Identification of Tumor Microenvironment-Related Prognostic Biomarkers for Ovarian Serous Cancer 3-Year Mortality Using Targeted Maximum Likelihood Estimation: A TCGA Data Mining Study

Lu Wang[1,2], Xiaoru Sun[1,2], Chuandi Jin[1,2], Yue Fan[3] and Fuzhong Xue[1,2]*

[1] Institute for Medical Dataology, Cheeloo College of Medicine, Shandong University, Jinan, China, [2] Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China, [3] Key Laboratory of Trace Elements and Endemic Diseases, National Health and Family Planning Commission, School of Public Health, Xi'an Jiaotong University, Xi'an, China

Ovarian serous cancer (OSC) is one of the leading causes of death across the world. The role of the tumor microenvironment (TME) in OSC has received increasing attention. Targeted maximum likelihood estimation (TMLE) is developed under a counterfactual framework to produce effect estimation for both the population level and individual level. In this study, we aim to identify TME-related genes and using the TMLE method to estimate their effects on the 3-year mortality of OSC. In total, 285 OSC patients from the TCGA database constituted the studying population. ESTIMATE algorithm was implemented to evaluate immune and stromal components in TME. Differential analysis between high-score and low-score groups regarding ImmuneScore and StromalScore was performed to select shared differential expressed genes (DEGs). Univariate logistic regression analysis was followed to evaluate associations between DEGs and clinical pathologic factors with 3-year mortality. TMLE analysis was conducted to estimate the average effect (AE), individual effect (IE), and marginal odds ratio (MOR). The validation was performed using three datasets from Gene Expression Omnibus (GEO) database. Additionally, 355 DEGs were selected after differential analysis, and 12 genes from DEGs were significant after univariate logistic regression. Four genes remained significant after TMLE analysis. In specific, ARID3C and FREM2 were negatively correlated with OSC 3-year mortality. CROCC2 and PTF1A were positively correlated with OSC 3-year mortality. Combining of ESTIMATE algorithm and TMLE algorithm, we identified four TME-related genes in OSC. AEs were estimated to provide averaged effects based on the population level, while IEs were estimated to provide individualized effects and may be helpful for precision medicine.

Keywords: targeted maximum likelihood estimation, ovarian cancer, tumor microenvironment, gene expression, prognostic biomarker

INTRODUCTION

Ovarian cancer is one of the most common cancers and the leading cause of death of all gynecological cancers among women across the world (Bray et al., 2018). Ovarian serous cancer (OSC) is the most common histologic subtype, and it accounts for about 90% of all ovarian tumors (Bell et al., 2011). In the United States, approximately 1 in 78 women will develop ovarian cancer in their lifetime while the all-stage 5-year survival rate of ovarian cancers is only 47% (Torre et al., 2018). Over 70% of ovarian cancers are diagnosed at advanced stages (stage III or IV) and usually exhibit poor prognosis with a 5-year overall survival rate of ~30% (Howlader et al., 2016; Torre et al., 2018). The high mortality may partly due to the non-specific symptoms, lack of specific screening tools, and non-specific clinical drugs (Diaz-Padilla et al., 2012). Therefore, the identification of biomarkers with a poor short time prognosis remains significant. In this study, we focused on OSC 3-year mortality to evaluate short survival time-related biomarkers.

The role of the tumor microenvironment (TME) in OSC has been discussed increasing in recent years. TME refers to the constitution of a complex network that includes different subsets of stromal cells, immune cells, fibroblasts, blood vessels, extracellular matrix (ECM), endothelial cell precursors, and secreted factors (Kenny et al., 2007). It is now established that TME plays an important role in cancer progression, metastasis, and resistance to the therapies through complex interactions (Luo et al., 2016; Sung et al., 2016; Wang et al., 2017; Finkernagel et al., 2019). In particular, a positive correlation was observed between cancer-associated fibroblasts (CAFs) and the clinical stage. The co-culture of CAFs with ovarian cancer cells further showed stimulation of cancer cell invasion and migration (Zhang et al., 2011). Moreover, biomarkers like CD163, CD206/CD68 ratio, B7-H4, and IL-10 produced by tumor-associated macrophages (TAMs) were associated with survival and clinical outcome of ovarian cancer patients (Bartlett et al., 2001; Lan et al., 2013). However, it may still lack a full understanding of the molecular interactions between immune and stromal components in TME and how they affect ovarian cancer progression. Therefore, a further and deeper understanding of TME is urgently needed to better understand how their stromal and immune components affect OSC survival and clinical outcome.

To identify such relationships, classic epidemiologic analysis such as the regression method has discovered numerous biomarkers. However, those associations can be biased by confounding factors or misspecification of a parametric outcome model. Targeted maximum likelihood estimation (TMLE) is an efficient, double robust, semi-parametric methodology that has been proposed for estimating marginal effects (van der Laan and Rubin, 2006; van der Laan et al., 2009). TMLE is developed under the causal inference framework and allows researchers to estimate the average effect (AE) for binary treatment (Rubin, 1974;

Greenland and Robins, 1986; Luque-Fernandez et al., 2018). Though TMLE usually shows advantages over the commonly used propensity score method or G-computation with both point and interval estimation, it has not been widely implemented in epidemiologic research including biomarkers discovering and omics data mining.

In this study, we aimed to discover TME associated biomarkers that have effects on the 3-year mortality of OSC patients. OSC patients were collected from The Cancer Genome Atlas (TCGA) database with RNA-sequencing data employed for gene expression data. We present our study using the ESTIMATE algorithm to access the immune and stromal components level in ovarian tumor samples. Differential expression genes possibly representing TME status were selected. Univariate analysis followed by covariates selection analysis was performed to determine candidate prognostic biomarkers and their confounding factors sets. The TMLE method was conducted to evaluate average and individual effects, respectively. Four genes, ARID3C, CROCC2, FREM2, and PTF1A, were successfully selected as potential OSC 3-year mortality-related prognostic biomarkers.

RESULTS

Clinical Characteristics of the Study Patients

Figure 1 demonstrates the flowchart for the identification of prognostic biomarkers. 285 OSC patients from the TCGA database were included for statistical analysis. The detailed clinical characteristics of the studying population were summarized in **Table 1**.

DEGs Shared by ImmuneScore and StromalScore in OSC

By comparing the gene expression profiles of patients with high immune scores against those with low immune scores, a total of 894 (580 upregulated and 314 downregulated) DEGs were identified (**Figures 2A,C,D**). A total of 723 (467 upregulated and 256 downregulated) DEGs were identified by comparing the high and low stromal score groups (**Figures 2B–D**). |Log2 (fold-change)| > 1.5 and FDR < 0.05 were used as criterions for screening DEGs. A total of 230 DEGs were in common among the high immune/stromal score groups. A total of 125 DEGs were in common among the low immune/stromal score groups. These DEGs (total 355 genes) were possibly determinate factors for the status of TME. Results from gene ontology (GO) enrichment analysis indicated that the DEGs were mainly enriched for the immune-related GO terms, such as chemokine signaling pathway and immunoglobulin binding (**Figure 2E**). Based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis, chemokine signaling pathway, cytokine–cytokine receptor interaction, and hematopoietic cell lineage were significantly enriched pathways (**Figure 2F**). In all, these DEGs mainly mapped to immune-related functions and

---

Abbreviation: TMLE, targeted maximum likelihood estimation; TME, tumor microenvironment; OSC, ovarian serous cancer; AE, average effect; IE, individual effect; ECM, extracellular matrix; TAMs, tumor-associated macrophages; DEG, differentially expressed gene; GEO, Gene Expression Omnibus; ROC, receiver operating characteristic; AUC, area under the curve.

FIGURE 1 | The flowchart for identification of prognostic biomarkers.
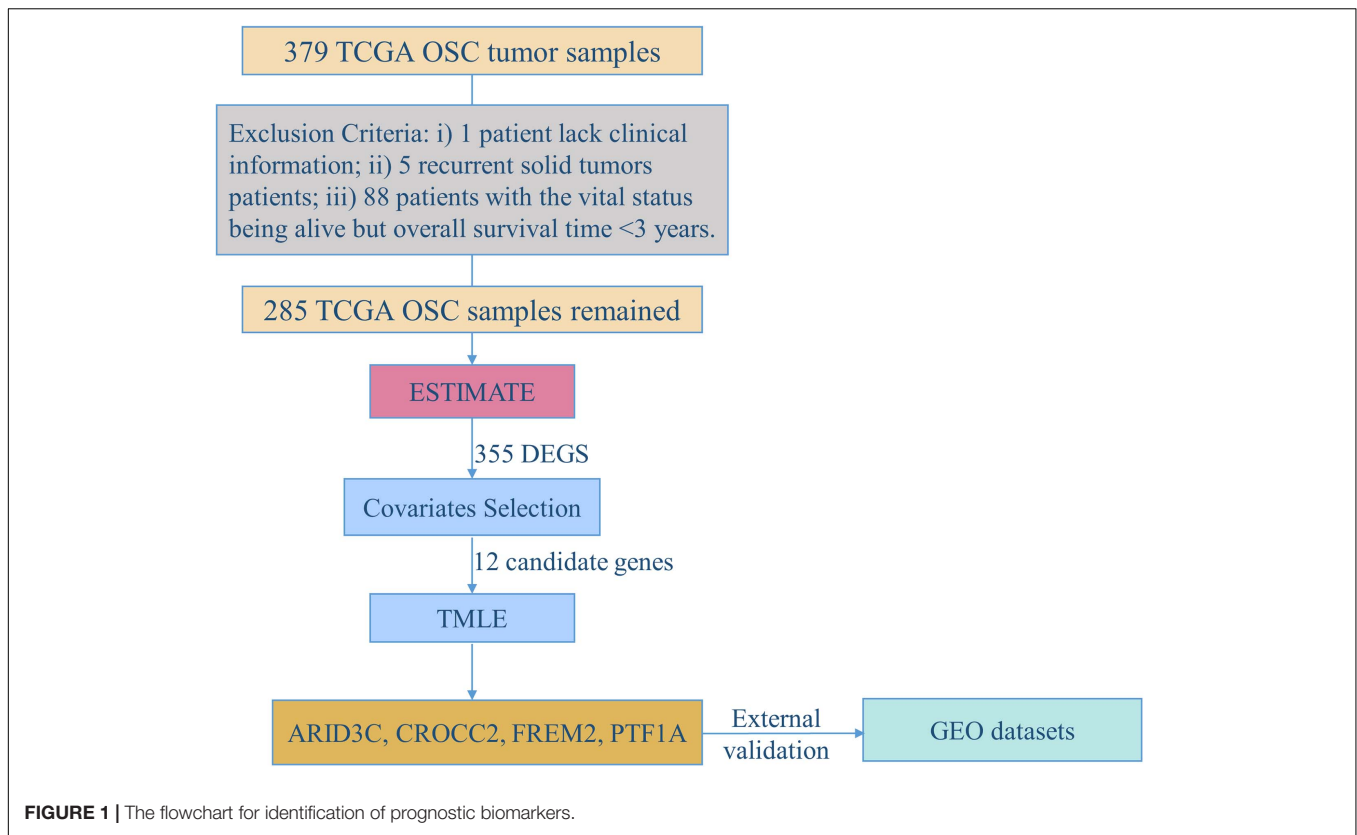
therefore indicate that immune components as the leading feature of TME in OSC.

Selection of Candidate Potential Prognostic Biomarkers and Confounding Factors

In the process of screening for prognostic-related biomarkers, the 355 DEGs shared by the high immune/stromal score and low immune/stromal score groups together with clinical variables were subjected to a univariate analysis. Out of these, 12 DEGs, age, and tumor residual disease were found to be significantly (with p-value < 0.05) correlated with the 3-year mortality of OSC patients from the TCGA database. The detailed information regarding the associations of univariate analysis is provided in **Table 2**. To determine the proper covariates set to control in the following TMLE analysis, each one of 12 candidate markers was subsequently undergoing a covariates selection procedure. **Supplementary Figure 1** shows the covariates set and their correlations for 12 candidate potential prognostic biomarkers. The number of confounding factors for each gene varied from 7 to 13, with FREM2 had maximum covariates set while SOHLH1 owning minimum covariates set. Age was the common confounding factor shared by all 12 candidate genes, indicating the essential role of age as a confounder. Besides, FREM2 had maximum confounders as well as the common confounder shared by the other 11 candidate genes, indicating that FREM2 may be a hub gene in the OSC prognosis regulation network.

TMLE Estimation

After TMLE analysis, the association of four genes (ARID3C, CROCC2, FREM2, and PTF1A) with OSC 3-year mortality remained significant ($p < 0.05$). Their marginal odds ratio (MOR), AE, and 95% CI estimates are shown in **Table 3**. TMLE analysis suggesting positive associations between a high expression level of ARID3C and FREM2 with promising prognosis and negative associations for CROCC2 and PTF1A. The Violin plot of individual effect (IE) estimates for all candidate genes was shown in **Figure 3**. The meaning of IE is different from AE for the former one representing individualized effect estimates and the minority may have opposing effects compare with the average effect. Those inconsistent was observed in the IE of FREM2, indicating that the effect of this gene may differ in different population subsets. Compared with AE, IE may be more meaningful in terms of personalized treatment for precision medicine.

Validation in the Gene Expression Omnibus (GEO) Dataset

To further validate the expression of four TMLE significant genes (ARID3C, CROCC2, FREM2, and PTF1A), we conducted external validation using Gene Expression Omnibus (GEO) database. Due to the fact that GEO datasets contain limited gene expression profiles, three GEO datasets (GSE53963, GSE26193, and GSE13876) were obtained and successfully validated the expression of three genes (PTF1A, FREM2, and CROCC)

TABLE 1 | Description of clinic pathologic factors for 285 ovarian cancer patients from TCGA.

Clinic pathologic variable*	Dead (N = 115)	Alive (N = 170)
Age (years)	63.89 (11.26)	59.03 (10.2)
FIGO Stage		
I	1 (0.87)	0 (0.00)
II	2 (1.74)	10 (5.88)
III	88 (76.52)	135 (79.41)
IV	23 (20.00)	25 (14.71)
Unknown	1 (0.87)	0 (0.00)
Grade		
GB	1 (0.87)	1 (0.59)
G2	8 (6.96)	25 (14.71)
G3	104 (90.43)	140 (82.35)
G4	0 (0.00)	1 (0.59)
Unknown	2 (1.74)	3 (1.76)
Venous invasion		
NO	9 (7.83)	17 (10.00)
YES	17 (14.78)	24 (14.12)
Unknown	89 (77.39)	129 (75.88)
Lymphatic invasion		
NO	9 (7.83)	19 (11.18)
YES	32 (27.83)	33 (19.41)
Unknown	74 (64.35)	118 (69.41)
Tumor residual disease		
No macroscopic disease	11 (9.57)	27 (15.88)
1–10 mm	53 (46.09)	85 (50.00)
11–20 mm	8 (6.96)	12 (7.06)
>20 mm	32 (27.83)	25 (14.71)
Unknown	11 (9.57)	21 (12.35)

Data are mean ± SD and frequency (percent) for numeric and category variables, respectively.

separately (**Supplementary Table 1**). The expression of two genes (CROCC and PTF1A) were upregulated while FREM2 was downregulated in OSC patients with poor prognosis compared to promising prognosis. These results were largely consistent with our results in TCGA data. The sensitivity and specificity of each verified gene were evaluated. The sensitivity and specificity of PTF1A calculated from the GSE53963 dataset were 0.74 and 0.56, respectively. The sensitivity and specificity of FREM2 calculated from the GSE26193 dataset were 0.61 and 0.62, respectively. The sensitivity and specificity of CROCC calculated from the GSE13876 dataset were 0.59 and 0.63, respectively. The receiver operating characteristic (ROC) curve analyses and the area under the curve (AUC) were used to assess the discriminatory ability of four genes (ARID3C, CROCC, FREM2, and PTF1A) among 67 OSC patients with promising prognosis and 90 patients with poor prognosis derived from GSE13876 dataset. The AUC of all four genes was 0.763 (**Figure 4**). The specificity and sensitivity were 0.746 and 0.689, respectively.
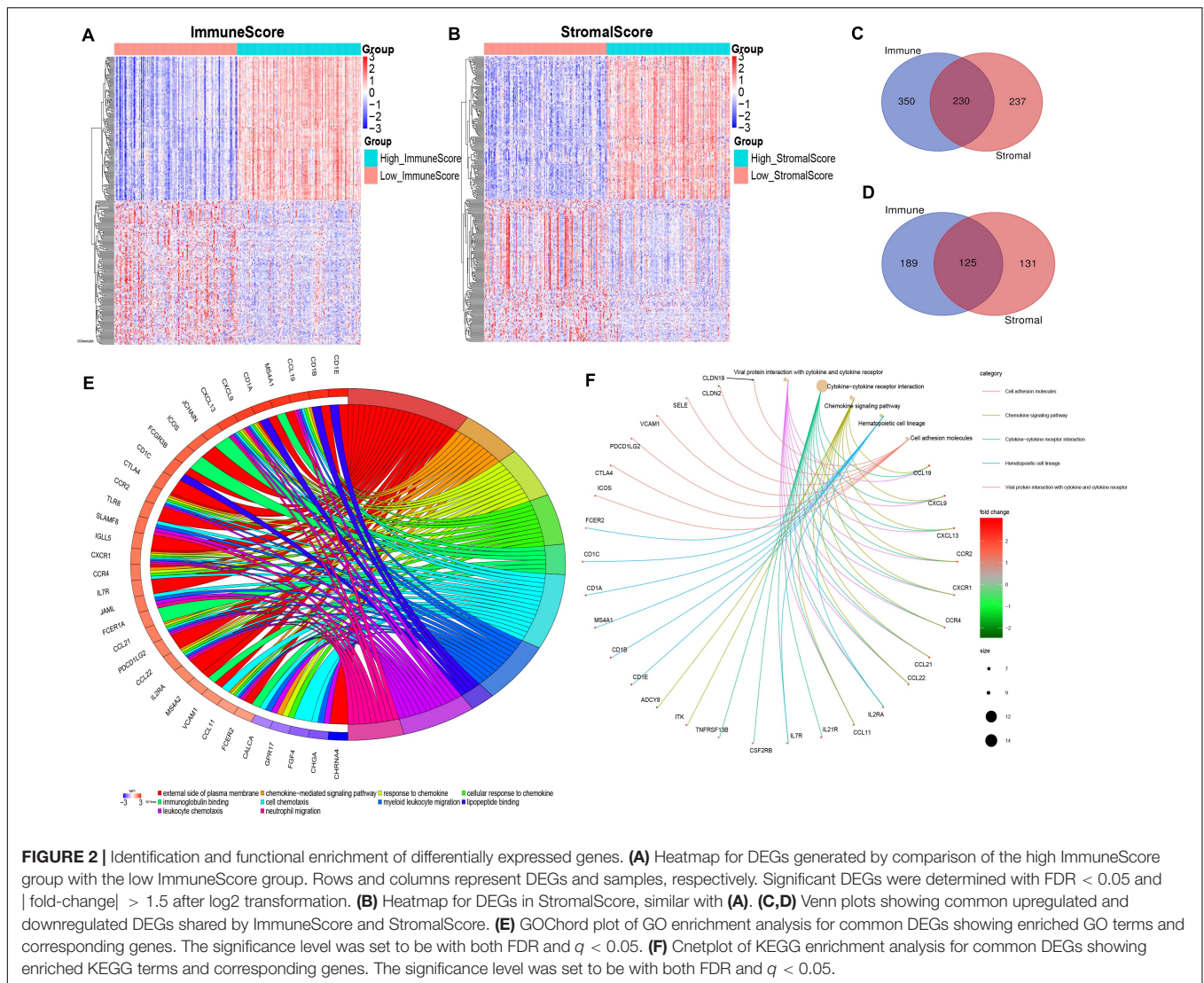
DISCUSSION

In the presented study, we attempted to identify TME-related genes that have an effect on the 3-year mortality of OSC patients

from the TCGA database. TMLE algorithm was performed to derive AE, IE, and MOR with doubly robust to model misspecification. ARID3C, CROCC2, FREM2, and PTF1A were identified as prognostic biomarkers for OSC patients. Two of them (FREM2 and PTF1A), alongside CROCC, were successfully validated in three GEO datasets. In addition to averaged effects estimation, individualized effects were estimated for each patient. Functional analysis indicated that ARID3C, CROCC2, and FREM2 might be involved in the immune status of TME.

In our research, increased expression of AT-rich interaction domain 3C (ARID3C) was found to be significantly associated with better OSC prognosis. ARID3C was recently characterized and belongs to the AT-rich interaction domain (ARID) family of proteins (Tidwell et al., 2011). The ARID3 subfamily consists of 3 members ARID3A, B, and C, which are expressed throughout most of hematopoietic development (Heng et al., 2008). The expression of ARID3A, the founding member of the subfamily, is tightly regulated during B cell differentiation (Webb et al., 1998; Heng et al., 2008). A recent study demonstrated that ARID3B also plays an important role in normal B cell development (Kurkewich et al., 2016). Similarly, ARID3C transcripts and proteins are expressed especially within B lineage lymphocytes and significantly co-activates ARID3A-dependent immunoglobulin heavy chain gene (IgH) transcription (Tidwell et al., 2011). In recent research, high IgH mRNA levels were reported positively associated with improved prognosis of breast cancer (Larsson et al., 2020). Indicating that ARID3C could modulate IgH activity to enhance immune response and improve the prognosis of cancers. The function of B lymphocytes in tumor development has contradictory opinions. Their presence has been found to be associated with a good prognosis in ovarian cancer through the priming of CD4+ and CD8+ T cells and the production of tumor-specific IgGs (Stumpf et al., 2009; Nielsen et al., 2012). Conversely, high B cell infiltration negatively correlates with patient survival in ovarian cancer (Gentles et al., 2015). In addition, ARID3A and ARID3B, two closely related paralogs of ARID3C, are reported to be important associated with tumorigenesis (Samyesudhas et al., 2014; Chien et al., 2015; Nakahara et al., 2017; Tu et al., 2019; Dausinas et al., 2020).

The FRAS1-related extracellular matrix 2 (FREM2) is an extracellular matrix protein that localizes in the lamina densa of epithelial basement membranes (Kiyozumi et al., 2006; Chiotaki et al., 2007). In the extracellular matrix (ECM), FREM2, and FRAS1 form a self-stabilizing complex with FREM1, which plays an important role in cell adhesion and intercellular signaling (Kiyozumi et al., 2006). ECM, as important content of TME, provides structural support as well as biochemical and biomechanical cues for cancer cell growth (De Palma et al., 2017). Previous studies have reported that mutations in this gene are associated with Fraser syndrome and play an important role in various tumorigenesis. In glioblastomas, FREM2 expression is positively associated with a favorable prognosis of IDH-WT glioblastomas, which is consistent with our research result (Jovčevska et al., 2019). In ovarian cancer, the differential expression of FREM2 has been reported in a few studies (Dakubo, 2019; Zhu et al., 2020). FREM2 has been identified as a target gene of transcription factor AP-2 gamma (TFAP2C, AP-2γ)

FIGURE 2 | Identification and functional enrichment of differentially expressed genes. **(A)** Heatmap for DEGs generated by comparison of the high ImmuneScore group with the low ImmuneScore group. Rows and columns represent DEGs and samples, respectively. Significant DEGs were determined with FDR < 0.05 and | fold-change| > 1.5 after log2 transformation. **(B)** Heatmap for DEGs in StromalScore, similar with **(A)**. **(C,D)** Venn plots showing common upregulated and downregulated DEGs shared by ImmuneScore and StromalScore. **(E)** GOChord plot of GO enrichment analysis for common DEGs showing enriched GO terms and corresponding genes. The significance level was set to be with both FDR and q < 0.05. **(F)** Cnetplot of KEGG enrichment analysis for common DEGs showing enriched KEGG terms and corresponding genes. The significance level was set to be with both FDR and q < 0.05.

(Woodfield et al., 2010), which is a sequence-specific DNA-binding transcription factor and belongs to the AP-2 family. AP-2γ has been established as a prognostic factor in human breast cancer (Cyr et al., 2015). Its paralog, AP-2α, has been observed to play a tumor-suppressive role in ovarian cancer (Sumigama et al., 2004).

Pancreas-associated transcription factor 1a (PTF1A) plays a critical role in controlling the development and physiological function of many organs, including the pancreas, brain, spinal cord, retina, and others (Jin and Xiang, 2019). Although there are limited studies about PTF1A outside pancreas, PTF1A protein abundance is highest in the ovary at the tissue level, indicating that it may have an important role in the ovary (Jin and Xiang, 2019). Furthermore, PTF1A is involved in the Notch-mediated HES/HEY network, which has been found by many studies being activated in ovarian cancer (Bell et al., 2011; Bocchicchio et al., 2019; Huang et al., 2019; Hubbard et al., 2019; Eoh et al., 2020). The Notch signaling has been found to be able to ensure the proliferation and migration of ovarian cancer cells and therefore

controls ovarian cancer cell survival (Bocchicchio et al., 2019). In addition, we found that one study showed that T follicular helper (Tfh) cell generation and function are reliant on Notch signaling (Dell'aringa and Lee Reinhardt, 2018).

CROCC2 is a newly annotated protein in Gascoigne et al. (2012) that has similarity to CROCC and belongs to the ciliary rootlet coiled-coil rootletin family. Rootletin is a structural component of the ciliary rootlet (Yang et al., 2002) and primary cilia that has been suggested to have an important role in carcinogenesis through facilitating cell signaling pathways platelet-derived growth factor receptor-a and Wnt pathways that are commonly dysregulated in many tumors (Yasar et al., 2017). Our study suggesting a positive effect between high expression of CROCC2 with poor 3-year OSC mortality. Further studies should focus on the biological function of CROCC2 on the prognosis of ovarian cancer.

To our knowledge, our work is the first to use the ESTIMATE algorithm combined with TMLE methodology to explore molecular markers that have effects on OSC 3-year prognosis.

TABLE 2 | Univariate logistic regression showing significant associations with 3-year mortality of OSC.

Variable*	Levels	OR/ Coefficient	95% CI	P-value
IGLV5-37	High expression level	1.82	(1.13; 2.93)	0.014
IGHJ3	High expression level	1.63	(1.01; 2.63)	0.046
IGHV1OR15-1	High expression level	1.82	(1.05; 3.15)	0.033
PTF1A	High expression level	1.73	(1.06; 2.84)	0.029
ARID3C	High expression level	0.59	(0.37; 0.96)	0.032
FREM2	High expression level	0.61	(0.38; 0.99)	0.046
CLDN19	High expression level	1.65	(1.02; 2.65)	0.040
CROCC2	High expression level	1.65	(1.02; 2.65)	0.040
SPAG6	High expression level	1.73	(1.07; 2.79)	0.025
MAGEA9B	High expression level	1.97	(1.13; 3.41)	0.016
ONECUT1	High expression level	1.71	(1.01; 2.91)	0.046
SOHLH1	High expression level	1.67	(1.03; 2.69)	0.036
Tumor residual disease	No macroscopic disease	Ref	–	–
	1–10 mm	1.53	(0.70; 3.34)	0.285
	11–20 mm	1.64	(0.53; 5.10)	0.396
	>20 mm	3.14	(1.31; 7.54)	0.010
	Unknown	1.29	(0.47; 3.54)	0.626
Age	year	0.01	(0.00; 0.01)	<0.001

Data are odds ratio and 95% confidence interval for category variables as well as coefficient (β) and 95% confidence interval for the numeric variable (Age).

The covariate selection procedure makes sure appropriate confounding factors are selected to reduce bias or loss of power. The prognostic biomarkers were driven from a causal inference framework-based TMLE algorithm. Such methodologies can be used to better inform future clinical therapy. Individual effect estimates were provided to display personal differences and may be helpful for precision medicine. There are also limitations to this study. To gain explainable effect estimation, we dichotomized the exposure (gene expression) and define the 3-year mortality of OSC instead of overall survival as the outcome variable. Therefore, we lost 88 samples due to the missing outcome variable. Further studies can be focused on the biological function role of identified genes in the network of TME.

MATERIALS AND METHODS

Data Source and Outcome Definition

Transcriptome RNA-seq read counts data of 379 OSC samples and the corresponding clinical data were downloaded from the TCGA database[1] using the R package "TCGAbiolinks" (Colaprico et al., 2016). The mRNA data was already processed using the GDC mRNA quantification analysis pipeline, which measures gene-level expression in HT-Seq raw read count. These values are generated through this pipeline by first aligning reads to the GRCh38 reference genome and then by quantifying the mapped reads. More information on the GDC pipeline can be found at: https://docs.gdc.cancer.gov/Data/Bioinformatics_ Pipelines/Expression_mRNA_Pipeline/. Gene mapping was performed using the gencode v22 annotation file. The samples were trimmed to 378 to include only patients with corresponding clinical information. The data from 373 primary solid tumors patients were further retained after removing five recurrent solid tumor samples. Finally, 88 samples were excluded with the vital status being alive but overall survival time <3 years. In total, 285 OSC patients were kept after quality control and were included in the following analysis. The outcome binary was defined as 3-year mortality of OSC, alive status was defined as patients with OS time ≥ 3 years, and dead status was defined as patients with the vital status being dead and OS time <3 years.

TME Construction

ESTIMATE algorithm was performed to estimate immune and stromal components in TME for each sample using the R package "estimate" (Yoshihara et al., 2013). The ImmuneScore, StromalScore, and ESTIMATEScore were calculated corresponding to the level of immune cells, stromal cells, and the sum of both, respectively. A higher ESTIMATEScore, StromalScore, and ImmuneScore, respectively, represent the lower tumor purity and higher infiltration levels of stromal and immune cells in tumor tissue.
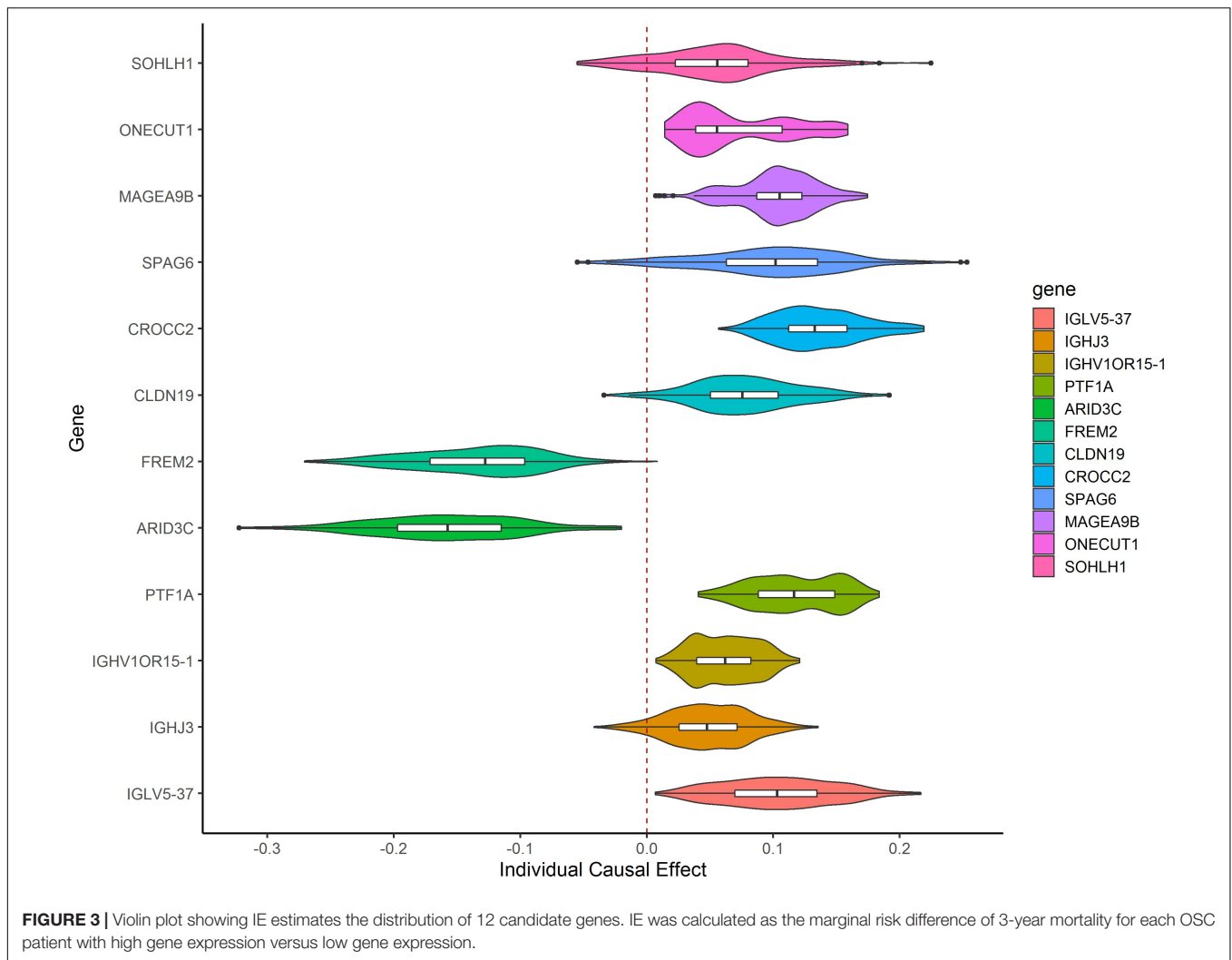
Differentiation Analysis Between High-Score and Low-Score Groups Regarding ImmuneScore and StromalScore

To better understand the correlation between gene expression profiles and immune and/or stromal scores, patients were divided into two groups based on the median value of the ImmuneScore and StromalScore, respectively. R package "DESeq2" was used to perform differentiation analysis. Genes where all counts were less than half sample numbers were filtered. The differentiation analysis was conducted in the high immune score group vs. the low immune score group and the high stromal score group vs. the low stromal score group, separately. For each DE analysis, we

---

[1]https://portal.gdc.cancer.gov/

TABLE 3 | Four genes shown significant association after TMLE analysis.

Gene	Chromosome	MOR	95% CI	P-MOR	AE	95% CI	P-AE
ARID3C	9	0.51	(0.32, 0.82)	0.005	−0.16	(−0.27, −0.05)	0.004
CROCC2	2	1.76	(1.12, 2.78)	0.015	0.14	(0.03, 0.24)	0.014
FREM2	13	0.57	(0.36, 0.90)	0.016	−0.14	(−0.24, −0.03)	0.015
PTF1A	10	1.63	(1.02, 2.60)	0.042	0.12	(0.00, 0.23)	0.043

FIGURE 3 | Violin plot showing IE estimates the distribution of 12 candidate genes. IE was calculated as the marginal risk difference of 3-year mortality for each OSC patient with high gene expression versus low gene expression.

specified | log2FC(fold-change)| > 1.5 (high immune/stromal score group vs. low immune/stromal score group) and false discovery rate (FDR) <0.05 as cutoffs to identify significant DEGs. The qualified DEGs that both upregulated immune and stromal scores or downregulated immune and stromal scores were finally selected for further analysis.

GO and KEGG Enrichment Analysis

For shared DEGs between high ImmuneScore and low ImmuneScore as well as high StromalScore and low StromalScore, GO and KEGG pathway enrichment analyses were conducted to investigate the shared biological function by using R packages "clusterProfiler," "enrichplot," "GOplot," and "ggplot2." Only terms with both FDR and *q*-value of <0.05 were considered significantly enriched.

Univariate Analysis of Associated Genes and Clinical Pathologic Factors

In the presented study, all OSC samples were grouped into a high-expression group and low-expression group compared

with the corresponding median expression level for each DEG. Univariate analysis of the association of binary DEGs and other clinical pathologic factors with 3-year mortality was evaluated using logistic regression. Statistical significance was defined as *p* < 0.05. For factors significantly associated with 3-year mortality in univariate analyses, a covariates selection procedure was conducted.

Selection of the Minimal Sets of Confounding Covariates

It is crucial to select a proper set of covariates for estimating unbiased effects between a treatment effect and an outcome variable in observational studies. Including unnecessary covariates may result in a loss of power or biased variance while lack of covariates may lead to unadjusted confounding effects. De Luna, Waernbaum, and Richardson (Luna et al., 2011) proposed a data-driven algorithm for the selection of minimal sets of covariates. Let $G = \{G_1, G_2, \ldots, G_p\}$ denote p binary candidate potential prognostic biomarkers, $X = \{X_1, X_2, \ldots, X_m\}$ denote m selected clinical covariates and Y denote OSC 3-year mortality.
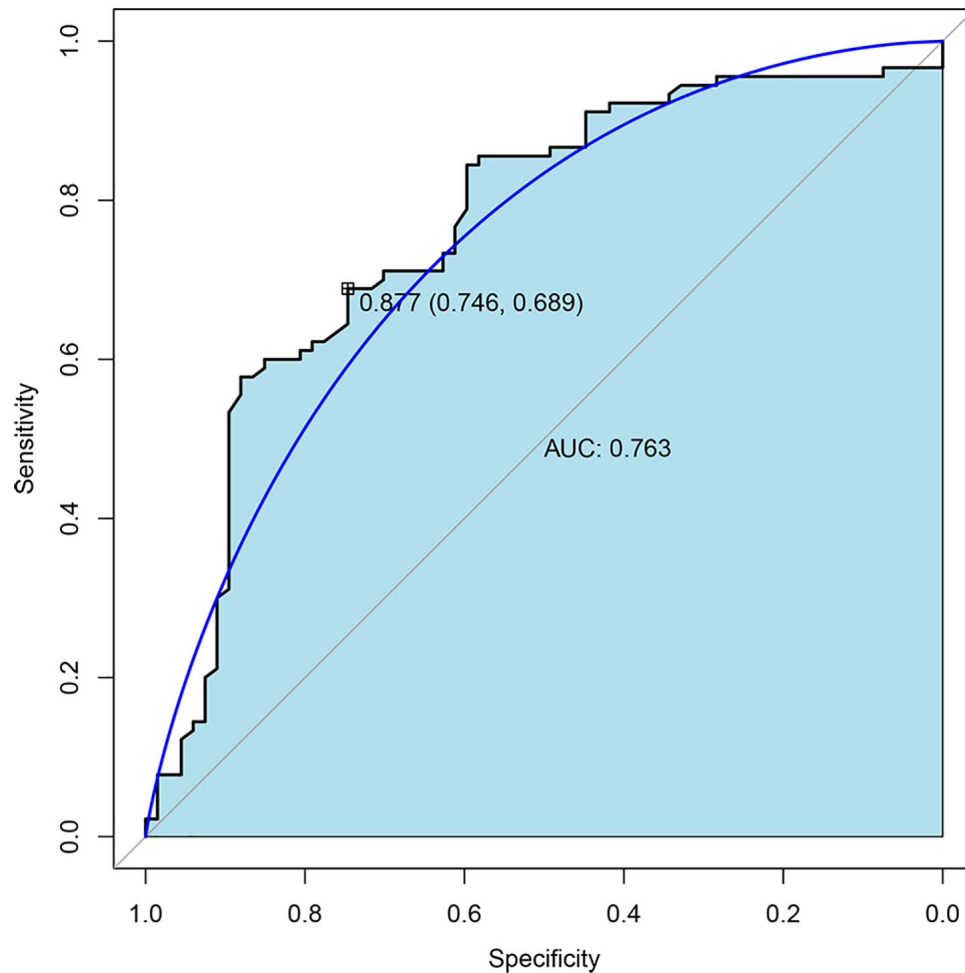
FIGURE 4 | Receiver operating characteristic (ROC) curves and area under the curve (AUC) of the combination of four gene biomarkers for 3-year mortality prediction of OSC patients in GEO dataset.

Then $C = \{(G\backslash G_j)+X\}$ denotes the complete covariate set for gene G_j and Y. We aimed to determine a minimum of confounding covariates of set W that satisfy 1) $Y \perp C| W$ and 2) $G_j \perp C| W$. To do so, we first found V to be a minimal set of C, making Y and the covariates not included in V conditionally independent, $Y \perp C\backslash V| V$. Then, W is determined as the minimal set of V, which means the G_j and the covariates are not included in the W that is conditionally independent: $G_j \perp V\backslash W| W$. R package *CovSel* was performed to select the minimal covariate sets V for each G_j.

TMLE Estimation Method

In this study, we aimed to estimate the effects of gene expression levels on OSC 3-year mortality. We are interested in three estimates of the effects: the AE, interpreted as the marginal risk difference of 3-year mortality for OSC patients with high gene expression versus low gene expression compared to median expression level; the IE, defined as the marginal risk difference of 3-year mortality for each OSC patient with high gene expression

versus low gene expression; and the MOR, interpreted as the odds ratio of death for patients with high gene expression versus low gene expression. Effects formalized by Rubin (1974) were developed under the potential outcome framework, and the casual parameters are therefore defined as the following:

$$AE = E\left[Y\left(1\right) - Y\left(0\right)\right] \tag{1}$$

$$IE = Y(1) - Y(0) \tag{2}$$

$$MOR = \{E[Y(1)] \times E[1 - Y(0)]\}/\{E[1 - Y(1)] \times E[Y(0)]\} \tag{3}$$

where $Y(1)$ represents OSC 3-year potential outcome they would have received had they been exposed to high gene expression level compared with median expression level ($G = 1$); and $Y(0)$ denotes OSC 3-year potential outcome had they been unexposed. In our study, OSC 3-year outcome Y has two possible values: 1 denotes death while 0 denotes alive.

Identification of the Effect Estimates

For the effect estimates to have a theoretical interpretation, several key assumptions are required: (1) there is the stable unit treatment value assumption (SUTVA), which assumes that the gene expression level of a given individual does not affect the potential OSC 3-year mortality of any other individuals (i.e., non-interference) and that the exposure level is the same for all individuals who were exposed at that level; (2) there are no unmeasured confounders, and this is formalized as $(Y(1), Y(0)) \perp G | W$, meaning that the gene expression level and potential mortality outcomes are independent after conditioning on the set of covariates (this assumption cannot be tested using the observed data); and (3) there is positivity, which requires that within strata of G, every individual has a non-zero probability of receiving either exposure condition; this is formalized as $0 < P(G = 1| W) < 1$ for a binary exposure. If the positivity assumption is violated, effects will not be identifiable.

Under these assumptions, the effect estimates from Eqn. (1–3) can be expressed as follows:

$$AE = E_W \left[E\left(Y | G = 1, W\right) - E\left(Y | G = 0, W\right)\right] \quad (4)$$

$$IE_i = E\left(Y | G = 1, W_i\right) - E\left(Y | G = 0, W_i\right) \quad (5)$$

$$MOR = \frac{E_W[E(Y|G = 1, W)] \times \{1 - E_W[E(Y|G = 0, W)]\}}{\{1 - E_W[E(Y|G = 1, W)]\} \times E_W[E(Y|G = 0, W)]} \quad (6)$$

where IE_i refers to the individual effect for the i-th individual, and W_i denotes the i-th individual's vector of covariates.

Estimation of Effect Estimates

In this article, the TMLE method was implemented to detect the potential prognostic biomarkers of OSC 3-year mortality and estimate their effects, including AE, IE and MOR. Note that, we estimate ATE and MOR from a population level to gain averaged effect and IE from individual level to demonstrate personalized information. In specific, effect estimation with TMLE begins with initial estimates of $E(Y| G, W)$ and $P(G = 1| W)$, then the following "targeting" step would optimize the bias-variance tradeoff for above estimates. Finally, the updated estimation of $E(Y| G, W)$ is used to generate above effect estimates. To avoid model misspecification and improve the robustness, Super Learner algorithm was constructed for parameter estimation. Five models (glm, glm.interaction, glmnet, xgboost, and randomforest) were selected in this study. R package "tmle," "randomForest," "glmnet," and "xgboost" were performed. All statistical analyses were performed using software R version 3.6.1 from CRAN[2] and $p < 0.05$ was considered statistically significant.

Validation in the Gene Expression Omnibus (GEO) Dataset

For the verification of TMLE significant genes, the GEO database was searched. Due to different experiment types

[2] http://cran.r-project.org/

and platforms, a single GEO dataset does not generally contain all TMLE significant genes. Hence, three ovarian cancer datasets were chosen, GSE53963 ($n = 174$), GSE26193 ($n = 107$) and GSE13876 ($n = 157$) as validation sets. Data of GSE53963 was obtained by GPL6480 Agilent-014850 Whole Human Genome Microarray 4x44K G4112F (Probe Name version). Data of GSE26193 was obtained by GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array. Data of GSE13876 was obtained by GPL7759 Operon human v3 ~35K 70-mer two-color oligonucleotide microarrays. All three datasets include complete survival information. The overlapped genes with 355 DEGs were first extracted from GEO datasets. The expression profile of CROCC2 was not available for most GEO datasets. We considered its family member CROCC as the replacement. We then used the average value when a duplicate sample was found and discretized the gene expression data according to the median value. Further, univariate logistic regression was performed between binary DEGs and other clinical pathologic factors with 3-year mortality. Covariates selection and TMLE analysis were followed among DEGs that were significant in univariate analyses. Sensitivity and specificity were calculated for verified TMLE significant genes separately. ROC curve analyses and AUC were used to assess the discriminatory ability of four genes (ARID3C, CROCC, FREM2, and PTF1A) using the GSE13876 dataset. Sensitivity and specificity were calculated for the combination of four genes.

CONCLUSION

In summary, with a combination of the ESTIMATE and TMLE algorithms, we identified four TME-related genes in OSC using the TCGA database. With the help of TMLE, effects were estimated for gene expression level on OSC prognosis. ARID3C and FREM2 were potential prognostic biomarkers for promising 3-year survival while CROCC2 and PTF1A might be biomarkers for poor 3-year survival. More importantly, in addition to AE, IE was estimated and provided to present a personalized effect on recognized genes, which may be useful for predicting individual therapy effects. Therefore, further investigation should be conducted to clarify the biological role in metastasis or recurrence of OSC.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.gdc.cancer.gov; https://www.ncbi.nlm.nih.gov/geo/.

AUTHOR CONTRIBUTIONS

LW and XS performed most of the analyses and wrote the manuscript. FX conceived and designed the study. CJ and YF performed some of the analyses and edited the manuscript.

All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.625145/full#supplementary-material

Supplementary Figure 1 | Heat map showing the correlation between 12 candidates gene and 14 confounding variables. The color shade of each color box represented the corresponding Pearson coefficient value between two variables. For each gene, the black point labeled confounding variables were selected confounders.

Supplementary Table 1 | Three genes showed significant association after GEO validation.

REFERENCES

Bartlett, J. M. S., Gastl, G., and Plante, M. (2001). "Bioactive interleukin-6 levels in serum and ascites as a prognostic factor in patients with epithelial ovarian cancer," in *Ovarian Cancer: Methods and Protocols*, ed. J. M. S. Bartlett (Totowa, NJ: Humana Press), 121–123. doi: 10.1385/1-59259-071-3:121

Bell, D., Berchuck, A., Birrer, M., Chien, J., Cramer, D. W., Dao, F., et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615. doi: 10.1038/nature10166

Bocchicchio, S., Tesone, M., and Irusta, G. (2019). Convergence of Wnt and Notch signaling controls ovarian cancer cell survival. *J. Cell. Physiol.* 234, 22130–22143. doi: 10.1002/jcp.28775

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 68, 394–424. doi: 10.3322/caac.21492

Chien, C. S., Wang, M. L., Chu, P. Y., Chang, Y. L., Liu, W. H., Yu, C. C., et al. (2015). Lin28B/Let-7 regulates expression of Oct4 and Sox2 and reprograms oral squamous cell carcinoma cells to a stem-like state. *Cancer Res.* 75, 2553–2565. doi: 10.1158/0008-5472.CAN-14-2215

Chiotaki, R., Petrou, P., Giakoumaki, E., Pavlakis, E., Sitaru, C., and Chalepakis, G. (2007). Spatiotemporal distribution of Fras1/Frem proteins during mouse embryonic development. *Gene Expr. Patterns* 7, 381–388. doi: 10.1016/j.modgep.2006.12.001

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., et al. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44:e71. doi: 10.1093/nar/gkv1507

Cyr, A. R., Kulak, M. V., Park, J. M., Bogachek, M. V., Spanheimer, P. M., Woodfield, G. W., et al. (2015). TFAP2C governs the luminal epithelial phenotype in mammary development and carcinogenesis. *Oncogene* 34, 436–444. doi: 10.1038/onc.2013.569

Dakubo, G. D. (2019). "Cancer biomarkers in interstitial fluids," in *Cancer Biomarkers in Body Fluids*, (Cham: Springer), 255–271. doi: 10.1007/978-3-030-24725-6_13

Dausinas, P., Pulakanti, K., Rao, S., Cole, J. M., Dahl, R., and Cowden Dahl, K. D. (2020). ARID3A and ARID3B induce stem promoting pathways in ovarian cancer cells. *Gene* 738:144458. doi: 10.1016/j.gene.2020.144458

De Palma, M., Biziato, D., and Petrova, T. V. (2017). Microenvironmental regulation of tumour angiogenesis. *Nat. Rev. Cancer* 17, 457–474. doi: 10.1038/nrc.2017.51

Dell'aringa, M., and Lee Reinhardt, R. (2018). Notch signaling represents an important checkpoint between follicular T-helper and canonical T-helper 2 cell fate article. *Mucosal Immunol.* 11, 1079–1091. doi: 10.1038/s41385-018-0012-9

Diaz-Padilla, I., Duran, I., Clarke, B. A., and Oza, A. M. (2012). Biologic rationale and clinical activity of mTOR inhibitors in gynecological cancer. *Cancer Treat. Rev.* 38, 767–775. doi: 10.1016/j.ctrv.2012.02.001

Eoh, K. J., Kim, H. J., Lee, J. W., Kim, L. K., Park, S. A., Kim, H. S., et al. (2020). E2f8 induces cell proliferation and invasion through the epithelial–mesenchymal transition and notch signaling pathways in ovarian cancer. *Int. J. Mol. Sci.* 21:5813. doi: 10.3390/ijms21165813

Finkernagel, F., Reinartz, S., Schuldner, M., Malz, A., Jansen, J. M., Wagner, U., et al. (2019). Dual-platform affinity proteomics identifies links between the recurrence of ovarian carcinoma and proteins released into the tumor microenvironment. *Theranostics* 9, 6601–6617. doi: 10.7150/thno.37549

Gascoigne, D. K., Cheetham, S. W., Cattenoz, P. B., Clark, M. B., Amaral, P. P., Taft, R. J., et al. (2012). Pinstripe: a suite of programs for integrating transcriptomic and proteomic datasets identifies novel proteins and improves differentiation of protein-coding and non-coding genes. *Bioinformatics* 28, 3042–3050. doi: 10.1093/bioinformatics/bts582

Gentles, A. J., Newman, A. M., Liu, C. L., Bratman, S. V., Feng, W., Kim, D., et al. (2015). The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* 21, 938–945. doi: 10.1038/nm.3909

Greenland, S., and Robins, J. M. (1986). Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* 15, 413–419. doi: 10.1093/ije/15.3.413

Heng, T. S. P., Painter, M. W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S. J., et al. (2008). The immunological genome project: networks of gene expression in immune cells. *Nat. Immunol.* 9, 1091–1094. doi: 10.1038/ni1008-1091

Howlader, N., Noone, A. M., Krapcho, M., Miller, D., Bishop, K., Altekruse, SF Eds, et al. (2016). *SEER Cancer Statistics Review 1975-2013*. Bethesda, MD: National Cancer Institute. http://seer.cancer.gov/csr/1975_2013/, based Novemb 2015 SEER data submission, posted to SEER web site, April 2016.

Huang, C. C., Cheng, S. H., Wu, C. H., Li, W. Y., Wang, J. S., Kung, M. L., et al. (2019). Delta-like 1 homologue promotes tumorigenesis and epithelial-mesenchymal transition of ovarian high-grade serous carcinoma through activation of Notch signaling. *Oncogene* 38, 3201–3215. doi: 10.1038/s41388-018-0658-5

Hubbard, N., Prasasya, R. D., and Mayo, K. E. (2019). Activation of notch signaling by oocytes and Jag1 in mouse ovarian granulosa cells. *Endocrinology* 160, 2863–2876. doi: 10.1210/en.2019-00564

Jin, K., and Xiang, M. (2019). Transcription factor Ptf1a in development, diseases and reprogramming. *Cell. Mol. Life Sci.* 76, 921–940. doi: 10.1007/s00018-018-2972-z

Jovčevska, I., Zottel, A., Šamec, N., Mlakar, J., Sorokin, M., Nikitin, D., et al. (2019). High FREM2 gene and protein expression are associated with favorable prognosis of IDH-WT glioblastomas. *Cancers (Basel)* 11:1060. doi: 10.3390/cancers11081060

Kenny, P. A., Lee, G. Y., and Bissell, M. J. (2007). Targeting the tumor microenvironment. *Front. Biosci.* 12:3468–3474. doi: 10.2741/2327

Kiyozumi, D., Sugimoto, N., and Sekiguchi, K. (2006). Breakdown of the reciprocal stabilization of QBRICK/Frem1, Fras1, and Frem2 at the basement membrane provokes Fraser syndrome-like defects. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11981–11986. doi: 10.1073/pnas.0601011103

Kurkewich, J. L., Klopfenstein, N., Hallas, W. M., Wood, C., Sattler, R. A., Das, C., et al. (2016). Arid3b is critical for B Lymphocyte development. *PLoS One* 11:e0161468. doi: 10.1371/journal.pone.0161468

Lan, C., Huang, X., Lin, S., Huang, H., Cai, Q., Wan, T., et al. (2013). Expression of M2-polarized macrophages is associated with poor prognosis for advanced epithelial ovarian cancer. *Technol. Cancer Res. Treat.* 12, 259–267. doi: 10.7785/tcrt.2012.500312

Larsson, C., Ehinger, A., Winslow, S., Leandersson, K., Klintman, M., Dahl, L., et al. (2020). Prognostic implications of the expression levels of different immunoglobulin heavy chain-encoding RNAs in early breast cancer. *NPJ Breast Cancer* 6:28. doi: 10.1038/s41523-020-0170-2

Luna, X., De Waernbaum, I., and Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika* 98, 861–875. doi: 10.1093/biomet/asr041

Luo, Z., Wang, Q., Lau, W. B., Lau, B., Xu, L., Zhao, L., et al. (2016). Tumor microenvironment: the culprit for ovarian cancer metastasis? *Cancer Lett.* 377, 174–182. doi: 10.1016/j.canlet.2016.04.038

Luque-Fernandez, M. A., Schomaker, M., Rachet, B., and Schnitzer, M. E. (2018). Targeted maximum likelihood estimation for a binary treatment: a tutorial. *Stat. Med.* 37, 2530–2546. doi: 10.1002/sim.7628

Nakahara, S., Fukushima, S., Yamashita, J., Kubo, Y., Tokuzumi, A., Miyashita, A., et al. (2017). AT-rich interaction domain-containing protein 3B is a new tumour marker for melanoma. *Acta Derm. Venereol.* 97, 112–114. doi: 10.2340/00015555-2449

Nielsen, J. S., Sahota, R. A., Milne, K., Kost, S. E., Nesslinger, N. J., Watson, P. H., et al. (2012). CD20+ tumor-infiltrating lymphocytes have an atypical CD27−memory phenotype and together with CD8+ T cells promote favorable prognosis in ovarian cancer. *Clin. Cancer Res.* 18, 3281–3292. doi: 10.1158/1078-0432.CCR-12-0234

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* 66, 688–701. doi: 10.1037/h0037350

Samyesudhas, S. J., Roy, L., and Cowden Dahl, K. D. (2014). Differential expression of ARID3B in normal adult tissue and carcinomas. *Gene* 543, 174–180. doi: 10.1016/j.gene.2014.04.007

Stumpf, M., Hasenburg, A., Riener, M. O., Jütting, U., Wang, C., Shen, Y., et al. (2009). Intraepithelial CD8-positive T lymphocytes predict survival for patients with serous stage III ovarian carcinomas: relevance of clonal selection of T lymphocytes. *Br. J. Cancer* 101, 1513–1521. doi: 10.1038/sj.bjc.6605274

Sumigama, S., Ito, T., Kajiyama, H., Shibata, K., Tamakoshi, K., Kikkawa, F., et al. (2004). Suppression of invasion and peritoneal carcinomatosis of ovarian cancer cells by overexpression of AP-2α. *Oncogene* 23, 5496–5504. doi: 10.1038/sj.onc.1207723

Sung, P. L., Jan, Y. H., Lin, S. C., Huang, C. C., Lin, H., Wen, K. C., et al. (2016). Periostin in tumor microenvironment is associated with poor prognosis and platinum resistance in epithelial ovarian carcinoma. *Oncotarget* 7, 4036–4047. doi: 10.18632/oncotarget.6700

Tidwell, J. A., Schmidt, C., Heaton, P., Wilson, V., and Tucker, P. W. (2011). Characterization of a new ARID family transcription factor (Brightlike/ARID3C) that co-activates Bright/ARID3A-mediated immunoglobulin gene transcription. *Mol. Immunol.* 49, 260–272. doi: 10.1016/j.molimm.2011.08.025

Torre, L. A., Trabert, B., DeSantis, C. E., Miller, K. D., Samimi, G., Runowicz, C. D., et al. (2018). Ovarian cancer statistics, 2018. *CA Cancer J. Clin.* 68, 284–296. doi: 10.3322/caac.21456

Tu, M. J., Ho, P. Y., Zhang, Q. Y., Jian, C., Qiu, J. X., Kim, E. J., et al. (2019). Bioengineered miRNA-1291 prodrug therapy in pancreatic cancer cells and patient-derived xenograft mouse models. *Cancer Lett.* 442, 82–90. doi: 10.1016/j.canlet.2018.10.038

van der Laan, M. J., Rose, S., and Gruber, S Eds (2009). "Readings in targeted maximum likelihood estimation," in *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 254*, (Berkeley, CA: University of California).

van der Laan, M. J., and Rubin, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* 2:11. doi: 10.2202/1557-4679.1043

Wang, M., Zhao, J., Zhang, L., Wei, F., Lian, Y., Wu, Y., et al. (2017). Role of tumor microenvironment in tumorigenesis. *J. Cancer* 8, 761–773. doi: 10.7150/jca.17648

Webb, C. F., Smith, E. A., Medina, K. L., Buchanan, K. L., Smithson, G., and Dou, S. (1998). Expression of bright at two distinct stages of B lymphocyte development. *J. Immunol.* 160, 4747–4754.

Woodfield, G. W., Chen, Y., Bair, T. B., Domann, F. E., and Weigel, R. J. (2010). Identification of primary gene targets of TFAP2C in hormone responsive breast carcinoma cells. *Genes Chromosomes Cancer* 49, 948–962. doi: 10.1002/gcc.20807

Yang, J., Liu, X., Yue, G., Adamian, M., Bulgakov, O., and Li, T. (2002). Rootletin, a novel coiled-coil protein, is a structural component of the ciliary rootlet. *J. Cell Biol.* 159, 431–440. doi: 10.1083/jcb.200207153

Yasar, B., Linton, K., Slater, C., and Byers, R. (2017). Primary cilia are increased in number and demonstrate structural abnormalities in human cancer. *J. Clin. Pathol.* 70, 571–574. doi: 10.1136/jclinpath-2016-204103

Yoshihara, K., Shahmoradgoli, M., Martínez, E., Vegesna, R., Kim, H., Torres-Garcia, W., et al. (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* 4:2612. doi: 10.1038/ncomms3612

Zhang, Y., Tang, H., Cai, J., Zhang, T., Guo, J., Feng, D., et al. (2011). Ovarian cancer-associated fibroblasts contribute to epithelial ovarian carcinoma metastasis by promoting angiogenesis, lymphangiogenesis and tumor cell invasion. *Cancer Lett.* 303, 47–55. doi: 10.1016/j.canlet.2011.01.011

Zhu, Q., Zhang, J., Chen, Y., Hu, Q., Shen, H., Huang, R. Y., et al. (2020). Whole-exome sequencing of ovarian cancer families uncovers putative predisposition genes. *Int. J. Cancer* 146, 2147–2155. doi: 10.1002/ijc.32545