016/j.bbamem.2014.08.016.

35. Akle S, Chun S, Jordan DM, Cassa CA. Mitigating False-Positive Associations in Rare Disease Gene Discovery, Hum Mutat. 36(10)(2015)998-1003. doi: 10.1002/humu.22847.

36. Barnoud T, Parris JLD, Murphy ME. Common genetic variants in the TP53 pathway and their impact on cancer, Journal of Molecular Cell Biology. 11(7)(2019)578-585. doi: 10.1093/jmcb/mjz052.

37. Mayakonda A, Lin DC, Assenov Y,et al. Maftools: Efficient and comprehensive analysis of somatic variants in cancer, Genome Res. 28(11)(2018)1747-1756. doi: 10.1101/gr.239244.118.

38. Matthias Scheffler, Michaela A Ihle, Rebecca Hein, et al. K-ras Mutation Subtypes in NSCLC and Associated Co-occuring Mutations in Other Oncogenic Pathways, J Thorac Oncol. 14(4)(2019)606-616 . doi: 10.1016/j.jtho.2018.12.013.

39. Cisowski J, Bergo MO. What makes oncogenes mutually exclusive? Small GTPases. 8(3)(2017)187-192. doi: 10.1080/21541248.2016.1212689.

40. O’Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer, Nature Reviews Genetics. 18(10)(2017)613-623. doi: 10.1038/nrg.2017.47.

41. Lash LH, Parker JC, Scott CS. Modes of action of trichloroethylene for kidney tumorigenesis, Environ Health Perspect. 2(Suppl 2)(2000)225-40. doi: 10.1289/ehp.00108s2225.