scientific reports

Check for updates

OPEN

Investigating prognostic features in high-grade serous ovarian cancer through gene regulatory network inference with single-cell transcriptomic profiles

Toshiyuki Itai[1], Yulin Dai[1], Wendao Liu[1,2], Dung-Fang Lee[1,2,3,4] & Zhongming Zhao[1,2,5]✉

This study aimed to identify prognostic features in high-grade serous ovarian cancer (HGSOC) through the application of gene regulatory network (GRN) inference with single-cell RNA-sequencing (scRNA-seq) profiles. To achieve this goal, we developed a workflow comprising scRNA-seq analysis, metacell construction, GRN inference, and a binary classification task for prognosis prediction. We curated 118,173 cells from HGSOC patients in three conditions (Before-chemotherapy, After-chemotherapy, and control samples) from previous studies, and then constructed 1,211 metacells. GRN inference analysis revealed 312 regulons, each consisting of one transcription factor and its targeted features. For prognosis evaluation, we used bulk RNA-seq data covering 342 HGSOC patients from The Cancer Genome Atlas (TCGA) and defined a binary outcome of overall survival ≥ 2 years from initial diagnosis, with censored cases at last follow-up assigned to the appropriate class by observed time. We prioritized the features of the TCGA data based on regulon information and differentially expressed features extracted from the metacell data. Our results demonstrated that regulon-based prognostic features were more effective than differential expression-based features in both Before-chemotherapy and After-chemotherapy groups. Our framework can be generalized to other types of cancer when single-cell data for GRN inference and bulk RNA-seq data with clinical outcomes are available.

Keywords Gene regulatory network, single-cell RNA-sequencing, metacell, high-grade serous ovarian cancer

Ovarian cancer (OC) ranks as the fifth most frequent cancer death in women, with approximately 239,000 newly diagnosed OC patients and 152,000 deaths reported annually worldwide[1,2]. OC is categorized into five primary subtypes according to pathology findings. Among them, high-grade serous ovarian cancer (HGSOC) accounts for 70 – 80% of all OC cases and has the worst prognosis[3]. Great challenges exist in early-stage diagnosis because of the anatomical location of the ovaries and the frequent development of chemoresistance after initial treatment. At advanced stages, approximately 30% to 40% of HGSOC patients survive for five years or more; thus, there is a strong need for a deep understanding of its pathophysiology to improve treatment outcomes.

Cellular function is orchestrated by highly organized expressions of tens of thousands of genes and non-coding RNAs (altogether, we refer to them as 'features' in this study). Their expression is controlled by dynamic and complex biological networks, often called gene regulatory networks (GRNs), which involve interactions among targeted features, transcription factors (TFs), and chromatin accessibility[4,5]. Several studies have focused on elucidating the relationship between HGSOC's cellular functions, TFs, and GRN[6–10]. However, there is still

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St. Suite 600, Houston, TX 77030, USA. [2]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA. [3]Department of Integrative Biology and Pharmacology, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [4]Center for Stem Cell and Regenerative Medicine, The Brown Foundation Institute of Molecular Medicine for the Prevention of Human Diseases, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [5]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA. ✉email: zhongming.zhao@uth.tmc.edu

limited knowledge about the relationship between GRN and prognostic features in HGSOC, especially in a cell-type-specific manner.

This study aimed to assess whether GRN is utilized to identify prognostic features in HGSOC. For this purpose, we developed a workflow consisting of GRN inference analysis and machine learning (ML) for predicting the prognosis of HGSOC patients with bulk RNA-seq data (Fig. 1). We collected paired single-cell RNA-sequencing (scRNA-seq) data of HGSOC, "Before-chemotherapy" and "After-chemotherapy", and extracted TF-target interactions (i.e., regulons) through GRN inference analysis. Our results showed that GRN inference analysis could effectively extract prognostic features in HGSOC. Furthermore, paired Before-chemotherapy and After-chemotherapy data enabled us to extract different prognostic features, and cell-type-specific information enabled specifying prognostic features at a cellular level, which is important considering the heterogeneous characteristics of the tumor microenvironment.

Materials and methods

Collection of HGSOC scRNA-seq and control data

We collected three public datasets from Gene Expression Omnibus (GEO: GSE165897, GSE191301, and GSE201047) that contained paired HGSOC scRNA-seq data Before-chemotherapy and After-chemotherapy[11–13]. Among the 15 patients, we used six patients' paired data, with each of the samples having more than 2,000 cells (Table S1). All these samples were diagnosed with HGSOC stage III or IV. For comparison, we collected scRNA-seq data from the normal fallopian tube (GSE151214) and ovary (GSE184880)[14,15].

scRNA-seq data analysis

We used Scanpy (version: 1.9.3) and scvi-tools (version: 0.20.3) in Python (version: 3.10.11) for doublet removal, basic filtering, data integration, and cell-type classification[16,17].To remove doublets, we used SOLO function in scvi-tools with default parameters. For basic filtering, we retained the cells with a minimum of 200 expressed genes, 1,000 unique molecular identifier (UMI) counts, and less than 20% mitochondrial gene expression. Batch correction was performed using scvi-tools with the default parameters. After the data integration, we manually clustered and annotated cell types using our marker gene sets (Table S2) curated from CellMarker 2.0[18] and a previous study of HGSOC scRNA-seq analysis[19].

Metacell construction

We constructed metacells using SEACells[20] in Python (version: 3.8.17). Briefly, SEACells uses a k-nearest neighbor graph and kernel archetypal analysis to aggregate single cells with a similar phenotype[20]. We used SEACells with default parameters, ensuring that one metacell included the 75 most similar single cells. After constructing 1,921 metacells, we kept 1,211 metacells comprising one cell type for GRN inference analysis.

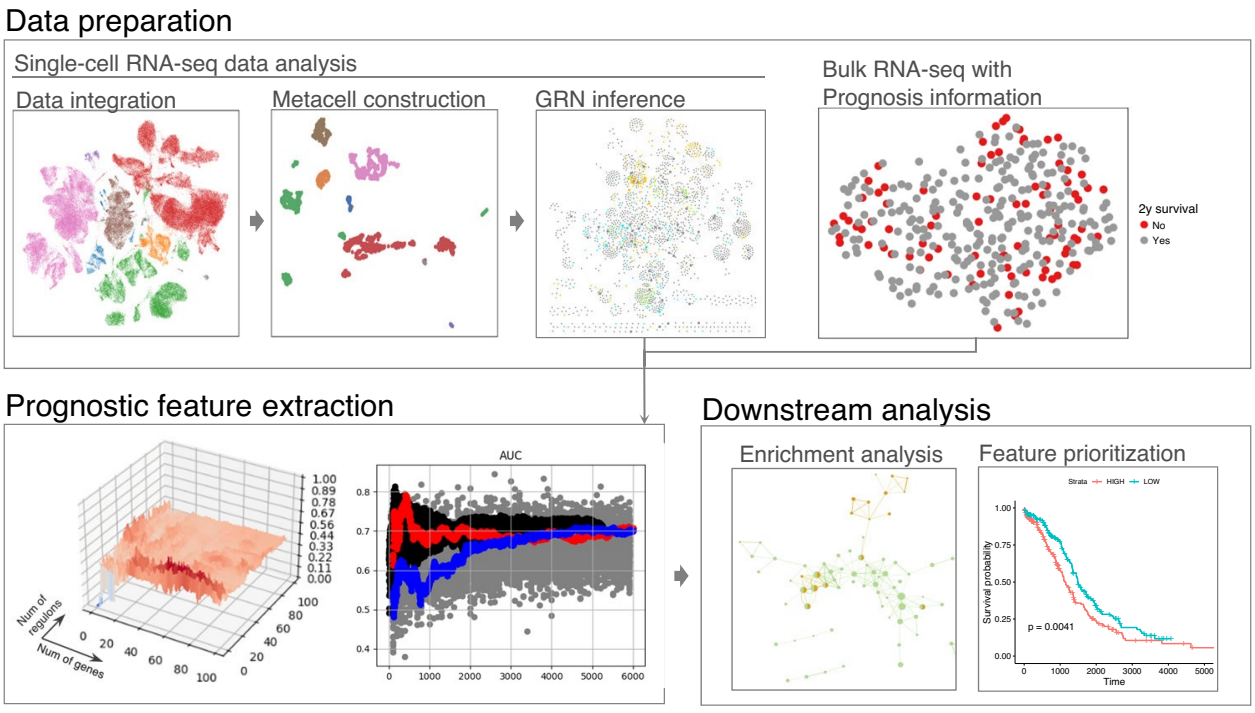

Fig. 1. Overview of our workflow. Single-cell RNA-seq data are utilized for gene regulatory network (GRN) inference analysis, while bulk RNA-seq data with clinical information are employed for a binary task to predict patients' longevity in high-grade serous ovarian cancer (HGSOC). The results obtained from the workflow can be used for further downstream analyses.

We employed the Wilcoxon rank-sum test to identify differentially expressed features by each cell type and treatment status.

GRN inference and downstream analyses

We used 1,211 metacells with 10,504 genes (expressed in more than 50% of the metacells) for GRN inference by running pySCENIC[21] on Python (version: 3.10.11). We also employed pySCENIC for downstream analyses, including the identification of specifically activated regulons by measuring regulon specificity scores, and selected the top 1% of regulons based on TF-target interaction importance scores. Cytoscape[22] was used for GRN visualization.

GRN-based prognostic feature extraction

To test whether GRN could be used for prognostic feature extraction, we conducted a binary classification task using bulk RNA-sequencing (RNA-seq) data of HGSOC from The Cancer Genome Atlas (TCGA) datasets.[23] We used R (version: 4.2.2) and TCGAbiolinks (version: 2.36.0)[24] to collect the bulk RNA-seq data and prognosis information from 342 HGSOC patients. We conducted a binary classification task to predict whether the patients lived over 2 years only based on the bulk RNA-seq data with linked clinical outcome data. We used 80% of the data for training and 20% for validation. We employed four machine learning methods, logistic regression, random forest, support vector machine (SVM), and XGBoost, which were implemented in the scikit-learn package (version: 1.3.0). The random forest model was run with default parameters (n_estimators = 100, criterion = "gini", max_depth = None, and max_features = "sqrt"). The XGBoost model used its defaults (n_estimators = 100, max_depth = 6, learning_rate = 0.3, and objective = "binary:logistic"). The SVM model was trained with default settings as well (C = 1.0, kernel = "rbf", and gamma = "scale"). We applied 'sklearn.metrics' function in scikit-learn package (version: 1.3.0) to calculate accuracy, F1 score, and area under the receiver operating characteristic curve (AUC) to evaluate ML model performance.

Permutation test

To assess the relationship between the number of features and ML performance, we tested feature set sizes from 100 to 60,000 in steps of 100 (i.e., 100, 200, 300, …, 60,000). This yielded 600 distinct feature-set sizes. For each feature-set size, we performed 100 randomized permutations consisting of two steps: 1) features were randomly selected, then 2) the logistic regression model was used for training and validation processes, resulting in 60,000 tests (600 feature sizes × 100 permutations). Accuracy, F1 score, and AUC were used to evaluate the model performance.

Gene set enrichment analysis

We used WebGestalt[25] (accessed October 30, 2023) for gene functional enrichment analysis. We conducted Over-Representation Analysis using the Gene Ontology (GO) database, including its three domains, Biological Process, Molecular Function, and Cellular Component, to identify enriched GO-terms.

Survival analysis

We used the bulk RNA-seq data of HGSOC from the TCGA datasets for survival analysis[23]. DEseq2 (version: 1.48.0)[26] and SummarizedExperiment (version: 1.38.1)[27] were used for data cleaning. We used survival (version: 3.8.3)[28] and survminer (version: 0.5.0)[29] to analyze and draw survival curves, respectively.

Results

HGSOC scRNA-seq data integration and metacell construction for GRN inference

We collected 31 samples from five datasets, comprising six patient samples and ten control samples (Table S1). After quality control and doublet removal, we obtained and annotated 118,173 cells with curated marker genes (Table S2), including 29,681, 49,233, and 39,259 cells in the groups "Before-chemotherapy", "After-chemotherapy", and "Control", respectively (Fig. 2A, Table S3). Marker gene expression in each cell type is presented in Fig. 2B.

The computational cost of GRN inference is generally high, which is problematic especially when a wide variety of data is used or computational resources are limited. To address this, we employed SEACells for metacell construction[20] so that every most similar 75 single cells were aggregated into one metacell. We constructed 1,921 metacells from the integrated dataset, and then removed 710 metacells consisting of more than one cell type. This resulted in 1,211 metacells for downstream analyses. The number of features shared among the metacells significantly outnumbered those in single cells. Specifically, at the median, metacells shared 10,504 expressed features (vs. 1,172 genes at the single-cell level; Fig. 2C), reflecting reduced sparsity after aggregation. Uniform Manifold Approximation and Projection (UMAP) visualizations of the metacells are presented in Fig. 2D and Figure S1, which did not show significant distinction from those with scRNA-seq data. Differentially expressed (DE) features by cell type and treatment status are provided in Supplementary Tables S4 and S5, respectively.

Each cell type and treatment status had a distinct regulon activity pattern

Using raw count data of the 10,504 features in 1,211 metacells, we obtained 312 regulons through GRN inference analysis using pySCENIC[21]. Each regulon typically consists of one TF and several targeted features (targets) that are assumed to be regulated by the TF. The number of genes regulated by a TF (regulon size) ranged from 2 to 3,953, with an average and median value of 749 and 409, respectively (Figure S2 and Table S6). TF-target interactions were quantified as importance scores for each regulon in metacells (Figure S3). We clustered the 1,211 metacells based on the 312 regulon-activity scores, indicating that each cell type and treatment status had distinct regulon-activity patterns (Fig. 2E). These 312 regulon-activity scores were also used to extract regulons specifically activated in each cell type and treatment status (Fig. 2 F, G). Among them, five TF genes (*ARID3A*,

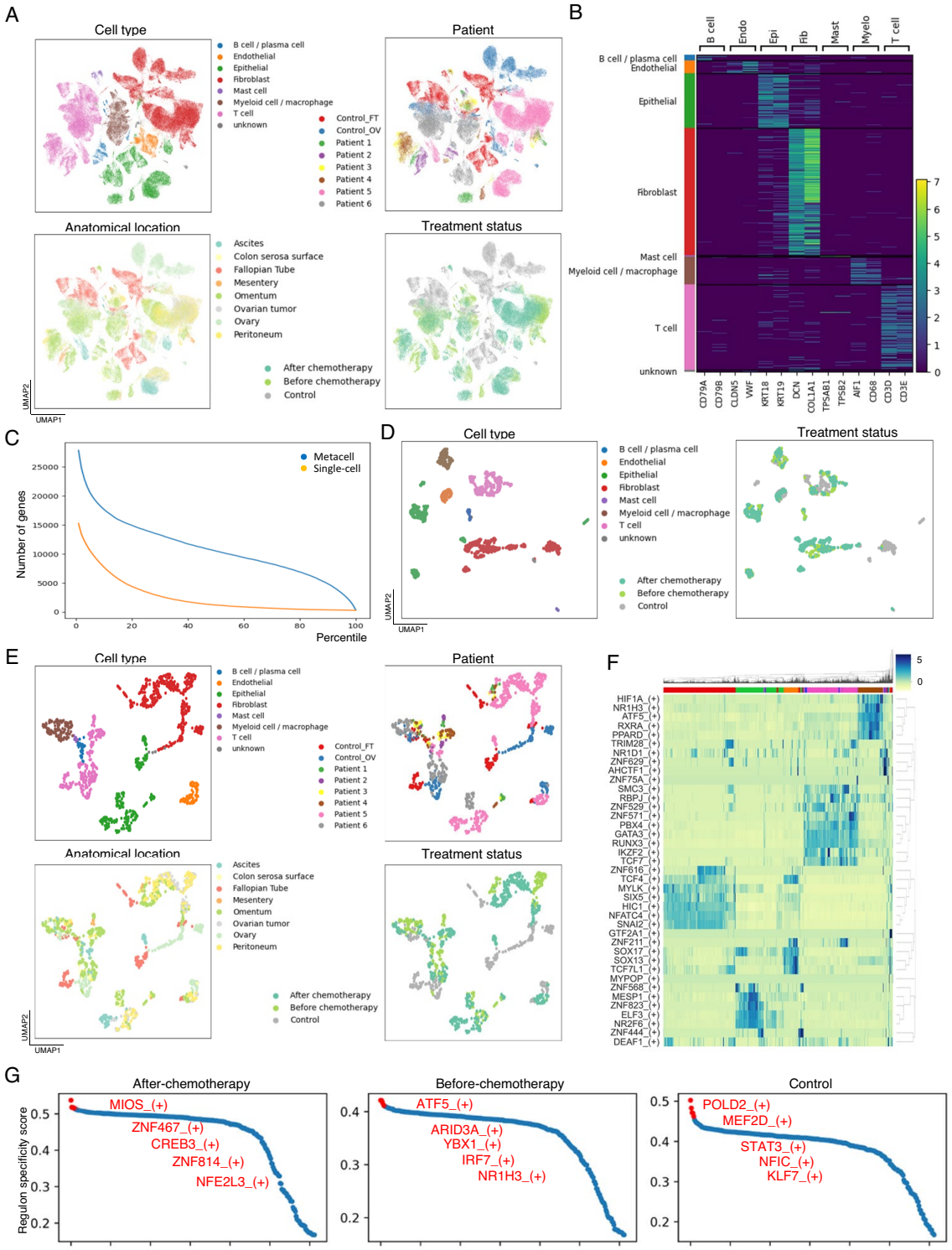

Fig. 2. Singe-cell RNA-seq (scRNA-seq) data integration, metacell construction, and gene regulatory network (GRN) inference of HGSOC data. (**A**) Uniform Manifold Approximation and Projection (UMAP) plots of the integrated scRNA-seq data. (**B**) Marker gene expressions in each cell type. (**C**) The number of shared features in single-cell data and metacell data. The x-axis represents percentile, and the y-axis represents the number of shared features (coding and noncoding genes). (**D**) UMAP plots of the metacell data clustered based on expression profiles. (**E**) UMAP plots of metacells, clustered based on regulon activity score. (**F**) Specific regulons in each cell type. (**G**) Specific regulons in each treatment status. The y-axis represents regulon specificity scores.

ATF5, *CREB3*, *NFE2L3*, and *YBX1*) have been implicated in the pathophysiology of ovarian cancer[30–34], underscoring the value of prioritizing regulons for the identification of features associated with HGSOC.

DE features in each treatment status were closely connected to specific regulons

Considering that each treatment status had distinct feature expression patterns (Fig. 2D) and regulon-activity patterns (Fig. 2E), we hypothesized that regulon activities and feature expressions were closely connected. To test the hypothesis, we evaluated the relationship between DE features and regulons. Firstly, we selected the top 1% TF-target interactions by importance score (Figure S4, Table S7), including 2,337 TF-target interactions constituting 241 regulons, which we referred to as the top 1% regulons. We counted the number of DE features (logFC > 1 and adjusted p-value < 0.05; FC: fold change) within these regulons for each treatment status. There were 442, 170, and 461 DE features in 'After-chemotherapy', 'Before-chemotherapy', and 'Control' groups, respectively (Table S5). Among them, 360 DE features (142 features in 'After-chemotherapy', 85 in 'Before-chemotherapy', and 133 in 'Control' group) were included in the 2,337 TF-target interactions (Table S8), showing that DE features significantly interacted with TFs ($p < 0.0001$, Figure S5A). There were 360 DE features included in the 135 regulons. Thirty-eight regulons (28.1%) had DE features in more than one treatment status, and 97 regulons had DE features in only one treatment status (Table S8), indicating that DE features in different treatment statuses were regulated by distinct TFs ($p = 0.0033$, Figure S5B). One notable example is an ELF3 regulon [denoted as ELF3_(+)]. This regulon consisted of *ELF3* as a TF and 42 TF-target interactions in the top 1% regulon, in which 23 features were DE features in the After-chemotherapy group. Collectively, these results indicated that DE features were closely connected to specific regulons.

Reported prognostic features in HGSOC did not show significant overlap with the regulons

Building on the previous studies that identified genes related to the prognosis of HGSOC, we evaluated the relationship between regulons and these genes. We collected 276 genes as prognostic marker genes, which have been previously validated through survival analysis in 3,769 women with HGSOC[35]. Fifty-seven genes were included in the top 1% 2,337 TF-target interactions (Table S9), which were not significantly related to TFs ($p = 0.66$, Figure S6). Among the 57 genes, four, one, and nine genes overlapped with the DE features in control, Before-chemotherapy, and After-chemotherapy, respectively. Seven out of the nine genes that overlapped between prognostic genes and DE features in the After-chemotherapy group were included in RUNX2_(+) regulon. These results suggested that some reported prognostic features were not regulated by TFs, and exploring treatment-status-specific regulons might facilitate further identification of prognostic features of HGSOC. This comparison also suggested our GRN approach may help identify novel cancer markers for further validation.

Machine-learning performance depends non-monotonically on feature set size

Before we examined whether regulons for feature extraction could improve ML performance for prognosis prediction, we evaluated the relationship between the number of features and ML performance. Bulk RNA-seq data and patients' clinical information from TCGA were collected and cleaned for this analysis[23]. From the TCGA data, we curated 256 HGSOC patients who lived for more than two years and 86 patients who died within two years from the initial diagnosis. Expression profiles of the RNA-seq data are presented in Figure S7A. We tested whether ML could predict patients' prognosis (live more than two years or not) using the bulk RNA-seq data. Considering the limited sample size in our dataset, we first applied logistic regression. To assess the relationship between the number of features and ML performance, we conducted a series of 100 permutation tests as ablation analyses at intervals of every 100 features, spanning the range from 100 to 60,000 features, resulting in a total of 60,000 permutations (Figure S7B). Consistent with the expectation, permutations of up to 10,000 features included both the top and worst performances. With this result, we concluded that feature selection had the potential to improve ML performance.

Treatment-status-specific regulons improve ML performance on prognosis prediction

To evaluate whether regulons improve ML performance, we first made three lists of the top 100 regulons in the treatment statuses based on the specificity scores (Table S10). Features were extracted from the lists using the following steps. First, we determined the number of regulons (N regulons) and features (M genes). Second, we extracted the top N regulons based on the specificity scores. Lastly, we extracted the top M genes based on the importance scores. This approach was iteratively applied across the range of 1 to 100 regulons and 1 to 100 genes, yielding a total of 10,000 combinations. We employed logistic regression first by using F1 score, AUC, and accuracy as performance measure.

Regardless of the lists used, the top 1% performance scores were achieved with up to one thousand features (Figure S8 and Table S11). The distribution patterns of regulon-gene combinations varied between treatment statuses. The top scores were achieved by using the combinations of several regulons with a few dozen genes in Before-chemotherapy specific regulons, or the combinations of a few dozen regulons with several genes in After-chemotherapy-specific regulons (Fig. 3A and Figure S9). As expected, the top 1% performance scores based on Before-chemotherapy and After-chemotherapy regulons outperformed those based on control samples (Fig. 3B and Figure S10), suggesting that the regulons activated in disease states and the treatment were associated with prognosis. Other ML models obtained comparable results, although they did not significantly outperform the logistic regression model, possibly due to the limited amount of data (Figures S11–S14). We compared the ML performances between regulon-based and DE feature-based methods. In the DE feature-based method, we extract the top N features based on the logFC scores or adjusted p-values. Interestingly, the regulon-based feature extraction method significantly outperformed the DE-feature-based feature extraction (Figs. 3C and S15). Taken together, regulon-based feature extraction could improve ML performance and outperformed DE-based approach.

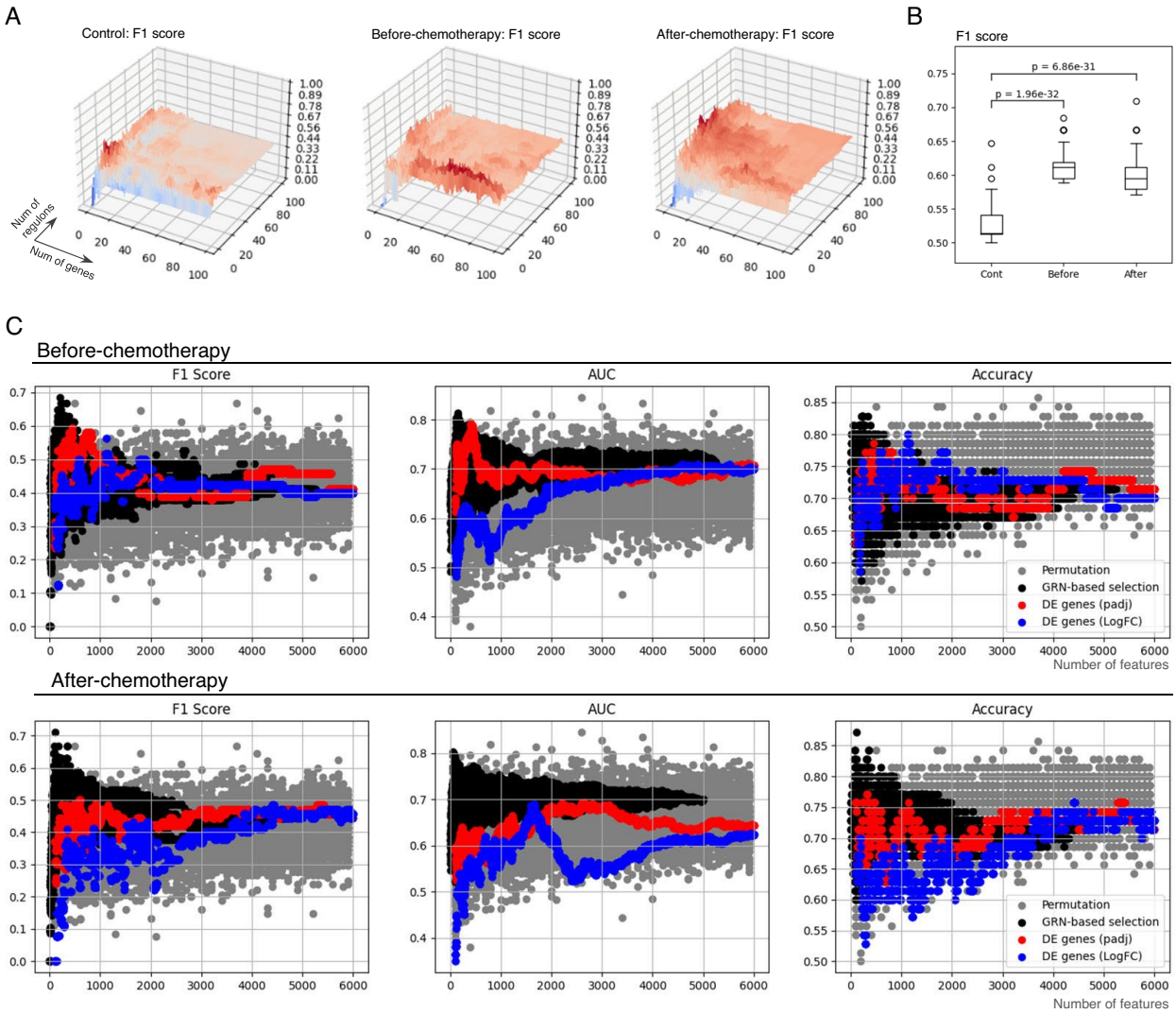

Fig. 3. Performance evaluation of gene regulatory network (GRN) and machine learning (ML) analysis. We used 80% and 20% of RNA-seq data of 342 HGSOC patients from the TCGA data for training and validation, respectively. (**A**) 3D plots of F1 scores calculated by logistic regression and the top 100 specific regulons in each treatment status. The x-axis and the y-axis represent the number of genes and the number of regulons, respectively. The z-axis represents F1 score. (**B**) Top 1% F1 scores in each treatment status. (**C**) The relationship between ML-performance and the number of features. The permutation test results (see Figure S7B) are shown as reference values.

Treatment-status-specific regulons showed overlapping and distinct cell functions

Because metacells in the Before-chemotherapy and After-chemotherapy groups had distinct regulon activity patterns (Fig. 2E), we hypothesized that extracted prognostic features using regulons in the Before-chemotherapy and After-chemotherapy groups had distinct cellular functions. To test it, we obtained the average numbers of regulons and genes among the top 1% scores in the two groups: 4 regulons and 70 genes in regulons specific to the Before-chemotherapy group, and 46 regulons and 5 genes in regulons specific to the After-chemotherapy group (Figure S16). Their TFs and genes are listed in Table S12. Enrichment analyses showed that they had shared and distinct functions (Fig. 4, Tables S13 and S14). Regulons specific to the Before-chemotherapy and After-chemotherapy groups were enriched with 105 and 45 GO terms, respectively. Among them, eight GO-terms were shared between them, which were related to cell leukocyte adhesion, interleukin-6 (IL-6) production, organism interferon production, and negative regulation of intracellular signal transduction. Six out of these eight shared GO-terms were related to cell leukocyte adhesion, which has been reported to affect the prognosis of ovarian cancer[36]. Furthermore, IL-6 is also associated with invasion and metastasis functions in ovarian cancer[37]. These results suggested that regulons in the Before-chemotherapy and After-chemotherapy groups had distinct cellular functions related to prognosis, although some functions are shared between the two groups.

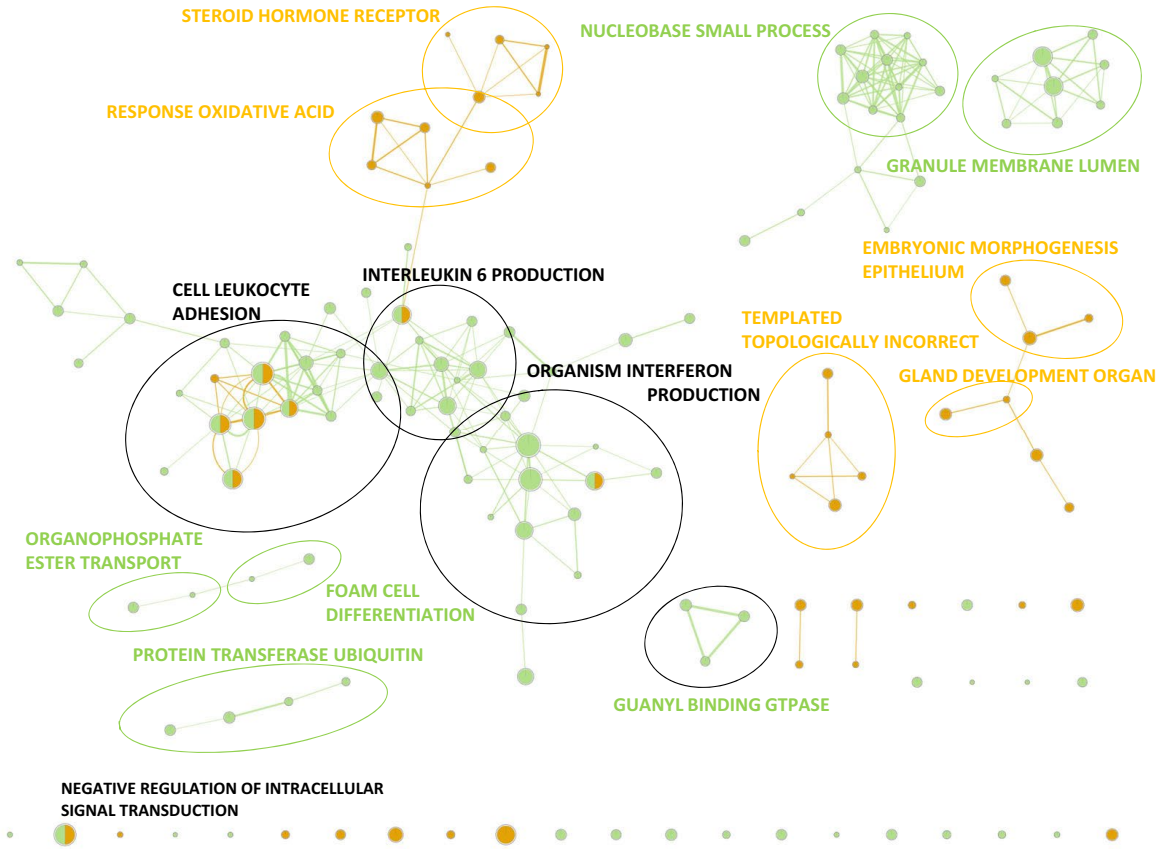

Fig. 4. Gene Ontology (GO) terms enriched with the prognostic features. Features obtained from the regulons in 'Before-chemotherapy' and 'After-chemotherapy' groups are labeled in orange and light green, respectively. The GO terms shared between the two groups are labeled in black. Node size reflects the number of features in each GO term, and edge width represents the number of shared features between the nodes.

Cell-type-specific GRNs did not generally outperform all-cell models

Our analyses thus far demonstrated that regulons improved ML performance for prognosis prediction. However, it was still unclear how those extracted features function in each cell type and how to prioritize the extracted features. To address this, we evaluated cell-type-specific GRN inference in each treatment status.

After extracting the top 100 regulons specifically activated in each treatment status under each cell type (Table S15), we performed GRN/ML performance analysis in different cell types (Figures S17 – S23). GRN inference in mast cells in the Before-chemotherapy group was not available because there were no metacells in this category. The top 1% F1 scores in each cell type under each treatment status are presented in Fig. 5A, B and Table S16. Within the Before-chemotherapy subgroups, no subgroup outperformed that with all metacells in the Before-chemotherapy group. Within the After-chemotherapy groups, only T-cell-specific GRN outperformed that of all metacells. With these results, we concluded that cell-type-specific regulons did not significantly improve ML performance.

To prioritize features for prognosis prediction, we explored the relation between cell-type-specific regulons and treatment-status-specific regulons. We counted the number of overlapped regulons specific to Before-chemotherapy group and After-chemotherapy group in each cell type, as well as in all metacells (Fig. 5C, and Tables S17 and S18). For example, 48.2% of treatment-status-specific regulons in total cells overlapped T cell-specific regulons in the Before-chemotherapy group, while 61.2% of treatment-status-specific regulons in total cells overlapped fibroblast-specific regulons in After-chemotherapy group. To identify key features that function in cell-type-specific manners, we extracted the features shared by Before-chemotherapy group and After-chemotherapy group in each cell type and also in all metacells (intersection of four sets in Fig. 5C, Table S19). Sorting nexin 8 encoded by gene *SNX8* was the only feature identified in more than half of the cell types. *SNX8* was identified in endothelial, epithelial, fibroblast, and myeloid/macrophage metacells. This finding was consistent with the expression patterns of *SNX8*, though myeloid/macrophage had the most abundant expression level (Figure S24A). Survival analysis with TCGA data of HGSOC suggested that the high expression of *SNX8* was associated with poor prognosis ($p = 0.0041$, Figure S24B). These results demonstrated that cell-type-specific regulons can be instrumental in prioritizing the features extracted from regulons for disease prognosis prediction.

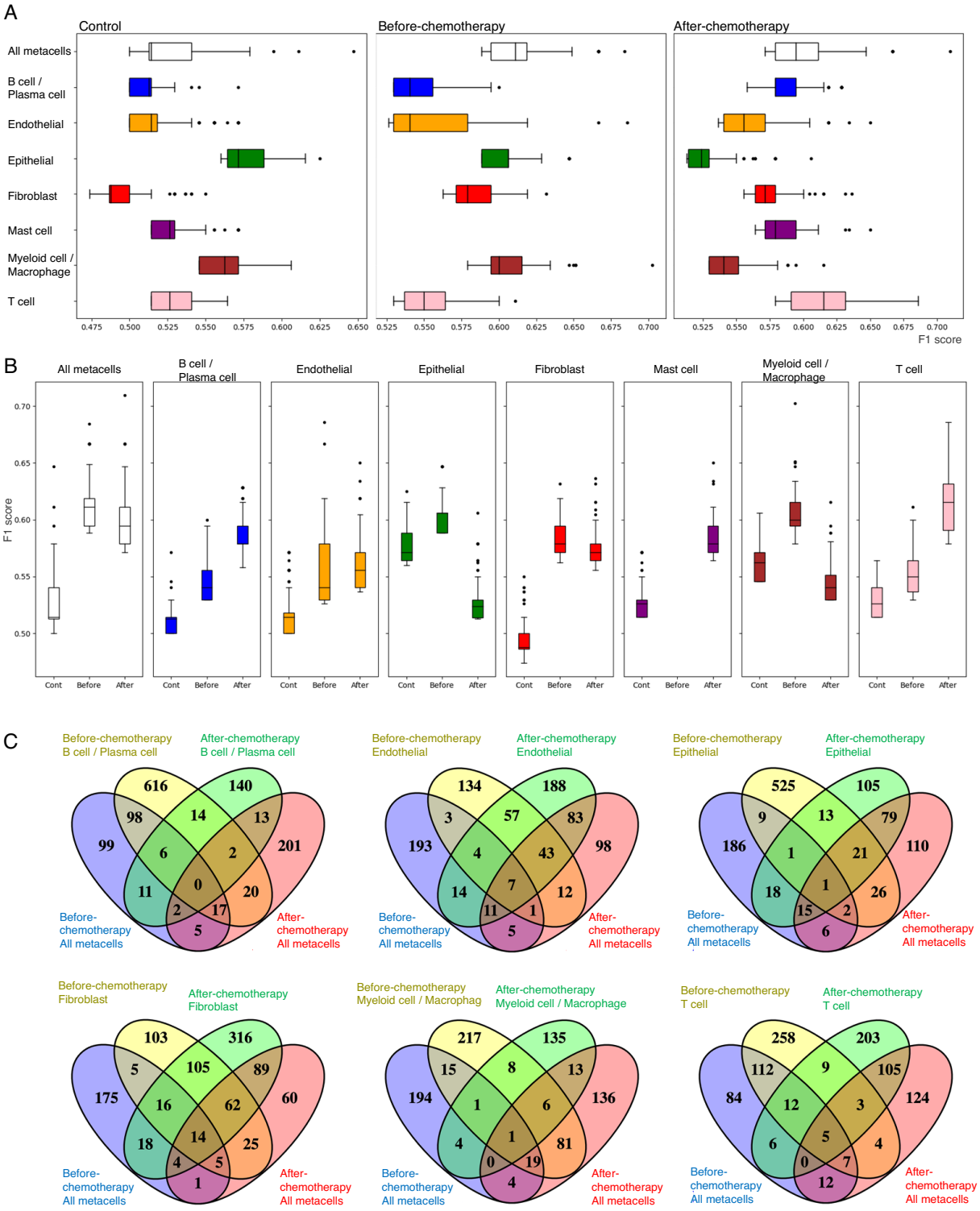

Fig. 5. Cell-type-specific gene regulatory network (GRN) inference in each treatment status. (**A**, **B**) Top 1% F1 scores obtained from each cell-type-specific GRN in each treatment status. Focusing on cell-type-specific regulons (colored boxes) did not achieve better performance than that by focusing on treatment-specific regulons, which encompassed all cell types. (**C**) Venn diagrams showing the shared features among each cell-type-specific GRN in each treatment status and regulons of the Before-chemotherapy and After-chemotherapy groups.

Discussion

In this study, we aimed to identify prognostic features in HGSOC with the application of GRN inference analysis. To do this, we integrated scRNA-seq data of HGSOC and constructed metacells, conducted GRN inference analysis, and tested whether regulons were utilized for prognostic feature extraction. Our findings indicate that regulons were more effective in extracting prognostic features in HGSOC compared to approaches that primarily focus on DE features. The use of paired data, encompassing samples before and After chemotherapy, helped extract distinct prognostic features. By utilizing cell-type-specific regulons, we prioritized the extracted features and identified a prognostic role for *SNX8* has been scarcely reported in HGSOC. Our approach can be used for other diseases with different scRNA-sequencing analysis workflows and ML algorithms.

Our results demonstrated that GRN inference analysis can be incorporated into clinical research or translational fields by identifying prognostic features, thereby enhancing ML performance for tasks related to clinical outcomes. GRN inference has been employed in various applications, such as identifying hub TFs for specific cellular functions, comparing GRNs between different conditions or cell types, and conducting in silico perturbation to infer the TFs crucial for cell development[5]. However, they are mostly applicable to basic research fields, while a few studies have addressed GRN inference for tasks related to clinical applications[38,39]. Our workflow expands the field to better explore clinical management through the GRN inference analysis to extract prognostic features of human diseases by leveraging the biological information from both scRNA-seq and bulk RNA-seq datasets.

One challenge in GRN inference using scRNA-seq is the computational cost due to the sparsity of single-cell data. It becomes more problematic when attempting to integrate many datasets, or when computational resources are limited. On the other hand, the integration of multiple datasets for GRN inference will provide a better understanding of diseases due to diverse cell types, such as cancers, which have strong tumor microenvironments. To address this, we performed GRN inference analysis using metacells constructed by SEACells[20]. Our results demonstrated that metacell-based GRN inference outperformed DE feature-based method in extracting prognostic features. This suggests that employing metacell construction could be a viable option, particularly when the computational cost is a concern.

Our results support the idea that targeting transcription factors for cancer treatment is feasible, although it is generally considered promising yet challenging. Some transcription factors are considered potential therapeutic targets in HGSOC, including YBX1[40], whose regulon was activated specifically in HGSOC Before-chemotherapy. Considering that the employment of a number of regulons improves the performance of predicting prognosis, targeting multiple transcription factors in HGSOC therapy might offer a promising strategy to disrupt broader tumor processes than single-target therapies, enabling tailor-made treatments.

Focusing on treatment status-specific and cell-type-specific regulons facilitated the identification of cellular functions associated with HGSOC prognosis Before-chemotherapy and After-chemotherapy, as well as the prioritization of the extracted features. We successfully identified *SNX8* as a prognosis factor for HGSOC. SNX8 is a member of the sorting nexin family proteins, which are classified into seven subtypes based on their functional domains[41]. SNX8 is known to have several functions, including endosome-to-Golgi transport[42] and the modulation of the innate immune response[43–45]. A few studies have reported that *SNX8* was related to human diseases, including nonalcoholic fatty liver disease[46], Alzheimer's disease[47], neuropathic pain[48], and neurodevelopmental delay[49]. However, to our knowledge, only one report has mentioned a potential connection between *SNX8* and HGSOC[50]. Our investigation suggests that the observed correlation between elevated *SNX8* expression and a poor prognosis in HGSOC may be attributed to several factors. First, SNX8 protein has the capability to activate IKKβ[45], a pivotal kinase in oncogenic NF-KB activation, as well as in the signaling pathways of mTORC1 and FOXO3a.[51–53] The activation of NF-κB and mTORC1, coupled with the inactivation of FOXO3a, could contribute to the progression, migration, and metastasis of HGSOC. Second, *SNX8* exhibits high expression in immune cells, particularly those associated with myeloid/macrophages, resulting in activated IKKβ/NF-κB signaling that produces numerous pro-tumorigenic factors, thereby facilitating the development of HGSOC. Therefore, Therapeutic targeting of the SNX8 protein may hold potential in the treatment of HGSOC.

This study has several limitations. First, the GRN inference analysis was conducted with only scRNA-seq data, leaving uncertainty regarding the potential enhancement of ML performance by incorporating single-cell multiome data. Single-cell multiome assay can examine both gene expression and regulation (often TF regulation) in one experiment which avoids the batch effect. It is especially useful for GRN inference[54]. Second, the enhancements observed in ML performance have not yet reached a level considered satisfactory for clinical application. This limitation may result from the relatively small sample size of HGSOC individuals (n = 342) used in this analysis and the insufficiency of detailed clinical information. We failed to make a validation cohort from other studies because they lack clinical information, including the patients' longevity data. More bulk RNA-seq data with detailed clinical information are warranted. Lastly, our results warrant further experimental validation. The underlying pathophysiology of the identified prognostic features, including the novel candidate *SNX8*, in the microenvironment of HGSOC remains unclear. Due to the scope of this bioinformatics project, we will extend this work for future validation.

In summary, we demonstrated that GRN inference analysis was used for prognostic feature extraction. We identified a novel prognostic gene, *SNX8*, by leveraging treatment status and cell-type-specific information. Our framework is applicable not only to cancer but to other diseases when both scRNA-seq data for GRN inference and bulk RNA-seq data with clinical outcomes are available. Future studies should prioritize the integration of additional modalities, such as single-cell multiome data, other ML models like deep neural networks with hyperparameter tuning, employing methods to analyze differentially expressed features while taking care of patient heterogeneity, and multi-modal clinical data (such as single-cell data, several clinical outcomes, and detailed clinical background). We also expect larger sample sizes to enable more sophisticated analyses,

including regression models, using survival-specific methods (e.g., Cox-based ML, survival-XGBoost/SVM) to take censored outcomes into account.

Data availability

We used five public datasets (GEO: GSE165897, GSE191301, GSE201047, GSE151214, and GSE184880). The metacell and regulon data generated from the five datasets are available upon request. Interested researchers should contact the corresponding author for access.

Code availability

All R and Python scripts to obtain the results presented in this manuscript are available on https://github.com/bsml320/HGSOC.

Received: 7 May 2025; Accepted: 3 October 2025
Published online: 10 November 2025

References

1. Board, P. D. Q. A. T. E. *PDQ Cancer Information Summaries* (National Cancer Institute, 2002).
2. Reid, B. M., Permuth, J. B. & Sellers, T. A. Epidemiology of ovarian cancer: a review. *Cancer Biol. Med.* **14**, 9–32. https://doi.org/10.20892/j.issn.2095-3941.2016.0084 (2017).
3. Bowtell, D. D. et al. Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat. Rev. Cancer* **15**, 668–679. https://doi.org/10.1038/nrc4019 (2015).
4. Macneil, L. T. & Walhout, A. J. Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Res.* **21**, 645–657. https://doi.org/10.1101/gr.097378.109 (2011).
5. Badia, I. M. P. et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754. https://doi.org/10.1038/s41576-023-00618-5 (2023).
6. Madhamshettiwar, P. B., Maetschke, S. R., Davis, M. J., Reverter, A. & Ragan, M. A. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. *Genome Med.* **4**, 41. https://doi.org/10.1186/gm340 (2012).
7. Lawrenson, K. et al. A Study of High-Grade Serous Ovarian Cancer Origins Implicates the SOX18 Transcription Factor in Tumor Development. *Cell Rep.* https://doi.org/10.1016/j.celrep.2019.10.122 (2019).
8. Di Palma, T. & Zannini, M. PAX8 as a Potential Target for Ovarian Cancer: What We Know so Far. *Onco. Targets Ther.* **15**, 1273–1280. https://doi.org/10.2147/OTT.S361511 (2022).
9. Li, H. et al. Development of a novel transcription factors-related prognostic signature for serous ovarian cancer. *Sci. Rep.* **11**, 7207. https://doi.org/10.1038/s41598-021-86294-z (2021).
10. Barger, C. J. et al. Co-regulation and function of FOXM1/RHNO1 bidirectional genes in cancer. *Elife* https://doi.org/10.7554/eLife.55070 (2021).
11. Zhang, K. et al. Longitudinal single-cell RNA-seq analysis reveals stress-promoted chemoresistance in metastatic ovarian cancer. *Sci. Adv.* https://doi.org/10.1126/sciadv.abm1831 (2022).
12. Loret, N. et al. Distinct Transcriptional Programs in Ascitic and Solid Cancer Cells Induce Different Responses to Chemotherapy in High-Grade Serous Ovarian Cancer. *Mol. Cancer Res.* **20**, 1532–1547. https://doi.org/10.1158/1541-7786.MCR-21-0565 (2022).
13. Shen, Y. et al. The impact of neoadjuvant chemotherapy on the tumor microenvironment in advanced high-grade serous carcinoma. *Oncogenesis* **11**, 43. https://doi.org/10.1038/s41389-022-00419-1 (2022).
14. Dinh, H. Q. et al. Single-cell transcriptomics identifies gene expression networks driving differentiation and tumorigenesis in the human fallopian tube. *Cell Rep.* https://doi.org/10.1016/j.celrep.2021.108978 (2021).
15. Xu, J. et al. Single-Cell RNA Sequencing Reveals the Tissue Architecture in Human High-Grade Serous Ovarian Cancer. *Clin. Cancer Res.* **28**, 3590–3602. https://doi.org/10.1158/1078-0432.CCR-22-0296 (2022).
16. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15. https://doi.org/10.1186/s13059-017-1382-0 (2018).
17. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166. https://doi.org/10.1038/s41587-021-01206-w (2022).
18. Hu, C. et al. Cell Marker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkac947 (2023).
19. Vazquez-Garcia, I. et al. Ovarian cancer mutational processes drive site-specific immune evasion. *Nature* **612**, 778–786. https://doi.org/10.1038/s41586-022-05496-1 (2022).
20. Persad, S. et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757. https://doi.org/10.1038/s41587-023-01716-9 (2023).
21. Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276. https://doi.org/10.1038/s41596-020-0336-2 (2020).
22. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504. https://doi.org/10.1101/gr.1239303 (2003).
23. Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. Nature 474, 609-615 https://doi.org/10.1038/nature10166 (2011).
24. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv1507 (2016).
25. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205. https://doi.org/10.1093/nar/gkz401 (2019).
26. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. https://doi.org/10.1186/s13059-014-0550-8 (2014).
27. Morgan, M., Obenchain, V., Hester, J. & Pagès, H. SummarizedExperiment: SummarizedExperiment container. https://bioconductor.org/packages/SummarizedExperiment https://doi.org/10.18129/B9.bioc.SummarizedExperiment, (2022).
28. Therneau, T. A Package for Survival Analysis in R. R package version 3.5–7. https://CRAN.R-project.org/package=survival (2023).
29. Kassambara, A., Kosinski, M. & Biecek, P. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.9. https://CRAN.R-project.org/package=survminer (2021).
30. Antony, F. et al. High-throughput assessment of the antibody profile in ovarian cancer ascitic fluids. *Oncoimmunology* https://doi.org/10.1080/2162402X.2019.1614856 (2019).
31. Dou, R., Wang, X. & Zhang, J. Prognostic Value and Immune Infiltration Analysis of Nuclear Factor Erythroid-2 Family Members in Ovarian Cancer. *Biomed. Res. Int.* https://doi.org/10.1155/2022/8672258 (2022).

32. Chen, A. et al. ATF5 is overexpressed in epithelial ovarian carcinomas and interference with its function increases apoptosis through the downregulation of Bcl-2 in SKOV-3 cells. *Int. J. Gynecol. Pathol.* **31**, 532–537. https://doi.org/10.1097/PGP.0b013e31824df26b (2012).
33. Dausinas, P. et al. ARID3A and ARID3B induce stem promoting pathways in ovarian cancer cells. *Gene* https://doi.org/10.1016/j.gene.2020.144458 (2020).
34. Tailor, D. et al. Y box binding protein 1 inhibition as a targeted therapy for ovarian cancer. *Cell Chem. Biol.* https://doi.org/10.1016/j.chembiol.2021.02.014 (2021).
35. Millstein, J. et al. Prognostic gene expression signature for high-grade serous ovarian cancer. *Ann. Oncol.* **31**, 1240–1250. https://doi.org/10.1016/j.annonc.2020.05.019 (2020).
36. Mezzanzanica, D. et al. Subcellular localization of activated leukocyte cell adhesion molecule is a molecular predictor of survival in ovarian carcinoma patients. *Clin. Cancer Res.* **14**, 1726–1733. https://doi.org/10.1158/1078-0432.CCR-07-0428 (2008).
37. Browning, L., Patel, M. R., Horvath, E. B., Tawara, K. & Jorcyk, C. L. IL-6 and ovarian cancer: inflammatory cytokines in promotion of metastasis. *Cancer Manag. Res.* **10**, 6685–6693. https://doi.org/10.2147/CMAR.S179189 (2018).
38. Xiong, Z. et al. Integrated Analysis of scRNA-Seq and Bulk RNA-Seq Reveals Metabolic Reprogramming of Liver Cancer and Establishes a Prognostic Risk Model. *Genes (Basel)* https://doi.org/10.3390/genes15060755 (2024).
39. Gong, X. et al. Integrated Analysis of Single-Cell and Bulk RNA-Seq Data reveals that Ferroptosis-Related Genes Mediated the Tumor Microenvironment predicts Prognosis, and guides Drug Selection in Triple-Negative Breast Cancer. *Biorxiv* https://doi.org/10.1101/2024.07.04.602021 (2025).
40. Meng, H. et al. YBX1 promotes homologous recombination and resistance to platinum-induced stress in ovarian cancer by recognizing m5C modification. *Cancer Lett.* https://doi.org/10.1016/j.canlet.2024.217064 (2024).
41. Hanley, S. E. & Cooper, K. F. Sorting Nexins in Protein Homeostasis. *Cells* https://doi.org/10.3390/cells10010017 (2020).
42. Dyve, A. B., Bergan, J., Utskarpen, A. & Sandvig, K. Sorting nexin 8 regulates endosome-to-Golgi transport. *Biochem. Biophys. Res. Commun.* **390**, 109–114. https://doi.org/10.1016/j.bbrc.2009.09.076 (2009).
43. Guo, W. et al. SNX8 modulates the innate immune response to RNA viruses by regulating the aggregation of VISA. *Cell Mol. Immunol.* **17**, 1126–1135. https://doi.org/10.1038/s41423-019-0285-2 (2020).
44. Wei, J. et al. SNX8 modulates innate immune response to DNA virus by mediating trafficking and activation of MITA. *PLoS Pathog.* **14**, e1007336. https://doi.org/10.1371/journal.ppat.1007336 (2018).
45. Wei, J. et al. SNX8 mediates IFNgamma-triggered noncanonical signaling pathway and host defense against Listeria monocytogenes. *Proc. Natl. Acad. Sci. U S A* **114**, 13000–13005. https://doi.org/10.1073/pnas.1713462114 (2017).
46. Hu, Y. et al. Fatty Acid Synthase-Suppressor Screening Identifies Sorting Nexin 8 as a Therapeutic Target for NAFLD. *Hepatology* **74**, 2508–2525. https://doi.org/10.1002/hep.32045 (2021).
47. Rosenthal, S. L. et al. Beta-amyloid toxicity modifier genes and the risk of Alzheimer's disease. *Am. J. Neurodegener. Dis.* **1**, 191–198 (2012).
48. Reyes-Gibby, C. C. et al. Genome-wide association study identifies genes associated with neuropathy in patients with head and neck cancer. *Sci. Rep.* **8**, 8789. https://doi.org/10.1038/s41598-018-27070-4 (2018).
49. Mastromoro, G. *et al.* Small 7p22.3 microdeletion: Case report of Snx8 haploinsufficiency and neurological findings. *Eur J Med Genet* **63**, 103772 (2020). https://doi.org/10.1016/j.ejmg.2019.103772
50. Lisowska, K. M. et al. Gene expression analysis in ovarian cancer - faults and hints from DNA microarray study. *Front Oncol.* **4**, 6. https://doi.org/10.3389/fonc.2014.00006 (2014).
51. Lee, D. F. et al. IKK beta suppression of TSC1 links inflammation and tumor angiogenesis via the mTOR pathway. *Cell* **130**, 440–455. https://doi.org/10.1016/j.cell.2007.05.058 (2007).
52. Hu, M. C. et al. IkappaB kinase promotes tumorigenesis through inhibition of forkhead FOXO3a. *Cell* **117**, 225–237. https://doi.org/10.1016/s0092-8674(04)00302-2 (2004).
53. Lee, D. F. & Hung, M. C. Advances in targeting IKK and IKK-related kinases for cancer therapy. *Clin. Cancer Res.* **14**, 5656–5662. https://doi.org/10.1158/1078-0432.CCR-08-0123 (2008).
54. Yan, F. et al. Single-cell multiomics decodes regulatory programs for mouse secondary palate development. *Nat. Commun.* **15**, 821. https://doi.org/10.1038/s41467-024-45199-x (2024).

Acknowledgements

We thank lab members of the Bioinformatics and Systems Medicine Laboratory (BSML) for the valuable discussion. T.I. and W.L. are CPRIT Postdoctoral Fellow and Predoctoral Fellow, respectively, in the Biomedical Informatics, Genomics, and Translational Cancer Research Training Program (BIG-TCR) funded by Cancer Prevention & Research Institute of Texas (CPRIT, RP210045). ZZ was partially supported by National Institutes of Health grants (R01CA276513 and R01LM012806). D.F.L. was partially supported by NIH R01CA246130. We thank the technical support from the CPRIT-funded Cancer Genomics Core (RP240610). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

T.I.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Visualization. Y. D.: Validation, Writing – Review & Editing W. L.: Validation, Writing – Review & Editing. D.F.L.: Writing—Original Draft, Writing – Review & Editing, Supervision, Project administration, Funding acquisition. Z.Z.: Conceptualization, Methodology, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

Declarations

Competing interests

Zhongming Zhao serves as an associate editor for Scientific Reports. The other authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-22937-9.

Correspondence and requests for materials should be addressed to Z.Z.