Cancer Cell International

RESEARCH

Open Access

Machine learning identified extrachromosomal DNA-related 12 gene signatures to predict cancer immunotherapy response

Yan Ju[1,2,3†], Jingwei Zhang[1,2,3†], Jiaming Deng[1,2,3], Xingyuan Wu[1,2,3], Haotian Yang[4], Changyu Tao[5,6*] and Xiao Li[1,2,3*]

Abstract

Extrachromosomal circular DNA (ecDNA) has emerged as a critical determinant of poor clinical outcomes and immune escape in tumors, but the high cost and technical complexity of current detecting techniques limit its broader investigation in cancer immunotherapy. Leveraging the combined machine learning algorithms including the least absolute shrinkage and selection operator (LASSO) regression, RandomForest (RF) and Recursive Feature Elimination (RFE), we developed a 12-gene transcriptomic score (EC_score) to predict the existence of ecDNA through RNA-seq. EC_score demonstrated reliable predictive performance in two independent cohorts (AUC > 0.70), validated by fluorescence in situ hybridization (FISH) in both cell lines and clinical samples. Next, we found that EC_score emerged as an independent adverse prognostic factor across multiple immunotherapy cohorts. Notably, high EC_score correlated with cell cycle activation and immunosuppression, characterized by reduced lymphocytes infiltration and upregulated immunosuppressive markers, including MHC molecules, co-inhibitory immune checkpoints and TGF-β signals. In general, we established and validated a 12-gene signature (EC_score) derived from RNA-seq, offering a novel computational tool for predicting the presence of extrachromosomal circular DNA and stratifying cancer immunotherapy response.

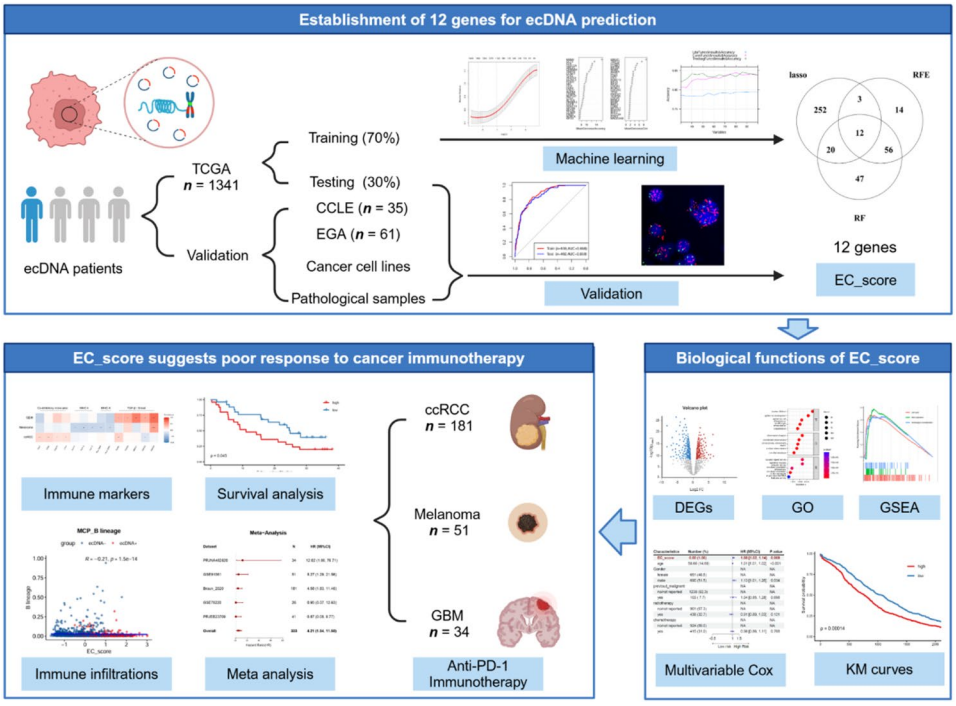
Keywords Extrachromosomal circular DNA (ecDNA), Oncogene amplification, Machine learning, Cancer immunotherapy, Immune infiltration

†Yan Ju, Jingwei Zhang these authors contributed equally to this work.

*Correspondence:
Changyu Tao
taochangyu@pku.edu.cn
Xiao Li
dr.lixiao@163.com

Full list of author information is available at the end of the article

Graphical abstract



Overview of the analytical framework integrating multi-omics data, machine learning, and experimental validation to establish EC_score and characterize its clinical implications.

Introduction

Extrachromosomal circular DNA (ecDNA), a double-stranded circular DNA molecule, is prevalently distributed across various tumors [1–3]. The circular structure of ecDNA results in enhanced chromatin accessibility and non-Mendelian inheritance, facilitating oncogene amplification [4, 5]. Prior researches have shown a correlation between ecDNA abundance and poor prognosis in pan-cancer [6], as well as its potential role in shaping an immunosuppressive tumor microenvironment [7, 8]. Nevertheless, the mechanistic underpinnings linking ecDNA to immunotherapy outcomes remains inadequately understood [1, 5, 9, 10].

The identification of ecDNA in immunotherapy cohorts poses significant challenges [11]. Although FISH offers high specificity for ecDNA detection, its clinical application on large scale cohorts is limited by low throughput and the requirement for gene-specific probes [3, 12]. To overcome these constraints, Circle-seq was developed by purifying circular DNA through exonuclease treatment, followed by rolling circle amplification and sequencing [13]. Although this approach circumvents the need for specific probes, its broader adoption remains challenged by experimental complexity and resource requirements.

Recent advances of next-generation sequencing (NGS) technologies coupled with computational algorithms have led to the development of diverse analytical flamework for ecDNA detection, including AmpliconArchitect (AA) using whole genome sequencing (WGS) data [14], GCAP utilizing whole exome sequencing (WES) data [7], ATACAmp leveraging assay for transposase-accessible chromatin using sequencing (ATAC-seq) profiles [15] and scEC&T-seq applying single cell sequencing data [16]. Among these, AA has emerged as the predominant approach for ecDNA prediction in WGS data [6, 11]. However, the less available WGS data and high cost render current methods impractical for population-scale screening [17]. Given the widespread accessibility of bulk RNA sequencing (RNA-seq) [18], we aimed to develop a cost-effective computational framework for ecDNA prediction through transcriptional profiles.

Here, we developed EC_score, a machine learning framework integrating LASSO, RF, and RFE to identify a 12-gene signature from RNA-seq for ecDNA prediction. The reliability and generalizability of the EC_score were validated through independent external datasets, metaphase cancer cells and pathological samples. Through Gene Set Enrichment Analysis (GSEA), high EC_score samples exhibited activation of cell cycle related

pathways, aligning with the biological characteristics of ecDNA.

Notably, our analysis demonstrated that EC_score significantly correlated with poor prognosis in patients receiving anti-PD-1 immunotherapy across multiple cancer types. The significant associations between EC_score and diverse immunosuppressive mechanisms were also discovered, while elevated EC_score was correlated with decreased infiltration of immune cell populations, including CD4 + T cells, CD8 + T cells, and B cells, indicating an immunosuppressive tumor microenvironment in high EC_score patients. In conclusion, we established and validated a novel transcriptome-based scoring system (EC_score) that effectively predicts ecDNA presence and systematically characterizes its relationship with tumor immune microenvironment and immunotherapy response.

Materials and methods

Data source

Clinical and transcriptomic data of TCGA tumors were obtained from The Cancer Genome Atlas (TCGA, https://portal.gdc.cancer.gov), covering 322 circular ecDNA tumors (ecDNA$^+$), 1019 no-focal tumors (ecDNA$^-$) and 513 chromosomal tumors according to the previous study [6]. The RNA-seq of small cell lung cancer (SCLC) was sourced from the European Genome Phenome Archive (EGA) (EGAS00001000925, www.ebi.ac.uk/ega/home) [19]. The expression profiles of cell lines in Cancer Cell Line Encyclopedia (CCLE) were downloaded from DepMap (https://depmap.org/portal/). Immunotherapy cohorts including Braun_2020 (nivolumab for ccRCC, n = 181), GSE91061 (nivolumab for melanoma, n = 51), GSE78220 (nivolumab or pembrolizumab for melanoma, n = 26), PRJNA482620 (nivolumab or pembrolizumab for glioblastoma, n = 34) and PRJEB73280 (nivolumab or pembrolizumab for melanoma, n = 41) were collected from the Tumor Immunotherapy Gene Expression Resource (TIGER, http://tiger.canceromics.org/) and the Gene Expression Omnibus (GEO, https://ncbi.nlm.nih.gov/geo). All expression-level comparisons and subsequent analyses of RNAseq were based on log-transformed transcripts per million (TPM) values to ensure cross-dataset comparability.

Functional analysis of differentially expressed genes

TCGA tumors were categorized into ecDNA$^+$ and ecDNA$^-$ groups based on AA-predicted ecDNA amplification from a previous study [6]. Differential expression analysis was performed using "DESeq2" R package (v1.38.3), with significant differentially expressed genes (DEGs) identified at |log2 Fold Change| >1 and adjust p value < 0.05 (Benjamini-Hochberg correction). Differential expression analysis was performed using DESeq2 on raw count data. Gene Ontology (GO) enrichment was conducted using "clusterProfiler" R package (v4.6.2), and results were visualized with "pheatmap" R package (v1.0.12).

Machine learning-based gene selection

To reduce batch effects, the TCGA cohorts were assigned randomly into training (70%) and testing sets (30%). We further compared the distribution of multiple confounding factors including age, gender, pathological stage, and ecDNA status across the groups (Supplemental_table S2), aiming to minimize the impact of these confounding factors on machine learning. Three machine learning methods, including LASSO regression, RF and RFE, were applied for feature gene selection in the training set. LASSO regression incorporated a regularization parameter (λ) to shrink coefficients of non-informative genes to zero, balancing model complexity and predictive performance. Hyperparameter λ was optimized using 10-fold cross-validation on the training set. The λ value that maximized the AUC values for ecDNA prediction was selected as optimal, resulting in λ = 0.710 (Supplemental_Figure1A-B). RFE iteratively removed the least important features and evaluated the model performance to refine the feature set. Three feature selection functions include Ldafuncs, CaretFuncs, and TreebagFuncs were compared, each implemented with 10-fold cross-validation to assess accuracy at each elimination step. The function and corresponding number of selected features that yielded the highest accuracy were chosen as optimal. Finally, the TreebagFuncs with 85 retained features was selected (Supplemental_Figure1C). RF constructed an ensemble of binary recursive classification trees by bootstrapping the training dataset and randomly selecting a subset of features for splitting at each tree node. The final classification result was determined by majority voting across all trees. Hyperparameters "ntree" (number of trees) and "mtry" (number of features sampled per tree) were optimized via 10-fold cross-validation in the training set. The RF model that achieved the highest AUC was designated as the final model, with optimal parameters identified as "ntree" = 1000 and "mtry" = 300 (Supplemental_Figure1D-E). The R packages used in the process included "glmnet" (v4.1.8), "randomForest" (v4.7.1.1), "caret" (v6.0.94) and "ggplot2" (v3.5.0).

EC_score development

EC_score was constructed using the single sample gene set enrichment analysis (ssGSEA) and validated via receiver operating characteristic curves (ROC) in training, testing and external validation sets. For the ssGSEA calculation, the kcdf argument was set to 'Poisson' for raw count matrices and 'Gaussian' for log-TPM transformed matrices according to the software recommendations

[20]. ROC curves were plotted by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across all possible threshold values of the EC_score. The area under the ROC curve (AUC) was calculated to quantify overall discriminative ability. The R packages "GSVA" (v1.46.0) and "pROC" (v1.18.5) were utilized.

Cell lines and cell culture

Cell lines NCI-H69 (SCLC, RRID: CCVL_1579) and NCI-H524 (SCLC, RRID: CCVL_1568) provided by Shanghai Fuheng Biotechnology Co., LTD. were cultured in RPMI-1640 medium with 10% FBS and 1% penicillin-streptomycin. Cell lines were maintained in a 5% CO2 incubator at 37 °C and were passaged every 2 to 4 days. Cell line identities were confirmed via short tandem repeat profiling, and all lines were certified to be free of mycoplasma infection.

RNA sequencing

This study was approved by the Institutional Review Board at Peking University People's Hospital (2023PHB216-001). This work was conducted in accordance with the Declaration of Helsinki. The inclusion standard was surgically resected lung samples and the postoperative diagnosis was small cell lung cancer by at least two experienced pathologists. Tissue specimens were fixed in 4% paraformaldehyde, washed with running water, dehydrated in a graded ethanol series, vitrified with dimethylbenzene, and embedded in paraffin. Formalin-fixed paraffin-embedded (FFPE) sections with 4 μm in thickness were prepared for further use with a microtome. Total RNA was extracted from the surgically resected FFPE tumor samples using the HiPure FFPE RNA kit, and quantified via the Qubit™ RNA HS Assay Kit (ThermoFisher Scientific, USA). The purity and integrity of the RNA were assessed respectively with a Take3 microvolume plate (BioTek, USA) and the RNA Cartridge kit of the Qseq100 Bio-Fragment Analyzer (Bioptic, China). RNA-seq libraries were prepared using the Hieff NGS MaxUp Human rRNA Depletion Kit (rRNA & ITS/ETS) and sequenced on the DNBSEQ T7 platform, resulting in sequencing data consisting of 100 bp paired-end reads in FASTQ format. The raw sequencing data was processed to eliminate low-quality reads and adapters using Trim Galore (v0.6.7). Clean reads were obtained from each sample, aligned to the hg38 reference using the STAR (v2.7.8a) and annotated with the Gencode (v37) database. Gene expression was computed in Rsem (v1.3.0). The sequencing coverage and quality statistics of each sample were summarized in Supplemental_table S1.

Fluorescence in situ hybridization (FISH)

Cell lines were treated with 10 ng/ml colcemid (MCE, HY-N0282) for 4 h. The cells were resuspended in 0.075 M KCl at 37 °C for 20 min and fixed using Carnoy's fixative (3:1 methanol/glacial acetic acid, v/v). The fixed cells were placed on pre-humidified slides and then treated with pepsin (Sigma-Aldrich) for 3 to 10 min. For FFPE samples, slides were deparaffinized with xylene and dehydrated with ethanol before being rehydrated in gradient ethanol. These slides were digested with pepsin at 37 °C for 3 min and washed with 2× saline-sodium citrate (SSC) buffer for 5 min. Subsequently, samples were treated with MYC or MYCN probes (Wuhan HealthCare Biotechnology Co., Ltd., Wuhan, China), denatured at 85 °C for 5 min and hybridized at 37 °C overnight. Following hybridization, the slides were immersed in 70% ethanol for 3 min and stained with 4,6-diamidino-2-phenylindole (DAPI) for 10 min. Images were captured on confocal microscope (Olympus, Tokyo, Japan) with a 63X oil lens.

Gene set enrichment analysis (GSEA)

DEGs between the high and low EC_score groups, as determined by the median EC_score, were subjected to GSEA [21] for pathway enrichment analysis using "clusterProfiler" R package (v4.6.2). The Enrichment Score (ES) indicated the degree to which the specified gene sets from KEGG were overrepresented at the top or bottom of a ranked list. The Normalized enrichment score (NES) quantified the density of modified genes in the dataset with random expectancies, normalized by the numbers of genes found in each gene cluster. Adjusted p values were calculated by comparing actual data with 1,000 Monte-Carlo simulations.

Immune infiltration analysis

The EPIC bioinformatics algorithm was employed to estimate the tumor-infiltration immune cell composition based on reference gene expression profiles [22]. The MCP-counter was applied to quantify the abundance of immune cell populations for each sample, utilizing the $\log_2$ geometric mean of the set of transcriptomic markers [23]. The R packages "EPIC" (v1.1.6) and "MCPcounter" (v 1.2.0) were utilized.

Statistical analysis

All analyses were performed in R software (v4.2.3, http://www.r-project.org). Statistical analyses are briefly described in the figure legends. The survival outcomes were evaluated through Kaplan-Meier curves, wherein patients were dichotomized into high and low EC_score groups based on the median value, with the log-rank test used to assess the statistical significance of the differences. The correlations were analyzed by Pearson's or

Spearman's tests. The calibration was visually assessed through the consistency of the predictive overall survival (OS) and the observed OS based on 1000 bootstrap resamples. The EC_score was treated as a continuous variable in univariate, multivariate Cox regression analyses, and the meta-analysis to calculate hazard ratios and confidence intervals for a precise evaluation of its prognostic impact. Univariate and multivariate Cox regression analyses were conducted to determine independent prognostic factors, with statistically significant variables included in the multivariate Cox proportional hazards model. Meta-analysis of immunotherapy cohorts was employed to combine the OS. Due to the low heterogeneity between the results included in the meta-analysis ($I^2 = 6.12\%$), the fixed effects model was applied to calculate the parameter changes between studies. Wilcoxon tests were conducted using wilcox.test function. A two-sided p value of 0.05 was considered statistically significant. Key R packages included "survival" (v3.5.3), "survminer" (v0.4.9), "Hmisc" (v5.1.3), "rms" (v6.8.1), "meta" (v7.0.0), "metafor" (v4.6.0) and "stats" (v4.2.3). The graphical abstract was created through BioRender (http://www.biorender.com).

Results

Establishment of a 12-Gene score for predicting EcDNA existence

To develop a transcriptomic model predicting the presence of extrachromosomal circular DNA in tumors, we identified 2,025 DEGs (FDR < 0.05, |log2FC| >1) between ecDNA$^+$ (n = 322) and ecDNA$^-$ (n = 1019) samples as defined by AmpliconArchitect (AA) from the TCGA database [6]. Through gene ontology (GO) analysis, 627 upregulated DEGs in ecDNA$^+$ tumors exhibited significant enrichment in critical cell cycle processes, such as nuclear division and chromosome segregation (Fig. 1A), consistent with the hyperproliferative phenotype driven by ecDNA-mediated oncogene amplification [24].

Next, we randomly partitioned the TCGA cohort (n = 1341) into training and testing sets in a 7:3 ratio, ensuring a balanced distribution of clinical characteristics (Supplemental_table S2). For feature selection, we employed machine learning algorithms including LASSO regression, RF and RFE (Fig. 1B, Supplemental_Fig.1 A-E), to refine the DEGs to a 12-gene signature (*DSCC1*, *MTBP*, *HOXC13*, *POU6F2*, *FAM72B*, *FAM72D*, *MAGED4B*, *NKX2-5*, *PLAC1*, *FOXI3*, *KRTAP4-1*, *OTX2*). Notably, all these genes were overexpressed in ecDNA$^+$ samples (Fig. 1C). Subsequently, we calculated the EC_score based on the enrichment of the genes and found that most ecDNA$^+$ samples exhibited high EC_score (Fig. 1D). The derived EC_score demonstrated great performance in predicting ecDNA existence in the training set (AUC = 0.868) and testing set (AUC = 0.859) (Fig. 1E). In addition, we found higher EC_score in ecDNA amplified samples compared to chromosomal amplification in TCGA tumors (Supplemental_Fig.1 F). We then examined the EC_score in each cancer type (sample size > 5) and observed an upward trend of EC_score as the ratio of ecDNA increased (Supplemental_Fig.2 A), suggesting an enriched ecDNA prevalence in cancers characterized by higher EC_score.

To further validate the generalizability of EC_score, we applied EC_score to independent databases including cancer cell lines from the CCLE database (ecDNA positive cases: 24/35) [12] and clinical samples (ecDNA positive cases: 13/61) from the EGA database (EGAS00001000925) [19] (Supplemental_table S3). The external validation in FISH-confirmed cell lines (AUC = 0.735) and AA-classified clinical specimens (AUC = 0.732) further supported the performance of EC_score in predicting ecDNA existence (Fig. 1F). Collectively, the EC_score maintained consistent performance (AUC >0.70) across various datasets, indicating its efficacy and generalizability in forecasting ecDNA existence across different experimental and clinical contexts.

EC_score indicated cell cycle activation and extrachromosomal oncogene amplification

To explore the biological relevance of EC_score, we analyzed all available RNA-seq (n = 10530) from pan-cancer samples in TCGA. Samples were stratified into high and low EC_score groups using the median value as the cutoff. A total of 3,176 DEGs were identified, with 1577 upregulated DEGs enriched in nuclear division and chromosomal region (Supplemental_Fig.2B). Further GSEA confirmed the activation of the Cell Cycle and DNA Replication, indicating enhanced cell cycle progression (Supplemental_Fig.2 C-D). Moreover, we also observed upregulation of pathways associated with ecDNA biogenesis in high EC_score tumors, including Homologous Recombination, Fanconi anemia and Mismatch repair (Supplemental_Fig.3 A) [25, 26].

Considering that ecDNA is mainly found in tumor cells [27], we next explore the biological features related to EC_score in cancer cell lines. All available cancer cell lines with RNA-seq from CCLE database (n = 1114) were stratified into high and low EC_score groups based on the median EC_score. A total of 450 upregulated DEGs in the high EC_score group were also primarily related to cell cycle processes, such as nuclear division, chromosome segregation and mitotic nuclear division (Supplemental_Fig.3B). These findings suggest that high EC_score tumors are characterized by comprehensive activation of cell cycle machinery, consistent with the reports of enhanced cellular proliferation in ecDNA$^+$ tumors [6, 24].

Given that *MYC* and *MYCN* oncogenes frequently locate on ecDNA and contribute to gene amplification

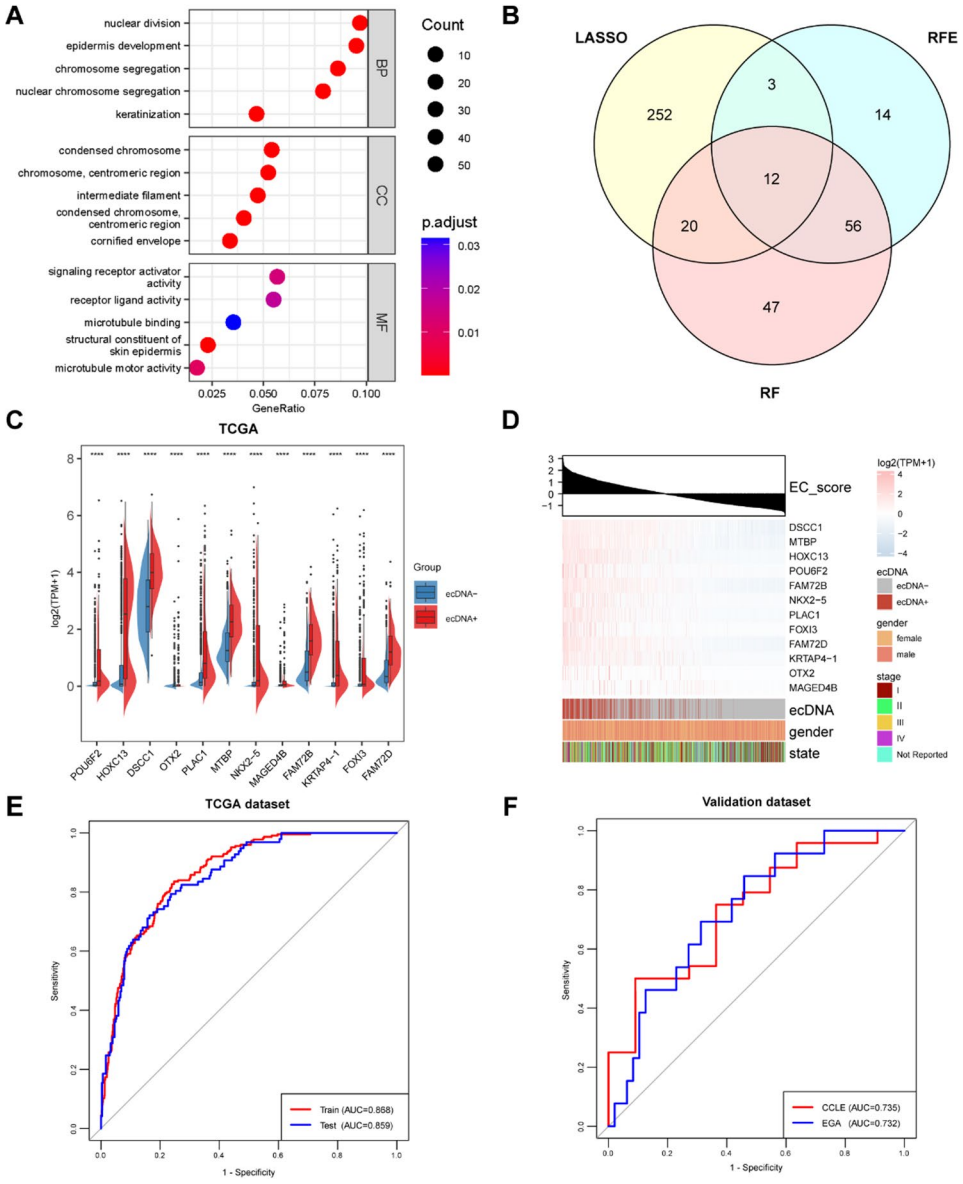


Fig. 1 Machine learning screening key genes for EC_score establishment. (A) Top enriched GO pathways in ecDNA$^+$ TCGA tumors by GO enrichment analysis. (B) Venn diagram identifying 12 consensus genes from LASSO regression, RF and RFE. (C) Expression levels of signature genes (log2[TPM + 1]) between ecDNA$^+$ and ecDNA$^-$ TCGA tumors (Wilcoxon test, ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$). (D) Unsupervised clustering heatmap of signature gene expression (log2[TPM + 1], Z-score normalized). (E) The prediction performance of EC_score in the training set ($n = 939$) and testing set ($n = 402$) assessed using ROC curves. (F) External validation of EC_score in EGA dataset ($n = 61$) and CCLE dataset ($n = 35$)

[14, 28], we combined EC_score with the expression levels of the oncogenes to infer potential ecDNA existence from CCLE database. Among all cancer cell lines, NCI-H524 exhibited high EC_score and elevated MYC expression (EC_score >2, log2(TPM + 1) >10), while NCI-H69 demonstrated the highest EC_score along with high MYCN expression (log2(TPM + 1) >9) (Fig. 2A). FISH staining of metaphase NCI-H524 displayed widespread MYC amplification with a MYC/CEP8 ratio >25. Moreover, multiple dispersed signals were not co-localized with linear chromosomes, illustrating substantial extrachromosomal MYC amplification in NCI-H524 (Fig. 2B). Similarly, extrachromosomal MYCN amplification was confirmed by the multiple scattered signals with a MYCN/LAF4 ratio exceeding 45, which were not co-localized with DAPI-stained chromosomes in metaphase of NCI-H69 cells (Fig. 2B). These results indicated that EC_score can effectively predict the cell cycle activation and ecDNA presence in cancer cell lines.

To further validate the performance of EC_score in clinical samples, we performed RNA-seq on pathological samples from a retrospective cohort comprising patients

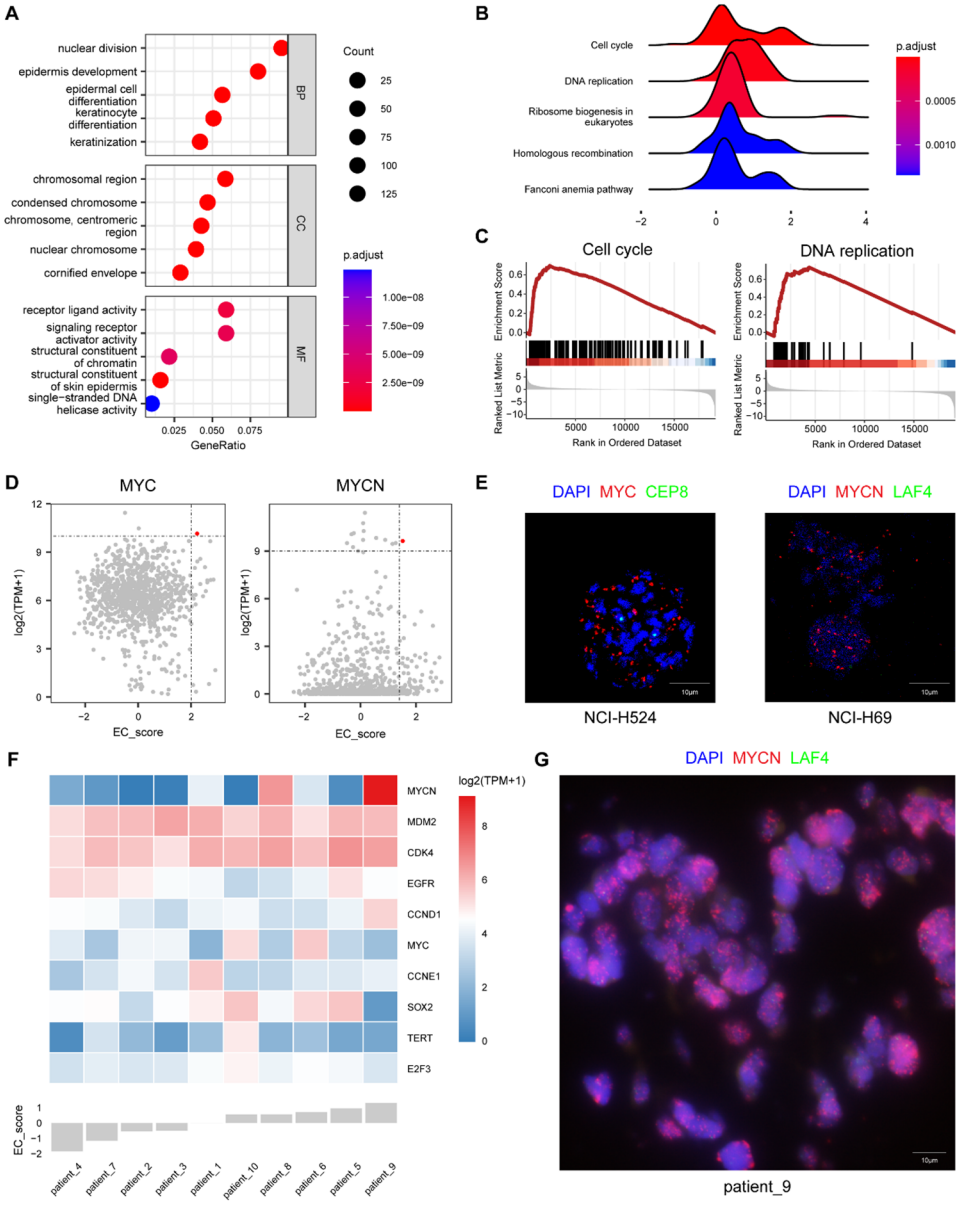


Fig. 2 Experimental Validation of EC_score. (A) Scatter plots correlating EC_score (x axis) with gene expression (y axis) to identify potential cell lines with extrachromosomal amplification of *MYC* and *MYCN*. Red dots represent selected cell lines from the CCLE database. (B) Representative FISH images of NCI-H524 and NCI-H69 metaphase cells. MYC and CEP8 were labeled with red and green probes. MYCN and LAF4 were labeled with red and green probes. The scale bar represents 10 μm. (C) Heatmap of EC_score and top frequent amplified genes reported in ecDNA-positive clinical samples (log2[TPM + 1], Z-score normalized) [5]. (D) Representative FISH images of patient_9. MYCN and LAF4 (control gene) were labeled with red and green probes, respectively. The scale bar represents 10 μm

diagnosed with SCLC by at least two experienced pathologists (Supplemental_table S4). Considering the most frequently amplified genes in ecDNA reported in recent study [5], we depicted the expression of these top oncogenes and EC_score and subsequently found that the patient_9 revealed both the highest EC_score and the expression of *MYCN* gene (Fig. 2C). Further FISH validation of patient_9 demonstrated the scattered *MYCN* signals in the pathological sample (MYCN/LAF4 ratio >10) (Fig. 2D), which has been reported the ecDNA-like sample [29], validating the utility of EC_score to predict ecDNA in clinical samples.

EC_score predicted cancer immunotherapy response

Given that ecDNA is associated with tumor prognosis [6, 11, 30], we investigated whether higher EC_score correlates with worse cancer survival. Analysis of TCGA samples revealed that patients with elevated EC_score exhibited poorer OS compared to those with lower

EC_score ($p < 0.0001$) (Supplemental_Fig.3 C). A calibration curve showed strong alignment between observed and predicted OS values, suggesting a close relationship between EC_score and patient prognosis (Supplemental_Fig.3D). Multivariate Cox regression analysis further validated EC_score as an independent prognostic factor (HR = 1.46; 95% CI, 1.29–1.65). Other significant risk factors included age (HR = 1.02; 95% CI, 1.01–1.03), male gender (HR = 1.62; 95% CI, 1.25–2.09), pathological stage Ⅲ (HR = 1.54; 95% CI, 1.08–2.19) and Ⅳ (HR = 2.68; 95% CI, 1.81–3.98) (Fig. 3A). In summary, our results established EC_score as a predictor of poor clinical outcome in tumors.

Based on established associations between ecDNA and immune evasion in tumors [6–8], we hypothesized that EC_score could predict cancer immunotherapy response. Through a meta-analysis of five anti-PD-1 immunotherapy cohorts receiving Nivolumab or Pembrolizumab (n = 333, I^2 = 6.12%) (Supplemental_table S5) [31–34], we observed that elevated EC_score correlated with worse survival in glioblastoma (GBM) (PRJNA482620; HR = 12.02; 95% CI, 1.88–76.71), melanoma (GSE91061; HR = 5.27; 95% CI, 1.29–21.58) and advanced clear cell renal cell carcinoma (ccRCC) (Braun_2020; HR = 4.58; 95% CI, 1.83–11.46). Overall, the meta-analysis indicated that elevated EC_score is associated with reduced effectiveness of immunotherapy (HR = 4.21; 95% CI, 1.54–11.50) (Fig. 3B). Notably, high EC_score ccRCC patients exhibited reduced median OS (20.2 vs. 22.4 months, p = 0.006) [32] (Fig. 3C), while high EC_score advanced melanoma patients also showed decreased median OS (12.7 vs. 16.1 months, p = 0.045) [33] (Fig. 3D). Collectively, these findings highlighted that EC_score serves as a valuable prognostic biomarker for forecasting anti-PD-1 treatment efficacy, potentially enabling improved patient stratification for immunotherapy.

High EC_score samples exhibited upregulated immunosuppressive pathways alongside decreased immune infiltrations

Following the observed association between elevated EC_score and poor immunotherapy response, we investigated its potential mechanisms. Known immune evasion mechanisms include the downregulation of antigen processing and presentation (APP) [35], overexpression of exhausted T cells molecules [36] and dysregulation of the transforming growth factor beta (TGF-β) signaling pathway [37]. In melanoma, EC_score inversely correlated with MHC-I (HLA-A/B/C) and MHC-II (HLA-DQ1/DM) expression (Fig. 4A), consistent with previous

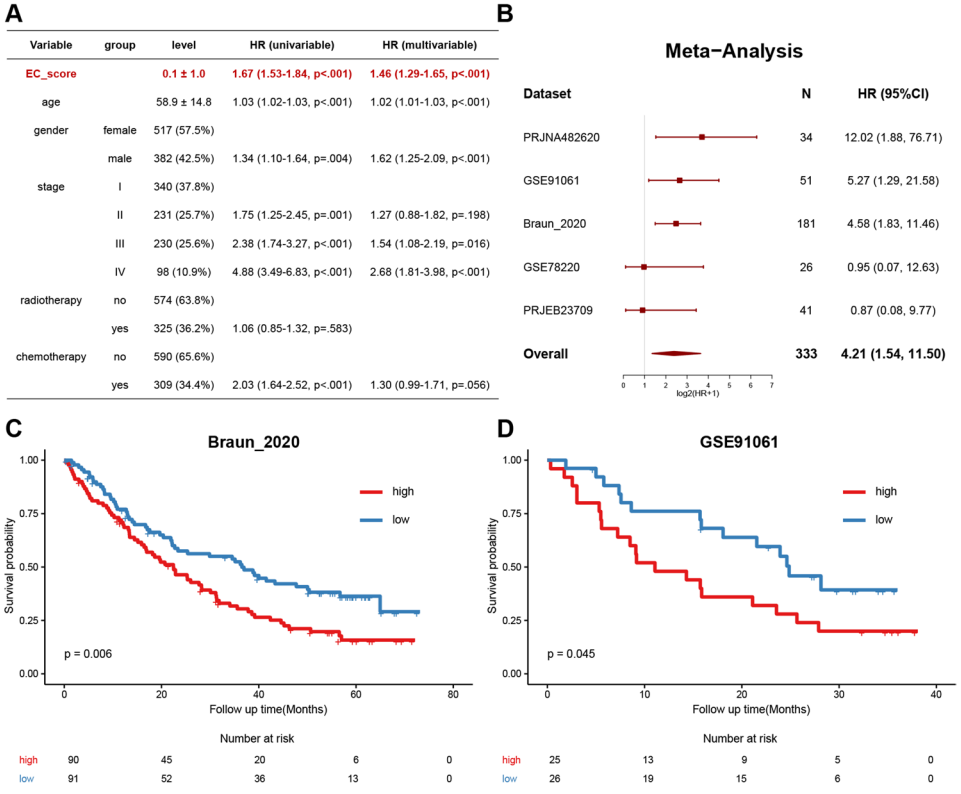


Fig. 3 EC_score serves as an indicator of poor prognosis in cancer immunotherapy. (A) Univariate and multivariate Cox regression analyses of EC_score in TCGA cohort. (B) Meta-analysis of five anti-PD-1 immunotherapy cohorts. (C-D) Kaplan-Meier survival curves for ccRCC and melanoma patients stratified by EC_score. *P* values were calculated using the log-rank test

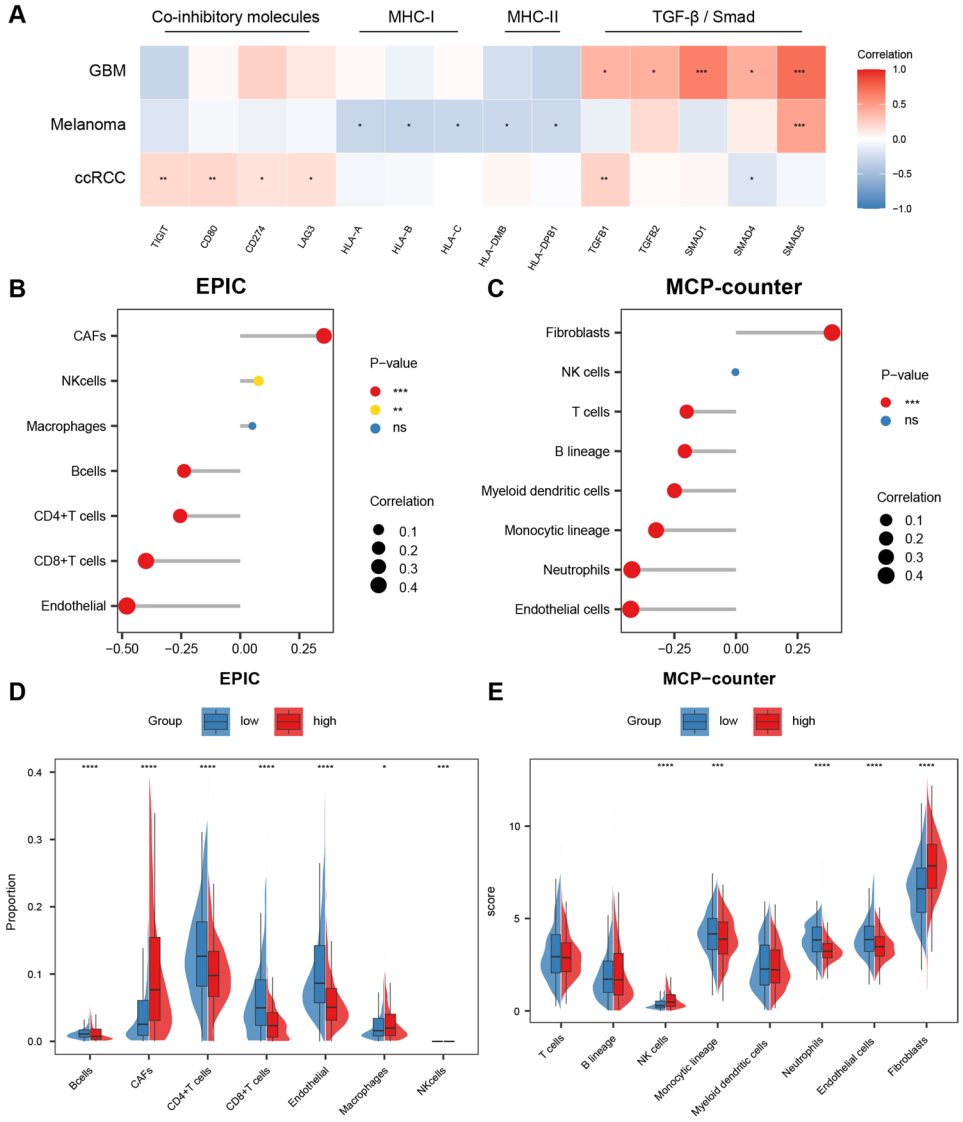


Fig. 4 Association between EC_score and tumor immunity. (A) Correlations between EC_score and immune markers in immunotherapy cohorts. *P* values were calculated using Spearman's test, ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$. (B) Correlations between EC_score and immune cell proportions predicted by EPIC in TCGA cohort. *P* values were calculated using Spearman's test. (C) Correlations between EC_score and immune cell signature scores predicted by MCP-counter in TCGA cohort. *P* values were calculated using Spearman's test. (D) Proportions of immune cells predicted by EPIC in the TCGA cohort, stratified by the median value of EC_score. *P* values were calculated with the Wilcoxon test, ns $p > 0.05$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$, **** $p < 0.001$. (E) Immune cell signature scores predicted by MCP-counter in the TCGA cohort, stratified by the median value of EC_score. *P* values were calculated with the Wilcoxon test

reports linking MHC deficiency to APP and anti-PD-1 [38–40]. Moreover, prolonged antigen exposure can lead to T cell functional exhaustion, characterized by the expression of co-inhibitory immune checkpoint receptors such as CTLA-4, TIGIT, B71 (CD80), B7H1 (CD274) [41]. In our analysis, ccRCC samples revealed positive associations between EC_score and exhaustion markers (Fig. 4A), aligning with prior evidence that low T-effector signatures predict reduced PFS in RCC patients receiving immune therapy [42]. Additionally, we observed a positive correlation between EC_score and TGF-β signaling in GBM patients (Fig. 4A), reinforcing the mediation role of TGF-β signaling in GBM-associated immunosuppression [43]. Collectively, these findings suggested that EC_score reflects diverse immunosuppressive mechanisms in different cancer types, including downregulation of antigen presentation, T cell exhaustion and activation of TGF-β signaling pathways.

Given the critical role of immune infiltration in antitumor immunity [41], we assessed immune cell proportions using EPIC algorithm in TCGA cohorts [22]. We found that EC_score was negatively correlated with the proportions of endothelial (R = -0.48, p < 2.2e-16), CD8 + T cells (R = -0.40, p < 2.2e-16), CD4 + T cells (R

= -0.25, p < 2.2e-16) and B cells (R = -0.24, p < 2.2e-16), while positively correlated with cancer-associated fibroblasts (CAFs) (R = 0.35, p < 2.2e-16) (Fig. 4B). These findings were corroborated by MCP-counter algorithm [23], which also indicated negative associations between EC_score and the infiltrations of immune cells such as myeloid dendritic cells (R = -0.25, p < 2.2e-16), B lineage (R = -0.21, p = 1.5e-14) and T cells (R = -0.20, p = 1.4e-13) (Fig. 4C). Besides, high EC_score tumors further demonstrated depleted B and T cell populations (Fig. 4D-E). Collectively, these findings underscore the association between high EC_score and immune infiltrations, particularly the deficiency of lymphocytes in tumors, potentially contributing to immunotherapy resistance.

Discussion

Extrachromosomal circular DNA has emerged as a crucial determinant in tumor evolution and therapy resistance [7, 11, 12]. However, the transcriptomic relationship between ecDNA and immunotherapy outcomes remains poorly elucidated. To address this gap, our study developed EC_score, a transcriptome-based 12-gene signature that reliably predicts ecDNA presence across diverse datasets, cell lines and clinical samples. Importantly, EC_score effectively demonstrated clinical utility in stratifying immunotherapy response and survival outcomes, while revealing tumor-specific immunosuppressive mechanisms associated with ecDNA-related immune evasion. This study advances our understanding of the relationship between ecDNA and cancer immunotherapy, offering novel perspectives for predicting cancer immunotherapy response.

Current ecDNA detection methodologies face substantial limitations [11]. While FISH provides high specificity [3, 6, 12], its low throughput and probe dependency restrict clinical scalability [12, 44]. Circle-seq enables genome-wide ecDNA screening but requires complex experimental workflows [45]. CRISPR/Cas9-based approaches like CRISPR-CATCH [46] and CRISPR-C [47] offer high-resolution analysis for targeting ecDNA [48], but remain high costs and technical complexity for large cohorts [7]. In terms of bioinformatic analysis, AA is currently the most recognized tool through WGS junction analysis [12]. This approach reconstructs large-scale ecDNA structures from the extending WGS-derived breakpoint data. GCAP, a machine learning tool, analyzes genomic characteristics in WES data [7]. ATAC-seq based tools like Circle-finder [49] and ATACAmp [15] exploit the enhanced chromatin accessibility that distinguishes ecDNA from linear DNA duplications [12]. At the single-cell level, scATAC-seq enables high-resolution identification of ecDNA amplification [50]. scEC&T-seq allows sequencing through Smart-seq2, but complex ecDNA enrichment procedures result in low throughput and high costs [16].

Among multi-omics data, RNA-seq has emerged as the most accessible and cost-effective resource, making it the predominant choice for biomarker discovery and tumor transcriptome characterization [51]. A recent study has introduced a "CorEx" gene set which consisted of more than 600 genes to predict ecDNA presence (AUC = 0.636) [24]. In our study, we developed EC_score for ecDNA prediction to achieve comparable predictive accuracy (AUC >0.70) with a 12-gene panel (*DSCC1*, *MTBP*, *HOXC13*, *POU6F2*, *FAM72B*, *FAM72D*, *MAGED4B*, *NKX2-5*, *PLAC1*, *FOXI3*, *KRTAP4-1*, *OTX2*). The association between EC_score and cell cycle activation aligns with established ecDNA biology [6], further validating the predictive ability of EC_score. Notably, although EC_score effectively predicts ecDNA, it does not incorporate canonical ecDNA-amplified oncogenes (e.g., MYC, MYCN, ERBB2). Instead, EC_score captures shared transcriptomic features of ecDNA-positive tumors, such as cell cycle activation, as established in prior studies [10, 24]. Despite its predictive power across cohorts, the EC_score is biologically uninterpretable for individual samples via ssGSEA. We therefore recommend embedding samples into a same-tumor-type or pan-cancer reference cohort, then estimating ecDNA likelihood using published type-specific frequencies [5]. Additionlly, due to the scarcity of samples simultaneously possessing RNAseq data and WGS data for ecDNA prediction, our external validation cohorts with small sample size may introduce potential bias when estimating the predictive AUC value [52]. To address this, we validated the performance of the EC_score across multiple datasets, aiming to mitigate the impact of small sample sizes on the ROC curve. In summary, the EC_score still requires validation in cohorts with a larger sample size, awaiting further investigation in future study.

To develop the EC_score signature, we employed a combination of LASSO, RF and RFE for feature selection. This multi-method approach was chosen to leverage their complementary strengths in handling high-dimensional transcriptomic data. LASSO provides effective regularization and variable shrinkage, which is critical for reducing redundancy in gene expression features [53]. RF contributes robustness to non-linear relationships and multicollinearity while offering intrinsic feature importance evaluation [54]. RFE further refines the feature set through iterative elimination, enhancing stability and generalizability [55]. While other algorithms, such as SVM and XGBoost are valuable for prediction tasks, they are less directly suited for interpretable feature selection in high-dimensional settings due to limitations in inherent feature ranking or increased overfitting risk [56, 57]. Future studies may benefit from other modeling methods

like nested cross-validation and advanced machine learning techniques to further optimize gene selection.

To validate the EC_score, we performed FISH focusing on MYC and MYCN [28, 58]—frequent ecDNA-associated oncogenes in SCLC—using selected cell lines and a clinical sample showing high expression of these genes. Although results supported EC_score prediction, limitations include small sample size and restricted gene coverage, underscoring the need for broader validation across more genes and larger cohorts.

Besides, another study has predicted the association between ecDNA presence and poor immunotherapy outcomes in gastric and nasopharyngeal carcinomas using WES data [7]. However, due to the limited identification of ecDNA in immunotherapy datasets, more evidence is needed to confirm the relationship between ecDNA and cancer immunotherapy outcomes. In this study, we applied EC_score through public immunotherapy cohorts to evaluate its potential as a biomarker for immunotherapy efficacy. Multivariate Cox regression analysis and meta-analysis identified EC_score as an independent risk factor for poor anti-PD-1 immunotherapy response, with high EC_score patients showing shorter OS in ccRCC and melanoma cohorts. In addition, to reduce potential bias from the limited sample sizes, we integrated multiple immunotherapy cohorts for validation (Supplemental_table S5). The low heterogeneity (I^2 = 6.12%) of this integration not only strengthened the validation of the EC_score's predictive performance and reduced the impact of outliers, but also yielded a significant result (HR = 4.21; 95% CI, 1.54–11.50), which has partially mitigated the bias caused by small sample sizes. However, the low heterogeneity of these integrated cohorts has also restricted the generalizability of the EC_score, necessitating further research to expand the sample size for enhanced validation.

Although the association between ecDNA and poor cancer immunotherapy responses has been observed, the underlying mechanisms remain to be explored [6]. Prior study proposed that ecDNA may contribute to immune evasion by downregulation of antigen presentation [8]. A recent study revealed that ecDNA harboring immunomodulatory genes is associated with depletion of T cell [5]. In addition to pan-cancer analysis, ecDNA identified in oesophageal adenocarcinoma or Barrett's oesophagus carries immunosuppressive genes such as SOCS1 and CIITA, which were closely associated with pathways such as cytotoxic T-cell immune evasion [59]. In urothelial carcinoma, ecDNA has been found to correlate with reduced or absent levels of HLA-I molecules [10], indicating downregulation of the antigen presentation pathway. In an analysis of the tumor microenvironment in pancreatic ductal adenocarcinoma, it was also found that the proportion of fibroblasts was significantly increased in tumors harboring ecDNA, suggesting a potential mechanism that promotes tumor immune evasion [60]. Here, our findings revealed tumor-specific mechanisms linking ecDNA to immune evasion. Specifically, high EC_score melanomas exhibited downregulation of APP, while ccRCC patients expressed co-inhibited immune checkpoint receptors and GBM demonstrated activation of TGF-β signaling pathways. These distinct molecular signatures suggest that ecDNA-mediated mechanisms of immunotherapy resistance vary across tumor types. Interestingly, neither previous literature reports nor the findings of this study have identified an association between ecDNA-mediated tumor immune evasion and specific genes carried by ecDNA. Instead, it has been observed that ecDNA-driven tumor immune evasion exhibits tumor heterogeneity, which may be related to the unique biological characteristics of different tumor types. For instance, previous studies have reported cancer-type-specific immune evasion mechanisms: ccRCC patients who had low scores of effector T-cell exhibited poor prognosis following immunotherapy [42]; downregulation of MHC II induced resistance to anti-PD-1 therapy in melanoma [39]; and the TGF-β signaling pathway mediates immunosuppression in glioblastoma [43]. All these findings correspond to the associations we observed between the EC_score and tumor-type-specific immune evasion pathways in these malignancies. However, additional experimental evidence and expanded cohort investigations are still required to validate these relationships.

Overall, our findings provide novel insights into the relationship between ecDNA and tumor immune evasion. We demonstrated that high EC_score correlates with multiple immunosuppressive mechanisms contributing to poor immunotherapy responses, including impaired antigen presentation, T cell exhaustion, and enhanced TGF-β signaling. Furthermore, we validated the immunosuppressive status in high EC_score tumors through immune deconvolution analyses EPIC and MCP-counter algorithms. These findings offer new directions for further research in ecDNA-mediated immune evasion and developing novel therapeutic strategies for cancer treatment.

However, several limitations of this study should be noted. First, while EC_score has demonstrated efficacy in predicting ecDNA across several cancers, its generalizability needs to be evaluated beyond the 22 tumor types currently represented in TCGA datasets. Second, although the efficacy of EC_score in ecDNA prediction was verified by FISH in cancer cell lines and pathological samples, more experimental evidence would further confirm its predictive ability. Third, while the biological functions of OTX2 and MAGED4B are consistent with the EC_score, their expression levels in the TCGA cohort

are relatively low. These low expressions may potentially compromise the accuracy of detection, highlighting the need for further investigation into the average expression levels in larger cohorts. Fourth, while we tried to reduce the risk of overfitting, the potential limitation of overfitting in machine learning are still unavoidable, requiring further technological innovations to address this issue. Furthermore, the association between EC_score, proliferative index and CAFs suggested a potential confounding factor for immunotherapy response, warranting further investigation at the single-cell level. Finally, while our study demonstrates correlations among EC_score, ecDNA existence and poor immunotherapy responses, the causal relationship between ecDNA and immunotherapy requires further investigation in clinical cohorts.

Abbreviations

AA	AmpliconArchitect
AUC	area under the ROC curve
APP	antigen processing and presentation
ATAC-seq	assay for transposase-accessible chromatin using sequencing
CCLE	Cancer Cell Line Encyclopedia
ccRCC	clear cell renal cell carcinoma
ecDNA	extrachromosomal circular DNA
EGA	European Genome-Phenome Archive
FFPE	formalin-fixed paraffin-embedded
FISH	fluorescence in situ hybridization
GBM	glioblastoma
GO	gene ontology
GSEA	gene set enrichment analysis
LASSO	Least Absolute Shrinkage and Selection Operator
OS	overall survival
RF	Random Forest
RFE	Recursive Feature Elimination
PFS	progression-free survival
RNA-seq	RNA sequencing
ROC	receiver operating characteristic curve
SCLC	small cell lung cancer
SVM	Support Vector Machine
ssGSEA	single sample gene set enrichment analysis
TCGA	The Cancer Genome Atlas Program
TGF-β	transforming growth factor beta
WGS	whole-genome sequencing
WES	whole-exome sequencing
XGBoost	eXtreme Gradient Boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12935-025-04056-7.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements
The authors would like to thank the researchers who provided open access to the RNA-seq data.

Author contributions
Y. J. and J. Z. designed and performed the experiments. Y. J. and J. D. conducted the bioinformatics analysis. Y. J., X. W. and H. Y. wrote the paper. X. L. and C. T. participated in the revision of the draft. All authors read and approved the final paper.

Funding
This study was funded by the Peking University People's Hospital Research and Development Funds (Grant number RDX2023-04).

Data availability
The data generated in this study were available in the supplementary files or from the corresponding authors upon reasonable request. Pan-cancer RNA-seq information from TCGA was retrieved from the Genomic Data Commons Data Portal (http://portal.gdc.cancer.gov/projects/). Publicly available SCLC RNA-seq data (EGAS00001000925) in this study were obtained from the EGA database (www.ebi.ac.uk/ega/home). Immunotherapy cohorts data were retrieved from Braun_2020 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7499153IF: 50.0 Q1 B1/), GSE91061 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE91061), GSE78220 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE78220), PRJNA482620 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6810613IF: 50.0 Q1 B1/), PRJEB73280(https://www.ebi.ac.uk/ena/browser/view/PRJEB23709). CCLE cell lines expression profiles (n = 1114) were downloaded from DepMap (https://depmap.org/portal/) with log2(TPM + 1) transformation.

Declarations

Ethics statement
This study was approved by the Institutional Review Board at Peking University People's Hospital (2023PHB216-001).

Informed consent
Explicit consent was obtained for experimental use of pathological specimens, with all personal identifiers removed.

Animal studies
N/A.

Competing interests
The authors declare no competing interests.

Author details
[1]Thoracic Oncology Institute, Department of Thoracic Surgery, Peking University People's Hospital, Beijing 100044, China
[2]Research Unit of Intelligence Diagnosis and Treatment in Early Non-small Cell Lung Cancer, Chinese Academy of Medical Sciences, Peking University People's Hospital, Beijing 2021RU002, 100044, China
[3]Institute of Advanced Clinical Medicine, Peking University, Beijing, China
[4]Peking University Health Science Center, Beijing 100191, China
[5]Department of Medical Bioinformatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China
[6]Department of Human Anatomy, Histology and Embryology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

Received: 10 April 2025 / Accepted: 22 October 2025
Published online: 18 December 2025

References

1. Lange JT, et al. The evolutionary dynamics of extrachromosomal DNA in human cancers. Nat Genet. 2022;54(10):1527–33.
2. Verhaak RGW, et al. Extrachromosomal oncogene amplification in tumour pathogenesis and evolution. Nat Rev Cancer. 2019;19(5):283–8.
3. Turner KM, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017;543(7643):122–5.
4. Morton AR, et al. Functional enhancers shape extrachromosomal oncogene amplifications. Cell. 2019;179(6):1330–e134113.

5. Bailey C, et al. Origins and impact of extrachromosomal DNA. Nature. 2024;635(8037):193–200.
6. Kim H, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. Nat Genet. 2020;52(9):891–7.
7. Wang S, et al. Machine learning-based extrachromosomal DNA identification in large-scale cohorts reveals its clinical implications in cancer. Nat Commun. 2024;15(1):1515.
8. Wu T, et al. Extrachromosomal DNA formation enables tumor immune escape potentially through regulating antigen presentation gene expression. Sci Rep. 2022;12(1):3590.
9. Wang Y, et al. EccDNAs are apoptotic products with high innate immunostimulatory activity. Nature. 2021;599(7884):308–14.
10. Lv W, et al. Spatial-Temporal diversity of extrachromosomal DNA shapes urothelial carcinoma evolution and Tumor-Immune microenvironment. Cancer Discov; 2025.
11. Yan X, et al. Extrachromosomal DNA in cancer. Nat Rev Cancer. 2024;24(4):261–73.
12. Wu S, et al. Circular EcDNA promotes accessible chromatin and high oncogene expression. Nature. 2019;575(7784):699–703.
13. Møller HD, et al. Genome-wide purification of extrachromosomal circular DNA from eukaryotic cells. J Vis Exp. 2016;(110):e54239.
14. Deshpande V, et al. Exploring the landscape of focal amplifications in cancer using ampliconarchitect. Nat Commun. 2019;10(1):392.
15. Cheng H, et al. ATACAmp: a tool for detecting ecDNA/HSRs from bulk and single-cell ATAC-seq data. BMC Genomics. 2023;24(1):678.
16. Chamorro González R, et al. Parallel sequencing of extrachromosomal circular DNAs and transcriptomes in single cancer cells. Nat Genet. 2023;55(5):880–90.
17. Schwarze K, et al. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. Genet Med. 2018;20(10):1122–30.
18. Hoogstrate Y, et al. Transcriptome analysis reveals tumor microenvironment changes in glioblastoma. Cancer Cell. 2023;41(4):678–92. .e7.
19. George J, et al. Comprehensive genomic profiles of small cell lung cancer. Nature. 2015;524(7563):47–53.
20. Hänzelmann S, et al. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7.
21. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
22. Racle J, et al. Simultaneous enumer cancer immune cell types bulk tumor gene expression data. Elife. 2017;6:e26476.
23. Becht E, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. 2016;17(1):218.
24. Lin MS, et al. Transcriptional immune suppression and up-regulation of double-stranded DNA damage and repair repertoires in ecDNA-containing tumors. Elife. 2024;12:RP88895.
25. Rose JC, et al. Disparate pathways for extrachromosomal DNA biogenesis and genomic DNA repair. Cancer Discov. 2025;15(1):69–82.
26. Engel JL, et al. The Fanconi anemia pathway induces chromothripsis and ecDNA-driven cancer drug resistance. Cell. 2024;187(21):6055–6070.e22.
27. van Leen E, et al. The genomic and Spatial mobility of extrachromosomal DNA and its implications for cancer therapy. Nat Genet. 2022;54(2):107–14.
28. Pongor LS, et al. Extrachromosomal DNA amplification contributes to small cell lung cancer heterogeneity and is associated with worse outcomes. Cancer Discov. 2023;13(4):928–49.
29. Gilbreath C, et al. Robust detection of gene amplification in Formalin-Fixed Paraffin-Embedded samples by fluorescence in situ hybridization. J Vis Exp. 2024;(209):10.3791/66978.
30. Wu S, et al. Extrachromosomal DNA: an emerging hallmark in human cancer. Annu Rev Pathol. 2022;17:367–86.
31. Chen Z, et al. TIGER: A web portal of tumor immunotherapy gene expression resource. Genomics Proteom Bioinf. 2023;21(2):337–48.
32. Braun DA, et al. Interplay of somatic alterations and immune infiltration modulates response to PD-1 Blockade in advanced clear cell renal cell carcinoma. Nat Med. 2020;26(6):909–18.
33. Riaz N, et al. Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell. 2017;171(4):934–e94916.
34. Zhao J, et al. Immune and genomic correlates of response to anti-PD-1 immunotherapy in glioblastoma. Nat Med. 2019;25(3):462–9.
35. Dersh D, et al. Genome-wide screens identify Lineage- and Tumor-Specific genes modulating MHC-I- and MHC-II-Restricted immunosurveillance of human lymphomas. Immunity. 2021;54(1):116–e13110.
36. Sharma P, et al. Immune checkpoint therapy-current perspectives and future directions. Cell. 2023;186(8):1652–69.
37. Yi M, et al. Targeting cytokine and chemokine signaling pathways for cancer therapy. Signal Transduct Target Ther. 2024;9(1):176.
38. Jhunjhunwala S, et al. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. Nat Rev Cancer. 2021;21(5):298–312.
39. James JL, et al. Polycomb repressor complex 2 suppresses interferon-responsive MHC-II expression in melanoma cells and is associated with anti-PD-1 resistance. J Immunother Cancer. 2023;11(11):e007736.
40. Rodig SJ, et al. MHC proteins confer differential sensitivity to CTLA-4 and PD-1 Blockade in untreated metastatic melanoma. Sci Transl Med. 2018;10(450):eaar3342.
41. Chow A, et al. Clinical implications of T cell exhaustion for cancer immunotherapy. Nat Rev Clin Oncol. 2022;19(12):775–90.
42. McDermott DF, et al. Clinical activity and molecular correlates of response to Atezolizumab alone or in combination with bevacizumab versus Sunitinib in renal cell carcinoma. Nat Med. 2018;24(6):749–57.
43. Rocha Pinheiro SL, et al. Immunotherapy in glioblastoma treatment: current state and future prospects. World J Clin Oncol. 2023;14(4):138–59.
44. Hung KL, et al. EcDNA hubs drive cooperative intermolecular oncogene expression. Nature. 2021;600(7890):731–6.
45. Prada-Luengo I, et al. Sensitive detection of circular DNAs at single-nucleotide resolution using guided realignment of partially aligned reads. BMC Bioinformatics. 2019;20(1):663.
46. Hung KL, et al. Targeted profiling of human extrachromosomal DNA by CRISPR-CATCH. Nat Genet. 2022;54(11):1746–54.
47. Møller HD, et al. CRISPR-C: circularization of genes and chromosome by CRISPR in human cells. Nucleic Acids Res. 2018;46(22):e131.
48. Jinek M, et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–21.
49. Kumar P, et al. ATAC-seq identifies thousands of extrachromosomal circular DNA in cancer and cell lines. Sci Adv. 2020;6(20):eaba2489.
50. Zhu Q, et al. Single cell multi-omics reveal intra-cell-line heterogeneity across human cancer cell lines. Nat Commun. 2023;14(1):8170.
51. Hutter C, Zenklusen JC. The cancer genome atlas: creating lasting value beyond its data. Cell. 2018;173(2):283–5.
52. Hanczar B, et al. Small-sample precision of ROC-related estimates. Bioinformatics. 2010;26(6):822–30.
53. Freijeiro-González L, et al. A critical review of LASSO and its derivatives for variable selection under dependence among covariates. Int Stat Rev. 2022;90(1):118–45.
54. Breiman L. Random Forests Mach Learn. 2001;45(1):5–32.
55. Guyon IM, et al. An introduction to variable and feature selection. The Journal of Machine Learning Research; 2003.
56. Cheng H, et al. Nonlinear Feature Selection by Relevance Feature Vector Machine. In: Proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition. Leipzig, Germany: Springer-Verlag; 2007. pp. 144–159.
57. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
58. Pal Choudhuri S, et al. Acquired Cross-Resistance in small cell lung cancer due to extrachromosomal DNA amplification of MYC paralogs. Cancer Discov. 2024;14(5):804–27.
59. Luebeck J, et al. Extrachromosomal DNA in the cancerous transformation of barrett's oesophagus. Nature. 2023;616(7958):798–805.
60. Fiorini E, et al. MYC EcDNA promotes intratumour heterogeneity and plasticity in PDAC. Nature. 2025;640(8059):811–20.

Publisher's note