



OPEN Aggressive serous ovarian cancer subtype defined by high centrality lncRNA profiles and master transcription factors

Seonhyang Jeong^{1,3,4}, Young Suk Jo^{1,4}, Sunmi Park¹, Hwayoung Lee², Eun Gyeong Park^{2,3}, Sang Geun Jung³✉ & Jandee Lee²✉

Long non-coding RNAs (lncRNAs) regulate the progression and metastasis of high-grade serous carcinoma ovarian cancer (HGSC). However, HGSC is yet to be classified based on these transcripts. In addition, the crosstalk between master transcriptional factors (MTFs) and lncRNAs remains unclear. Therefore, we aimed to classify HGSC based on lncRNA expression and identify the integrated MTFs for highly correlated mRNAs and lncRNAs. Unsupervised clustering was conducted using highly expressed lncRNAs derived from 367 HGSC samples obtained from The Cancer Genome Atlas. DNA mutations, somatic copy number alterations, microRNA expression, and DNA methylome were analyzed to identify the genetic and epigenetic factors affecting unsupervised clustering. Multiple Sample Virtual Inference of Protein-activity by Enriched Regulon analysis (msViper) was conducted to identify transcription factors simultaneously exhibiting positive correlation with lncRNAs and mRNAs in each cluster. In vitro analyses were performed to determine if these lncRNAs regulate both the MTFs and target genes. Functional analysis enabled the lncRNA-based classification of HGSC into five groups: "Immune," "EMT," "Estrogen response," "EMT-Androgen response," and "Differentiation" groups. The EMT-Androgen response group showed poor prognosis in the oncologic outcome. Of the transcription factors selected in this group, three MTFs with the highest eigenvector centrality scores were identified (MSC, AEBP1, CREB3L1). However, seven lncRNAs exerted a higher centrality than the selected MTFs. Our results suggest that HGSC can be classified based on lncRNA expression and characterized using molecular features. Therefore, lncRNAs and MTFs may synergistically contribute to molecular features of HGSC that could be indicators for personalized medicine.

Keywords Ovarian cancer, High-grade serous ovarian cancer, Long Noncoding RNA, Biomarker, Multi-omics data

Ovarian cancer (OVCA) is an aggressive gynecological malignancy that is responsible for >130,000 cancer-related deaths worldwide. The overall 5-year survival rate is approximately 40% after surgery and systemic chemotherapy. Epithelial ovarian cancer (EOC) is composed of five histological subtypes: high-grade serous, low-grade serous, mucinous, clear-cell, and endometrioid. High-grade serous carcinoma (HGSC), the most common histological subtype, constitutes 70% of EOC cases. HGSC is typically diagnosed at an advanced stage, when the tumor has spread to the abdomen or outside the abdominal cavity, because it has no specific symptoms^{1,2}. This underscores the need for new diagnostic and therapeutic targets.

Long non-coding RNAs (lncRNAs) are transcripts that were identified in genomic studies in the late 1990s and 2000s. They are longer than 200 nucleotides and are presumed not to encode proteins. lncRNAs may regulate gene expression through chromatin remodeling, transcription, alternative splicing, and generation of microRNAs (miRNAs) or short biologically active peptides^{3,4}. Several lncRNAs affect the biological behavior of

¹Department of Internal Medicine, Yonsei University College of Medicine, Seoul, South Korea. ²Department of Surgery, Open NBI Convergence Technology Research Laboratory, Severance Hospital, Yonsei Cancer Center, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, South Korea. ³Department of Gynecological Oncology, Comprehensive Gynecologic Cancer Center, Bundang CHA Medical Center, CHA University, 59, Yatap-ro, Bundang-gu, Seongnam, Gyeonggi-do 13496, South Korea. ⁴Seonhyang Jeong and Young Suk Jo have contributed equally to this work. ✉email: sgoncol@chamc.co.kr; jandee@yuhs.ac

ovarian cancer, such as ANRIL, CCAT1, FAL1, H19, HOTAIR, and MALAT1^{5,6}. However, OVCA is yet to be classified based on lncRNAs.

Transcription factors and their regulatory networks govern gene expression programs that drive essential cellular processes and contribute to disease development. Master transcription factors (MTFs) often sit at the top of these regulatory hierarchies, directing the activity of other transcription factors and playing key roles in defining cell identity^{7,8}. Similar to other RNAs, lncRNAs are likely to be regulated by transcription factors. Conversely, lncRNAs can regulate the function of transcription factors. However, the crosstalk between MTF and lncRNAs has not been fully elucidated.

Although lncRNAs and transcription factors are known to regulate each other, the crosstalk between master transcription factors (MTFs) and lncRNAs in ovarian cancer remains poorly understood. Despite their emerging biological importance, lncRNAs have not yet been utilized for the molecular classification of high-grade serous ovarian cancer (HGSC), the most common and aggressive subtype of epithelial ovarian cancer. Previous stratification efforts have predominantly focused on protein-coding genes, potentially overlooking regulatory programs governed by non-coding elements. Since lncRNAs may reflect distinct transcriptional and epigenetic regulatory mechanisms, we hypothesized that lncRNA-based clustering could uncover novel HGSC subtypes that are not captured by coding gene expression alone.

In this study, we aimed to classify HGSC through unsupervised clustering based on lncRNA expression. We further aimed to identify the MTF representing each group. Finally, we investigated the interrelationship between lncRNA and MTF in the most clinically aggressive group. Our findings advance the understanding of how lncRNAs and MTFs cooperatively influence gene regulatory programs associated with tumor progression and aggressiveness in HGSC.

Results

HGSC is classified into five clusters based on lncRNA expression

The lncRNAs and mRNAs were extracted from total RNA-seq data (TCGA OVCA) based on GENCODE and HUGO probemap, respectively (Fig. 1A). To determine the optimal number of clusters (k) in our cNMF analysis, we evaluated clustering results for k values ranging from 2 to 6 using three widely accepted validation metrics: cophenetic correlation coefficient, silhouette score, and total within sum of squares (WSS). For each k , consensus matrices and sample correlation matrices were generated, and clustering with $k=5$ yielded the most consistent and stable patterns in both matrices (Fig. S1A). The cophenetic coefficient remained consistently high across all k values tested, suggesting overall robustness of clustering. Meanwhile, the average silhouette width indicated optimal clustering at $k=2, 4, 5$, and 6 , implying multiple plausible solutions. In contrast, WSS analysis suggested $k=3$ as optimal due to minimized intra-cluster variance (Fig. S1B). Taken together, and prioritizing cluster stability and interpretability, we selected $k=5$ as the most biologically and computationally appropriate solution.

However, clustering performance metrics alone do not capture clinical relevance. Since the primary goal of clustering in this study was to stratify patients based on prognosis, we additionally performed survival analyses across different k -factors to evaluate whether the identified clusters reflected meaningful differences in patient outcomes. survival analyses for $k=2$ to 4 showed no significant prognostic differences (Fig. S1C-D), and $k=6$ lacked consistent support across validation metrics. In contrast, $k=5$ demonstrated the most significant prognostic separation, with distinct OS and DFS differences between clusters (OS log-rank $p=0.034$, DFS log-rank $p=0.023$; Fig. 1B). The number of patients per cluster was as follows: J1, $n=92$; J2, $n=66$; J3, $n=86$; J4, $n=57$; and J5, $n=66$. The number of lncRNAs in the J2 cluster was more than twice that of the other clusters (J1 cluster, $n=284$; J2 cluster, $n=536$; J3 cluster, $n=297$; J4 cluster, $n=212$; J5 cluster, $n=171$; Fig. 1C). The expression status of the 1,500 selected mRNAs was evaluated according to the five lncRNA clusters (Fig. 1D). Similarly, the number of mRNAs in the J2 cluster was twice as high as that in other clusters (J1 cluster, $n=331$; J2 cluster, $n=577$; J3 cluster, $n=208$; J4 cluster, $n=201$; and J5 cluster, $n=183$). In summary, HGSC was classified into five clusters according to lncRNA expression. The J4 cluster exhibited poor survival whereas the J2 cluster harbored numerous lncRNAs and mRNAs.

The five HGSC clusters showed group-specific function

The function of lncRNAs in each cluster was predicted using the funcpred analysis tool. The lncRNAs in all five clusters were commonly related to spermatogenesis (Fig. S2A-E). In contrast, each cluster was associated with specific gene sets related to cluster-specific lncRNAs (Fig. 2). The J1 cluster was associated with immune-related gene sets such as interferon alpha/gamma and inflammatory responses. The J2 cluster was related to epithelial to mesenchymal transition (EMT), whereas the J3 cluster was related to early and late estrogen response. In the J4 cluster, apical junction, EMT, adipogenesis, myogenesis, hypoxia, and androgen response were predicted as the function of J4-specific lncRNAs. Although EMT was the common function of lncRNAs between J2 and J4 clusters, J4 was predicted to have diverse lncRNA function compared to J2. In contrast, J5 was related to bile acid metabolism and peroxisomes.

We further performed functional analysis using mRNAs that exhibited upregulated expression in each cluster. As shown in Fig. S3A-E, functional prediction results similar to those of lncRNAs were observed. For example, the J1 cluster showed interferon alpha/gamma and inflammatory responses. However, functional prediction using lncRNAs did not identify allograft rejection or upregulation of KRAS signaling. In addition to the EMT gene set, mRNAs with upregulated expression in the J2 cluster were related to KRAS signaling. The J3 cluster mRNAs were related to estrogen-responsive and immune-related genes such as those related to inflammatory response and TNFalpha signaling. The J4 cluster showed identical results to those observed during lncRNA functional prediction. In contrast, the downregulation of KRAS signaling and late estrogen response were related to mRNA function in the J5 cluster. This result was different from that observed in lncRNA functional prediction.

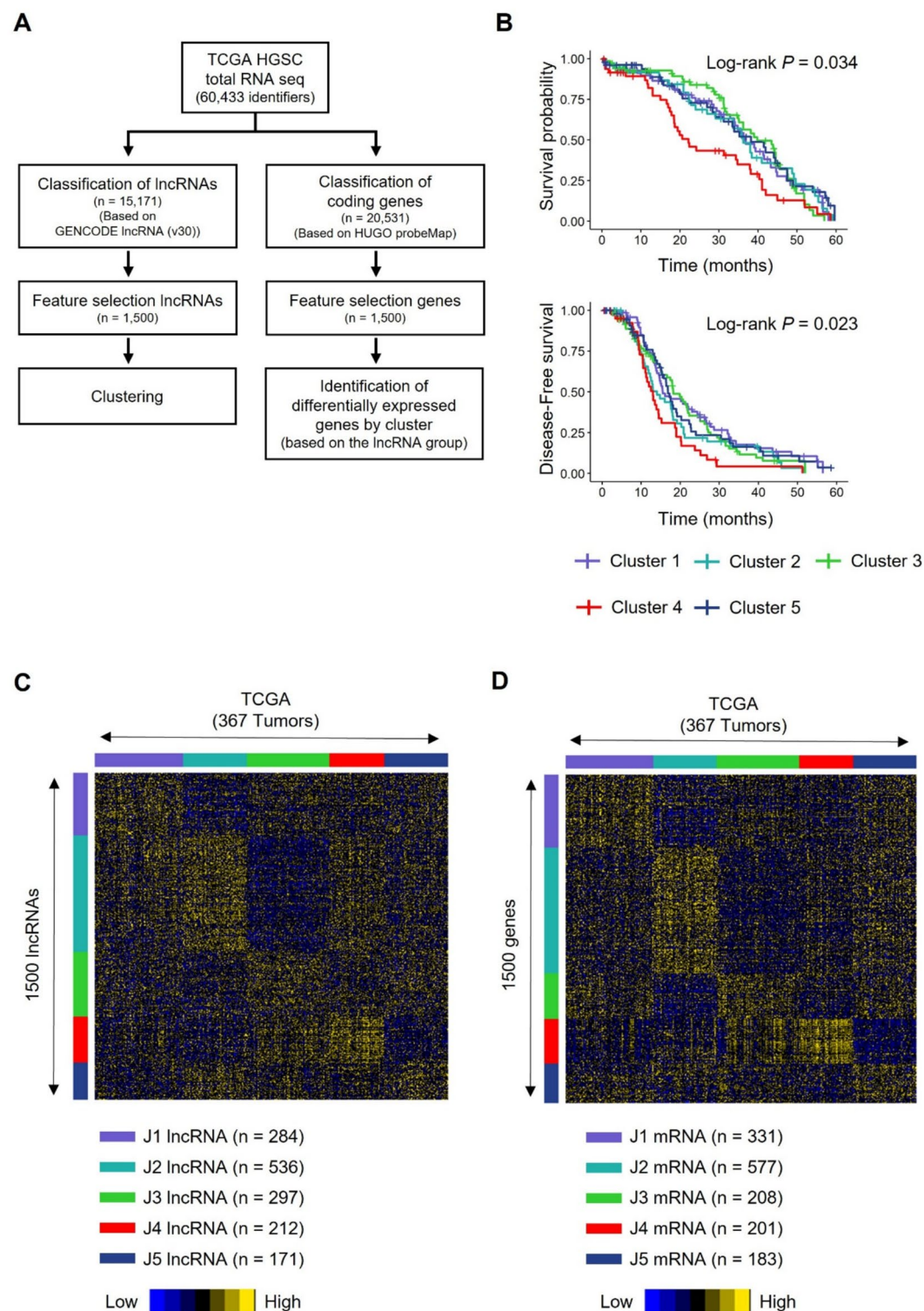


Fig. 1. Unsupervised clustering of TCGA HGSC based on lncRNA expression. (A) Schematic flow sheet of clustering in this study. (B) Comparison of overall and disease-free survival, as determined according to unsupervised clustering based on lncRNA expression. (C) Unsupervised clustering of 367 samples from TCGA HGSC based on lncRNA expression using total RNA-seq data from TCGA HGSC. (D) Heatmap of mRNA expression from TCGA HGSC, as determined using unsupervised clustering based on lncRNA expression.

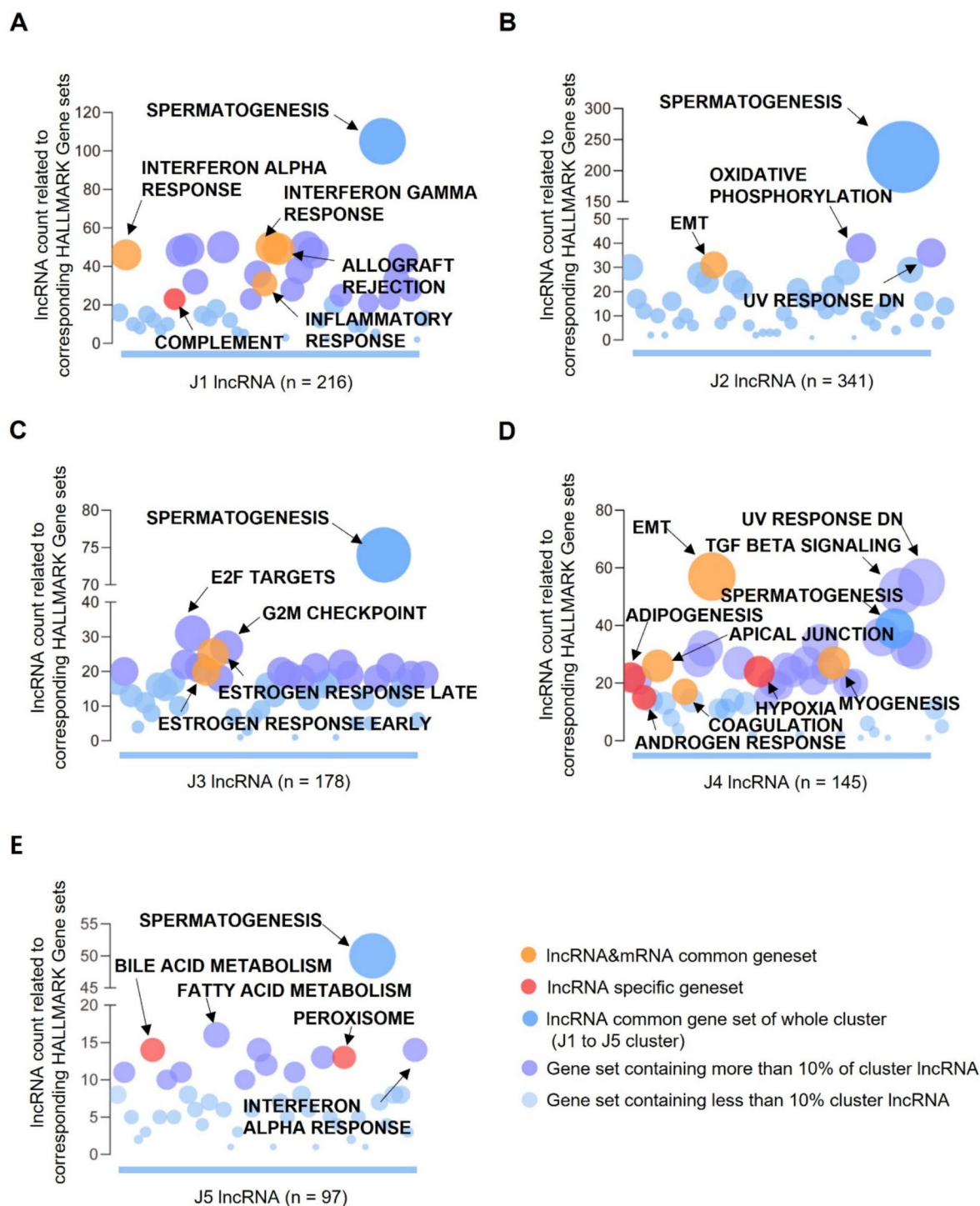


Fig. 2. Functional prediction of lncRNA and mRNA for five clusters. (A–E) Results of the prediction of lncRNA (red)-related gene sets, including simultaneously predicted gene sets of lncRNA and mRNA (orange) from the J1 to the J5 clusters, respectively. Sky blue circles indicate gene sets functionally predicted by all five clusters. Purple and light blue circles indicate gene sets containing 10% and < 10% lncRNA in each cluster, respectively.

The five groups were named as follows based on the lncRNA and mRNA functional prediction results: J1, Immune group; J2, EMT group; J3, Estrogen response group; J4, EMT-androgen response group; J5, Differentiated group. Although we only identified a small number of gene sets in the J5 group, the upregulated lncRNAs were similar to those reported in the differentiated group through TCGA OVCA analysis (Fig. S4 and Table. S2).

Representative transcription factors of the five groups were individually identified

To investigate the characteristics of groups classified based on lncRNA expression patterns, mutation profiles, miRNA expression patterns, and DNA methylation were analyzed (Fig. S5–6, Table S3–4). However, no differences were observed between the groups. This suggests that the differences in expression observed across the lncRNA groups are caused by factors other than epigenetic changes. Therefore, we identified transcription factors as one of the factors regulating the expression of large gene groups. First, we selected mRNAs whose expression showed a positive correlation with lncRNA expression in each group to identify transcription factors in the lncRNA-based groups (Fig. 3A, Fig. S7A–B). Since the transcription factors regulating lncRNA expression are largely unknown, we identified specific transcription factors for each group using these highly correlated mRNAs. The number of lncRNAs and mRNAs in the J2 group was more than twice as high as that in the other groups. However, the proportion of positively correlated lncRNAs and mRNAs was higher in the J4 group than in the other groups, suggesting a stronger coordination between their upregulation in this group.

Transcription factor inference analysis was conducted using msViper to identify unique transcription factors for each group. Three transcription factors were identified in J1: CDKN2A, ETV7, and IFI27 (Fig. 3B, Table S5). Twenty-one transcription factors were identified in J2, including DACH1, DACT2, and ZNF423 (Fig. 3C and Table S6). Moreover, the transcription factors identified in J3 were: RUNX1, TFAP2C, and TRPS1 (Fig. 3D and Table S7). RUNX2 was identified as a representative transcription factor of J4. Similarly, ZFH4, SNAI2, TWIST1, and TWIST2 were related to J4 (Fig. 3E and Table S8). Furthermore, the transcription factor EHF was identified in the J5 cluster (Fig. 3F and Table S9).

We identified lncRNAs and mRNAs whose expression showed a positive correlation with that of known transcriptional target genes of representative transcription factors in the five groups. In the J1 cluster, 46.0% of lncRNAs had a positive correlation with transcription factor target genes: J2, 30.2%; J3, 18.8%; J4, 88.7%; and J5, 7.4% (Fig. S7C). Furthermore, 27.9%, 11.5%, 17.8%, 69.2%, and 7.5% of mRNAs in the J1, J2, J3, J4, and J5 clusters, respectively, showed a positive correlation (Fig. S7D). Based on this analysis, we hypothesized that MTFs drive the coordinated upregulation of lncRNAs and mRNAs in the J4 group.

MSC, AEBP1, and CREB3L1 are master regulators that govern transcriptional programs in the J4 group

Co-expression network analysis was conducted using eigen centrality to identify the most influential transcriptional factors (Fig. 4A–E). We further eliminated every mRNA and lncRNA shown in Fig. 4D except representative transcription factors to show co-expression network analysis results among transcription factors more clearly. The node closest to 1 was the lncRNA LINC01614 in J4. COL3A1 exhibited the second highest eigen centrality (eigen centrality = 0.997), whereas MSC (eigen centrality = 0.918) was identified as the top node among representative transcription factors (Table S10). As AEBP1 and CREB3L1 also showed significant eigen centrality (> 0.5) in addition to MSC, COL3A1, LUM, DCN, FNDC1, THBS2, COL1A1, MMP2, FN1, and FAP were simultaneously regulated by these three transcription factors. In addition, these simultaneously regulated genes exhibited high eigenvector centrality (Fig. S8A–C). MSC, AEBP1, and CREB3L1 may be related to extracellular matrix, cell adhesion, and collagen remodeling, which are representative features of EMT^{9–11}. Therefore, we generated scores reflecting these EMT features to verify these hypotheses. J4 showed higher scores than the other groups (Fig. 5A). Consistent with this result, nine simultaneously regulated genes were functionally related to EMT, as determined through gProfiler (Fig. 5B–C).

lncRNAs Regulate EMT-Associated MTFs in the J4 group

As shown in Fig. 6A, MSC, AEBP1, and CREB3L1 were identified as MTFs in the J4 group. These MTFs were predicted to simultaneously regulate six downstream transcription factors: SNAI2, RUNX2, PRRX1, ZFH4, ETV1, and TWIST1. Co-expression network analysis revealed that LINC01614 has the highest centrality. However, lncRNAs harboring high centrality were not included in the target list of MTFs. Therefore, we postulated that these lncRNAs might be central regulators of the transactivation of MTFs. To verify this hypothesis, we first selected lncRNAs exhibited a positive correlation ($R \geq 0.4$) with MSC, AEBP1, and CREB3L1. Fifteen of the selected lncRNAs commonly included in the three MTFs were finally collected (Fig. 6B). All 15 lncRNAs showed eigen centrality ≥ 0.76 (Fig. 6C). Therefore, we hypothesized that these lncRNAs may regulate MTFs in J4 to increase the expression of EMT-related genes (Fig. 6D).

MTF silencing affects the expression of EMT-related genes but not that of lncRNAs exhibiting high centrality

We performed in vitro analyses based on CCLE data to select the cell line suitable to represent the J4 group (Fig. S9A–B). The 59 M cells showed an increased expression of the MTF target genes in the J4 group compared to other cell lines (Fig. S9C–D). The EHF scores of the target genes in the J5 group also increased in these cells. However, the lncRNA scores were exclusively increased in J4. Consistent with our hypothesis, ssGSEA revealed a significant increase in the expression of the transcriptional target genes *MSC*, *AEBP1*, and *CREB3L1* in 59 M cells (Fig. S9E–G). Next, we silenced each of the three MTFs and performed qRT-PCR to estimate changes in the expression of simultaneously regulated genes and lncRNAs exhibiting high centrality (> 0.9) in 59 M cells. As hypothesized, treatment with appropriate siRNAs downregulated the expression of simultaneously regulated genes whereas that of lncRNAs exhibiting high centrality was not affected (Fig. 7A–I). These results support the hypothesis that high-centrality lncRNAs may participate in the regulation of MTF activity, potentially contributing to the upregulation of EMT-related genes.

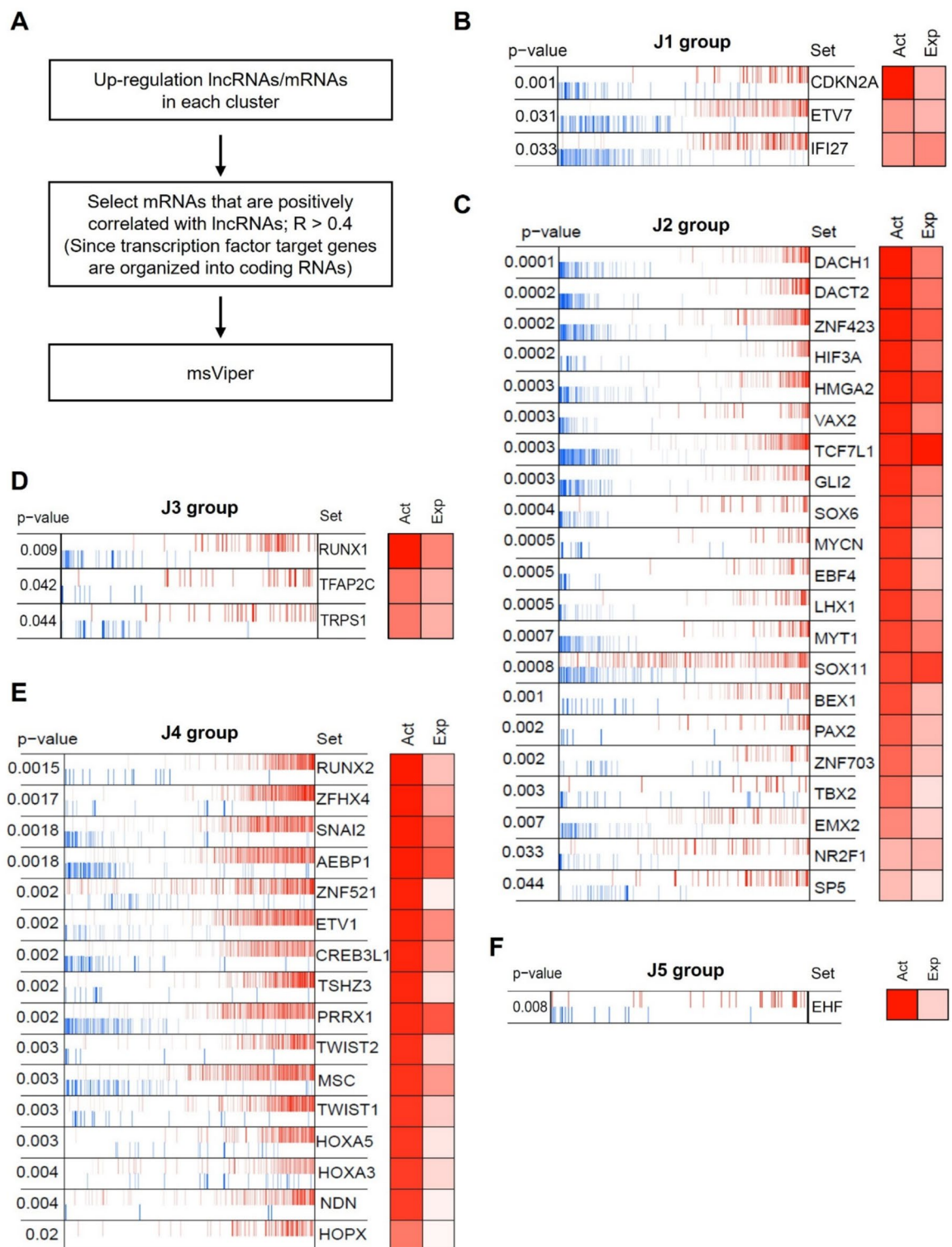


Fig. 3. Prediction of cluster-specific transcription factors. (A) A flow sheet of the prediction of transcription factors using msViper. (B and F) The msVIPER plot shows the transcriptional activity of the most significantly differentially active transcription factors in each group. Each plot shows the expression of transcription factor targets of one group compared to other groups. Vertical lines resembling barcodes represent transcriptional factor target genes. These lines are sorted from left to right based on expression levels, from the most downregulated to the most upregulated expression across the groups. Each blue and red bar indicates negative and positive regulation of the TF target genes shown in the third column, respectively. The “Act” and “Exp” columns indicate the normalized enrichment score of the expected activity (Act) of the entire transcriptional factor network and the expression level (Exp) of the transcriptional factor itself. P -values were determined using the enrichment method in the VIPER algorithm. $P < 0.05$ was considered statistically significant.

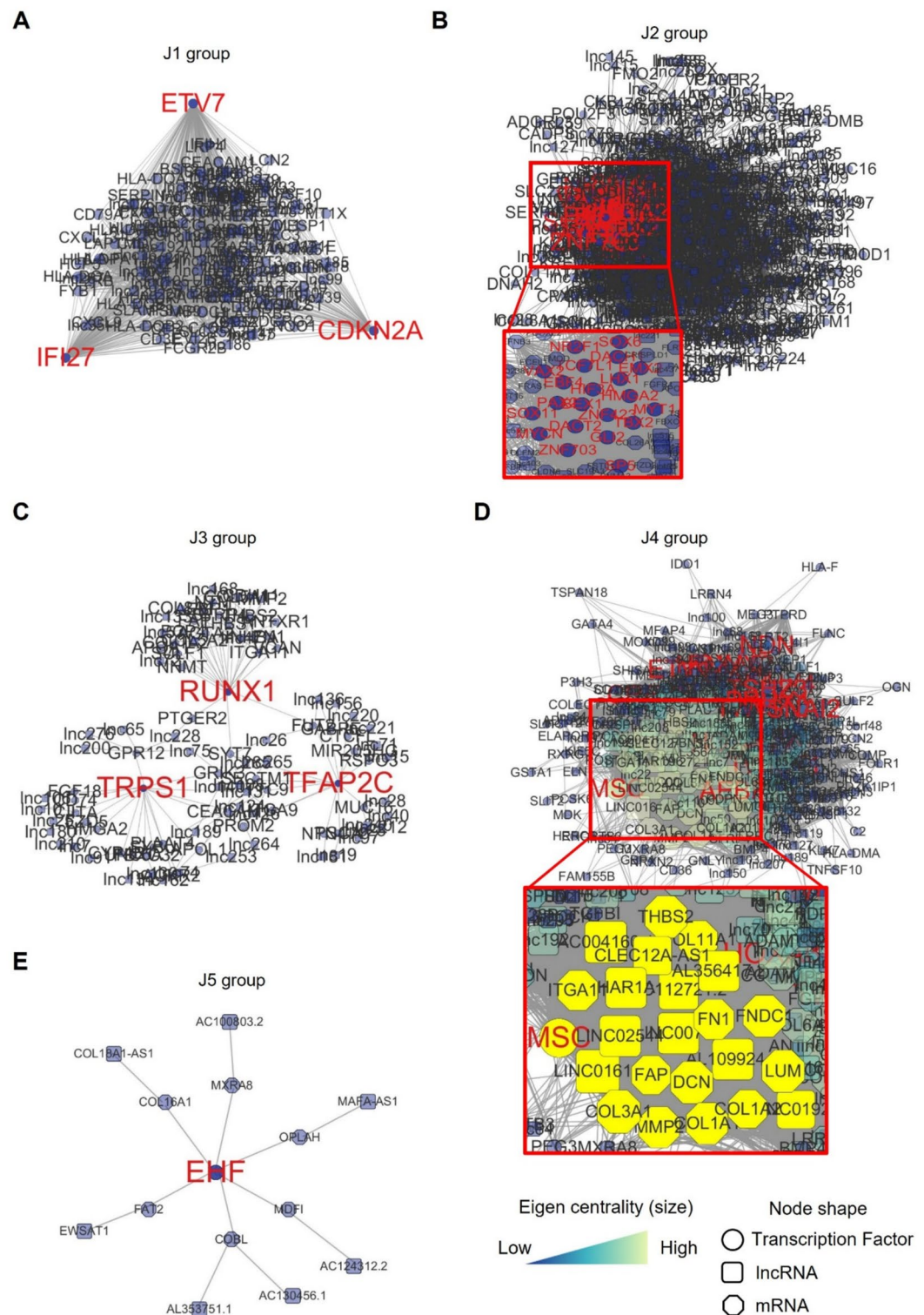


Fig. 4. Construction of the TF-mRNA-lncRNA co-expression network. (A and E) TF-mRNA-lncRNA co-expression networks for each group, as determined based on eigen centrality. Higher scores represent larger nodes and are closer to yellow, whereas lower scores represent smaller nodes and are closer to blue. The names of transcription factors are indicated in red. The red box is a magnified view of the denser parts of the network, as indicated. The eigenvector centrality of the yellow node is ≥ 0.9 .

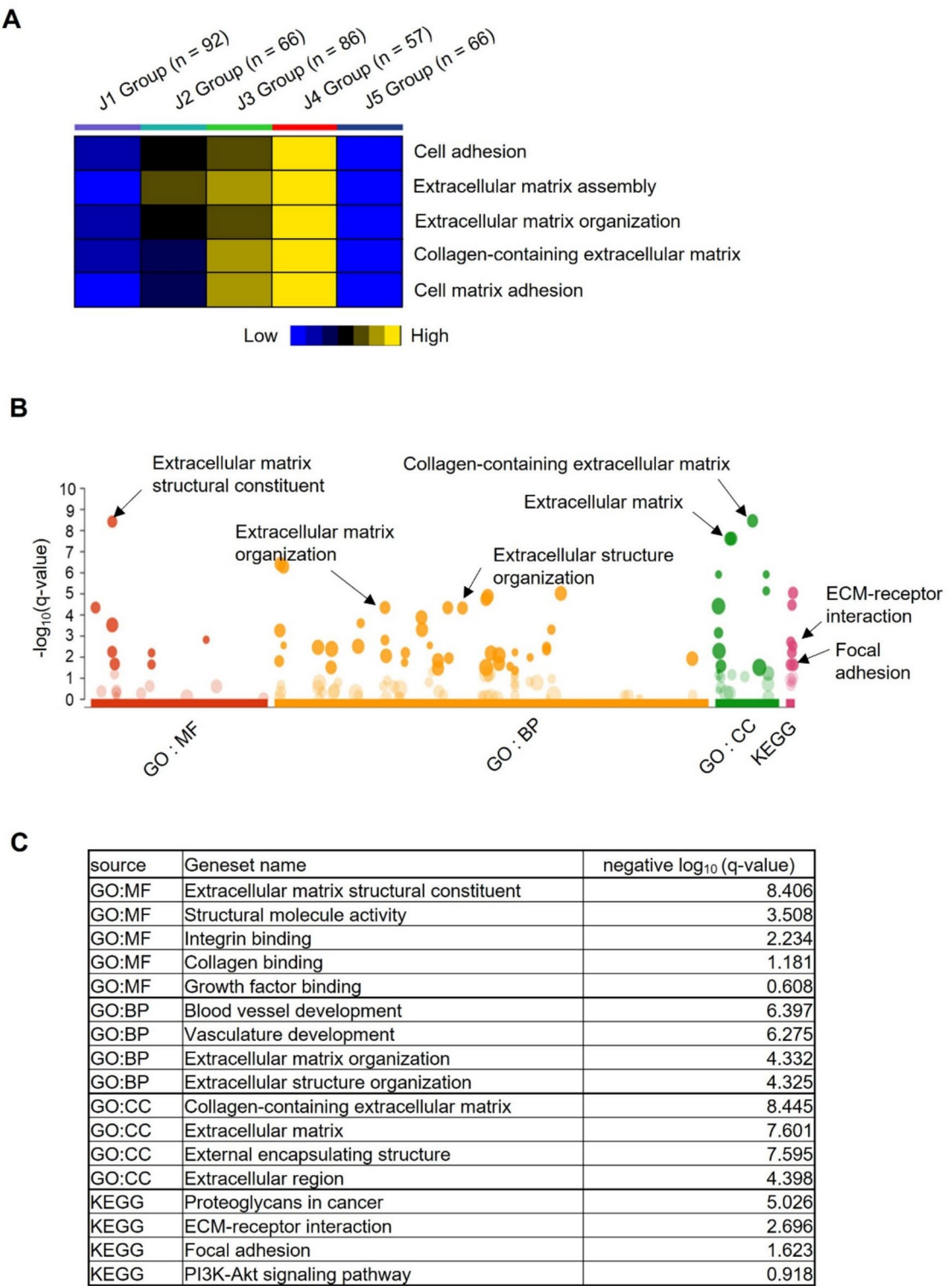


Fig. 5. Comparison of target gene expression of predicted MTFs in the J4 group. **(A)** Comparison of the expression of gene sets related to MTFs in each cluster using ssGSEA. **(B)** Simultaneous functional prediction of nine genes identified through network analysis using gProfiler. **(C)** Results summarizing the top five gene sets for each gene set source.

Discussion

Molecular subtypes of OVCA were first identified using microarray data from 285 Australian patients¹². TCGA later reported four subtypes classified as “Mesenchymal,” “Differentiated,” “Proliferative,” and “Immunoreactive” based on RNA-seq data. Most subtypes have been classified based on microarray or mRNA expression data¹³.

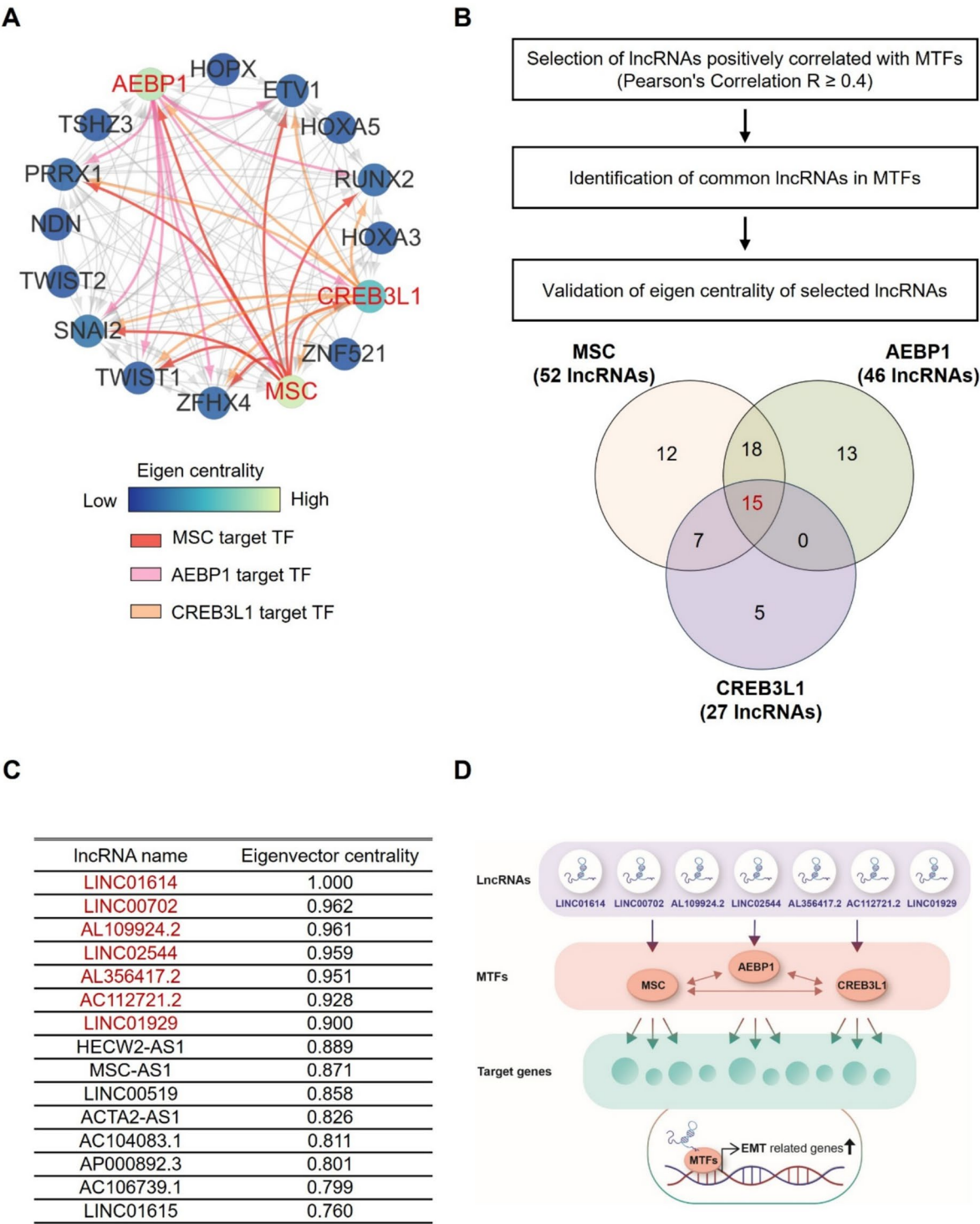
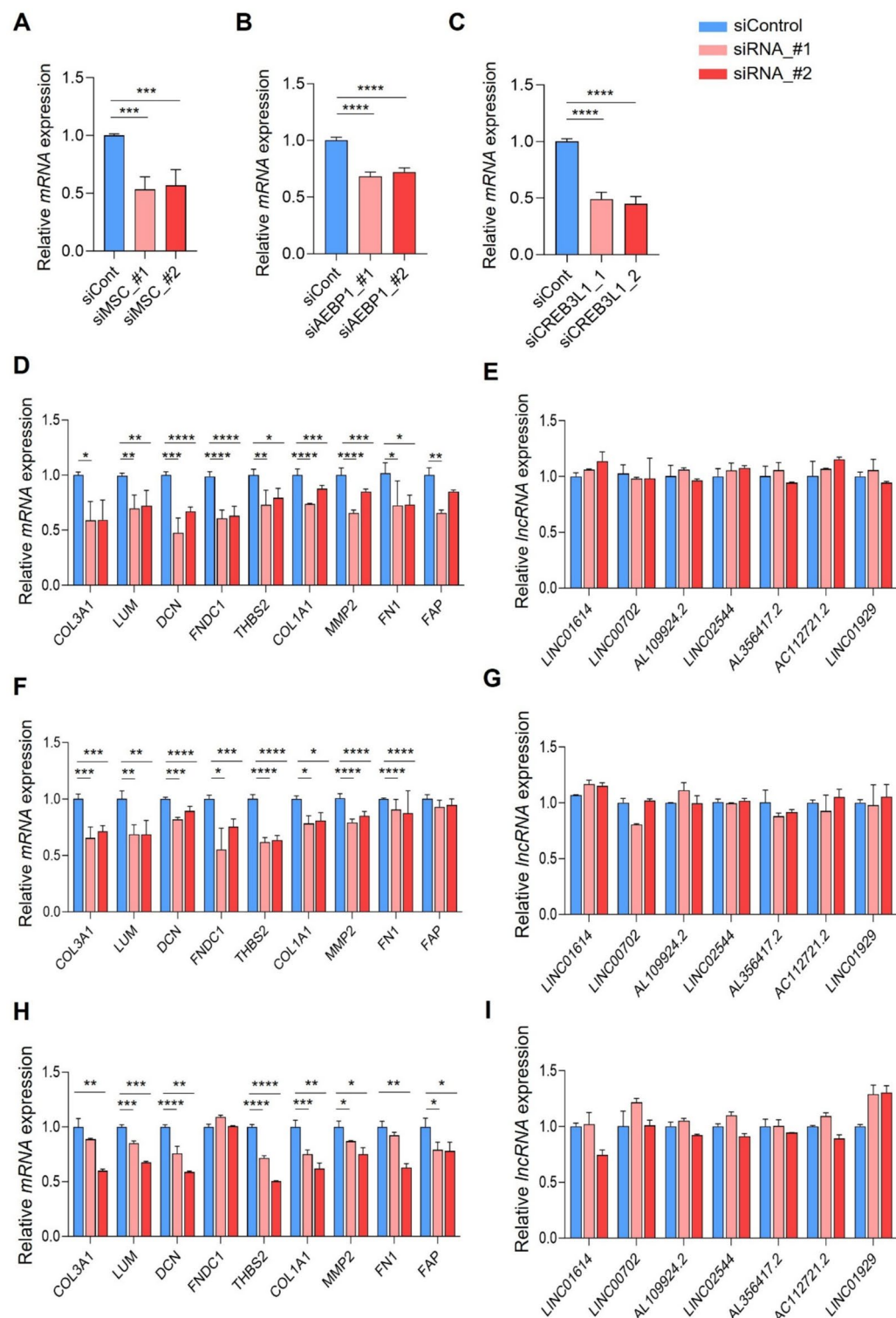


Fig. 6. Discovery of MTF-regulating lncRNAs. **(A)** Target transcription factors of top MSC, AEBP1, and CREB3L1 in the J4 group. Yellow and blue indicate high and low eigenvector centrality scores, respectively. **(B)** The scheme for analytic flow (left panel) and Venn diagram (right panel) presenting the number of lncRNAs simultaneously exhibiting positive correlation with three representative MTFs based on Pearson's correlation ($R \geq 0.4$). **(C)** A list of lncRNAs simultaneously correlated with three MTFs. **(D)** A proposed model of EMT gene regulation of MTFs through lncRNAs in HGSC.



In 2016, HGSC was classified based on DNA methylation, protein, microRNA, and gene expression data for OVCA samples obtained from TCGA¹⁴. However, lncRNAs remain considerably understudied. In addition, the classification of HGSC based on lncRNA expression has not been investigated.

In this study, we performed CNMF clustering using HGSC RNA-seq data and classified HGSC into five clusters based on lncRNA expression. In addition, we predicted lncRNA function using FuncPred. Although lncRNAs have been actively investigated since 2010, a direct functional prediction tool has not been established¹⁴. Therefore, similar to “FuncPred,” lncRNA functions are indirectly inferred by matching coding genes. Alternatively, functions are suggested through *cis*-acting by evaluating neighboring genes. We also conducted functional prediction of mRNAs based on lncRNA expression using gProfiler to mitigate the limitation of indirect functional prediction in the five clusters. The results were consistent with those obtained in FuncPred. This approach enabled the classification of HGSC into five groups. Each group was classified as follows based

◀ **Fig. 7.** The effect of MTF silencing on target gene expression, as determined using in vitro analyses. 59 M cells were treated with control siRNA (siControl) or three MTF siRNAs for 48 h. q-PCR results showing the expression of MTFs (MSC, AEBP1, and CREB3L1) (**A** and **C**), mRNA target genes (**D** and **F**), and lncRNAs (**G** and **I**) in transfected cells are shown. Data are presented as the mean \pm S.E.M. *P*-values were calculated using the Mann–Whitney U test, and all experiments were performed in triplicate. **p* < 0.05, ***p* < 0.01, ****p* < 0.001, *****p* < 0.0001. Statistically insignificant results are not shown. Abbreviations: MSC, Musculin; AEBP1, AE binding protein 1; CREB3L1, CAMP responsive element binding protein 3 like 1; COL3A1, Collagen type III alpha 1 chain; LUM, Lumican; DCN, Decorin; FNDC1, Fibronectin type III domain containing 1; THBS2, Thrombospondin-2; COL1A1, Collagen, type I, alpha 1; MMP2, matrix metalloproteinase-2; FN1, Fibronectin 1; FAP, fibroblast activation protein alpha; LINC01614, Long intergenic non-protein coding RNA 1614; LINC00702, Long intergenic non-protein coding RNA 702; LINC02544, Long intergenic non-protein coding RNA 2544; AL356417.2, novel transcript, antisense to FNDC1; AC112721.2, novel protein (LOC728009); LINC01929, Long intergenic non-protein coding RNA 1929.

on functional prediction: “Immune group,” “EMT group,” “Estrogen response group,” “EMT-androgen response group,” and “Differentiated group”.

We analyzed HGSC multi-omics data obtained from TCGA to investigate differences in molecular characteristics among the five groups. First, comparison of mutation profiles revealed no significant differences in the frequency of driver mutations among the groups. Similarly, comparison of SCNA revealed no significant differences in amplification and deletion among the five groups. This is contrary to the representative amplification of 1q21.3 chromosomes frequently observed in the five groups in previous studies^{15,16}. Despite extensive reports of lncRNA-miRNA interaction, comparison of miRNA expression patterns among the five groups also revealed no significant differences in median miRNA expression in the present study. Methylation is another important regulatory mechanism of gene expression^{17,18}. However, the methylation β -values were not relatively low in the five HGSC groups.

Although multi-omics analyses did not reveal dominant alterations that could explain our observations, these results prompted us to investigate transcriptional regulatory mechanisms in the five groups. lncRNAs can regulate transcriptional factors at multiple steps (transcription factors accessing DNA, mRNA synthesis, processing, stability, and translation)¹⁹. The transcription factors of each cluster were identified using ARACNe and Viper. In our analysis, transcription factors were identified independently within each cluster based on differentially expressed mRNAs and lncRNAs that were specific to each group. As a result, no overlapping TFs were observed across the five clusters. This outcome was not due to the intentional exclusion of shared TFs but rather reflected the nature of the analysis design, where each cluster was evaluated using a distinct set of input genes. The target genes of transcription factors were compatible with our functional prediction of mRNAs and lncRNAs. Furthermore, co-expression network analysis identified lncRNAs with high eigenvector centrality that were associated with MSC, AEBP1, and CREB3L1 in the J4 group²⁰. While these MTFs also exhibited high centrality, this metric was used to prioritize candidates for further investigation, as it reflects network influence rather than direct biological causality.

MSC is a member of the helix-loop-helix family of transcription factors and was first reported in mouse skeletal muscle precursors. MSC and LEF1 functions are related to EMT-related extracellular matrix (ECM) organization and cell-ECM interactions. AEBP1 promotes tumorigenesis through the NF- κ B pathway and EMT in colon and gastric cancers. Similarly, CREB3L1 contributes to cancer onset and progression and may be a promising clinical biomarker for cancers involved in the initiation of EMT^{9–11}. These results indicate that the three MTFs contribute to EMT-related gene regulation, which is a predicted representative function of the J4 group. In addition, lncRNAs whose expression was highly correlated with the three MTFs had high centrality. Thus, in silico and in vitro experiments suggest that they could regulate MTFs. Consistent with our hypothesis, lncRNAs act as signaling molecules to regulate the transcription of downstream genes. LINC01614, an lncRNA exhibiting the highest centrality, promotes the development of lung and breast cancers. Additionally, most of the genes co-expressed with LINC01614 have been associated with EMT in human cancers^{21–23}.

Several lncRNAs have been identified as potential clinical biomarkers for response to therapy or prognosis in breast cancer. Although their clinical utility is yet to be clearly demonstrated, the use of lncRNAs as predictive biomarkers in response to treatment has advantages over protein- and mRNA-based biomarkers because lncRNAs reveal remarkable tissue- and stage-specific expression patterns^{24,25}. LINC00702 is upregulated in ovarian cancer cells, promoting tumor development. In breast cancer, LINC00702 knockdown significantly reduces cell growth and induces apoptosis. Similarly, suppression of LINC02544 has been shown to reduce proliferation, invasion, and migration of LUSC cells^{26,27}. LINC01929 has been implicated in regulating key cancer cell behaviors, including proliferation, migration, invasion, and apoptosis, across multiple cancer types. In oral squamous cell carcinoma, LINC01929 knockdown reduced proliferation and invasion while promoting apoptosis, whereas in breast cancer, its suppression inhibited cell proliferation and induced cell cycle arrest. These findings highlight the potential of LINC01929 as a therapeutic target^{28,29}. For the remaining lncRNAs (AL109924.2, AL356417.2, and AC112721.2), no prior studies have been reported to date. However, our functional prediction (funcPred) analysis suggested that these lncRNAs may be associated with EMT and MYC target gene regulation.

In line with recent trends in lncRNA research, our study also elucidates the potential role of lncRNAs as a prognostic marker or therapeutic target in HGSC. Our lncRNA-based clustering suggests that the J4 group has a worse prognosis. In addition, J4 was primarily characterized by EMT- and androgen-response signatures, although additional pathways may contribute to its aggressive phenotype. Finally, our co-expression network

analysis suggested the mechanistic role of high-centrality lncRNAs in the transactivation of MTFs. This result was supported by in vitro silencing experiments.

Conclusions

This study provides a novel classification of HGSC based on lncRNA expression and presents functional prediction of five groups according to lncRNA and mRNA expression: “Immune,” “EMT,” “Estrogen response,” “EMT-Androgen response,” and “Differentiation” groups. Of these, the EMT-Androgen response group showed strong centrality mediated by MTFs, including MSC, AEBP1, and CREB3L1, as well as group-specific lncRNAs. These findings revealed the biology of a special subgroup of HGSC and provided potential diagnostic and therapeutic targets for the EMT-Androgen response group.

Methods

Public databases

Data derived from 367 OVCA samples were obtained from The Cancer Genome Atlas (TCGA, <https://www.cancer.gov/tcga>). The downloaded dataset included total RNA-seq, miRNA, methylation, and clinical data. The TCGA-OV GISTIC2 gene-level copy number and somatic mutation data was downloaded from The Broad Institute TCGA GDAC Firehose with no further processing³⁰.

We identified 47 cell lines representing major subtypes of OVCA within the Cancer Cell Line Encyclopedia project (CCLE, <https://sites.broadinstitute.org/ccle/>). We analyzed these cell lines using RNA-seq data. The cell lines were then classified using the widely accepted OVCA classification paradigm based on clinico-pathological and molecular evidence: Type I (n=28) and Type II (n=19) cell lines^{31,32}. As all Type II OVCA cell lines correspond to HGSC, RNA-seq data of the 19 cell lines were analyzed.

RNA-seq data processing and subtype discovery

In total, 60,433 identifiers from the RNA-seq data were downloaded using TCGA GDC; lncRNAs and coding RNAs were classified using the GTF file (v30). Furthermore, 15,171 lncRNAs were identified using GENCODE as described previously³³. Of the remaining 45,262 genes, 20,531 coding genes were identified using the HUGO probe map.

The top 1500 lncRNAs were selected through the Median Absolute Deviation (MAD) method in the order of high expression. Unsupervised clustering was conducted using consensus nonnegative matrix factorization (CNMF) for the selected lncRNAs (ConsensusClusterPlus R package; parameters: max $K=6$, reps=100). Next, multiple k-factor decompositions of expression matrices were computed and their stability evaluated³⁴. Consensus values ranging from 0 (never clustered) to 1 (always clustered) were marked by blue to red in the consensus matrix. The sample correlation matrix ranges from -1.0 to 1.0 and is colored white to blue. Clustering results were verified using the cophenetic coefficient, average silhouette width, and total within sum of square (WSS). Cophenetic coefficients and average silhouette widths closer to 1 indicate a suitable cluster. Total WSS is used to determine the point where the WSS rapidly decreases based on the sum of squares of distances within the cluster as the appropriate number of clusters. The appropriate number of clusters was determined based on the results of the three test methods. Next, the mRNA expression pattern was classified based on our cluster according to lncRNA expression. The top 1,500 of the 20,531 coding genes were selected using the MAD method, similar to lncRNA. Differentially expressed genes were subsequently identified according to the group.

Functional enrichment analysis

The function of lncRNAs was predicted using the funcpred database³⁵. Funcpred predicts lncRNA function using tissue-specific and evolutionarily conserved expression. The lncRNAs matching the coding genes were reclassified into gene sets to which the coding genes belonged. Functional analysis was performed using the hallmark gene annotation source, and significant gene set results satisfying p -value < 0.05 were selected. The number of lncRNAs contained in each gene set was subsequently counted. mRNA functional analysis was performed using gprofiler³⁶, a public web server used to characterize and manipulate gene lists resulting from high-throughput mining of genomic data. mRNAs expressed specifically in each cluster were selected, and functional analysis was conducted using hallmark gene sets.

Identification of transcription factors

Master regulator inference analysis was performed based on the ARACNe and viper algorithms^{37,38}. The analysis was conducted based on the ARACNe-AP package with default parameters as described on GitHub (<https://github.com/califano-lab/ARACNe-AP>). TRRUST³⁹ and GTRD⁴⁰ lists were collected for transcription factors, GO transcription factors, and active genes. In total, 2,192 transcription factors were selected⁴¹. Next, lncRNAs and mRNAs with upregulated expression were selected in each cluster; those showing positive correlation (Pearson's Correlation, $R \geq 0.4$) were then collected. As the transcription factor targeting the lncRNA was unknown, an indirect method through simultaneously upregulated mRNA expression was used. We used a t-test and generated a null model through sample permutations and subsequent t-tests to compare gene expression changes between groups. Master regulator inference analysis was conducted via the msiviper function from the viper package (in R package v.1.40.0) using t-statistics, corresponding p -values, and a null model. $P < 0.05$ was considered statistically significant.

Construction of a network of master transcription factors

The eigen centrality of each transcription factor was identified using the tidygraph R package based on the results of the selected MTF⁴². lncRNAs with a positive correlation ($R \geq 0.4$) with the target gene of the MTF were

selected to evaluate centrality. This approach enabled the linking of MTF—mRNA—lncRNA, and the result was visualized using Cytoscape (version 3.9.1)⁴³. In addition, the topological properties of the regulatory network were visualized, and the direct crosstalk of the target gene was validated using the edge bundle function, a Cytoscape plug-in.

Enrichment analysis of transcription factors targeting related genes

The target gene of each transcription factor was selected based on the result confirming the master transcription factor (MTF). A single-sample gene set enrichment analysis (ssGSEA) was performed using the CCLE dataset to confirm the expression of selected target gene sets at the cellular level. The ssGSEA is a rank-based algorithm that calculates a score illustrating the level of absolute enrichment of a particular gene set in each sample. The ssGSEA was analyzed through the GenePattern (<https://www.genepattern.org/>), and the results were visualized through the heatmap R package.

Cell culture and transfections

The Caov3 (cat#30,075), OVCAR3 (cat#30,161), and SNU8 (cat#00,008) cell lines were purchased from the Korean Cell Line Bank (Seoul, Korea), whereas the 59 M (cat#89,081,802) cell line was purchased from the European Collection of Authenticated Cell Cultures (UK). All cell lines were authenticated using short tandem-repeat profiling. The OVCAR3, SNU8, and 59 M cells were cultured in Roswell Park Memorial Institute medium (cat# 10040CV; Corning, NY, USA) supplemented with 10% fetal bovine serum (cat# 35015CV; Corning), 1% penicillin, and streptomycin. Similarly, CAOV3 cells were cultured in Dulbecco's modified Eagle's medium (cat# 10013CV; Corning) supplemented with 10% fetal bovine serum (cat# 35015CV; Corning) and 1% penicillin and streptomycin (cat# 15,140,122; Thermo Fisher Scientific, Waltham, MA, USA).

Three different small interfering RNAs (siRNAs) targeting Musculin (*MSC*, cat# 9242–1, 9242–2, 9242–3), AE binding protein 1 (*AEBP1*, cat# 165–1, 165–2, 165–3), and CAMP responsive element binding protein 3 like 1 (*CREB3L1*, cat# 90,993–1, 90,993–2, 90,993–3) were purchased from Bioneer (Daejeon, Korea). Non-targeting siRNA was used as a negative control. The RNAi oligonucleotide or RNAi negative control was transfected into the cells using a Lipofectamine RNAiMAX Transfection Reagent (Thermo Fisher Scientific) according to the manufacturer's instructions.

RNA isolation and real-time PCR

Total RNA was extracted from cells using TRIzol[®] reagent (Invitrogen, Waltham, MA, USA), and cDNA was synthesized from total RNA using the SuperScript[®] III First Strand Synthesis Kit (Invitrogen). Real-time PCR (qPCR) was performed using Power SYBR[®] Green Master Mix (Applied Biosystems, Foster City, CA, USA) on the Step One[™] Real-Time PCR System (Applied Biosystems). The comparative cycle threshold (CT) method was used to evaluate relative quantification. The primers used in quantitative reverse transcription PCR (RT-qPCR) are listed in Supplemental Table 1. Each experiment was performed in triplicate. GAPDH was used as an internal control.

Statistical analyses

SPSS statistics software (Version 26.0; IBM, Armonk, NY, USA) or GraphPad Prism (Version 10.4.1; GraphPad Software, San Diego, CA, USA) were used for statistical analyses. Overall and disease-free survival were calculated as the number of years between the year of diagnosis and the year of all-cause death, the date of last follow-up, or 5-year censored survival data. Survival analysis was assessed by comparing overall and disease-free survival according to clusters classified using Kaplan–Meier curves and Log-rank tests⁴⁴. The survival curve was plotted using the R package 'survminer' (v.0.5.0)⁴⁵. Mean values were compared using Student's t-test (two-tailed) and analysis of variance as appropriate. Pearson's correlation coefficient was used to examine the relationship between lncRNAs and the genes of interest. Data are expressed as the mean \pm SD, and all *p*-values are two-sided. *p* < 0.05 was considered statistically significant.

Data availability

We obtained data from 367 ovarian cancer (OVCA) samples in The Cancer Genome Atlas (TCGA, <https://www.cancer.gov/tcga>) and identified 47 cell lines representing major OVCA subtypes from the Cancer Cell Line Encyclopedia (CCLE, <https://sites.broadinstitute.org/ccle>).

Received: 10 March 2025; Accepted: 6 June 2025

Published online: 01 July 2025

References

1. Prat, J. Ovarian carcinomas: Five distinct diseases with different origins, genetic alterations, and clinicopathological features. *Virchows Arch.* **460**, 237–249 (2012).
2. Punzon-Jimenez, P., Lago, V., Domingo, S., Simon, C. & Mas, A. Molecular management of high-grade serous ovarian carcinoma. *Int. J. Mol. Sci.* **23**, 13777 (2022).
3. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **43**, 904–914 (2011).
4. Quinodoz, S. & Guttman, M. Long noncoding RNAs: An emerging link between gene regulation and nuclear organization. *Trends Cell Biol.* **24**, 651–663 (2014).
5. Akrami, R., Hoell, J., Schultz, N., Sander, C. & Larsson, E. Comprehensive analysis of long non-coding RNAs in ovarian cancer reveals global patterns and targeted DNA amplification. *PLoS ONE* <https://doi.org/10.1371/journal.pone.0080306> (2013).
6. Jeong, S. et al. Long non-coding RNA-based functional prediction reveals novel targets in notch-upregulated ovarian cancer. *Cancers (Basel)* **14**, 1557 (2022).
7. Libermann, T. A. & Zerbini, L. F. Targeting transcription factors for cancer gene therapy. *Curr. Gene Ther.* **6**, 17–33 (2006).

8. Garraway, L. A. & Sellers, W. R. Lineage dependency and lineage-survival oncogenes in human cancer. *Nat. Rev. Cancer* **6**, 593–602 (2006).
9. Zhang, S., Li, M., Ji, H. & Fang, Z. Landscape of transcriptional deregulation in lung cancer. *BMC Genom.* **19**, 435 (2018).
10. Liu, J. Y. et al. AEBP1 promotes epithelial-mesenchymal transition of gastric cancer cells by activating the NF-kappaB pathway and predicts poor outcome of the patients. *Sci. Rep.* **8**, 11955 (2018).
11. Lin, Z. et al. Pan-cancer analysis of CREB3L1 as biomarker in the prediction of prognosis and immunotherapeutic efficacy. *Front Genet* **13**, 938510 (2022).
12. Tothill, R. W. et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14**, 5198–5208 (2008).
13. Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
14. Zhang, Z. et al. Molecular subtyping of serous ovarian cancer based on multi-omics data. *Sci. Rep.* **6**, 26001 (2016).
15. Hu, X. et al. A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer Cell* **26**, 344–357 (2014).
16. Jeong, S. et al. Relationship of focally amplified long noncoding on chromosome 1 (FAL1) lncRNA with E2F transcription factors in thyroid cancer. *Medicine (Baltimore)* **95**, e2592 (2016).
17. Mercer, T. R., Dinger, M. E. & Mattick, J. S. Long non-coding RNAs: Insights into functions. *Nat. Rev. Genet.* **10**, 155–159 (2009).
18. Huang, W., Li, H., Yu, Q., Xiao, W. & Wang, D. O. LncRNA-mediated DNA methylation: An emerging mechanism in cancer and beyond. *J. Exp. Clin. Cancer Res.* **41**, 100 (2022).
19. Wei, J. W., Huang, K., Yang, C. & Kang, C. S. Non-coding RNAs as regulators in epigenetics (Review). *Oncol. Rep.* **37**, 3–9 (2017).
20. Sikdar, S. & Datta, S. A novel statistical approach for identification of the master regulator transcription factor. *BMC Bioinform.* **18**, 79 (2017).
21. Wang, D., Zhang, H., Fang, X., Cao, D. & Liu, H. Pan-cancer analysis reveals the role of long non-coding RNA LINC01614 as a highly cancer-dependent oncogene and biomarker. *Oncol. Lett.* **20**, 1383–1399 (2020).
22. White, N. M. et al. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. *Genome Biol.* **15**, 429 (2014).
23. Liu, A. N., Qu, H. J., Yu, C. Y. & Sun, P. Knockdown of LINC01614 inhibits lung adenocarcinoma cell progression by up-regulating miR-217 and down-regulating FOXF1. *J. Cell Mol. Med.* **22**, 4034–4044 (2018).
24. Yan, X. et al. Comprehensive genomic characterization of long non-coding RNAs across human cancers. *Cancer Cell* **28**, 529–540 (2015).
25. Iyer, M. K. et al. The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
26. Chai, D. et al. Knockdown of LINC00702 inhibits the growth and induces apoptosis of breast cancer through the Wnt/beta-catenin pathway. *Heliyon* **9**, e20651 (2023).
27. Wei, W., Xu, T., Zhang, Y., Huang, Y. & Wang, X. Upregulation of long noncoding RNA linc02544 and its association with overall survival rate and the influence on cell proliferation and migration in lung squamous cell carcinoma. *Discov. Oncol.* **13**, 41 (2022).
28. Salimian, N., Peymani, M., Ghaedi, K., Mirzaei, S. & Hashemi, M. Diagnostic and therapeutic potential of LINC01929 as an oncogenic lncRNA in human cancers. *Pathol. Res. Pract.* **244**, 154409 (2023).
29. Gu, Y., Feng, Z., Xu, X., Jin, L. & Jiang, G. LINC01929 Is a prognostic biomarker for multiple tumours and promotes cell proliferation in breast cancer through the TNF/STAT3 axis. *J. Cell Mol. Med.* **28**, e70227 (2024).
30. Deng, M., Bragelmann, J., Kryukov, I., Saraiva-Agostinho, N. & Perner, S. FirebrowseR: An R client to the Broad Institute's Firehose Pipeline. *Database (Oxf.)* **2017**, baw160 (2017).
31. Rojas, V., Hirshfield, K. M., Ganesan, S. & Rodriguez-Rodriguez, L. Molecular characterization of epithelial ovarian cancer: Implications for diagnosis and treatment. *Int. J. Mol. Sci.* **17**, 2113 (2016).
32. Shih, I. M., Wang, Y. & Wang, T. L. The origin of ovarian cancer species and precancerous landscape. *Am. J. Pathol.* **191**, 26–39 (2021).
33. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
34. Xu, T. et al. CancerSubtypes: An R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics* **33**, 3131–3133 (2017).
35. Perron, U., Provero, P. & Molineris, I. In silico prediction of lncRNA function using tissue specific and evolutionary conserved expression. *BMC Bioinform.* **18**, 144 (2017).
36. Raudvere, U. et al. g:Profiler: A web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
37. Alvarez, M. J. et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.* **48**, 838–847 (2016).
38. Lachmann, A., Giorgi, F. M., Lopez, G. & Califano, A. ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* **32**, 2233–2235 (2016).
39. Han, H. et al. TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
40. Yevshin, I., Sharipov, R., Kolmykov, S., Kondrakhin, Y. & Kolpakov, F. GTRD: A database on gene transcription regulation-2019 update. *Nucleic Acids Res.* **47**, D100–D105 (2019).
41. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
42. Pedersen, T. L. tidygraph: A Tidy API for Graph Manipulation. *1.2.3 edn* (2023).
43. Shannon, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
44. Agarwal, G. G. Statistics for surgeons—understanding survival analysis. *Indian J. Surg. Oncol.* **3**, 208–214 (2012).
45. Alboukadel Kassambara, M. K., Przemyslaw Biecek survminer: Drawing Survival Curves using 'ggplot2' 0.4.9 edn R package (2021).

Author contributions

SGJ and JL designed the research. SHJ, YSJ, SP, HL, and EGP performed the experiments and analyzed the data. SHJ and YSJ created the figures. SHJ, YSJ, SGJ, and JL wrote and edited the paper. All authors read and approved the final manuscript.

Funding

This work was supported by the National Research Foundation of Korea (NRF) funded by the Korean government MSIT RS-2023-NR076402 (Y.S.J.), RS-2022-NR073487 (S.G.J.), and RS-2020-NR047825 (J.L.).

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-06262-9>.

Correspondence and requests for materials should be addressed to S.G.J. or J.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025