Whole Transcriptome Analyses Identify Pairwise Gene Circuit Motif in Serous Ovarian Cancer

Lixin Cheng
Shenzhen People's Hospital https://orcid.org/0000-0002-9427-383X
Xubin Zheng
The Chinese University of Hong Kong
Rennan Ling
Shenzhen People's Hospital
Jing Gao
University of Helsinki and Helsinki University Hospital
Kwong-Sak Leung
The Chinese University of Hong Kong
Man-Hon Wong
The Chinese University of Hong Kong
Shu Yang
Shenzhen People's Hospital
Yakun Liu
The Fourth Hospital of Hebei Medical University
Ming Dong
Bioland Laboratory
Huimin Bai
Beijing Chaoyang Hospital
Lin Kang (✉ kang.lin@szhospital.com)
Shenzhen People's Hospital
Haili Li
Shenzhen People's Hospital

Research

Keywords: Serous ovarian cancer, Transcriptome, ceRNA, Biomarker, Diagnosis, Prognosis

Posted Date: May 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-422831/v1

Whole transcriptome analyses identify pairwise gene circuit motif in serous ovarian cancer

Lixin Cheng[1,#], Xubin Zheng[1,2,#], Rennan Ling[1], Jing Gao[3,4], Kwong-Sak Leung[2], Man-Hon Wong[2], Shu Yang[1], Yakun Liu[5], Ming Dong[6], Huimin Bai[7], Lin Kang[1]*, Haili Li[1,8]*

[1]Shenzhen People's Hospital, The Second Clinical Medical College of Jinan University, The First Affiliated Hospital of Southern University of Science and Technology, Shenzhen, Guangdong, China

[2]Department of Computer Science, The Chinese University of Hong Kong, Hong Kong

[3]Department of Pulmonary Medicine, University of Helsinki and Helsinki University Hospital, Finland

[4]Respiratory Medicine Unit, Department of Medicine, Karolinska Institute, Solna, Stockholm, Sweden

[5]Department of Gynecology, The Fourth Hospital of Hebei Medical University, Shijiazhuang, Hebei, China

[6]Bioland Laboratory (Guangzhou Regenerative Medicine and Health Guangdong Laboratory), Guang zhou, Guangdong, China

[7]Department of Obstetrics and Gynecology, Beijing Chaoyang Hospital, Capital Medical University, Beijing, China

[8]Department of Gynecology of Shenzhen People's Hospital, The Second Clinical Medical College of Jinan University, The First Affiliated Hospital of Southern University of Science and Technology, Shenzhen, Guangdong, China

[#]These authors contributed equally to this study.

[*]Correspondence: Lin Kang (kang.lin@szhospital.com) or Haili Li (lihaili0311@hotmail.com)

Keywords: Serous ovarian cancer, Transcriptome, ceRNA, Biomarker, Diagnosis, Prognosis

Words: 4237 words, no abstract

ABSTRACT (221 words)

Background: Ovarian cancer is the most lethal gynaecological malignancy, resulting in approximately 14,000 deaths annually in the United States. Transcriptome data are emerging as an effective tool possibly leading to clinical applications for various cancers.

Methods: We collected eight serous ovarian carcinomas (SOCs) and eight normal ovary samples, and generated a whole transcriptome profile of human ovarian cancer using microarrays, including mRNAs, lncRNAs, and circRNAs. We constructed a competing endogenous RNA (ceRNA) network involving these three types of RNAs and identified immune-related circRNAs from the network. Moreover, we proposed a gene-pair filtering method to identify significant expression reversals from integrated multi-cohorts, which mitigates the technical variation and improves the statistical power.

Results: Three pairs of mRNAs (BIRC5:PRKCQ, PTK2B:OGN, and S100A14:NR2F1) were identified as promising biomarkers and were fused as an indicator (SOC index) for diagnostic prediction. Validation in three independent cohorts demonstrated that the SOC index carries a very high predictive capacity (average ROC, 0.99; sensitivity, 0.98; specificity, 1.00). Additionally, the SOC index exhibited its prognostic potential to discriminate SOC patients between early and late stage disease.

Conclusions: Our findings elucidate the repertoire of RNA expressions in SOC and identified three gene pairs for the primary screening of SOC. Further biological experiments of the three gene pairs are warranted in order to investigate the underlying function mechanisms involved in ovarian cancer development.

BACKGROUND

Ovarian cancer is the most lethal malignancy worldwide in gynaecology [1]. According to estimates from the American Cancer Society, 1 in 78 women will suffer from ovarian cancer [2]. Furthermore, around 21,750 will be newly diagnosed and 13,940 will die from ovarian cancer in 2020 [2]. Epithelial ovarian carcinomas account for 90% of ovarian cancer cases and serous ovarian carcinoma (SOC) is thus far the most common subtype in epithelial ovarian carcinomas [3, 4]. Although Cancer antige 125 (CA125) and human epididymis protein 4 (HE4) were approved by the Food and Drug Administration (FDA) to screen for ovarian cancer, neither shows a high sensitivity (less than 80%) [5, 6]. The immune system plays a key role in cancer initiation and progression, and a variety of studies demonstrated that ovarian cancer is a type of immunogenic tumour [7]. Previous research has preliminarily confirmed the prognostic value of the immune system in ovarian cancer [8, 9]. Thus, the transcriptional characteristics of immune-related RNAs are vital for ovarian cancer diagnosis and prognostics [10, 11].

Competing endogenous RNAs (ceRNAs) represent a regulation mode in which ceRNAs interact with other RNAs by competing for the shared target microRNAs [10, 12]. Chiu et al. validated the ceRNA regulatory network in prostate and breast adenocarcinomas [13]. Liang et al. constructed a ceRNA regulatory network for mesenchymal ovarian cancer and identified the downregulation of lncRNA pro-transition associated R (PTAR) s potentially inhibiting cancer metastasis by sponging miR-101 [14]. PTAF is a pivotal regulator of the epithelial-to-mesenchymal transition promoting the invasion–metastasis cascade of ovarian cancer. Furthermore, they demonstrated that the overexpression of PTAF can upregulate snail family zinc finger 2 (SNAI2) by directly sponging miR-25, leading to the promotion of ovarian cancer epithelial-to-mesenchymal transition and invasion [14]. Nevertheless, current research on the ceRNA regulatory network in ovarian cancer merely focuses on the competing relationship between lncRNAs and mRNAs [15-19], given that no circular RNA (circRNA) expression data are available for ovarian cancer. As a layer of the gene regulatory network, circRNAs feature a variety of biological processes, including tumour cell proliferation, migration, and invasion [20]. Investigating circRNA expression and correlation between circRNAs and other types of RNAs will shed light on the underlying molecular mechanisms of ovarian cancer.

Although several previous works emphasised the capability of RNA as cancer biomarkers, these transcriptome markers are not considered sufficient for clinical utility. The primary reason stems from the lack of stability or consistency across different cohorts or measurement platforms. Specifically, the limitations include (1) a low statistical significance and an instability of prediction models resulting from the small size of normal ovarian samples; (2) inconsistent transcriptome size, the amount of total RNA per sample from the tumour and from normal tissues; and (3) the datasets highly susceptible to technical variations such as batch effects. To overcome these challenges, several recent studies took advantage of the relative rank of genes and developed gene panels for cancer diagnosis and prognostics [21, 22]. This strategy concentrates on the relative expression ordering of genes, rather than the gene expression abundance, which is not comparable across samples in many cases. However, the relative expression ordering ignores the size effect difference between the expressions of two genes. A large expression difference of two genes contribute equally to a small one, probably generating false positive discoveries.

In this study, we generated expression profiles from eight SOCs and eight normal ovary samples to characterise the full RNA expression pattern of human SOC. To comprehensively understand the alteration of RNAs, we first assessed the differentially expressed mRNAs, long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs), which are immune-related. Based upon these, we then constructed a competing endogenous RNA (ceRNA) network affecting the expression of the protein-coding genes and selected them as candidate biomarkers due to their topological importance. We proposed a novel gene-pair filtering method to draw out significant expression reversals from multiple integrated cohorts, thereby mitigating the technical variation and improving the statistical power. Next, we developed a diagnostic indicator, called the SOC index using a machine-learning method, based on the gene pairs for the accurate discrimination between SOCs and normal controls. SOC index training on the collected cohort and two large cohorts achieved a high sensitivity and specificity across three independent cohorts. The SOC index also shows a significant correlation with tumour infiltration (CD4+, CD8+) and associated with clinical stage and prognosis. These findings suggest that the SOC index has the potential to detect SOC and monitor disease severity of patients.

MATERIALS AND METHODS

Patients and samples

Table 1 summarises the characteristics of the eight female patients with SOC and eight control samples included in this study. Patients with ovarian cancer and patients with cervical cancer (normal ovarian tissue specimens) were recruited sterilely at the Fourth hospital of Hebei Medical University between 2015 and 2017. The studies were approved by the institutional review board of Hebei Medical University. Both cancerous and normal ovarian tissues were quickly excised and snap-frozen in a liquid nitrogen tank at −80°C until further use.

RNA extraction

Tissues were homogenised in the TRIZOL reagent (Invitrogen, USA) using a Qiagen Tissuelyser. Total RNA was extracted in accordance with the manufacturer's protocol and then quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA). The RNA integrity of each sample was assessed by denaturing agarose gel electrophoresis.

Microarray experiments

To identify deregulated RNAs associated with SOC patient outcomes, we conducted a microarray study and profiled mRNAs, lncRNAs, and circRNAs, respectively. We performed Arraystar Human LncRNA Microarray V2.0 and Arraystar Human circRNA Array V2.0 analyses on all 16 samples. The expressions of lncRNAs and mRNA were quantified using the first platform, while the expressions of circRNAs were measured using the second. Total RNA from each sample was measured using NanoDrop ND-1000. Sample preparation and microarray hybridisation were performed based on the standard protocols of Arraystar (Agilent Technology, USA). Processing RNA was different between the two platforms. For the lncRNA platform, rRNA was removed from total RNA using the mRNA-ONLY™ Eukaryotic mRNA Isolation Kit (Epicentre Biotechnologies, USA). For the circRNA platform, total RNAs were digested with Rnase R (Epicentre, Inc.) to remove linear RNAs and enrich circular RNAs. Then, the two platforms followed the same steps below.

Each sample was amplified and transcribed into fluorescent cRNA utilising a random priming method

(Arraystar Super RNA Labeling Kit; Arraystar). The labelled cRNAs were purified using the RNeasy Mini Kit (Qiagen, Germany). The concentration and specific activity of the labelled cRNAs (pmol Cy3/μg cRNA) were measured using NanoDrop ND-1000. Here, 1 μg of each labelled cRNA was fragmented by adding 5-μl 10 × blocking agent and 1-μl 25 × fragmentation buffer, heating the mixture to 60°C for 30 min, and then adding 25-μl 2 × hybridization buffer to dilute the labelled cRNA. Next, 50-μl hybridisation solution was dispensed into a gasket slide and assembled on the circRNA expression microarray slide. The slides were incubated for 17 hr at 65°C in an Agilent Hybridisation Oven. The hybridised arrays were washed, fixed, and scanned using the Agilent Scanner G2505C. We used the Agilent Feature Extraction software (version 11.0.1.1) to analyse the acquired array images.

Differential and functional analysis

The expression profiles of SOC patients and normal controls were analysed to identify differentially expressed mRNAs, lncRNAs, and circRNAs, respectively. Quantile normalisation was used to normalise the profiles to render the data comparable across samples. The significance levels were estimated using the Student's t-test and, then, adjusted using the Benjamini and Hochberg (BH) multiple testing correction method. The effect size was also taken into account using the twofold change [23, 24]. In total, we obtained 1881 downregulated mRNAs and 1995 upregulated mRNAs, 1849 downregulated lncRNAs and 718 upregulated lncRNAs, as well as 122 downregulated circRNAs and 69 upregulated circRNAs. The differentially expressed mRNAs were subjected to Gene Set Enrichment Analysis (GSEA) [25] for functional enrichment analysis. We performed all analyses using R (R x64, version 3.5.2).

ceRNA network construction

We used the Pearson's correlation test to calculate the expression correlation among mRNAs, lncRNAs, and circRNAs. Only RNAs positively correlated with each other (Pearson correlation coefficient (PCC) > 0.9 and $P < 0.01$) were considered for the construction of the competing endogenous regulatory network [10, 26]. The final network was illustrated using Cytoscape [27].

Denoising individualised pairwise analysis

The abundance of genes may vary across different detection platforms or preprocessing methods, but the relative ranking is stable in a pair of genes. Herein, we proposed denoising individualised pairwise analysis to identify differentially expressed gene pairs (DEPs). The gene expression profiles from different discovery cohorts were concatenated directly as a training set. Then, we performed pairwise subtraction for all genes to form gene pairs ($g_i - g_j$) in a single sample (**Figure 2A**).

However, the gene expression value may vary due to technical noise and can be expressed as $g =$

$g' + \varepsilon$, where g' is the true value of the gene expression and $\varepsilon \in (-\infty, +\infty)$ represents the error of measurement. When subtraction is performed, it becomes $\left(g_i' + \varepsilon_i\right) - \left(g_j' + \varepsilon_j\right) = g_i' - g_j' + \varepsilon_i - \varepsilon_j$. The difference within a pair may not only result from the ground truth of the gene expression $g_i' - g_j'$, but also from the technical variation $\varepsilon_i - \varepsilon_j$ when counting RNAs from reads. To remove the relative expression caused by the technical variation, we introduced difference threshold Δ, where only the difference $g_i' - g_j' + \varepsilon_i - \varepsilon_j$ exceeding Δ was regarded as effective. To this end, we applied the function $f(g_i, g_j) = \begin{cases} 1, g_i - g_j > \Delta \\ -1, g_i - g_j < -\Delta \\ 0, otherwise \end{cases}$ to each gene pair (**Figure 2B**).

Following the intrasample analysis, we conducted a population analysis to obtain the significance level of the gene pairs. For each gene pair, a contingency table was calculated without considering the samples within the difference threshold, given that those samples are regarded as noise (**Figure 2C**). We then calculated the Fisher's exact test, and the significantly different pairs (P < 0.001) were preserved as potential biomarkers.

SOC index generation

To further select fewer genes capable of achieving the optimal discrimination for SOC, we utilised the forward selection on gene pairs with AUC for performance evaluation. All possible combinations from one to ten pairs of genes were evaluated and fivefold cross-validation was applied to avoid bias within the data. When using more than three pairs, the best combinatorial pairs could reach an AUC of 1.00 for all five folds. Therefore, the three pairs (BIRC5:PRKCQ, PTK2B:OGN, and S100A14:NR2F1) were identified as biomarkers for SOC.

The SOC index was calculated using the least absolute shrinkage and selection operator (LASSO) [28-30], a modified linear regression model with regularisation that can avoid overfitting and improve the generalisability of a model. The three pairs (BIRC5:PRKCQ, PTK2B:OGN, and S100A14:NR2F1) were extracted from the training set. The linear model was optimised using the following loss function with an L1 penalty: $\mathcal{L}\left(\mathrm{W}; \alpha\right) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i * W) + \alpha \sum_{j=1}^{3}\left|w_j\right|$, where n is the number of samples, Y is the label for each sample, $X = [f(\mathrm{BIRC5}, \mathrm{PRKCQ}) \quad f(\mathrm{PTK2B}, \mathrm{OGN}) \quad f(\mathrm{S100A14}, \mathrm{NR2F1}) \quad 1]$ is the vector of the three pairs and the constant term, $W = [w_3 \quad w_2 \quad w_1 \quad w_0]^T$ is the vector of weights, and α is the coefficient for the L1 penalty. Due to the unbalanced size of the SOC samples and normal controls, resampling was applied to the training dataset. The L1 coefficient α was chosen through fivefold cross-validation and calculated as 4.2×10^{-4}, which had the best performance (**Figure 3E**). Based on this, we trained and finalised the SOC index $\mathrm{R} = 0.154 * f(\mathrm{BIRC5}, \mathrm{PRKCQ}) + 0.179 * f(\mathrm{PTK2B}, \mathrm{OGN}) + 0.248 * f(\mathrm{S100A14}, \mathrm{NR2F1}) + 0.645$.

Correlation analysis with tumour infiltration

To explore the association with immune cell infiltration, Spearman's rank correlation coefficient was adopted to estimate the correlation between the SOC index and immune cells in cancer ($|R| > 0.3$ and $P < 0.05$). For each sample, we calculated the levels of immune cell infiltration using the Tumour Immune Estimation Resource (TIMER) [31].

RESULTS

Clinical characteristics and differential analysis of the samples

Aged between 41 and 66 years, eight serous ovarian carcinoma (SOC) patients and eight patients without SOC were included in this study. Patients with SOC had lymphatic metastasis and ascites, while patients without SOC did not. The clinical tumour marker cancer antigen 125 (CA125) in two of eight SOC patients was within normal range (<35 U/mL) and only one patient had an abnormal alpha fetoprotein (AFP) level (>7 ng/mL), which reflected the low sensitivity (75% for CA125 and 12.5% for AFP) of the existing clinical protein tumour biomarkers.

We analysed the samples by microarray using the Agilent platform (**Table 1**), yielding comprehensive expression profiles for 17,972 mRNAs, 19,394 lncRNAs, and 5594 circRNAs. Among these, 3876 mRNAs, 2567 lncRNAs, and 191 circRNAs were screened as differentially expressed genes (DEGs) with an absolute fold change >2 and a false discovery rate (FDR) adjusted *P*-value from a Student's t-test <0.10 (**Figure 1A**). Using GSEA [25] Kyoto Encyclopedia of Genes and Genomes (KEGG) [32] functional enrichment analyses showed that DEGs are remarkably involved in the pathways related to cancer immunity, including T-cell activation, cytokine production, immune regulation, and tumour-infiltrating lymphocyte differentiation and migration (**Figure 1B and Figure S1**). According to ImmPort [33] and ImmLnc [34], 270 mRNAs and 426 lncRNAs are immune-related (**Figure 1C and 1D**).

Identification of mRNAs that participate as ceRNAs in SOC

The competitive endogenous RNA (ceRNA) hypothesis revealed an intrinsic mechanism within RNAs that regulate biological processes. LncRNAs, circRNAs, and mRNAs act as miRNA sponges or ceRNAs by competing for the shared microRNAs (miRNAs) (**Figure 1E**). For instance, in the mRNA–

miRNA–lncRNA interaction, changes in the expression of lncRNA alter the number of unbound miRNAs, thereby affecting the expression abundance of the target mRNA. To determine if RNAs participate as ceRNAs, we identified 255 mRNAs, 362 lncRNAs, and 91 circRNAs positively correlated with each other (Pearson's correlation coefficient (PCC) >0.9, $P < 0.01$) and established the competing endogenous regulatory network among these immune-related RNAs (**Figure 1F**). Notably, 91 circRNAs in the network were identified as immune-related circRNAs in SOC, providing an opportunity for immunology studies of ovarian cancers involving circRNAs (**Supplementary Table S1; Figures S2 and S3**).

Following topological network analysis [17, 35-38], we found that mRNAs have a significantly higher network degree and betweenness, while a lower transitivity than non-coding RNAs (**Figure 1G**), indicating that mRNAs tend to lie at the centre in star-like structures. Hence, they exert a greater interactive influence in the regulatory network. Beyond this, given that no circRNA and few lncRNA expression datasets are available for independent validation, we merely selected these mRNAs for further analysis and potential biomarker identification.

Identification of gene-pair markers

In addition to our cohort, we collected all publicly available tissue expression cohorts for SOC from The Cancer Genome Atlas (TCGA) [39] and the Gene Expression Omnibus (GEO) [40], including TCGA-Affy, TCGA-Agilent, GSE18520, GSE6008, and GSE40595 (**Table 2**). mRNA from the six cohorts were quantified using three platforms: the Affymetrix HG-U133A Array, the Affymetrix HG-U133 Plus 2.0 Array, and the Agilent G4502A. In total, 255 genes in the ceRNA regulatory network were extracted from our cohort, from which 219 genes existed in all five public cohorts. We combined the two largest cohorts—TCGA-Affy and GSE18520—with our cohort to comprise the discovery set, consisting of 646 patients and 26 normal controls (**Table 2**). The remaining cohorts served as independent validation sets.

A preliminary study showed that the six cohorts shared few differentially expressed mRNAs (**Supplementary Figure S4**), and using absolute gene abundance was not suitable to identify stable biomarkers across different platforms. Specifically, only one gene was commonly selected as differentially expressed across the six cohorts. Therefore, we proposed a denoising individualised pairwise analysis framework (**Figure 2**; see the Methods section) that selects differentially expressed gene pairs (DEPs) and constructed a SOC index by summarising these DEPs using the least absolute shrinkage and selection operator (LASSO). First, we performed subtraction between every possible pair of genes within each sample (**Figure 2A**), converting the difference between a pair into a 'greater' signal (1) or 'smaller' signal (0) with a noise interval of 0.5 expected to filter out most of the false discoveries

(**Figure 2B**). The gene expression value may include technical variation and is expressed by $g + \varepsilon$, where g is the true gene expression and $\varepsilon \in (-\infty, +\infty)$ represents the error of measurement. Therefore, we used the noise interval to filter out the difference induced by technical variation. If the difference within a pair did not exceed 0.5, the pairing signal was assigned 0. The signals for all possible gene pairs were defined as the relative expression level of the sample, taking into account the technical variation bias caused by data integration. We derived a pairwise spectrum of 23,871 gene pairs. For each pair in the pairwise spectrum, we obtained a contingency table across the population (**Figure 2C**) and 52 DEPs composed of 52 genes were identified (P < 0.001, FDR corrected Fisher's exact test). A heat map displays the effect size (difference between two genes in a pair) of each gene pair (**Figure 3A**). Notably, the denoised relative expression shows a clearer distinction between SOC patients and normal controls (**Figure 3B**).

Next, we applied sequential forward selection on the 52 gene pairs and obtained three pairs (BIRC5:PRKCQ, PTK2B:OGN, and S100A14:NR2F1; **Table 3**). Together, these pairs exhibit an optimal classification performance in a fivefold cross-validation (accuracy = 1.00; **Figure 3C**). The genome position of these genes in different chromosomes appear in **Figure 3D**. The box plot shows the reversal relationship in the training and validation datasets (**Supplementary Figure S5**). To this end, we developed a composite SOC index ranging from 0 to 1.20 on 672 samples in the training set using the LASSO regression model. The optimal L1 penalty coefficient $\alpha = 4.2 \times 10^{-4}$ was determined by fivefold cross-validation with a minimal mean-square error (**Figure 3E**). After training, the SOC index was expressed as

$$\mathrm{R} = 0.154 * f(\mathrm{BIRC5}, \mathrm{PRKCQ}) + 0.179 * f(\mathrm{PTK2B}, \mathrm{OGN}) + 0.248 * f(\mathrm{S100A14}, \mathrm{NR2F1}) + 0.645,$$

$$where\ \ f(g_i, g_j) = \begin{cases} 1, g_i - g_j > \Delta \\ -1, g_i - g_j < -\Delta, \\ 0, otherwise \end{cases} \ \Delta = 0.5.$$

The model analysis of the SOC index demonstrated that each pair provided an independent signal toward the identification of SOC with a varying area under the curve (AUC; **Figure 3F** and **G**), highlighting that multiple weighted pairs exhibited a complementary value (the area under the receiver operating characteristic (AUROC) = 1.00; the area under the precision-recall curve (AUPRC) = 1.00) in discriminating SOC compared to using a single pair or the direct combination of pairs. The SOC index (**Figure 4;** training set, middle panel) distinguished between SOC patients and normal controls in the training set with a high sensitivity (99.1%) and specificity (100%; **Figure 4;** training set: upper and lower panels).

Validation of independent cohorts

To determine if the SOC index obtained from the discovery set is reproducible in other SOC cohorts, we applied the same locked model to three independent validation cohorts (GSE6008, GSE40595, and TCGA-Agilent) measured using three different platforms (Affymetrix HG-U133A Array, Affymetrix HG-U133 Plus 2.0 Array, and Agilent G4502A). Three-dimensional (3D) plots of the three pairs and the box plots of the SOC index in both GSE6008 and GSE40595 (Affymetrix HG-U133A Array platform) indicated that our denoising relative expression model established from the training set carried a sensitivity of 100%, a specificity of 100%, an AUROC of 1.00, and AUPRC of 1.00 (**Figure 4**: GSE6008 and GSE40595; **Figure 5**: GSE6008 and GSE40595). In TCGA-Agilent, we assessed the performance of the three pairs and the SOC index to again distinguish SOC from normal using the Agilent platform (**Figure 4**: TCGA-Agilent). We found that the SOC index resulted in a 93.8% sensitivity and 100% specificity with an AUROC of 0.98 and an AUPRC of 0.99 (**Figure 4**: TCGA-Agilent; **Figure 5**: TCGA-Agilent). Taken together, we observed a highly accurate discrimination in the three independent cohorts, indicating that the SOC index carries a very high predictive value in assisting SOC diagnosis.

SOC index indicates tumour progression

Given that the SOC index of the three identified gene pairs potentially discriminated between SOC and normal controls, we questioned whether the score correlated with the tumour status. To address this, we performed a tumour-infiltrating immune cell analysis using TIMER2.0 [41], and decomposed the bulk mRNA expression into cell-type proportion for our cohort and the other four cohorts (GSE18520, GSE40595, GSE 6008, and TCGA-Affy). We evaluated the correlation between the SOC index and the population of different cell types on five cohorts (**Figure 6A**). For GSE18520, the SOC index has a significantly positive correlation with CD4+ (**Figure 6A** and **Supplementary Figure S6**; SCC = 0.373, $P < 2.613E-03$, Spearman's correlation) and negatively correlated with CD8+ (**Figure 6A** and **Supplementary Figure S6**; SCC = -0.419, $P < 6.278E-04$), resulting in a significant negative correlation with CD8+/CD4+ (SCC = -0.413, $P < 1.043E-03$). The same trend can also be observed in the other cohorts (**Figures 6B–E**).

Furthermore, the SOC index associated with the clinical stage of SOC. Patients with stages IB, IC, and II disease tended to have a high SOC index reaching 1.10, while the SOC index for stages III and IV disease were substantially lower in the TCGA-Affy cohort (**Figure 6F**). When stratifying patients into early and late stage disease, the SOC index for stage III and IV was significantly lower than that for stage I and II (**Figure 6G**; $P < 0.005$, Student's t-test). Moreover, patients were stratified into high-risk (n = 455) and low-risk (n=128) groups according to the median SOC index. We found that patients with a higher SOC index (>1.10) tended to experience better clinical outcomes with a longer survival when compared to those with a lower SOC index (**Figure 6H**; $P < 0.004$, log-rank test). Thus, our findings demonstrated that a higher SOC index corresponded to better overall survival and early stage SOC, illustrating the prognostic power of our gene-pair markers.

DISCUSSION

To obtain comprehensive RNA profiles for serous ovarian carcinomas (SOC), we used microarrays to measure mRNAs, lncRNAs, and circRNAs from eight SOC samples and eight normal ovarian samples. We constructed a competing endogenous RNA regulatory network on the basis of the immune-related and differentially expressed RNAs. Owing to the topological importance and sufficient resources, mRNAs in the competing endogenous regulatory network from our cohort and the two largest public SOC cohorts were extracted to construct a diagnostic model. We identified three gene pairs based on their relative expression ordering and, then, combined them to form a SOC index using LASSO. Validation in three independent cohorts reveals that the model is strongly reproducible and accurate (average AUC close to 1).

The proposed SOC index is superior to the commonly used liquid biopsy biomarkers for ovarian cancer, such as CA125 and HE4. Among all studies regarding CA125 and HE4, CA125 reaches an AUC between 0.78 and 0.93 and HE4 achieves an AUC from 0.82 to 0.96 [42], while our SOC index accomplished 1.00 in two validation cohorts and 0.98 in the other one. Moreover, the machine-learning methods conducted by Yeganeh et al. only obtained an AUC of 0.86 to 0.93 [43].

Additionally, the proposed SOC index reflects tumour infiltration and appears effective for determining SOC prognosis and stage stratification. The SOC index inversely correlated with CD8+/CD4+, which is lower in SOC patients and higher among normal controls. A higher SOC index score implies early stage SOC and an improved prognosis compared to a lower score (**Figures 6G** and **6H**), which seems like a paradox but not rare in tumour biomarkers. For instance, the isocitrate dehydrogenase (IDH) mutation is oncogenic in gliomas, but is prevalent in lower-grade gliomas and the presence of the IDH mutation indicates a longer overall survival in patients with gliomas [44-46].

Normal ovarian samples are crucial for expression profiling studies relying on comparisons with malignant ovarian tissues. However, tissues from normal donors are rare for ovarian cancer because of the invasive procedure. In total, only 33 normal ovarian samples for gene expression were publicly available among TCGA and GEO before this study. We profiled the expression of eight normal ovarian samples from cervical cancer patients not metastasised to the ovaries, representing important support and complements to ovarian research.

In addition to the mRNA and lncRNA expressions, we measured genome-wide circRNAs using microarray previously unexamined, thereby providing an opportunity to investigate the regulatory role of non-coding RNAs in ovarian cancer. This research may also facilitate studies related to the mechanism and therapeutics involving circRNAs in ovarian cancer.

Given the rarity of ovarian samples, data integration is necessary for a large-scale study capable of obtaining accurate biomarkers [47]. The primary challenge lies in integrating cohorts, related to the technical variation between platforms and the batch effect from different experiments. To address this issue, the relative expression of gene pairs in each sample rather than the absolute expression value of a single gene was taken into account. Although this might lose some quantitative information from the expression data, datasets from different resources were integrated for model training, thereby substantially increasing the sample size and improving the statistical power of detecting reversal gene pairs. More importantly, the reversal gene pairs can be easily applied to independent individuals, since they do not require any extra preprocessing of population samples. Notably, neglecting the size effect difference of two genes for relative expression ordering result in a number of false positives among significant results, given that a large expression difference of two genes contribute equally to a small one. To address this, we used a parameter, difference threshold Δ, to reduce the rate of false positive discoveries.

CONCLUSION

In this study, we generated a whole transcriptome profile of human ovarian cancer by analysing eight serous ovarian carcinomas (SOCs) and eight normal ovary samples using microarrays, including mRNAs, lncRNAs, and circRNAs. We constructed a competing endogenous RNA (ceRNA) network involving these three types of RNAs and identified immune-related circRNAs from the network. Based upon that, we proposed filtering gene-pair method to integrate multiple cohorts and obtained SOC index. Our results elucidate the repertoire of RNA expressions in SOC and suggest that only three mRNA pairs are sufficient for the primary screening of SOC. Further biological experiments of the three gene pairs are warranted in order to investigate the underlying function mechanisms involved in ovarian cancer development. Moreover, we measured genome-wide circRNA expressions previously unexamined, providing an opportunity to investigate the mechanism and therapeutics involving circRNAs in ovarian cancer.

ABBREVIATIONS

SOC: serous ovarian carcinomas; ceRNA: competing endogenous RNA; CA125: Cancer antige 125; HE4: human epididymis protein 4; FDA: Food and Drug Administration; PTAR: pro-transition associated R; SNAI2: nail family zinc finger 2; circRNA: circular RNA; ncRNA: long non-coding

RNA; GSEA: Gene Set Enrichment Analysis; PCC: Pearson correlation coefficient; DEP: differentially expressed gene pairs; LASSO: least absolute shrinkage and selection operator; TIMER: Tumour Immune Estimation Resource; AFP: bnormal alpha fetoprotein; DEG: differentially expressed genes; KEGG: Kyoto Encyclopedia of Genes and Genomes; miRNA: microRNA; TCGA: The Cancer Genome Atlas; GEO: the Gene Expression Omnibus; DEP: differentially expressed gene pair; AUROC: the area under the receiver operating characteristic; IDH: isocitrate dehydrogenase

DECLARATIONS

CONSENT FOR PUBLICATION

Not applicable.

DATA AVAILABILITY

Data are available upon request.

CONFLICTS OF INTEREST

The authors declare no competing interests.

AUTHOR CONTRIBUTIONS

LC and XZ conceived the idea and drafted the manuscript. HL and LK collected the data. XZ and LC carried out the data analysis. HL and LK supervised this project. RL, JG, KL, and MW helped interpret the results and provided suggestions. SY, YL, MD, and HB revised the manuscript. All authors read and approved the final manuscript.

FUNDING

This work was supported by Guangdong-Shenzhen Joint Fund of China (Grant no. 2019A1515110097).

ACKNOWLEDGEMENTS

The results are in part based upon data generated by the TCGA Research Network (https://www.cancer.gov/tcga). We are also grateful to Vanessa L Fuller, from Language Services at the University of Helsinki, Finland, for assistance in the English-language revision of this work.

REFERENCES

1. Guo W, Zhu L, Yu M, Zhu R, Chen Q, Wang Q: **A five-DNA methylation signature act as a novel prognostic biomarker in patients with ovarian serous cystadenocarcinoma**. *Clinical Epigenetics* 2018, **10**(1):142.
2. **Key Statistics for Ovarian Cancer** [https://www.cancer.org/cancer/ovarian-cancer/about/key-statistics.html]
3. Nam EJ, Yoon H, Kim SW, Kim H, Kim YT, Kim JH, Kim JW, Kim S: **MicroRNA expression profiles in serous ovarian carcinoma**. *Clin Cancer Res* 2008, **14**(9):2690-2695.
4. Hao D, Li J, Wang J, Meng Y, Zhao Z, Zhang C, Miao K, Deng C, Tsang BK, Wang L *et al*: **Non-classical estrogen signaling in ovarian cancer improves chemo-sensitivity and patients outcome**. *Theranostics* 2019, **9**(13):3952-3965.
5. Rojas V, Hirshfield KM, Ganesan S, Rodriguez-Rodriguez L: **Molecular Characterization of Epithelial Ovarian Cancer: Implications for Diagnosis and Treatment**. *Int J Mol Sci* 2016, **17**(12).
6. Muinao T, Deka Boruah HP, Pal M: **Multi-biomarker panel signature as the key to diagnosis of ovarian cancer**. *Heliyon* 2019, **5**(12):e02826.
7. Hao D, Liu J, Chen M, Li J, Wang L, Li X, Zhao Q, Di L-j: **Immunogenomic Analyses of Advanced Serous Ovarian Cancer Reveal Immune Score is a Strong Prognostic Factor and an Indicator of Chemosensitivity**. *Clinical Cancer Research* 2018, **24**(15):3560.
8. Zhang L, Zhu P, Tong Y, Wang Y, Ma H, Xia X, Zhou Y, Zhang X, Gao F, Shu P: **An immune-related gene pairs signature predicts overall survival in serous ovarian carcinoma**. *Onco Targets Ther* 2019, **12**:7005-7014.
9. Shen S, Wang G, Zhang R, Zhao Y, Yu H, Wei Y, Chen F: **Development and validation of an immune gene-set based Prognostic signature in ovarian cancer**. *EBioMedicine* 2019, **40**:318-326.
10. Cheng L, Nan C, Kang L, Zhang N, Liu S, Chen H, Hong C, Chen Y, Liang Z, Liu X: **Whole blood transcriptomic investigation identifies long non-coding RNAs as regulators in sepsis**. *J Transl Med* 2020, **18**(1):217.
11. Liu X, Zheng X, Wang J, Zhang N, Leung K-S, Ye X, Cheng L: **A long non-coding RNA signature for diagnostic prediction of sepsis upon ICU admission**. *Clin Transl Med* 2020, **10**(3):e123.
12. Sanchez-Mejias A, Tay Y: **Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics**. *J Hematol Oncol* 2015, **8**:30.
13. Chiu HS, Martinez MR, Bansal M, Subramanian A, Golub TR, Yang X, Sumazin P, Califano A: **High-throughput validation of ceRNA regulatory networks**. *BMC Genomics* 2017, **18**(1):418.
14. Liang H, Yu T, Han Y, Jiang H, Wang C, You T, Zhao X, Shan H, Yang R, Yang L *et al*: **LncRNA PTAR promotes EMT and invasion-metastasis in serous ovarian cancer by competitively binding miR-101-3p to regulate ZEB1 expression**. *Mol Cancer* 2018, **17**(1):119.
15. Cheng L, Leung KS: **Quantification of non-coding RNA target localization diversity and its application in cancers**. *J Mol Cell Biol* 2018, **10**(2):130-138.
16. Liu X, Xu Y, Wang R, Liu S, Wang J, Luo Y, Leung K-S, Cheng L: **A network-based algorithm for the identification of moonlighting noncoding RNAs and its application in sepsis**. *Briefings in Bioinformatics* 2020.
17. Cheng L, Liu P, Leung KS: **SMILE: a novel procedure for subcellular module identification with localisation expansion**. *IET Syst Biol* 2018, **12**(2):55-61.
18. Wang X, Han L, Zhou L, Wang L, Zhang LM: **Prediction of candidate RNA signatures for recurrent ovarian cancer prognosis by the construction of an integrated competing endogenous RNA network**. *Oncol Rep* 2018, **40**(5):2659-2673.
19. Li W, Ma S, Bai X, Pan W, Ai L, Tan W: **Long noncoding RNA WDFY3-AS2 suppresses tumor progression by acting as a competing endogenous RNA of microRNA-18a in ovarian cancer**. *Journal of Cellular Physiology* 2020, **235**(2):1141-1154.
20. Geng Y, Jiang J, Wu C: **Function and clinical significance of circRNAs in solid tumors**. *J Hematol Oncol* 2018, **11**(1):98.

21. Li B, Cui Y, Diehn M, Li R: **Development and Validation of an Individualized Immune Prognostic Signature in Early-Stage Nonsquamous Non-Small Cell Lung Cancer**. *JAMA Oncol* 2017, **3**(11):1529-1537.

22. Sun J, Bao S, Xu D, Zhang Y, Su J, Liu J, Hao D, Zhou M: **Large-scale integrated analysis of ovarian cancer tumors and cell lines identifies an individualized gene expression signature for predicting response to platinum-based chemotherapy**. *Cell Death Dis* 2019, **10**(9):661.

23. Cheng L, Wang X, Wong PK, Lee KY, Li L, Xu B, Wang D, Leung KS: **ICN: a normalization method for gene expression data considering the over-expression of informative genes**. *Mol Biosyst* 2016, **12**(10):3057-3066.

24. Liu X, Li N, Liu S, Wang J, Zhang N, Zheng X, Leung KS, Cheng L: **Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review**. *Front Bioeng Biotechnol* 2019, **7**:358.

25. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP: **GSEA-P: a desktop application for Gene Set Enrichment Analysis**. *Bioinformatics* 2007, **23**(23):3251-3253.

26. Nan CC, Zhang N, Cheung KCP, Zhang HD, Li W, Hong CY, Chen HS, Liu XY, Li N, Cheng L: **Knockdown of lncRNA MALAT1 Alleviates LPS-Induced Acute Lung Injury via Inhibiting Apoptosis Through the miR-194-5p/FOXP2 Axis**. *Front Cell Dev Biol* 2020, **8**:586869.

27. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks**. *Genome Res* 2003, **13**(11):2498-2504.

28. Wang J, Xiang X, Bolund L, Zhang X, Cheng L, Luo Y: **GNL-Scorer: A generalized model for predicting CRISPR on-target activity by machine learning and featurization**. *J Mol Cell Biol* 2020.

29. Wang J, Zhang X, Cheng L, Luo Y: **An overview and metanalysis of machine and deep learning-based CRISPR gRNA design tools**. *RNA Biol* 2020, **17**(1):13-22.

30. Tibshirani R: **Regression Shrinkage and Selection Via the Lasso**. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996, **58**(1):267-288.

31. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS: **TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells**. *Cancer Res* 2017, **77**(21):e108-e110.

32. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes**. *Nucleic Acids Res* 2000, **28**(1):27-30.

33. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V, Smith T, Campbell J *et al*: **ImmPort: disseminating data to the public for the future of immunology**. *Immunol Res* 2014, **58**(2-3):234-239.

34. Li Y, Jiang T, Zhou W, Li J, Li X, Wang Q, Jin X, Yin J, Chen L, Zhang Y *et al*: **Pan-cancer characterization of immune-related lncRNAs identifies potential oncogenic biomarkers**. *Nat Commun* 2020, **11**(1):1000.

35. Cheng L, Fan K, Huang Y, Wang D, Leung K-S: **Full Characterization of Localization Diversity in the Human Protein Interactome**. *Journal of Proteome Research* 2017, **16**(8):3019-3029.

36. Cheng L, Liu P, Wang D, Leung KS: **Exploiting locational and topological overlap model to identify modules in protein interaction networks**. *BMC Bioinformatics* 2019, **20**(1):23.

37. Li L, Liu M, Yue L, Wang R, Zhang N, Liang Y, Zhang L, Cheng L, Xia J, Wang R: **Host-Guest Protein Assembly for Affinity Purification of Methyllysine Proteomes**. *Anal Chem* 2020.

38. Cheng L, Zeng Y, Hu S, Zhang N, Cheung KCP, Li B, Leung K-S, Jiang L: **Systematic prediction of autophagy-related proteins using Arabidopsisthaliana interactome data**. *The Plant Journal* 2020, **n/a**(n/a).

39. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**(10):1113-1120.

40. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M *et al*: **NCBI GEO: archive for functional genomics data sets--update**. *Nucleic Acids Res* 2013, **41**(Database issue):D991-995.

41. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, Li B, Liu XS: **TIMER2.0 for analysis of**

tumor-infiltrating immune cells. *Nucleic Acids Res* 2020, **48**(W1):W509-W514.

42. Dochez V, Caillon H, Vaucel E, Dimet J, Winer N, Ducarme G: **Biomarkers and algorithms for diagnosis of ovarian cancer: CA125, HE4, RMI and ROMA, a review**. *J Ovarian Res* 2019, **12**(1):28.

43. Yeganeh PN, Mostafavi MT: **Use of Machine Learning for Diagnosis of Cancer in Ovarian Tissues with a Selected mRNA Panel**. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): 3-6 Dec. 2018 2018*; 2018: 2429-2434.

44. Hartmann C, Hentschel B, Simon M, Westphal M, Schackert G, Tonn JC, Loeffler M, Reifenberger G, Pietsch T, von Deimling A *et al*: **Long-Term Survival in Primary Glioblastoma With Versus Without Isocitrate Dehydrogenase Mutations**. *Clinical Cancer Research* 2013, **19**(18):5146.

45. Chen J-R, Yao Y, Xu H-Z, Qin Z-Y: **Isocitrate Dehydrogenase (IDH)1/2 Mutations as Prognostic Markers in Patients With Glioblastomas**. *Medicine* 2016, **95**(9):e2583-e2583.

46. Sanson M, Marie Y, Paris S, Idbaih A, Laffaire J, Ducray F, El Hallani S, Boisselier B, Mokhtari K, Hoang-Xuan K *et al*: **Isocitrate Dehydrogenase 1 Codon 132 Mutation Is an Important Prognostic Biomarker in Gliomas**. *Journal of Clinical Oncology* 2009, **27**(25):4150-4154.

47. Liu S, Zhao W, Liu X, Cheng L: **Metagenomic analysis of the gut microbiome in atherosclerosis patients identify cross-cohort microbial signatures and potential therapeutic target**. *FASEB J* 2020, **34**(11):14166-14181.

Table and Figure Legends

Tables

Table 1. Clinical and laboratory characteristics of patients with ovarian cancer and control samples.

Table 2. Gene expression datasets used in this study.

Table 3. Genome characteristics of the three gene pairs.

Figures

Figure 1. Construction of the competitive endogenous RNA regulatory network.

Figure 2. Workflow for computing the SOC index.

Figure 3. DEP identification and model construction.

Figure 4. Performance evaluation.

Figure 5. ROC and PRC curves in the independent GSE6008, GSE40595, and TCGA-Agilent validation cohorts.

Figure 6. The SOC index associated with tumour infiltration and prognosis.

Table 1 Clinical and laboratory characteristics of patients with ovarian cancer and control samples.

ID	Age	FIGO staging system	Tissue	Lymphatic metastasis	Ascites	AFP	CA125	Tumour size	Tumour size
						(ng/mL)	(U/mL)	(left; cm)	(right; cm)
A1	53	Cervical squamous carcinoma stage IB1	NOET	no	no	–	–	–	–
A2	52	Uterine sarcoma	NOET	unknown	unknown	–	–	–	–
A3	41	Endometrial carcinoma stage IA	NOET	no	no	–	522	–	–
A4	54	Cervical squamous cell carcinoma stage IB1	NOET	no	no	–	–	–	–
A5	53	Cervical squamous cell carcinoma stage IB1	NOET	no	no	–	–	–	–
A6	61	Cervical squamous cell carcinoma IIA2	NOET	unknown	unknown	–	–	–	–
A7	55	Cervical squamous carcinoma stage IB1	NOET	no	no	–	–	–	–
A8	51	Endometrial carcinoma stage IA	NOET	no	no	–	–	–	–
C1	66	Papillary serous cystadenocarcinomas stage IIIC	SOC	yes	yes	6.37	122.5	5*2.5*2.5	2.5*2.5*1.2
C2	66	Papillary serous cystadenocarcinomas stage IV	SOC	yes	yes	3.44	15.53	8*7*7	4*3*3
C3	50	High-grade papillary serous cystadenocarcinomas stage IIIC	SOC	yes	yes	2.77	3239	11*10*5	5.5*4*2.5
C4	51	Poorly differentiated papillary serous cystadenocarcinomas stage IIIC	SOC	yes	yes	3.84	18.9	9*7*4	3*1.5*1,5
C5	47	Papillary serous cystadenocarcinomas stage IIIC	SOC	yes	yes	2.38	1818	15*14*5	
C6	41	Ovarian mucinous cystadenocarcinomas stage IIIC	SOC	yes	yes	8.6	115	3*2*1	4*3*2
C7	55	High-grade papillary serous cystadenocarcinomas stage IIIC	SOC	yes	yes	2.96	730.9	1.8*1.3*0.9	1.8*1.2*0,7
C8	47	Papillary serous cystadenocarcinomas stage IIIC	SOC	yes	yes	2.15	1278	15*2*10	3*2.5*2.5

Note: FIGO, International Federation of Gynaecology and Obstetrics; NOET, normal ovarian epithelial tissue; AFP, alpha fetoprotein; SOC, serous ovarian carcinomas.

Table 2 Gene expression datasets used in this study.

Dataset	Cohorts	Platform	SOC	Control	Total
Training	TCGA-Affy	Affymetrix HG-U133A Array	585	8	593
	GSE18520	Affymetrix HG-U133 Plus 2.0 Array	53	10	63
	Own data	Agilent-033010	8	8	16
Test 1	GSE6008	Affymetrix HG-U133A Array	41	4	45
Test 2	GSE40595	Affymetrix HG-U133 Plus 2.0 Array	32	6	38
Test 3	TCGA-Agilent	Agilent G4502A	32	5	37

Table 3 Genome characteristics of the three gene pairs.

Symbol	Chrom	Strand	Start	End	EntrezID	Description
BIRC5	chr17	+	76210276	76221716	332	Baculoviral IAP repeat containing 5
PRKCQ	chr10	-	6469104	6622263	5588	Protein kinase C, theta
S100A14	chr1	-	153586731	153588808	57402	S100 calcium binding protein A14
NR2F1	chr5	+	92919042	92930315	7025	Nuclear receptor subfamily 2 group F member 1
OGN	chr9	-	95146248	95166937	4969	Osteoglycin
PTK2B	chr8	+	27179918	27316908	2185	Protein tyrosine kinase 2 beta

Figures

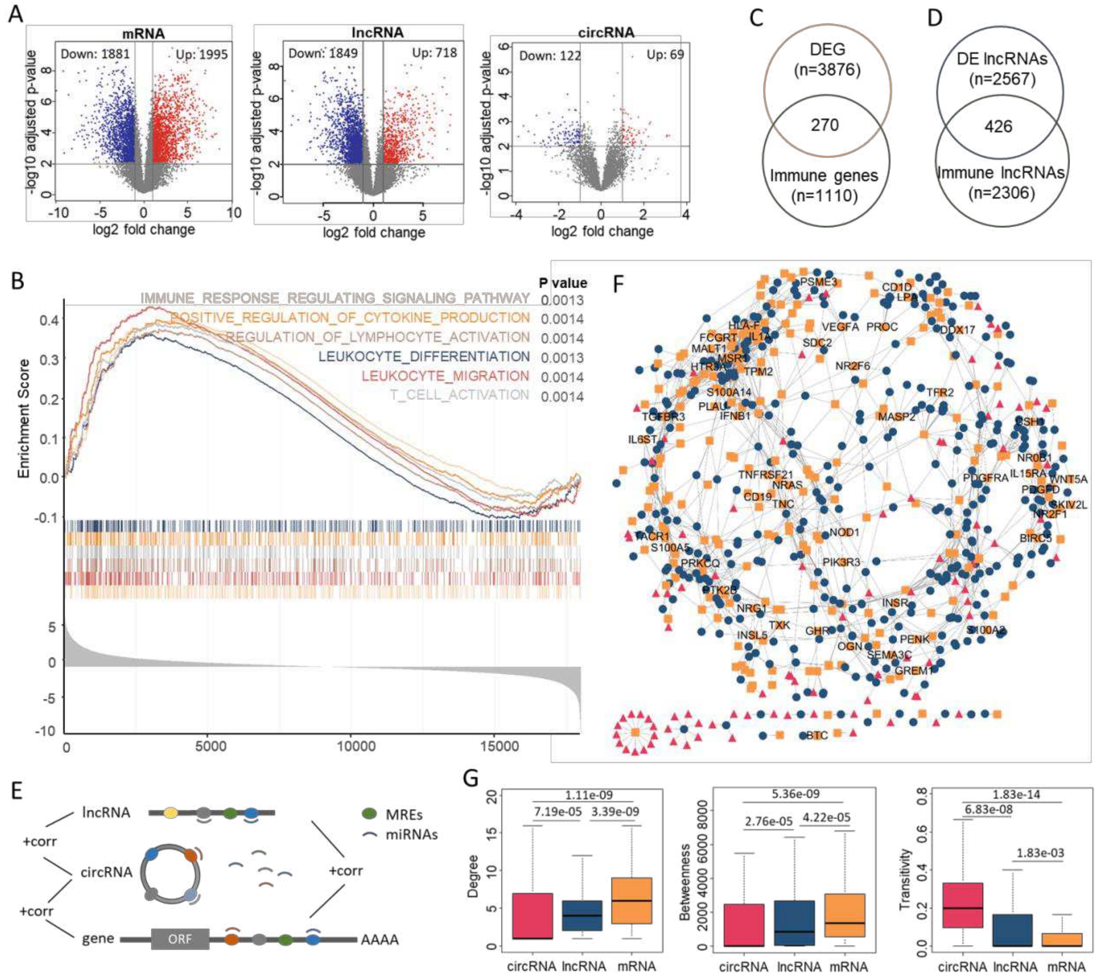
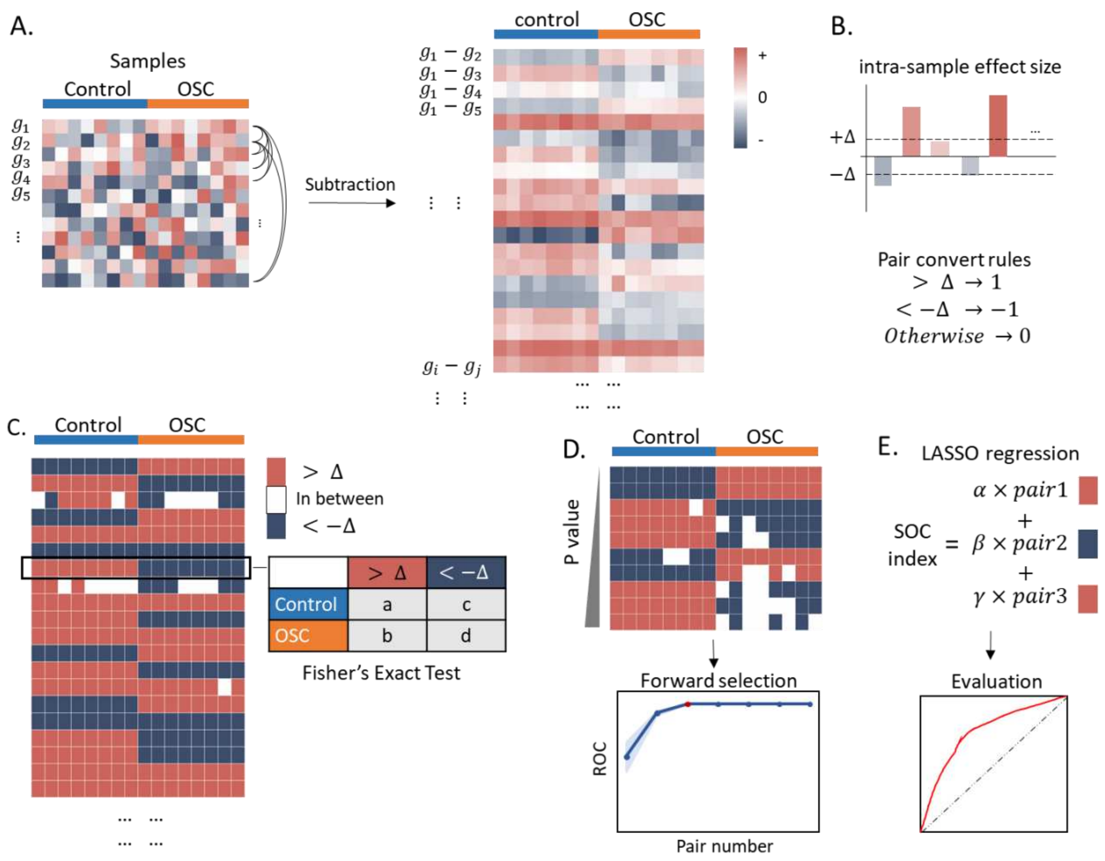


Figure 1 Construction of the competitive endogenous RNA regulatory network. A) Volcano plot of differentially expressed mRNAs, lncRNAs, and circRNAs. B) Functional analysis of differentially expressed RNAs using GSEA. A majority of the enriched biological processes relate to immunity. C) Venn diagram of the differentially expressed genes and immune-related genes. D) Venn diagram of the differentially expressed lncRNAs and immune-related lncRNAs. E) Schematic diagram of the competitive endogenous RNA (ceRNA) hypothesis. F) The immune-related ceRNA regulatory network composed of mRNAs, lncRNAs, and circRNAs. G) Topological analysis of circRNAs, lncRNAs, and mRNAs in the ceRNA network.



	$> \Delta$	$< -\Delta$
Control	a	c
OSC	b	d

Figure 2 Workflow for computing the SOC index. A) Subtraction between every two genes within each sample of the expression matrix. B) Gene pairs with a large effect size (Δ) were converted to intrasample relative expression indicators (1 or -1). Gene pairs with a small effect size possibly introduced from technical variation were filtered out (0). C) Based on the intrasample rank, a Fisher's exact test was performed across the population without considering a small effect size for each gene pair. D) DEPs were derived and further selected using forward selection. E) Linear combination of DEPs. The LASSO regression model was trained based on the selected pairs and evaluated using independent cohorts.

Note: DEPs, differentially expressed gene pairs; LASSO, least absolute shrinkage and selection operator; AUROC, area under the receiver operating characteristic curve.

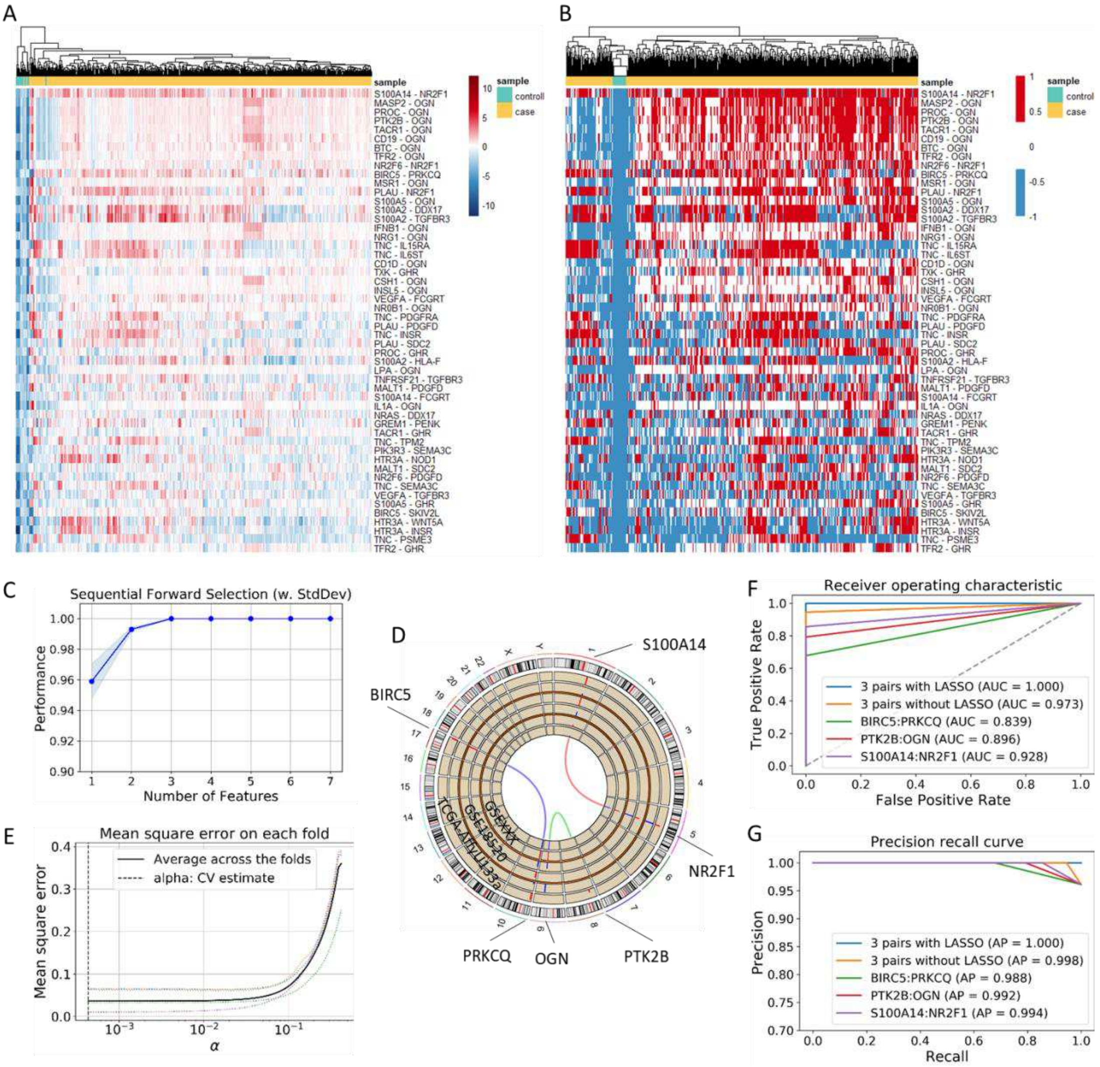
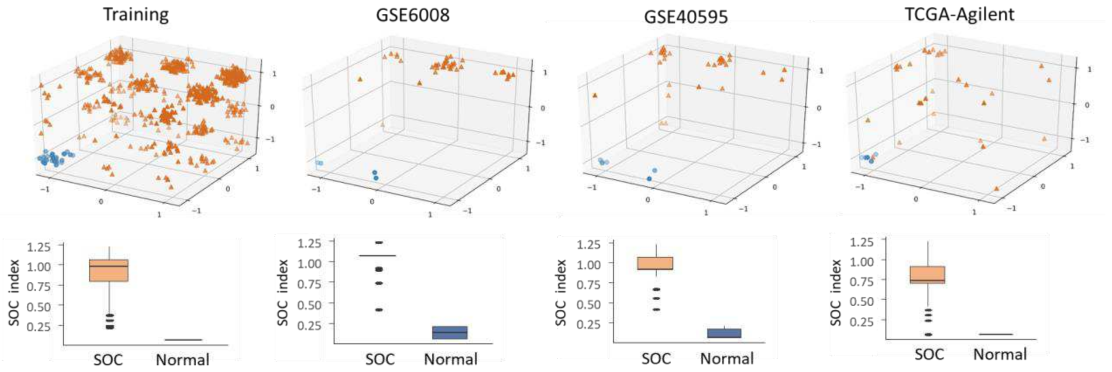


Figure 3 DEP identification and model construction. A) Heat map of the subtraction results for 52 DEPs. B) Heat map of the relative rank indicator for 52 DEPs. C) AUROC for the forward selection procedure with fivefold cross-validation. D) Position of the three gene pairs with the best performance in the chromosome and mean expression levels in three training cohorts. E) Mean square error of the LASSO regression for the L1 penalty coefficient in fivefold cross-validation. F) The ROC curves for different combinations (LASSO, three gene pairs without weights, and single pairs) in the training dataset. G) The PRCs curves for the different combinations in the training dataset.

Note: PRC, precision recall curves.



Training dataset	SOC	Normal	
Predict OSC	640	0	
Predict normal	6	26	Totals
Totals	646	26	672
Correct	640	26	666
Sensitivity (%)	99.1		
Specificity (%)		100	99.1

GSE6008	SOC	Normal	
Predict OSC	41	0	
Predict normal	0	4	Totals
Totals	41	4	45
Correct	41	4	45
Sensitivity (%)	100		
Specificity (%)		100	100

GSE40595	SOC	Normal	
Predict OSC	32	0	
Predict normal	0	6	Totals
Totals	32	6	38
Correct	32	6	38
Sensitivity (%)	100		
Specificity (%)		100	100

G4502A	SOC	Normal	
Predict OSC	32	0	
Predict normal	2	5	Totals
Totals	34	5	39
Correct	32	5	37
Sensitivity (%)	93.8		
Specificity (%)		100	94.9

Training set = TCGA + GSE18520 + Own data; Test1 = GSE6008; Test2 = GSE40595; Test3 = TCGA-Agilent

Figure 4 Performance evaluation. Three-dimensional representation of the three gene pairs (upper panel), box plot of the SOC index calculated by LASSO (middle panel), and confusion matrix (lower panel) in the training set and the three independent validation cohorts.

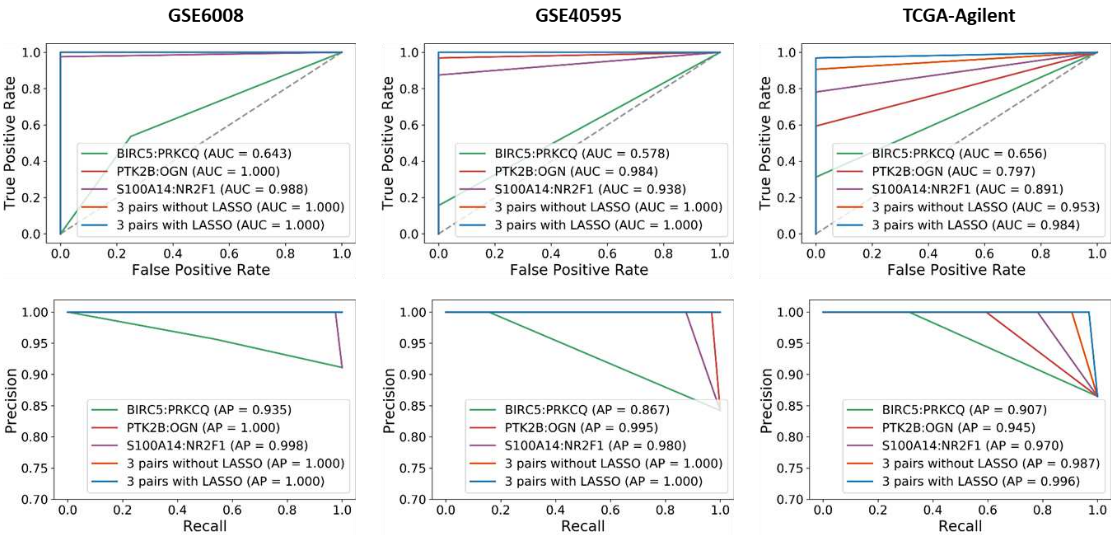


Figure 5 ROC and PRC curves in the independent GSE6008, GSE40595, and TCGA-Agilent validation cohorts.

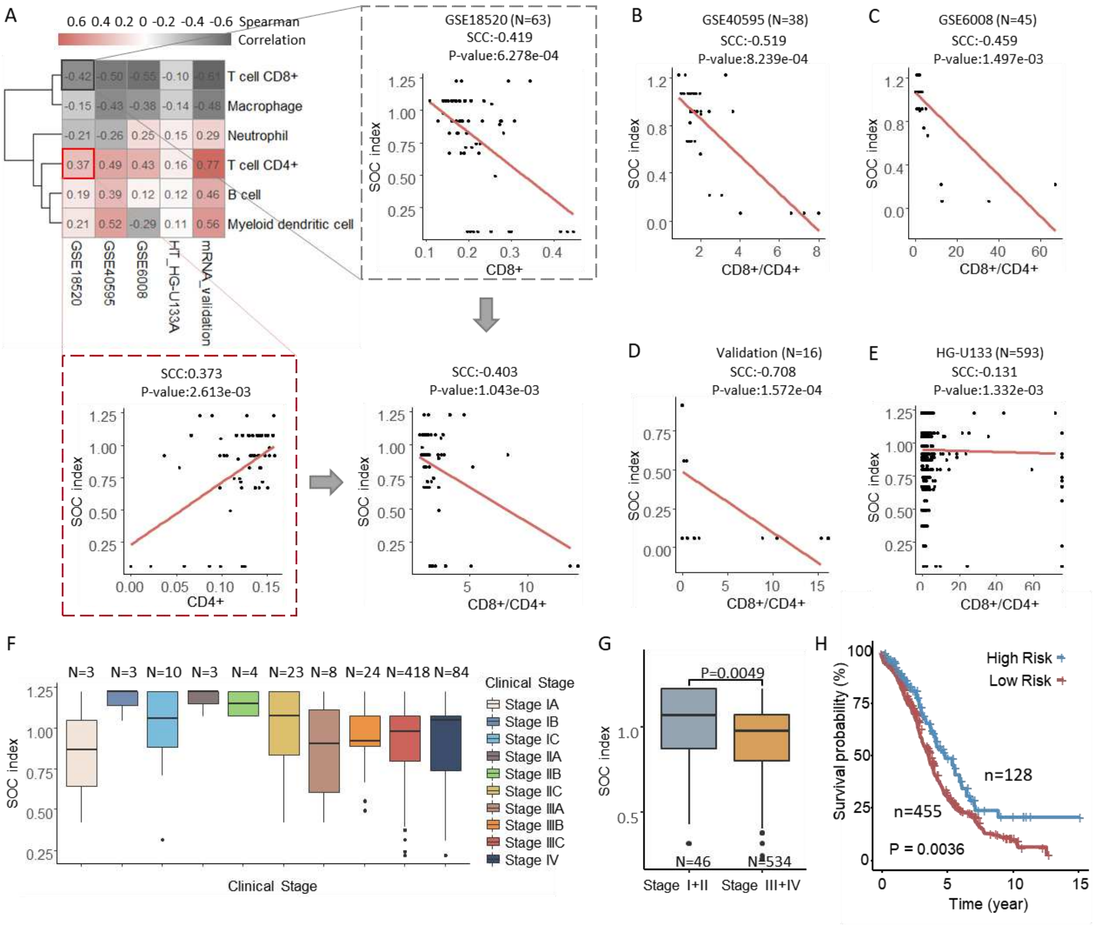


Figure 6 The SOC index association with tumour infiltration and prognosis. The SOC index correlated with tumour infiltration in five cohorts (A, B, C, D, and E) and SOC severity in the TCGA cohort (F, G, and H). (A) The SOC index negatively correlated with CD8+ and positively correlated with CD4+ in GSE18520. The SOC index also inversely correlated with CD8+/CD4+. In the other four cohorts, the SOC index and CD8+/CD4+ consistently negatively correlated (B, C, D, and E). (F) Box plot of the SOC index among different clinical stages. (G) Box plot of the SOC index between stages I and II versus stage III and IV. (H) Survival analysis for patients with a high and low SOC index.

Figures

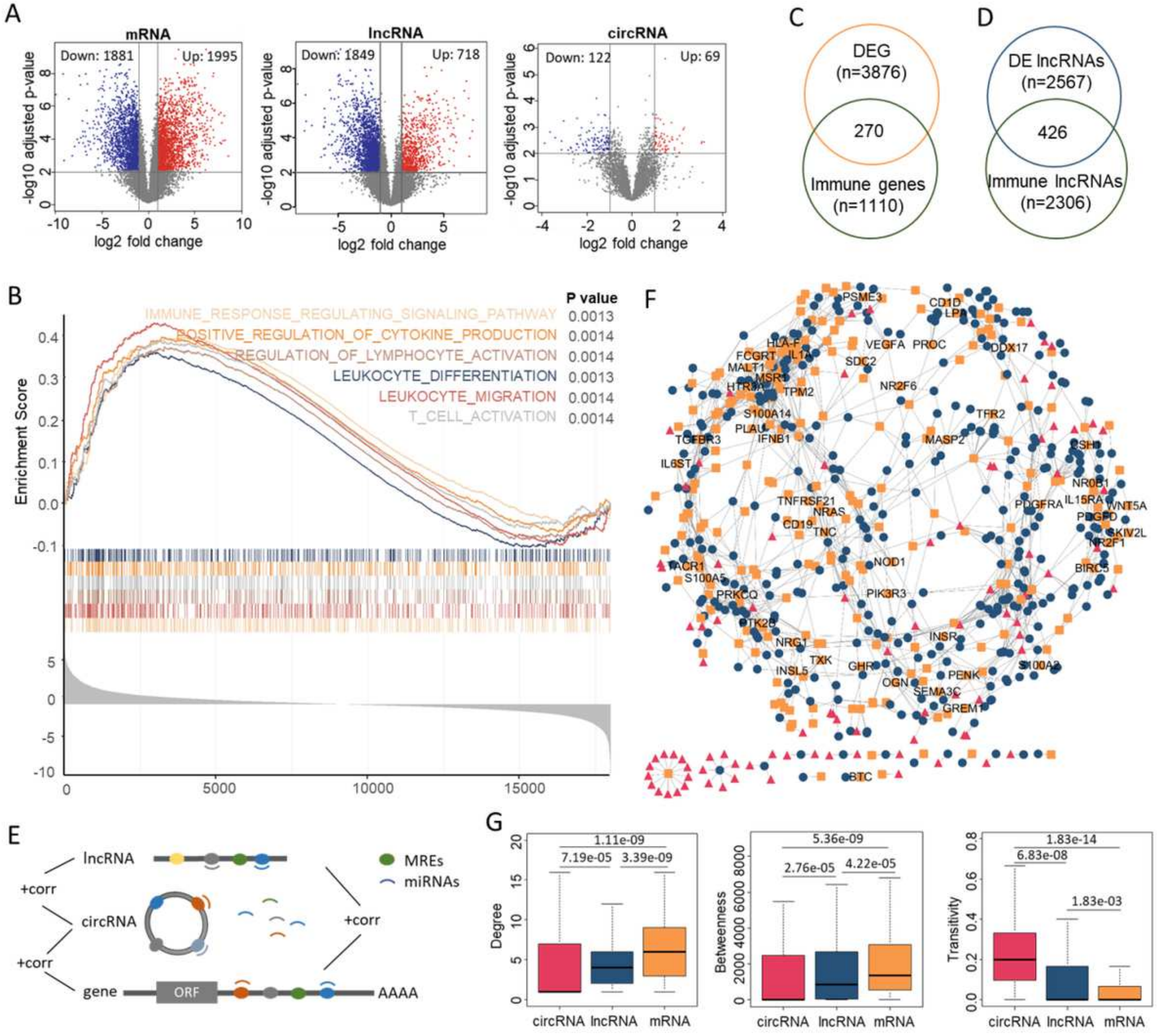


Figure 1

Construction of the competitive endogenous RNA regulatory network. A) Volcano plot of differentially expressed mRNAs, lncRNAs, and circRNAs. B) Functional analysis of differentially expressed RNAs using GSEA. A majority of the enriched biological processes relate to immunity. C) Venn diagram of the differentially expressed genes and immune-related genes. D) Venn diagram of the differentially expressed lncRNAs and immune-related lncRNAs. E) Schematic diagram of the competitive endogenous RNA (ceRNA) hypothesis. F) The immune-related ceRNA regulatory network composed of mRNAs, lncRNAs, and circRNAs. G) Topological analysis of circRNAs, lncRNAs, and mRNAs in the ceRNA network.

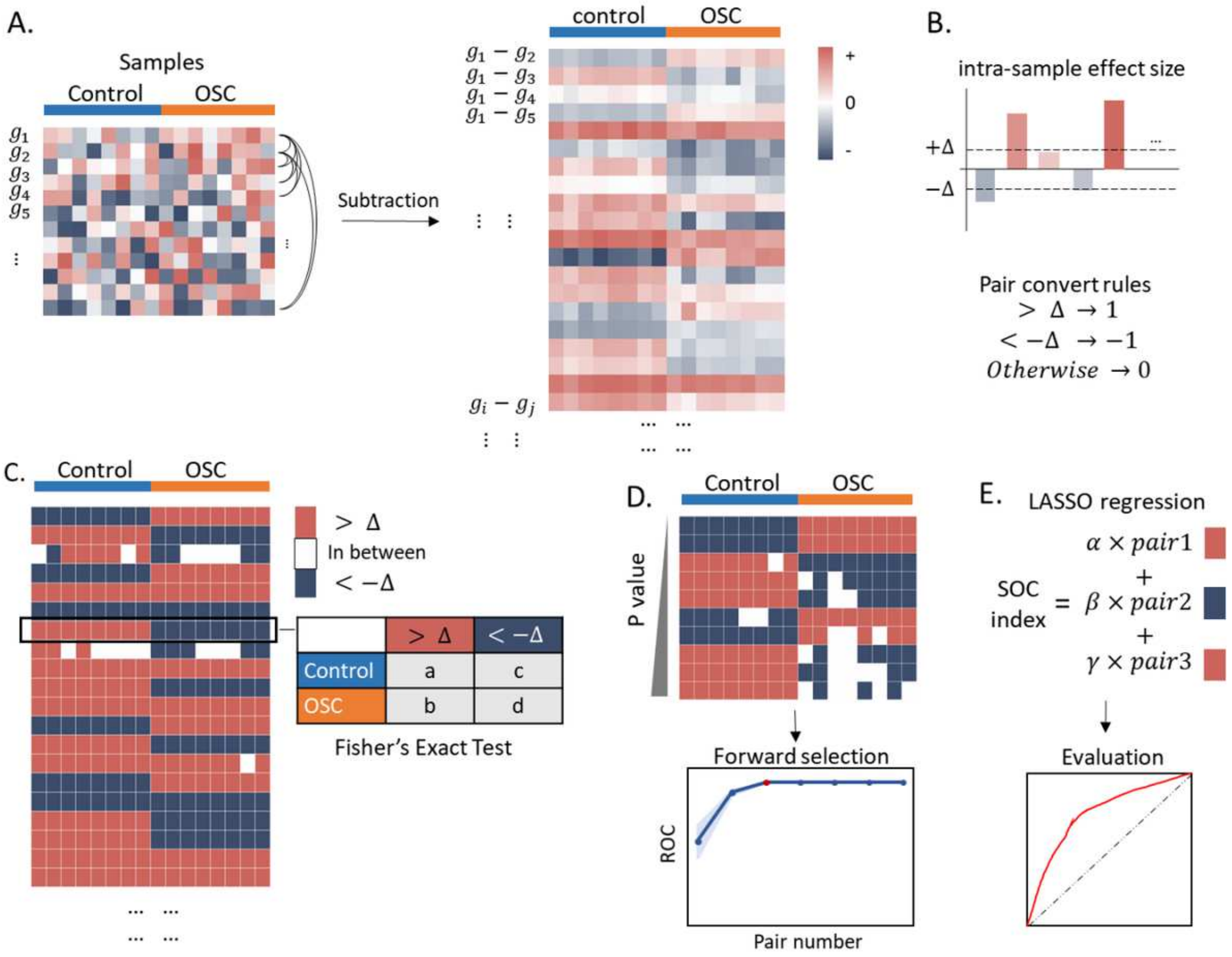


Figure 2

Workflow for computing the SOC index. A) Subtraction between every two genes within each sample of the expression matrix. B) Gene pairs with a large effect size (Δ) were converted to intrasample relative expression indicators (1 or -1). Gene pairs with a small effect size possibly introduced from technical variation were filtered out (0). C) Based on the intrasample rank, a Fisher's exact test was performed across the population without considering a small effect size for each gene pair. D) DEPs were derived and further selected using forward selection. E) Linear combination of DEPs. The LASSO regression model was trained based on the selected pairs and evaluated using independent cohorts. Note: DEPs, differentially expressed gene pairs; LASSO, least absolute shrinkage and selection operator; AUROC, area under the receiver operating characteristic curve.

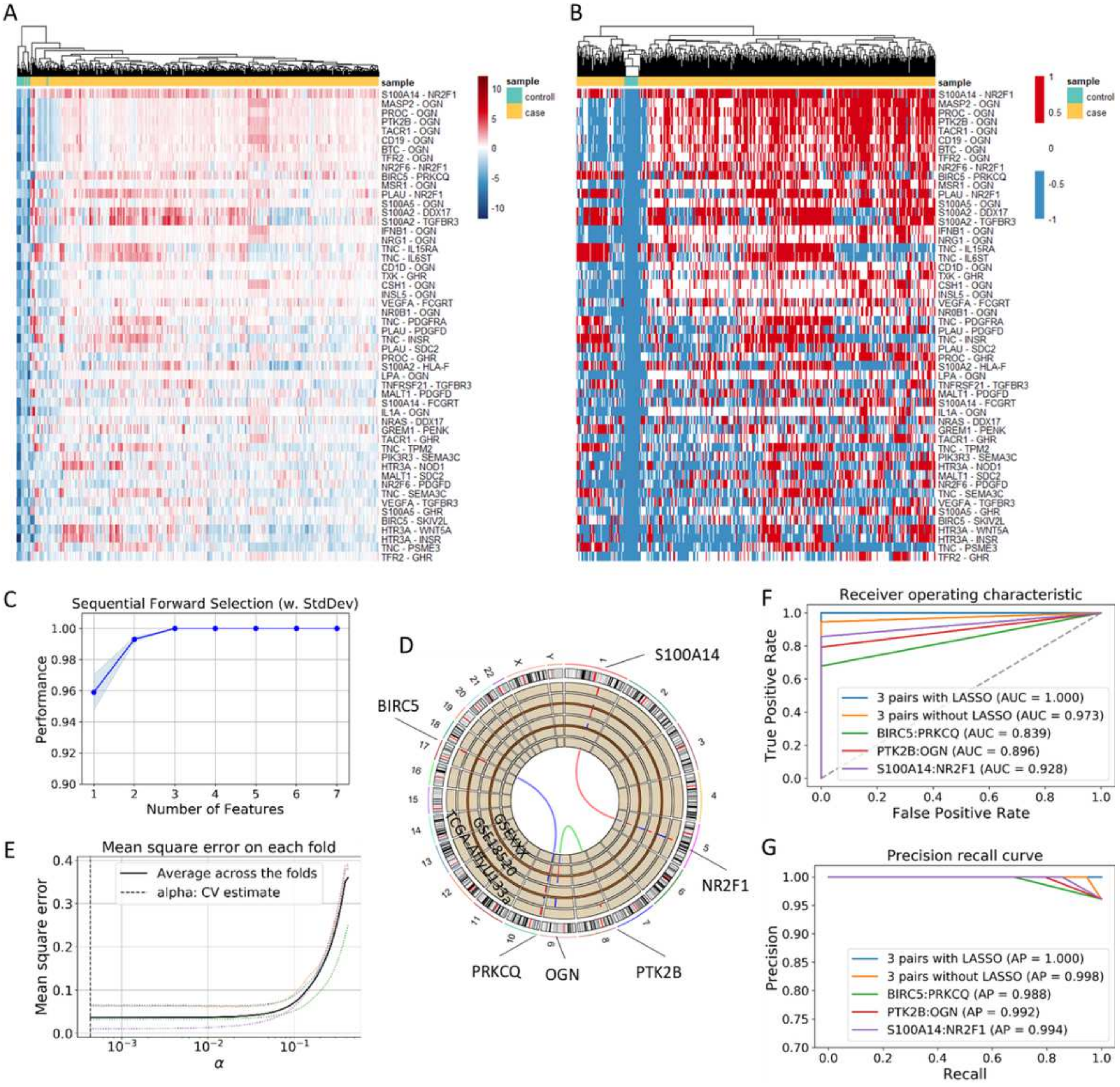
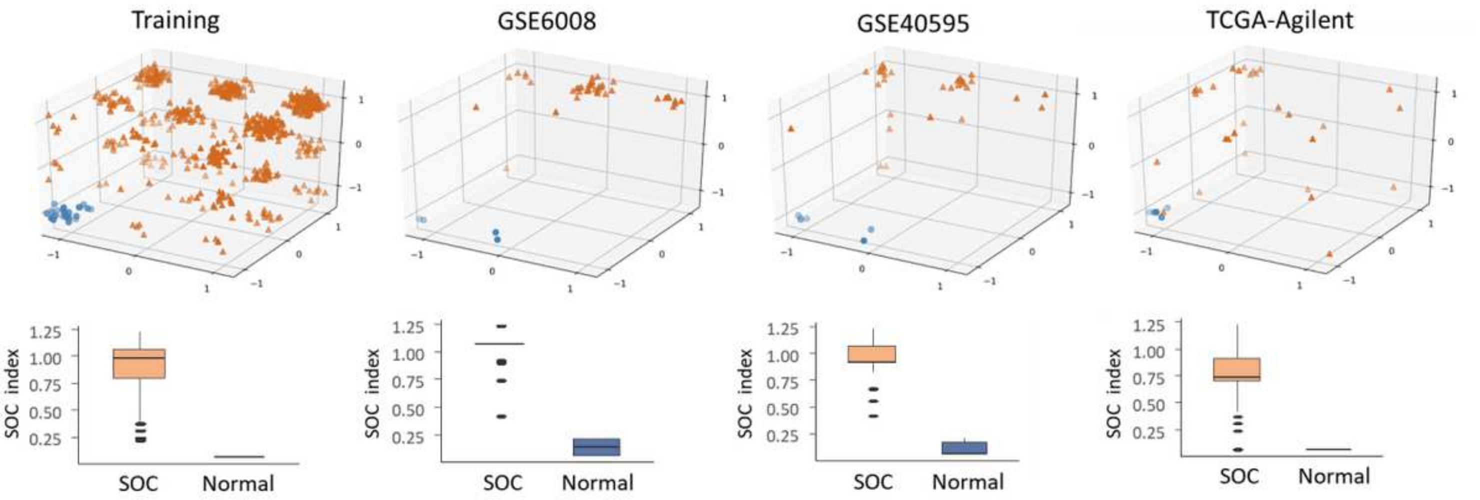


Figure 3

DEP identification and model construction. A) Heat map of the subtraction results for 52 DEPs. B) Heat map of the relative rank indicator for 52 DEPs. C) AUROC for the forward selection procedure with fivefold cross-validation. D) Position of the three gene pairs with the best performance in the chromosome and mean expression levels in three training cohorts. E) Mean square error of the LASSO regression for the L1 penalty coefficient in fivefold cross-validation. F) The ROC curves for different combinations (LASSO, three gene pairs without weights, and single pairs) in the training dataset. G) The PRCs curves for the different combinations in the training dataset. Note: PRC, precision recall curves.



Training dataset	SOC	Normal	
Predict OSC	640	0	
Predict normal	6	26	Totals
Totals	646	26	672
Correct	640	26	666
Sensitivity (%)	99.1		
Specificity (%)		100	99.1

GSE6008	SOC	Normal	
Predict OSC	41	0	
Predict normal	0	4	Totals
Totals	41	4	45
Correct	41	4	45
Sensitivity (%)	100		
Specificity (%)		100	100

GSE40595	SOC	Normal	
Predict OSC	32	0	
Predict normal	0	6	Totals
Totals	32	6	38
Correct	32	6	38
Sensitivity (%)	100		
Specificity (%)		100	100

G4502A	SOC	Normal	
Predict OSC	32	0	
Predict normal	2	5	Totals
Totals	34	5	39
Correct	32	5	37
Sensitivity (%)	93.8		
Specificity (%)		100	94.9

Training set = TCGA + GSE18520 + Own data; Test1 = GSE6008; Test2 = GSE40595; Test3 = TCGA-Agilent

Figure 4

Performance evaluation. Three-dimensional representation of the three gene pairs (upper panel), box plot of the SOC index calculated by LASSO (middle panel), and confusion matrix (lower panel) in the training set and the three independent validation cohorts.

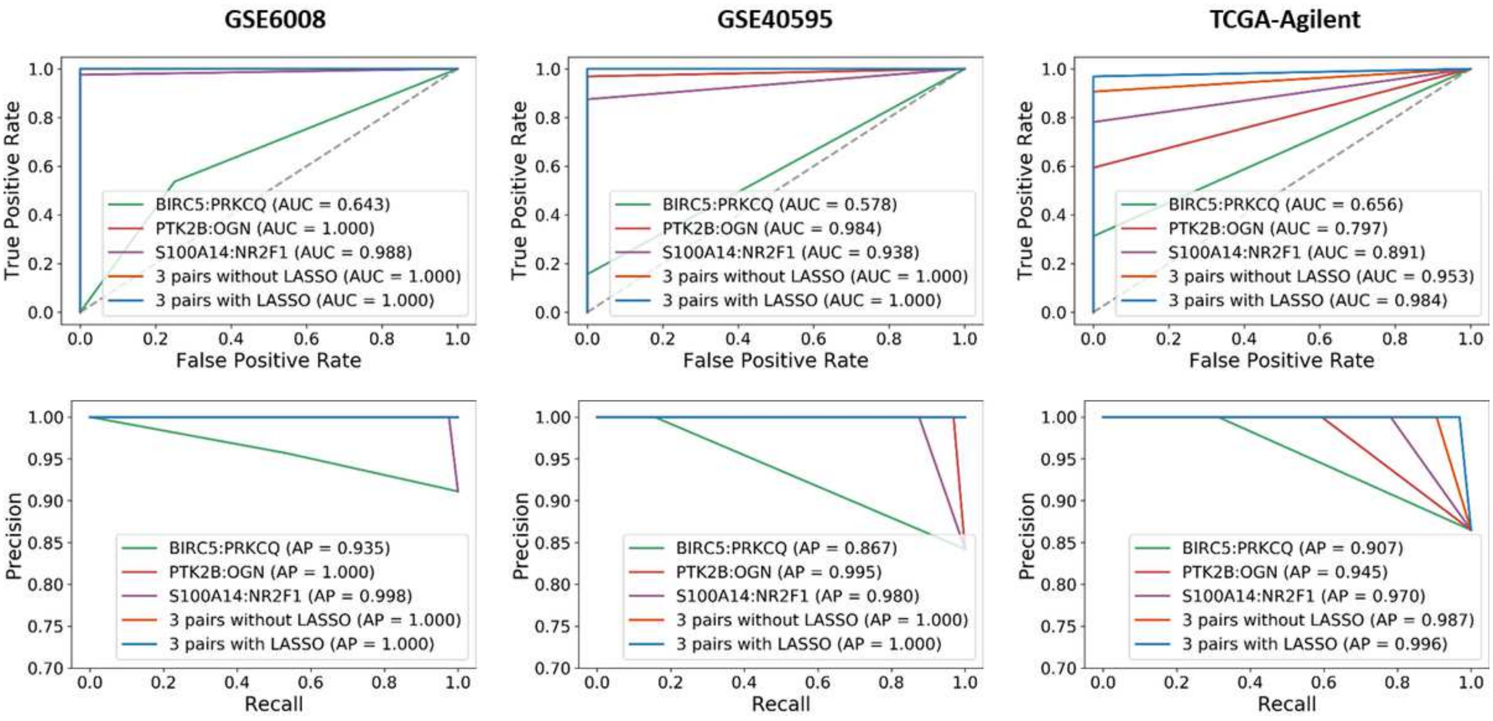


Figure 5

ROC and PRC curves in the independent GSE6008, GSE40595, and TCGA-Agilent validation cohorts.

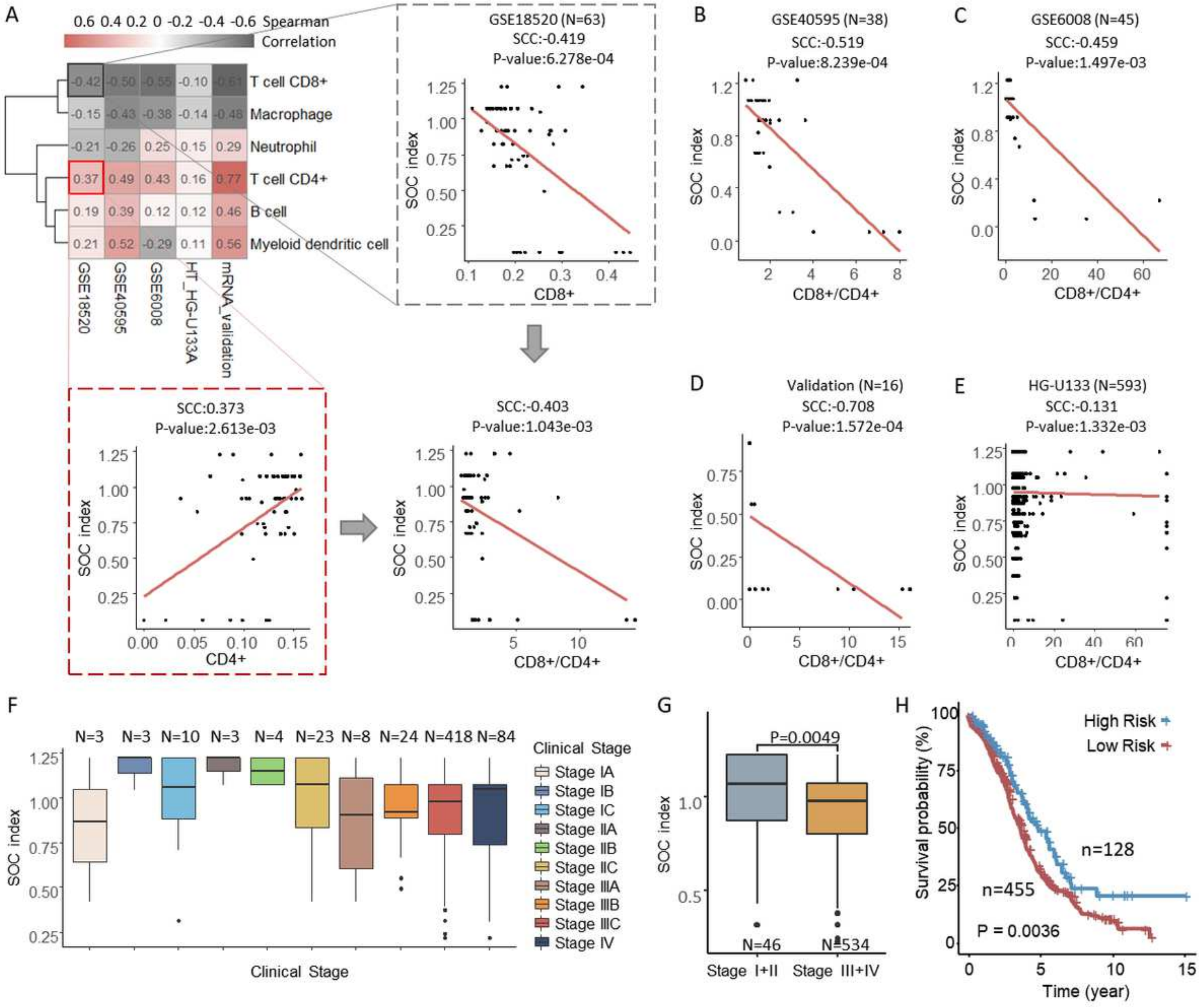


Figure 6

The SOC index association with tumour infiltration and prognosis. The SOC index correlated with tumour infiltration in five cohorts (A, B, C, D, and E) and SOC severity in the TCGA cohort (F, G, and H). (A) The SOC index negatively correlated with CD8+ and positively correlated with CD4+ in GSE18520. The SOC index also inversely correlated with CD8+/CD4+. In the other four cohorts, the SOC index and CD8+/CD4+ consistently negatively correlated (B, C, D, and E). (F) Box plot of the SOC index among different clinical stages. (G) Box plot of the SOC index between stages I and II versus stage III and IV. (H) Survival analysis for patients with a high and low SOC index.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementaryfigurestables.docx