A multi-gene predictive model for the radiation sensitivity of nasopharyngeal carcinoma based on machine learning

Kailai Li[1†], Junyi Liang[1†], Nan Li[2†], Jianbo Fang[1†], Xinyi Zhou[1], Jian Zhang[1], Anqi Lin[1*], Peng Luo[1*] and Hui Meng[1*]

[1] Department of Oncology, Zhujiang Hospital, Southern Medical University, Guangzhou, Guangdong, China.

[2] Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, Guangdong, China.

* Correspondence:

Anqi Lin: smulinanqi0206@i.smu.edu.cn ; Peng Luo: luopeng@smu.edu.cn ; Hui Meng: menghui8816@163.com

† Joint Authors. These authors have contributed equally to this work and share first authorship.

Abstrct

Background: Radiotherapy resistance is a major obstacle to the long-term survival of nasopharyngeal cancer patients, as it is a primary cause of recurrence and metastasis. Identifying radiotherapy-associated biomarkers can help improve the survival prognosis of nasopharyngeal cancer patients. Consequently, discovering biomarkers associated with radiosensitization is crucial.

Methods: We evaluated 113 combinations of machine learning algorithms and ultimately selected 48 to construct a radiotherapy sensitivity score (NPC-RSS) that can predict radiosensitivity in nasopharyngeal cancer patients. Furthermore, we analyzed the relationship between NPC-RSS and the expression of genes associated with immune and radiotherapy sensitivity profiles. We employed GSEA and ssGSEA to investigate the connection between NPC-RSS and signaling pathways.

Results: We selected the combined model glmBoost+NaiveBayes, which had the best AUC among 48 models, for our subsequent study. The NPC-RSS, built based on the 18 genes included in this model, can predict the results of the public dataset and the in-house dataset of Zhujiang Hospital, Southern Medical University, with considerable efficiency. The key genes of NPC-RSS are closely associated with immune characteristics, including chemokine and chemokine receptor families, and histocompatibility complex (MHC), and show more active immune processes. Meanwhile, these key genes were significantly associated with the expression of radiosensitization-related genes. Furthermore, GSVA and GSEA analyses demonstrated that different expression levels of key NPC-RSS genes influenced signaling pathways, such

as the Wnt/β-catenin signaling pathway, JAK-STAT signaling pathway,NF-kappa B signaling pathway and T cell receptor signaling pathway, which are associated with immunity and disease progression. The consistency of the expression of key genes SMARCA2 and CD9 with NPC-RSS was validated in in-house cell lines. The radiosensitive group, classified according to NPC-RSS, exhibited a more enriched and activated state of immune infiltration compared to the radioresistant group. Moreover, in single-cell samples, NPC-RSS was higher in the radiotherapy-sensitive group, with immune cells playing a predominant role.

Conclusions: In this study, we used machine learning to construct a predictive score, called NPC-RSS, associated with radiosensitivity in nasopharyngeal carcinoma patients; moreover, NPC-RSS is strongly associated with immune characteristics, expression of radiosensitivity-related genes, and signaling pathways related to disease progression. We hope that the NPC-RCC will enable more precise selection of the NPC population of potential beneficiaries of radiation therapy.

Keywords: Nasopharyngeal Cancer; Machine Learning; Radiotherapy Sensitivity; Resistance; Biomarker

Introduction

Nasopharyngeal carcinoma (NPC), which originates from the nasopharyngeal mucosal epithelium, is a malignant tumor with distinct epidemiological, histopathological, clinical, and therapeutic features among head and neck tumors[1-3]. According to the most recent global cancer statistics, a total of 133,354 new cases of nasopharyngeal cancer were reported, accounting for only 0.7% of all new cancer cases that year[4]. Although the incidence of nasopharyngeal cancer is relatively low compared with other cancer types, it has a unique geographic distribution. About 70% of new nasopharyngeal cancer cases occur in Eastern and Southeastern Asia, while the rest are mainly in South-Central Asia, North Africa, and East Africa[5]. Currently, radiotherapy is used as the primary curative treatment for patients with nasopharyngeal carcinoma. With the continuous advancement of radiotherapy technology, recent studies have shown that hyperfractionated intensity-modulated radiotherapy increases the 3-year overall survival rate in these patients from 55.0% to 74.6% compared to conventional fractionated intensity-modulated radiotherapy[6]. However, approximately 20-30% of patients experience locoregional recurrence of the tumor due to radiation resistance[7-9]. Studies have shown that radiotherapy resistance is the main reason for treatment failure in nasopharyngeal cancer patients [10, 11]. Given the importance of radiotherapy in nasopharyngeal cancer treatment, it is imperative to investigate the molecular mechanisms of radiotherapy resistance

and develop strategies to enhance the radiosensitivity of nasopharyngeal cancer cells.

In recent years, artificial intelligence (AI) has been deeply integrated into the medical field[12]. Machine learning (ML), an important branch of AI, has the ability to process nonlinear data, discover new patterns, and generate predictive models after learning from existing data. Machine learning plays a crucial role in developing predictive markers for tumor treatment and prognosis. Recent studies have focused on utilizing machine learning techniques to construct predictive models for these purposes. These studies employ various methods and algorithms, including integrated learning, immune-derived feature extraction, and AI network-guided signatures. For instance, one study utilized machine learning through tumor MR image feature mapping for patient stratification in adjuvant chemotherapy for locally advanced nasopharyngeal carcinoma[13]. Another study employed machine learning based on MRI/DWI radiomics signatures to predict the prognosis of nasopharyngeal carcinoma[14]. Additionally, machine learning-based algorithms have been developed to create predictive models for identifying lncRNAs associated with tumor-infiltrating immune cells, aiming to improve prognosis and immunotherapy response in cancer patients[15, 16]. These studies aim to enhance patient outcomes and prognosis while providing superior predictive power for personalized clinical treatment strategies. However, there is still a lack of machine learning-based biomarkers for predicting radiotherapy sensitivity in nasopharyngeal carcinoma.

In this study, a combination of multiple machine learning methods and transcriptomic data from nasopharyngeal cancer patients was used to construct a biomarker that can predict the sensitivity to radiotherapy in nasopharyngeal cancer patients, termed the Nasopharyngeal Carcinoma Radiosensitivity Score (NPC-RSS). In addition, we explored the biological mechanisms underlying the relationship between NPC-RSS and radiotherapy response in nasopharyngeal cancer patients. We anticipate that the construction of NPC-RSS will guide the selection of radiotherapy strategies for nasopharyngeal carcinoma patients, thereby improving the clinical outcomes and prognosis of these patients.

Materials and Methods

1.NPC patient data sources and data processing

We downloaded chemotherapy-related gene expression data from the Gene Expression Omnibus (GEO) database for 20 nasopharyngeal carcinoma (NPC) patients (GSE32389). Additionally, we collected samples from 34 NPC patients treated at Zhujiang Hospital of Southern Medical University between March 2017 and July 2021 (in-house cohort). The clinical characteristics of the

NPC patients in the in-house cohort are detailed in Supplementary Table 1. Written informed consent was obtained from the NPC patients whose nasopharyngeal carcinoma tissues were collected, and the study was approved by the Ethics Committee of Zhujiang Hospital of Southern Medical University. In this study, the in-house cohort was used as the training set, and GSE32389 was used as the validation set. The raw transcriptome sequencing data from the in-house cohort underwent quality control and preprocessing, which included data quality assessment, removal of low-quality reads, and removal of adapter or primer sequences using tools such as Trimmomatic and FastQC. Subsequently, the pre-processed reads were aligned to a reference genome or transcriptome using the alignment software HISAT2 to identify and localize RNA sequences. The aligned reads were then quantified using featureCounts to obtain the number of reads per gene or transcript. Gene expression levels were then quantified by normalizing the read counts to the length of each gene and the total number of reads, calculating the Fragments Per Kilobase Million (FPKM) values. The data for the GEO database were obtained from the Affymetrix® GPL570 platform. Microarray data from the Affymetrix platform were background corrected, quantile normalized, and log2 transformed using the R package "affy" with the RMA algorithm[17].

2.Construction of NPC-RSS

In the training set, we performed differential analysis of the gene expression data between nasopharyngeal carcinoma patients who were sensitive or resistant to radiotherapy using the limma R package. This analysis identified 71 differentially expressed genes (DEGs) ($P < 0.05$ & $|LogFC| > 0.585$, Supplementary Table 2)[18]. Next, we used these DEGs to develop a binary classification model using machine learning algorithms to predict radiotherapy sensitivity in nasopharyngeal carcinoma patients. In the data preprocessing stage, we performed robust multi-array average (RMA) normalization on the gene expression data to ensure that it met the input requirements of the machine learning algorithms, which assume a normal distribution.

(1) First, we integrated a combination of 12 machine learning algorithms, including Least Absolute Shrinkage and Selection Operator (LASSO), Ridge, Elastic Net (Enet), stepwise generalized linear models (Stepglm), support vector machines (SVM), gradient boosting with component-wise linear models (glmBoost), linear discriminant analysis (LDA), partial least squares generalized linear models (plsRglm), random forest (RF), gradient boosting machine (GBM), extreme gradient boosting (XGBoost), and naive Bayes. Among these, RF, LASSO, glmBoost, Stepglm with both forward and backward selection (Stepglm[both]), and Stepglm with backward selection (Stepglm[backward]) can perform feature selection and dimensionality reduction. We combined these algorithms with the others to form 113 combinations of machine learning algorithms. In each of these 113

combinations, one algorithm was used for feature selection, while the other was used to construct a binary classification model. The performance of these models was evaluated using cross-validation to ensure robustness. (2) Next, in the training set, we used these 113 combinations to construct predictive models using the expression data of the 71 DEGs. A total of 48 models were successfully constructed. (3) In the validation set, we used these predictive models to calculate a radiotherapy sensitivity score for each nasopharyngeal carcinoma (NPC) patient. (4) We evaluated the predictive performance of these models in the training and validation sets using receiver operating characteristic (ROC) curves and the average area under the curve (AUC). The final model was named the Nasopharyngeal Carcinoma Radiosensitivity Score (NPC-RSS).

3.Validating key genes of NPC-RSS at the cellular level

3.1 Cell culture

The CNE2 nasopharyngeal carcinoma cell line, previously preserved by the Pearl River Hospital Center for Translational Medicine, was cultured in RPMI 1640 medium supplemented with 10% fetal bovine serum and 1% penicillin/streptomycin. The cells were maintained in a humidified incubator at 37°C with 5% CO_2. Cell passaging and division ratios were carefully monitored and maintained within optimal ranges to ensure cell health and experimental consistency. Regular cell line authentication was performed to ensure the use of the correct cell line. Stringent quality control measures were implemented for the culture medium and fetal bovine serum to minimize experimental variability. Precise addition of penicillin/streptomycin was ensured to maintain bacterial resistance while minimizing potential adverse effects on cell growth and viability. Appropriate cryopreservation and thawing protocols were followed to ensure optimal cell preservation and recovery. Strict adherence to aseptic techniques was maintained throughout the cell culture process to prevent bacterial and fungal contamination. Continuous monitoring of the incubator's internal environment was performed to maintain stable CO_2 concentration and temperature. Meticulous execution of these steps is crucial for maintaining the stability and reliability of the cell culture system.

3.2 Cell irradiation and construction of CNE2-RS radiation-resistant cells.

A 6 MV photon beam was generated at room temperature using a linear accelerator (Elekta, SYNERGY PL, Switzerland) and delivered to the cells at a 180° gantry angle with a dose rate of approximately 600 cGy/min. The bottom of the cell culture vessel was covered with a 1.5 cm thick tissue-equivalent bolus. The source-to-surface distance (SSD) was 100 cm.

Each T25 culture flask was seeded with approximately 2x105 cells in the exponential growth phase and sequentially irradiated with increasing doses (2,

4, 6, 8, and 10 Gy) twice, followed by a single 10 Gy dose, for a total of 9 irradiations and a cumulative dose of 50 Gy. Following the initial 2 Gy irradiation, the cells were cultured and passaged twice before being irradiated with another 2 Gy. The surviving cells were then subjected to the aforementioned irradiation schedule, and the resulting population was designated as the radiation-resistant cell line CNE-2-RS (Fig 2E). Parental cells were used as controls and underwent the same treatment regimen without irradiation.

To maintain radioresistance, the CNE-2-RS cells used in the experiments were exposed to a low dose of approximately 1-2 Gy per month. Cells within 15 passages following the low-dose irradiation were used for subsequent experiments, and all experiments were conducted using cells in the exponential growth phase. The radioresistance of CNE-2-RS cells was subsequently validated using clonogenic assays, with results consistently confirming the maintained radioresistant phenotype.

3.3 RT-qPCR

Total RNA was extracted from CNE2 cells using TRIzol reagent (Invitrogen). Following the measurement of RNA concentration with a Nanodrop 2000 instrument (Thermo Scientific), cDNA synthesis was performed using the Color Reverse Transcription Kit (EZBioscience, USA). The relative mRNA expression levels of each gene were calculated using the $2^{(-\Delta\Delta CT)}$ method. The primer sequences of the genes are presented in Supplementary Table 3.

4.Annotation of immune-related features of NPC-RSS key genes

The CIBERSORT algorithm was applied to analyze the expression data of nasopharyngeal cancer patients to infer the relative proportions of 22 immune-infiltrating cells. Additionally, a correlation analysis was performed between gene expression and immune cell content using Spearman's method. Results with a P value < 0.05 were considered statistically significant. After grouping the nasopharyngeal cancer data based on the previously extracted NPC-RSS, the CIBERSORT method was applied to quantify the immune cell infiltration. The objective was to thoroughly investigate whether there was a significant difference in immune infiltration between the sensitive and resistant groups after stratifying nasopharyngeal cancer patients according to their NPC-RSS. To further elucidate the potential association of immune features in nasopharyngeal cancer, we also investigated the correlation of NPC-RSS key genes with immune-related feature genes obtained from the TISIDB database, including chemokine-related genes, immunosuppressant-related genes, MHC-related genes, immune stimulation-related genes, and receptor-related genes[19].

5.Further validation of NPC-RSS at the single-cell level.

The single-cell RNA sequencing (scRNA-seq) data were processed and quantified using the Cell Ranger (version 2.0.1) pipeline. Initially, reads obtained from 10x Genomics were aligned to the human reference genome hg19.The FASTQ sequence files of four samples (Supplementary Table 4) were aligned to the human reference genome hg19 using STAR software and the Cell Ranger "count" module, generating files containing barcode tables and gene tables. The raw gene expression matrices for each sample, generated by the CellRanger software, were merged in R and converted to Seurat objects for downstream analysis using the Seurat R package. Subsequently, cells with the following characteristics were removed: (1) number of unique molecular identifiers (UMIs) less than 201, (2) fewer than 100 or more than 6,000 expressed genes (potential cellular duplicates), and (3) percentage of mitochondrial genes greater than 20%. Additionally, genes expressed in fewer than three cells were removed from the dataset. After quality control, 28,957 high-quality nasopharyngeal carcinoma cells were retained for further analysis. Subsequently, the gene expression matrix was normalized to the total cell reads and mitochondrial reads using linear regression with Seurat's RegressOut function. Following data normalization, highly variable genes were identified across all individual cells while controlling for the relationship between mean expression and dispersion. Prior to the joint analysis of the four-sample scRNA-seq dataset, we employed the R package Harmony to mitigate batch effects and minimize their impact on downstream analyses[20]. Furthermore, we utilized Seurat's "CellCycleScoring" function to identify the expression of cell cycle-related genes across various cell subpopulations. We employed a previously defined core set of 43 G1/S and 54 G2/M cell cycle genes[21]. Cells were categorized based on the maximum average expression ("cycle score") across both gene sets. Following cell cycle analysis, no bias attributable to cell cycle genes was observed.

6.Cell clustering analysis and identification of marker genes for cellular subpopulations.

Normalization, clustering, differential gene expression analysis, and visualization were performed using the Seurat 3.1 R package. To reduce the dimensionality of the dataset, principal component analysis (PCA) was performed on the 2,000 variably expressed genes. The top 20 principal components were then used as input for the t-distributed stochastic neighbor embedding (t-SNE) algorithm, which was run using the default settings of the RunTSNE function in Seurat. Cell clustering was performed using the 'FindClusters' function in Seurat with a resolution of 0.3, resulting in the identification of 17 distinct cell clusters (clusters 0-16). Subsequently, the FindMarkers function was employed to identify specific marker genes for each cell subpopulation. The process was conducted as follows: (1) Differentially expressed genes (DEGs) for each cell subpopulation were identified by comparing the cells of that subpopulation to all other cells using the Wilcoxon

rank-sum test; (2) Bonferroni-corrected p-values less than 0.05 were used as the cut-off for determining statistically significant DEGs; (3) Genes with an average expression level more than 2-fold higher in a specific cell subpopulation compared to other subpopulations were selected as marker genes. Cell type annotation for each subpopulation was performed based on marker genes reported in previous studies[22], the CellMarker database, and the R package 'SingleR'[23].

7.GSVA and GSEA analysis of NPC-RSS key genes

GSVA and GSEA analysis of key NPC-RSS genes Based on the expression levels of key NPC-RSS genes, we stratified the patients into high and low expression groups and thoroughly investigated the signaling pathway differences between these groups using Gene Set Enrichment Analysis (GSEA). We utilized the Molecular Signatures Database (MSigDB) gene sets as the annotated gene sets for the subtype pathways to conduct differential expression analysis between subtypes. Gene consistency was assessed using a concordance score calculated by sorting the genes according to their LogFC (adjusted $P < 0.05$). Furthermore, to investigate the association of NPC-RSS with known radiosensitivity-related genes, we searched for "Radiosensitivity" in the GeneCards database (https://www.genecards.org/) to obtain radiosensitization-related genes, and examined the expression levels of the top 20 genes based on their Relevance scores and their correlation with the top-5 genes in NPC-RSS.

8.Statistical Analysis

The differentially expressed genes between NPCs of the sensitive and resistant groups in the dataset were identified using the R package "limma". For normally distributed variables, statistical differences between two groups were determined by two-tailed t-test, while one-way ANOVA was used for multiple group comparisons. For non-normally distributed variables, statistical differences between two groups were determined by the Wilcoxon rank-sum test, while the Kruskal-Wallis test was used for multiple group comparisons. All statistical analyses were conducted using R software (version 4.1.2). Each experiment was performed independently in triplicate unless otherwise specified. Spearman's rank correlation analysis was employed to assess the correlation between two variables.

In the single-cell sequencing data processing section, comparisons between two groups were performed using the unpaired two-tailed t-test (implemented in the R package limma) or the Mann-Whitney U test. All statistical analyses and graph visualizations were conducted using the R programming language.

For all experiments, samples from the same patient were processed in parallel. Cells from each patient sample were subjected to single-cell RNA

sequencing (10x Genomics) simultaneously but in different passages and vials. Boxplots were generated using the R base software package with default parameters. The boxes span the interquartile range (IQR; from the 25th to the 75th percentile), with the center line representing the median. The lower whiskers extend to the smallest value no further than 1.5×IQR from the 25th percentile. The upper whiskers extend to the largest value no further than 1.5×IQR from the 75th percentile. Due to the large number of data points represented in the box-and-whisker plots, individual data points were not shown to maintain clarity in the overall distribution.

Results

1.Construction of the NPC-RSS Model

As shown in Fig.1, 48 models were successfully constructed, and the area under the curve (AUC) of each model was calculated in both the training and validation sets. The best model combination was found to be glmBoost+NaiveBayes after sample weighting: the glmBoost algorithm was used to screen out the most valuable gene features, and then the NaiveBayes algorithm was employed to identify the most reliable model (Fig. 2A). The model identified 18 genes, which were used as candidate genes for subsequent studies. The NPC-RSS was constructed using logistic regression, and the model prediction formula after secondary modeling was: P(Y=1) = -15.3906 + 2.7369 * C1QTNF3 + 3.8292 * CA11 + 7.6856 * CD9 - 15.9445 * CDK5RAP3 - 11.2475 * CLDN1 + 14.3984 * DMC1 - 6.2474 * EYA1 + 1.6897 * IFI44L + 10.0473 * KNG1 + 0.9464 * KREMEN1 - 6.5371 * NT5DC2 - 10.4192 * NTRK3 + 8.1306 * PSG4 - 0.5755 * RFX4 - 2.1728 * RHOBTB3 - 7.1597 * SLC1A2 + 22.6640 * SMARCA2 - 0.3858 * TRIM58. The top 5 genes characterized in the NPC-RSS, in order of their contribution to radiotherapy sensitivity and weighting coefficients, are SMARCA2, DMC1, CD9, PSG4, and KNG1. The genes with absolute values of weight coefficients greater than 10 were SMARCA2, DMC1, KNG1, NTRK3, CLDN1, and CDK5RAP3 (Fig. 2B).

The prediction model constructed using these 18 genes demonstrated good diagnostic efficacy, with an area under the receiver operating characteristic (ROC) curve (AUC) of 0.996 for the training set. The GSE32389 dataset was employed as an external validation set, yielding an AUC of 0.823, which demonstrated the model's robustness. The overall sample-weighted AUC for the combined training and validation sets was 0.932 (Fig. 2C). The genes used to construct the model were derived from our center's NPC transcriptome expression data and underwent differential expression analysis, resulting in 2,932 differentially expressed genes (DEGs) (Fig. 2D).

Our previous analysis revealed that 18 gene features were consistently expressed in the NPC dataset (Supplementary Fig. 1). Furthermore, these 18

gene features exhibited significant differential expression between the sensitive and resistant groups (Supplementary Fig. 1).

2.In vitro validation of NPC-RSS.

The induction process of CNE2-RS radiation-resistant cells is shown in Fig. 2E. RNA was transcribed into cDNA, sequenced, and analyzed upstream by a high-throughput sequencer to obtain count data, which was further transformed into FPKM values for subsequent analysis. Differential analysis was performed between the parental and resistant groups using the limma package. The analysis results showed that SMARCA2 and CD9, among the key genes of NPC-RSS, were significantly highly expressed in the sensitive group ($p < 0.001$, t-test), which was consistent with the expression trend of the key genes in the previously constructed NPC-RSS (Fig. 2F). The heatmap of the differential analysis is shown in Fig. 2G. We also clustered these data based on the 18 genes in NPC-RSS after Z-score normalization for the heatmap. The results were consistent with the expression trend in NPC-RSS (Fig. 2H). Meanwhile, we calculated the radiotherapy sensitivity scores of three pairs of replicated CNE2-P and CNE2-RS samples using NPC-RSS and used the mean values for statistical testing. The results showed that the sensitivity scores of the sensitive group were significantly higher than those of the resistant group, further confirming the reliability of NPC-RSS (Fig. 2I).

3.Annotation of key NPC-RSS genes and immune-related features

To thoroughly explore the potential mechanism by which the NPC-RSS predicts radiosensitivity, we stratified the local NPC data into groups based on the NPC-RSS developed in this study and examined the association between these groups and immune infiltration. The proportion of immune cell content in each patient and the correlations among immune cells were visualized using various formats. The results demonstrated a significant elevation in the levels of T follicular helper cells, regulatory T cells (Tregs), and activated NK cells in the samples from the radiosensitive group compared to those from the radioresistant group (Fig. 3A), suggesting that the radiosensitive group exhibited a more immunologically active infiltration profile.

In this study, we further explored the relationship between NPC-RSS key genes and immune cells. We found that several key genes were highly correlated with immune cells: SMARCA2 was significantly positively correlated with resting CD4+ memory T cells; DMC1 was significantly positively correlated with resting NK cells; CD9 was significantly positively correlated with resting CD4+ memory T cells; and KNG1 was significantly negatively correlated with M1 macrophages, γδ T cells, monocytes, eosinophils, neutrophils, and other immune cell types (Fig. 3B). Furthermore, all these immune cell types were significantly correlated with each other (Fig. 3C). Correlations between these five key genes and various immune factors, including immunomodulators, chemokines, and cellular receptors, were

obtained from the TISIDB database (Fig. 3D). These analyses suggest that the NPC-RSS key genes are closely associated with the level of immune cell infiltration and play a crucial role in the immune microenvironment.

4.Validation of NPC-RSS at the single-cell level

The expression patterns of key NPC-RSS genes at the single-cell level exhibited a stronger correlation with model predictions. We analyzed the single-cell transcriptome profiles of four patients. Among the four patients, three were sensitive to radiotherapy, while one was resistant. Following single-cell suspension preparation, library construction, sequencing analysis, and quality control of the obtained data, we acquired transcriptome data for a total of 28,957 cells. Among these, 22,796 cells were derived from the radiotherapy-sensitive group, and 6,161 cells were from the radiotherapy-resistant group. Unsupervised clustering was performed on the 28,957 cells obtained from sequencing to identify cell subpopulations based on the similarity of their gene expression profiles. As illustrated in Figure 4, 17 cell subpopulations were identified through clustering. By examining the expression profiles of cell type-specific marker genes, we determined that these 17 cell subpopulations belonged to 7 distinct cell types: T cells, myeloid cells, B cells, epithelial cells, neutrophils, fibroblasts, and mast cells (Figure 4A).

T cells expressed CD3D, CD8A, and CD4 genes. Myeloid cells expressed C1QB, LYZ, and AIF1 genes. B cells expressed MS4A1 and CD79A genes. Epithelial cells specifically expressed KRT5 and EPCAM genes. Neutrophils expressed specific marker genes. Fibroblasts expressed DCN and COL1A1 genes. Mast cells expressed CPA3, TPSAB1, and MS4A2 as marker genes. We used NPC-RSS to calculate the radiotherapy sensitivity scores of each cell. The predicted scores of cells in the sensitive group were significantly higher than those in the resistant group, especially in epithelial cells, myeloid cells, fibroblasts, and mast cells (Fig. 4B). The sensitivity scores of all cells are shown in Fig. 4C. Moreover, most of the 18 genes in the NPC-RSS were related to immune cells (Fig. 4D). Interestingly, the radiotherapy-sensitive group showed significant dominance in terms of immune cells, whereas the radiotherapy-resistant group was mainly composed of malignant epithelial and stromal cells (Fig. 4E). It is noteworthy that T cells, myeloid cells, epithelial cells, mast cells, and fibroblasts consisted of multiple cell subpopulations (Fig. 4F), suggesting that these cell types may be heterogeneous.

5.Pathway enrichment analysis findings

Next, we investigated the specific signaling pathways associated with the key genes of NPC-RSS and explored the potential molecular mechanisms by which these genes influence radiosensitivity in nasopharyngeal carcinoma. GSVA results demonstrated that high expression of SMARCA2 enriched signaling pathways such as WNT_BETA_CATENIN_SIGNALING,

PI3K_AKT_MTOR_SIGNALING, TGF_BETA_SIGNALING, and NOTCH_SIGNALING (Fig5A). Similarly, high expression of DMC1 enriched INTERFERON_GAMMA_RESPONSE and IL2_STAT5_SIGNALING pathways (Fig5B).

Moreover, GSEA results showed that SMARCA2 was enriched in pathways such as the JAK-STAT signaling pathway, NF-kappa B signaling pathway, and T cell receptor signaling (Fig. 5C). DMC1 was enriched in pathways such as cell adhesion molecules, Hedgehog signaling pathway, and phosphatidylinositol signaling system (Fig. 5D). These findings suggest that NPC-RSS key genes may influence radiotherapy sensitivity in nasopharyngeal carcinoma patients through these pathways.

6.Correlation between NPC-RSS key genes and radiosensitization genes

Radiosensitivity-related genes were obtained by searching the GeneCards database (https://www.genecards.org/) for the term "Radiosensitivity." The local NPC data were grouped and analyzed based on NPC-RSS. The expression levels of the top 20 genes, ranked by their Relevance scores, were correlated. The results, presented in Supplementary Figure 2, showed significant differences in the expression of several radiosensitivity-related genes between the radiotherapy-resistant and sensitive groups.

Spearman correlation analysis indicated that the expression of SMARCA2 was positively correlated with high expression levels of TP53, EGFR, BRCA2, and ATM. Similarly, DMC1 expression was positively correlated with high expression levels of BRCA1, APTX, ATM, and CCR6, while CD9 expression was positively correlated with high expression levels of EGFR, TP53, ERBB2, and RAD51. The expression of KNG1 was correlated with high expression levels of HULC and significantly negatively correlated with high expression levels of BADA (Fig. 5E).

Discussion

Nasopharyngeal carcinoma (NPC), a malignant tumor, is prevalent in East and Southeast Asia, particularly in South China. Radiotherapy is the primary treatment modality for NPC; however, treatment failure occurs in approximately 20-30% of patients[1]. Radiotherapy sensitivity is a crucial factor influencing treatment outcomes in NPC patients. Thus, it is imperative to comprehensively investigate the mechanisms underlying radiotherapy sensitivity in NPC and identify effective therapeutic targets. Such investigations will facilitate the elucidation of therapeutic resistance mechanisms and provide a foundation for personalized treatment strategies. In the present study, we developed an integrative approach to establish a consistent nasopharyngeal carcinoma radiotherapy sensitivity signature

(NPC-RSS) using NPC samples obtained from Zhujiang Hospital of Southern Medical University (Figure 1). In summary, we fitted 113 models to the training dataset using the leave-one-out cross-validation (LOOCV) framework. Further validation using the independent dataset GSE32389 from the Gene Expression Omnibus (GEO) database demonstrated that the glmBoost+NaiveBayes model exhibited the best performance.

The integrated procedure's advantage lies in its utilization of various machine learning algorithms and their combinations to fit models for NPC radiotherapy sensitivity, achieving a consensus performance. Furthermore, the combination of algorithms can reduce the dimensionality of the variables, resulting in a more simplified and easily translatable model. The predictive validity of NPC-RSS has been further validated using NPC cell lines generated at our local center. Moreover, we performed a related validation on local single-cell data, which revealed that the radiotherapy-sensitive group exhibited a significant predominance of immune cells, whereas the radiotherapy-resistant group primarily comprised malignant epithelial and stromal cells.

Previous studies have demonstrated that the core genes of NPC-RSS, including SMARCA2, DMC1, CD9, PSG4, and KNG1, are associated with tumor radiosensitization to varying degrees[24-28]. The protein encoded by the SMARCA2 gene plays a crucial role in various cellular processes, including cytokine response, DNA repair, regulation of cell proliferation, lineage specification and development, and cell adhesion[29-33]. These processes are intimately linked to the radiosensitivity of tumor cells. Cell proliferation and DNA repair are critical determinants of tumor cell response to radiotherapy, and the involvement of SMARCA2 may influence this response[34-38]. Furthermore, SMARCA2 is implicated in processes such as cell adhesion and cytokine response, which are tightly coupled to the interactions of tumor cells within the microenvironment. Cell adhesion influences the low-dose hyper-radiosensitivity of a broad spectrum of tumor cells, potentially further modulating their sensitivity to radiotherapy[39].

Radiation is capable of inducing DNA double-strand breaks, and cells must repair this damage through DNA repair mechanisms to maintain genomic stability. DMC1, a key factor in homologous recombination, is involved in the repair process of DNA double-strand breaks[40-42]. Consequently, abnormal function or altered expression levels of DMC1 may influence cellular sensitivity to radiation and, consequently, the efficacy of radiation therapy. CD9 is believed to play a role in various stages of cancer development[43]. As a cell surface protein, CD9 is involved in regulating cell-cell adhesion and interactions, which may affect the response of tumor cells to radiation therapy. Moreover, CD9 may impact tumor sensitivity to radiation therapy by modulating the metastatic and invasive properties of tumor cells[44-47]. It has been shown that all human PSGs initiate TGF-β1. In vivo, PSGs may increase

the availability of active TGF-β1 in both soluble and matrix-bound forms of this potential cytokine[48]. Furthermore, TGF-β1 levels accurately predict radiosensitization in elderly patients with unresectable NSCLC undergoing 3D-CRT[49].

Pregnancy-specific glycoproteins (PSGs) may influence radiosensitivity by regulating transforming growth factor-beta 1 (TGF-β1) activity. In radiation therapy, radiation can induce an inflammatory response in tissues and affect angiogenesis. Kininogen-1 (KNG1), a factor associated with coagulation and inflammatory response[50], may play a regulatory role in radiosensitivity and the effects of radiation therapy. Furthermore, we found that the core genes of the nasopharyngeal carcinoma radiosensitivity signature (NPC-RSS), namely SMARCA2, DMC1, CD9, PSG4, and KNG1, had significant correlations with previously published genes related to radiosensitization (Fig. 5E). The roles of these genes in regulating radiosensitization further support the importance and reliability of the NPC-RSS developed in this study for predicting radiosensitization in nasopharyngeal carcinoma.

Previous studies have shown that SMARCA2 has a significant effect on the responsiveness to radiotherapy in non-small cell lung cancer, and its depletion can increase the sensitivity to radiotherapy, suggesting that SMARCA2 may be a potential therapeutic target for sensitization to radiotherapy[24]. In addition, the DMC1 gene was used in the construction of the radiosensitivity index (RSI) to assess the sensitivity and prognosis of radiotherapy, especially in endometrial cancer providing an important perspective on the DNA damage repair process[51]. For CD9, its study in glioblastoma multiforme (GBM) revealed that the increase of CD9 and CD81 in extracellular vesicles (EVs) after radiotherapy was closely related to the cytotoxic response after treatment, which provided a new biomarker for radiotherapy monitoring[27]. In contrast, the relationship between PSG4 and KNG1 and radiotherapy resistance has not been reported. In the future, the expression pattern of SMARCA2 in nasopharyngeal carcinoma patients and its relationship with the effect of radiotherapy can be analyzed by clinical samples to determine its potential as a therapeutic target for sensitization to radiotherapy. For DMC1, studying the correlation between its expression level and radiotherapy sensitivity in nasopharyngeal carcinoma may be helpful in predicting the efficacy and adjusting the treatment regimen. Meanwhile, by analyzing extracellular vesicles, especially CD9-containing vesicles, in nasopharyngeal carcinoma patients after radiotherapy, we assessed its feasibility as a biomarker to monitor radiotherapy response. These studies not only contribute to an in-depth understanding of the mechanism of these genes in nasopharyngeal carcinoma radiotherapy, but may also open up new strategies to improve the efficacy of radiotherapy.

Radiotherapy plays a pivotal role in cancer treatment, and there is growing evidence of the potential for synergistic effects between radiation therapy and

components of the immune response[52, 53]. However, there are limited data depicting the tumor immune microenvironment (TIME) associated with radiotherapy sensitivity. Indeed, recent preclinical studies have demonstrated that irradiated tumors can act as adjuvants to the immune system, promoting local tumor control and inducing regression of distant tumor deposits through abscopal effects[54]. In addition to potential systemic effects, radiation significantly alters the distribution of tumor antigens, recruits anti-tumor T cells, and induces a distinct inflammatory microenvironment[55]. Some studies have shown that tumors with high immune cell infiltration are more sensitive to radiation[56]; however, the relationship between radiation and immune response remains complex and context-dependent, and the optimal approach to integrating radiation and immunotherapy remains uncertain.

Studies have demonstrated that over one-third of tumor types in the radiotherapy-sensitive group of solid tumors exhibit an enrichment of CD8+ T cells, M1 macrophages, and NK+ cells[56]. To further investigate this phenomenon, we analyzed the immune infiltration between different subgroups of NPC-RSS and annotated the key NPC-RSS genes with immune-related features. The results demonstrated that the NPC radiosensitized group exhibited a higher level of immune infiltration and immune initiation, characterized by a greater abundance of activated NK cells, regulatory T cells (Tregs), and follicular helper T cells. The NPC-RSS key genes also exhibited a significant correlation and synergism with immune-related features. Notably, this synergistic correlation was most prominent in chemokines and chemokine receptor families, the major histocompatibility complex (MHC), immune cell interactions, and signaling pathways. These findings suggest that the radiosensitivity estimated by the NPC-RSS may be associated with a distinct tumor immune microenvironment, which could potentially be exploited through a combinatorial immunotherapy approach.

Notably, the results obtained from the collected single-cell data were consistent with the predicted trends of our constructed model. The radiotherapy-sensitive group exhibited a significant predominance of immune cells, while the radiotherapy-resistant group primarily comprised malignant epithelial and stromal cells. Furthermore, utilizing NPC-RSS to calculate the radiotherapy sensitivity score for each cell revealed that the predicted scores of cells in the sensitized group were significantly higher than those in the resistant group, particularly in epithelial cells, myeloid cells, fibroblasts, and mast cells. This finding suggests that the presence of immune cells might have influenced the tumor's responsiveness to treatment to some extent prior to radiation therapy. Tumors with a predominance of immune cells may be more likely to exhibit a positive response to radiation therapy, either due to the involvement of immune cells in tumor surveillance and clearance or because

the responsiveness of tumors to radiation therapy is modulated by immune cells.

Furthermore, pathway enrichment analysis suggested that the Wnt/β-catenin signaling pathway may be associated with the radiosensitivity of nasopharyngeal carcinoma (NPC). Gene Set Variation Analysis (GSVA) and Gene Set Enrichment Analysis (GSEA) indicated that the differential expression levels of SMARCA2, DMC1, CD9, PSG4, and KNG1 in the core of the NPC radiosensitivity signature (NPC-RSS) might influence various signaling pathways associated with disease progression, involving multiple biological processes such as DNA damage repair, cell death, immune regulation, cell proliferation, metabolism, tumor microenvironment regulation, and cell survival. For instance, the SMARCA2 high-expression group, which was the most significant, was enriched in the UV response, Wnt/β-catenin signaling pathway, NOTCH signaling pathway, TGFβ signaling pathway, PI3K-AKT-mTOR signaling pathway, Interferon Alpha response, Interferon Gamma response, and other signaling pathways. Previous studies have demonstrated that proteins related to the Wnt/β-catenin signaling pathway (DKK-3, β-catenin, and c-MYC) are involved in the development of NPC, and that knockdown of FOXO3a promotes radioresistance in NPC through the Wnt/β-catenin signaling pathway[57, 58]. Therefore, we hypothesized that SMARCA2 may regulate the radiosensitivity of NPC via the Wnt/β-catenin signaling pathway.

Incorporating the NPC-RSS scoring system into clinical decision-making and patient management involves several key steps: first, making genetic testing a standard part of nasopharyngeal cancer diagnosis and ensuring that physicians have timely access to scoring results to guide treatment planning. Second, physicians need to use the scoring results to customize individualized treatment plans and discuss them in a multidisciplinary team to optimize decision-making. At the same time, physicians should explain the clinical significance of the scores in detail and communicate effectively with patients to achieve shared decision-making. In addition, continuously monitoring the relationship between scoring and treatment outcomes, optimizing the scoring model based on actual data, and ensuring the integration of technology platforms and regulatory compliance are essential to safeguard the effective operation of the scoring system and the security of patient information.

The novelty of this study lies in the utilization of nasopharyngeal carcinoma expression data collected from Zhujiang Hospital of Southern Medical University, combined with multiple machine learning models, to develop NPC-RSS, a tool capable of effectively identifying the patient population more sensitive to radiotherapy. The tool was validated using single-cell data and cell lines from the in-house cohort, with the validation results aligning with the predictions of our constructed model. However, this study has several limitations. First, this study focuses on specific types of

nasopharyngeal carcinoma (NPC), and thus may not be directly generalizable to other subtypes or related head and neck cancers. Second, the sample size of the dataset is small, which may affect the generalization and extrapolation ability of the model. To minimize the impact of the small sample size, we used advanced statistical techniques, such as cross-validation, to enhance the stability and credibility of the results. Nonetheless, we realize that larger datasets are necessary, so we are collaborating with other research centers to expand our sample size to make our findings more robust and widely applicable. Further, while our study relies on bioinformatics methods to identify and analyze key genes, the lack of direct experimental validation of function is a glaring deficiency. We are currently seeking additional funding support and plan to collaborate with laboratories to conduct the necessary functional validation experiments, which will help to further confirm the specific role of these genes in radiotherapy. In addition, we recognize that there may be a risk of overfitting when using machine learning algorithms to process biomedical data. For this reason, we not only implemented regularization techniques during the model development stage, but also adopted a rigorous cross-validation strategy during model validation. These measures help ensure that our models maintain good predictive performance on unseen data.

Despite these limitations, our study provides new insights into understanding the molecular mechanisms of radiotherapy sensitivity and points to possible directions for future research. Future work will focus on expanding the dataset, performing experimental validation, and further optimizing our predictive model.

Conclusion

In this study, we developed a robust Nasopharyngeal Carcinoma Radiotherapy Susceptibility Score (NPC-RSS) using multiple bioinformatics and machine learning algorithms. The expression of key genes SMARCA2, DMC1, CD9, and KNG1 was validated to be consistent with the NPC-RSS in our in-house cell lines. Compared to the resistant group, the sensitive group classified by NPC-RSS exhibited a more abundant and active state of immune infiltration. Furthermore, in single-cell samples, NPC-RSS was elevated in the radiotherapy-sensitive group, where immune cells played a predominant role. Additional analysis uncovered substantial correlations between key NPC-RSS genes and genes previously demonstrated to be involved in the regulation of radiosensitization. Single-gene pathway enrichment analysis revealed that the top 2 NPC-RSS genes were associated with several signaling pathways, including the Wnt/β-catenin signaling pathway, JAK-STAT signaling pathway, NF-kappa B signaling pathway, and T cell receptor signaling pathway. These pathways may influence the immune microenvironment of nasopharyngeal carcinoma, thereby affecting its susceptibility to radiotherapy. Our findings provide a novel perspective for understanding NPC radiosensitivity.

Author Contributions

KL wrote the article; PL,KL and JZ designed the research; KL,JL,NL,JF and XZ performed the research; KL,JY, NL,JF,XZ, PL, and JZ Writing-review and editing.

Acknowledgments

None.

Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics Statement

The patients/participants provided their written informed consent to participate in this study and the research presented here has been performed in accordance with the Declaration of Helsinki and has been approved by the ethics committee of the Zhujiang Hospital of Southern Medical University.

Funding

None

Data avaliability

The datasets generated and/or analysed during the current study are available from the corresponding author on reasonable request.

1. Chua, M.L.K., et al., *Nasopharyngeal carcinoma.* Lancet, 2016. **387**(10022): p. 1012-1024.
2. Ferlay, J., et al., *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.* Int J Cancer, 2015. **136**(5): p. E359-86.
3. Wei, W.I. and J.S. Sham, *Nasopharyngeal carcinoma.* Lancet, 2005. **365**(9476): p. 2041-54.
4. Sung, H., et al., *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
5. Chen, Y.P., et al., *Nasopharyngeal carcinoma.* Lancet, 2019. **394**(10192): p. 64-80.
6. You, R., et al., *Hyperfractionation compared with standard fractionation in intensity-modulated radiotherapy for patients with locally advanced recurrent nasopharyngeal carcinoma: a multicentre, randomised, open-label, phase 3 trial.* Lancet, 2023. **401**(10380): p. 917-927.
7. Lee, A.W., et al., *Management of Nasopharyngeal Carcinoma: Current Practice and Future Perspective.* J Clin Oncol, 2015. **33**(29): p. 3356-64.
8. Ma, B.B. and A.T. Chan, *Recent perspectives in the role of chemotherapy in the management of advanced nasopharyngeal carcinoma.* Cancer, 2005. **103**(1): p. 22-31.
9. Agulnik, M. and J.B. Epstein, *Nasopharyngeal carcinoma: current management, future directions and dental implications.* Oral Oncol, 2008. **44**(7): p. 617-27.
10. Linkous, A.G. and E.M. Yazlovitskaya, *Novel radiosensitizing anticancer therapeutics.* Anticancer Res, 2012. **32**(7): p. 2487-99.
11. Ribassin-Majed, L., et al., *What Is the Best Treatment of Locally Advanced Nasopharyngeal Carcinoma? An Individual Patient Data Network Meta-Analysis.* J Clin Oncol, 2017. **35**(5): p.

498-505.

12. Khan, S., *Artificial Intelligence and Machine Learning in Clinical Medicine.* N Engl J Med, 2023. **388**(25): p. 2398.
13. Teng, X., et al., *Explainable machine learning via intra-tumoral radiomics feature mapping for patient stratification in adjuvant chemotherapy for locoregionally advanced nasopharyngeal carcinoma.* Radiol Med, 2023. **128**(7): p. 828-838.
14. Hu, Q., et al., *Machine Learning Based on MRI DWI Radiomics Features for Prognostic Prediction in Nasopharyngeal Carcinoma.* Cancers (Basel), 2022. **14**(13).
15. Liu, Z., et al., *Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer.* Nat Commun, 2022. **13**(1): p. 816.
16. Zhang, H., et al., *Machine learning-based tumor-infiltrating immune cell-associated lncRNAs for predicting prognosis and immunotherapy response in patients with glioblastoma.* Brief Bioinform, 2022. **23**(6).
17. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.
18. Liu, J., et al., *A comprehensive analysis of candidate genes and pathways in pancreatic cancer.* Tumour Biol, 2015. **36**(3): p. 1849-57.
19. Ru, B., et al., *TISIDB: an integrated repository portal for tumor-immune system interactions.* Bioinformatics, 2019. **35**(20): p. 4200-4202.
20. Korsunsky, I., et al., *Fast, sensitive and accurate integration of single-cell data with Harmony.* Nat Methods, 2019. **16**(12): p. 1289-1296.
21. Tirosh, I., et al., *Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq.* Science, 2016. **352**(6282): p. 189-96.
22. Chen, Y.P., et al., *Single-cell transcriptomics reveals regulators underlying immune cell diversity and immune subtypes associated with prognosis in nasopharyngeal carcinoma.* Cell Res, 2020. **30**(11): p. 1024-1042.
23. Aran, D., et al., *Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.* Nat Immunol, 2019. **20**(2): p. 163-172.
24. Zernickel, E., et al., *Targeting of BRM Sensitizes BRG1-Mutant Lung Cancer Cell Lines to Radiotherapy.* Mol Cancer Ther, 2019. **18**(3): p. 656-666.
25. Xu, M.D., et al., *The radiotherapy-sensitization effect of cantharidin: Mechanisms involving cell cycle regulation, enhanced DNA damage, and inhibited DNA damage repair.* Pancreatology, 2018. **18**(7): p. 822-832.
26. Erovic, B.M., et al., *Mcl-1, vascular endothelial growth factor-R2, and 14-3-3sigma expression might predict primary response against radiotherapy and chemotherapy in patients with locally advanced squamous cell carcinomas of the head and neck.* Clin Cancer Res, 2005. **11**(24 Pt 1): p. 8632-6.
27. Jennrich, S., et al., *CD9- and CD81-positive extracellular vesicles provide a marker to monitor glioblastoma cell response to photon-based and proton-based radiotherapy.* Front Oncol, 2022. **12**: p. 947439.
28. Xu, J., et al., *Overexpression of the Kininogen-1 inhibits proliferation and induces apoptosis of glioma cells.* J Exp Clin Cancer Res, 2018. **37**(1): p. 180.
29. Helming, K.C., X. Wang, and C.W.M. Roberts, *Vulnerabilities of mutant SWI/SNF complexes in cancer.* Cancer Cell, 2014. **26**(3): p. 309-317.

30. Wang, X., J.R. Haswell, and C.W. Roberts, *Molecular pathways: SWI/SNF (BAF) complexes are frequently mutated in cancer--mechanisms and potential therapeutic insights.* Clin Cancer Res, 2014. **20**(1): p. 21-7.
31. Zhu, Y., et al., *Brahma regulates the Hippo pathway activity through forming complex with Yki-Sd and regulating the transcription of Crumbs.* Cell Signal, 2015. **27**(3): p. 606-13.
32. Tian, W., et al., *Brahma-related gene 1 bridges epigenetic regulation of proinflammatory cytokine production to steatohepatitis in mice.* Hepatology, 2013. **58**(2): p. 576-88.
33. Dykhuizen, E.C., et al., *BAF complexes facilitate decatenation of DNA by topoisomerase IIalpha.* Nature, 2013. **497**(7451): p. 624-7.
34. Lee, S.B., et al., *2,3,6-Trisubstituted quinoxaline derivative, a small molecule inhibitor of the Wnt/beta-catenin signaling pathway, suppresses cell proliferation and enhances radiosensitivity in A549/Wnt2 cells.* Biochem Biophys Res Commun, 2013. **431**(4): p. 746-52.
35. Su, H., et al., *MicroRNA-301a targets WNT1 to suppress cell proliferation and migration and enhance radiosensitivity in esophageal cancer cells.* Oncol Rep, 2019. **41**(1): p. 599-607.
36. Wang, Z., et al., *IGF-1R Inhibition Suppresses Cell Proliferation and Increases Radiosensitivity in Nasopharyngeal Carcinoma Cells.* Mediators Inflamm, 2019. **2019**: p. 5497467.
37. Toulany, M., et al., *Blockage of epidermal growth factor receptor-phosphatidylinositol 3-kinase-AKT signaling increases radiosensitivity of K-RAS mutated human tumor cells in vitro by affecting DNA repair.* Clin Cancer Res, 2006. **12**(13): p. 4119-26.
38. Hadjipanayis, C.G. and N.A. DeLuca, *Inhibition of DNA repair by a herpes simplex virus vector enhances the radiosensitivity of human glioblastoma cells.* Cancer Res, 2005. **65**(12): p. 5310-6.
39. Mathur, A., et al., *Induction of epithelial-mesenchymal transition in thyroid follicular cells is associated with cell adhesion alterations and low-dose hyper-radiosensitivity.* Tumour Biol, 2023. **45**(1): p. 95-110.
40. Dunne, A.L., et al., *Relationship between clonogenic radiosensitivity, radiation-induced apoptosis and DNA damage/repair in human colon cancer cells.* Br J Cancer, 2003. **89**(12): p. 2277-83.
41. Zhang, X., et al., *Bufalin enhances radiosensitivity of glioblastoma by suppressing mitochondrial function and DNA damage repair.* Biomed Pharmacother, 2017. **94**: p. 627-635.
42. Xie, X., et al., *CDK4/6 Inhibitor Palbociclib Amplifies the Radiosensitivity to Nasopharyngeal Carcinoma Cells via Mediating Apoptosis and Suppressing DNA Damage Repair.* Onco Targets Ther, 2019. **12**: p. 11107-11117.
43. Murayama, Y., K. Oritani, and S. Tsutsui, *Novel CD9-targeted therapies in gastric cancer.* World J Gastroenterol, 2015. **21**(11): p. 3206-13.
44. Hori, H., et al., *CD9 expression in gastric cancer and its significance.* J Surg Res, 2004. **117**(2): p. 208-15.
45. Herr, M.J., et al., *Tetraspanin CD9 promotes the invasive phenotype of human fibrosarcoma cells via upregulation of matrix metalloproteinase-9.* PLoS One, 2013. **8**(6): p. e67766.
46. Yin, M., et al., *Osteopontin promotes the invasive growth of melanoma cells by activating integrin alphavbeta3 and down-regulating tetraspanin CD9.* Am J Pathol, 2014. **184**(3): p. 842-58.
47. Rappa, G., et al., *Tetraspanin CD9 determines invasiveness and tumorigenicity of human breast cancer cells.* Oncotarget, 2015. **6**(10): p. 7970-91.

48. Warren, J., et al., *Activation of latent transforming growth factor-beta1, a conserved function for pregnancy-specific beta 1-glycoproteins.* Mol Hum Reprod, 2018. **24**(12): p. 602-612.
49. Fu, Z.Z., et al., *Relationship of serum levels of VEGF and TGF-beta1 with radiosensitivity of elderly patients with unresectable non-small cell lung cancer.* Tumour Biol, 2014. **35**(5): p. 4785-9.
50. Abdullah-Soheimi, S.S., et al., *Patients with ovarian carcinoma excrete different altered levels of urine CD59, kininogen-1 and fragments of inter-alpha-trypsin inhibitor heavy chain H4 and albumin.* Proteome Sci, 2010. **8**: p. 58.
51. Yang, H., et al., *Uncovering a novel DNA repair-related radiosensitivity model for evaluation of radiotherapy susceptibility in uterine corpus endometrial cancer.* Heliyon, 2024. **10**(8): p. e29401.
52. Ngwa, W., *Systemic immune effects boost radiotherapy.* Nat Biomed Eng, 2018. **2**(8): p. 562-563.
53. Binder, D.C., Y.X. Fu, and R.R. Weichselbaum, *Radiotherapy and immune checkpoint blockade: potential interactions and future directions.* Trends Mol Med, 2015. **21**(8): p. 463-5.
54. Ngwa, W., et al., *Using immunotherapy to boost the abscopal effect.* Nat Rev Cancer, 2018. **18**(5): p. 313-322.
55. Demaria, S., et al., *Radiation dose and fraction in immunotherapy: one-size regimen does not fit all settings, so how does one choose?* J Immunother Cancer, 2021. **9**(4).
56. Grass, G.D., et al., *The Radiosensitivity Index Gene Signature Identifies Distinct Tumor Immune Microenvironment Characteristics Associated With Susceptibility to Radiation Therapy.* Int J Radiat Oncol Biol Phys, 2022. **113**(3): p. 635-647.
57. Pang, Q., et al., *Wnt/beta-Catenin Signaling Pathway-Related Proteins (DKK-3, beta-Catenin, and c-MYC) Are Involved in Prognosis of Nasopharyngeal Carcinoma.* Cancer Biother Radiopharm, 2019. **34**(7): p. 436-443.
58. Luo, M., et al., *FOXO3a knockdown promotes radioresistance in nasopharyngeal carcinoma by inducing epithelial-mesenchymal transition and the Wnt/beta-catenin signaling pathway.* Cancer Lett, 2019. **455**: p. 26-35.

List of abbreviations

NPC:List of abbreviations
NPC-RSS:Nasopharyngeal Carcinoma Radiosensitivity score
AI:Artificial Intelligence
ML:Machine Learning
MR:Magnetic Resonance
MRI/DWI:Magnetic Resonance Imaging/Diffusion Weighted Imaging
FPKM:Fragments Per Kilobase Million
RMA:robust multi-array average
DEGs:Differentially expressed genes
GSVA: Gene Set Variation Analysis
GSEA: Gene Set Enrichment Analysis
GEO: Gene Expression Omnibus
MRI: Magnetic Resonance Imaging
CT: Computed Tomography
PET-CT: Positron Emission Tomography-Computed Tomography
AJCC: American Joint Committee on Cancer
GTV: Gross Tumor Volume
CBCT: Cone Beam Computed Tomography
PTV: Planning Target Volume
FPKM: Fragments Per Kilobase of transcript per Million mapped reads
cDNA: Complementary DNA
HISAT2: Hierarchical Indexing for Spliced Alignment of Transcripts 2
LASSO: Least Absolute Shrinkage and Selection Operator Regression
Enet: Elastic Net
StepglM: Stepwise Generalized Linear Model
SVM: Support Vector Machine
glmBoost: Generalized Linear Model Boosting
LDA: Linear Discriminant Analysis
plsRglm: Partial Least Squares Regression Generalized Linear Model
RF: Random Forest
GBM: Gradient Boosting Machine
XGBoost: Extreme Gradient Boosting
CNE2: Cell line name (Nasopharyngeal carcinoma)
CNE2-R: Radiation-resistant CNE2 cell line
qRT-PCR: Quantitative Reverse Transcription Polymerase Chain Reaction
CIBERSORT: Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts

TISIDB: Tumor and Immune System Interaction Database

AUC: Area Under the Curve

UMIs:unique molecular identifiers

PCA:principal component analysis

t-SNE:t-distributed stochastic neighbor embedding

Figure1 Flowchart for constructing a predictive model for NPC radiotherapy sensitivity

Differentially expressed genes obtained from local NPC transcriptome data, grouped according to radiosensitivity and radioresistance, were used to predict radiotherapy sensitivity scores (NPC-RSS) of NPC patients using 12 machine learning algorithms, including Lasso, Ridge, Enet, Stepglm, SVM, glmBoost, LDA, plsRglm, RandomForest, GBM, XGBoost, and NaiveBayes. Additionally, 48 other combinations of validated frameworks were constructed to predict the radiotherapy sensitivity score (NPC-RSS) of nasopharyngeal carcinoma patients. The most effective NPC-RSS was finally constructed based on the combination of glmBoost+NaiveBayes, which yielded the best AUC. The role and biological significance of NPC-RSS in NPC radiotherapy sensitivity were comprehensively explored through tumor immune microenvironment analysis, pathway enrichment analysis, and single-cell transcriptomic analysis.

Figure2 Consensus NPC-RSS construction and validation using an integrated machine learning approach

(A) Based on 48 combined validation frameworks, including 12 machine learning algorithms (Lasso, Ridge, Enet, Stepglm, SVM, glmBoost, LDA, plsRglm, RandomForest, GBM, XGBoost, and NaiveBayes), the area under the curve (AUC) of each model was calculated for both the training and validation sets. Heatmaps were generated based on the sample-weighted AUC values. (B) Demonstration of NPC-RSS genes with absolute values of weight coefficients greater than 10, based on the glmBoost and NaiveBayes combination. (C) Receiver operating characteristic (ROC) curves for the training and validation sets. (D) Volcano plot depicting differentially expressed genes between the sensitive and resistant groups of the CNE-2 cell line. (E) Timeline of resistance strain induction in the CNE-2 cell line. (F) Expression of the top 5 weighted genes in the NPC-RSS signature for the CNE-2 cell line. (G) Heatmap of differentially expressed genes between the sensitive and resistant groups of the CNE-2 cell line. (H) Heatmap of z-scores for NPC-RSS genes in the CNE-2 cell line. (I) Analysis of differences in sensitivity scores among CNE-2 cell lines.

Figure3 Annotation analysis of in-house NPC cohorts based on NPC-RSS predictive grouping with immune-related features. (A) Comparison of immune cell infiltration between NPC-sensitive and resistant tissue groups. *P < 0.05, P < 0.01. (B) Bubble plot depicting the correlation between the top 5 weight-ranked gene features in the NPC-RSS (SMARCA2, DMC1, CD9, PSG4, KNG1) and tumor-immune-infiltrating cells in the radiotherapy-sensitive group. Bubble size represents the proximity of the P-value to zero, with orange and blue colors indicating the strength of positive and negative correlations, respectively. (C) Analysis of interactions among 22 different immune cell types in patients from the NPC-sensitive group. *P < 0.05, P < 0.01, *P < 0.001. (D)

Correlation analysis of the top 5 weight-ranked genes in the NPC-RSS (SMARCA2, DMC1, CD9, PSG4, KNG1) with functionally diverse immune genes in the radiotherapy-sensitive group. *P < 0.05, P < 0.01, ***P < 0.001.

Figure4 Biological characterization of key NPC-RSS genes at the single-cell level. (A) Clustered UMAP plot of three radiotherapy-sensitive and one radiotherapy-resistant sample with a total of 28,957 cells. Each color represents a cellular subpopulation (see cellular subpopulation annotations on the right). (B) Myeloid cells, epithelial cells, fibroblasts, and mast cells were significantly more abundant in samples from the radiotherapy-sensitive group compared to the radiotherapy-resistant group. (C) NPC-RSS scores displayed on all cells, with redder colors indicating higher scores. (D) NPC-RSS model gene expression in all cell subpopulations. Redder colors indicate higher expression, while bluer colors indicate lower expression. (E) Histogram showing the percentage of seven cell subpopulations in the radiotherapy-sensitive and radiotherapy-resistant groups. Different colors indicate different cell subpopulations. (F) Expression of marker genes used to annotate various cell subpopulations.

Figure5 GSVA and GSEA analyses of NPC-RSS key genes and their correlation with radiosensitization genes. (A) GSVA of SMARCA2. (B) GSVA of DMC1. (C) GSEA of SMARCA2. (D) GSEA of DMC1. (E) Pearson correlation bubble plot of the top five NPC-RSS genes (SMARCA2, DMC1, CD9, PSG4, and KNG1) with radiosensitization-related genes. The larger the bubble, the closer the p-value is to zero; the more orange the color, the stronger the positive correlation; the more green the color, the stronger the negative correlation.

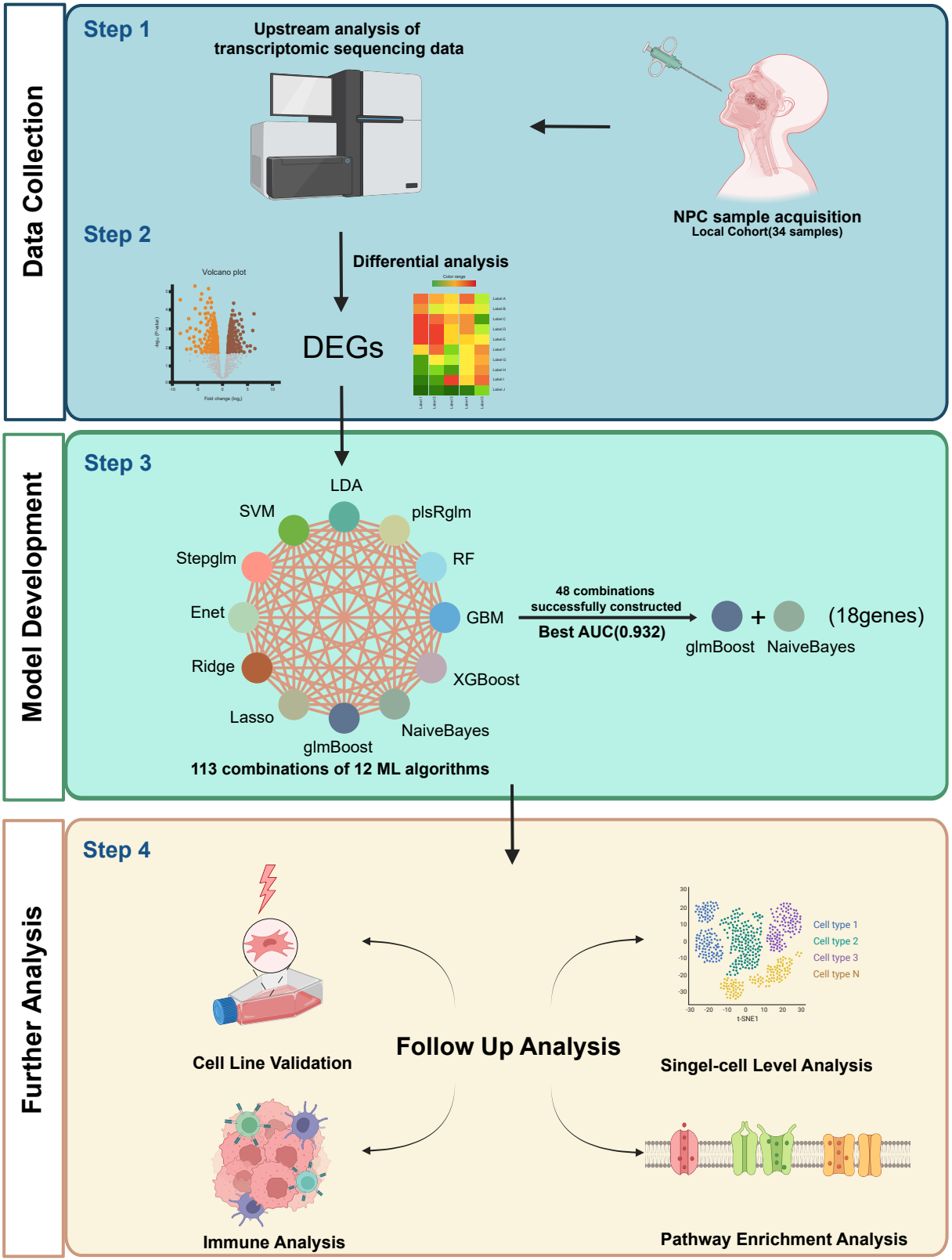
Supplementary Figure1. Expression of 18 model genes grouped according to NPC-RSS in our center's transcriptome data.

Supplementary Figure2. The top 20 radiosensitivity-related genes were identified using GeneCards (https://www.genecards.org/), and their expression was obtained from our center's transcriptome data, grouped according to NPC-RSS.

Supplementary Table1. Baseline characteristics of the 34 NPC samples.

Supplementary Table2. Baseline characteristics of the 4 NPC samples.

Supplementary Table3. Results of differential analysis of differential genes for which machine learning was performed.


Data Collection
Step 1
Upstream analysis of transcriptomic sequencing data
NPC sample acquisition
Local Cohort(34 samples)
Step 2
Differential analysis
Volcano plot
DEGs
Model Development
Step 3
LDA
SVM
plsRglm
Stepglm
RF
Enet
GBM
Ridge
XGBoost
Lasso
NaiveBayes
glmBoost
113 combinations of 12 ML algorithms
48 combinations successfully constructed
Best AUC(0.932)
glmBoost
NaiveBayes
(18genes)
Further Analysis
Step 4
Cell Line Validation
Follow Up Analysis
Singel-cell Level Analysis
Cell type 1
Cell type 2
Cell type 3
Cell type N
t-SNE1
Immune Analysis
Pathway Enrichment Analysis

A

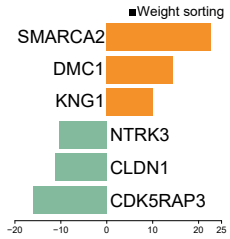
0.932	0.996	0.823	glmBoost+NaiveBayes
0.895	0.992	0.729	RF+plsRglm
0.89	1.000	0.703	RF+XGBoost
0.884	1.000	0.688	Stepglm[forward]
0.879	0.992	0.688	NaiveBayes
0.877	1.000	0.667	RF+Stepglm[forward]
0.873	1.000	0.656	glmBoost+Enet[alpha=0.7]
0.873	1.000	0.656	RF+Enet[alpha=0.3]
0.869	1.000	0.646	glmBoost+SVM
0.869	1.000	0.646	glmBoost+Lasso
0.869	1.000	0.646	RF+Enet[alpha=0.2]
0.866	0.996	0.646	RF+Enet[alpha=0.1]
0.865	1.000	0.635	glmBoost+Enet[alpha=0.8]
0.865	1.000	0.635	RF+LDA
0.861	1.000	0.625	glmBoost+Enet[alpha=0.5]
0.861	1.000	0.625	glmBoost+Enet[alpha=0.6]
0.857	1.000	0.615	glmBoost+Enet[alpha=0.3]
0.857	1.000	0.615	glmBoost+Enet[alpha=0.9]
0.857	1.000	0.615	RF+Enet[alpha=0.4]
0.857	1.000	0.615	glmBoost+RF
0.857	1.000	0.615	glmBoost+XGBoost
0.855	0.996	0.615	RF+Ridge
0.853	1.000	0.604	glmBoost+Enet[alpha=0.2]
0.853	1.000	0.604	glmBoost+Enet[alpha=0.4]
0.853	1.000	0.604	RF+Lasso
0.853	1.000	0.604	RF
0.85	0.909	0.750	plsRglm
0.85	1.000	0.594	glmBoost
0.85	1.000	0.594	RF+Enet[alpha=0.6]
0.846	1.000	0.583	glmBoost+Ridge
0.846	1.000	0.583	glmBoost+Enet[alpha=0.1]
0.843	0.996	0.583	LDA
0.84	1.000	0.568	glmBoost+Stepglm[forward]
0.838	1.000	0.562	RF+glmBoost
0.83	1.000	0.542	Enet[alpha=0.6]
0.83	1.000	0.542	Enet[alpha=0.7]
0.83	1.000	0.542	glmBoost+LDA
0.821	0.988	0.536	RF+NaiveBayes
0.819	1.000	0.510	Enet[alpha=0.2]
0.815	1.000	0.500	Enet[alpha=0.3]
0.811	1.000	0.490	Enet[alpha=0.1]
0.807	1.000	0.479	Enet[alpha=0.4]
0.803	1.000	0.469	Enet[alpha=0.5]
0.799	0.980	0.490	Ridge
0.774	0.850	0.646	RF+Enet[alpha=0.5]
0.77	0.919	0.516	XGBoost
0.763	0.955	0.438	SVM
0.763	0.955	0.438	RF+SVM

0.5 1

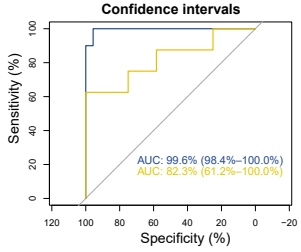
AUC 1 0.8 0.6 0.4

Cohort Test Train

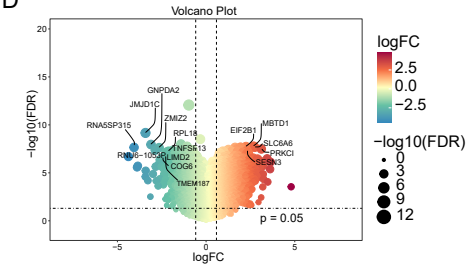
B



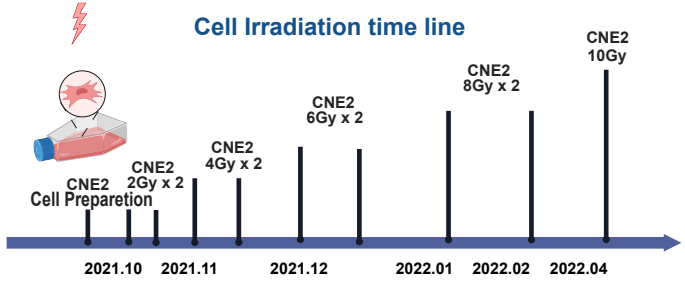
C



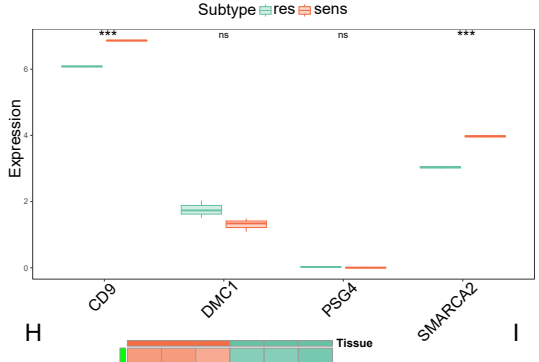
D



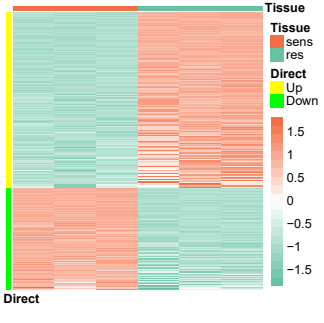
E



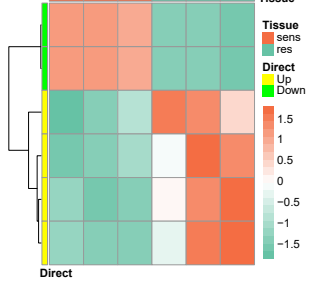
F



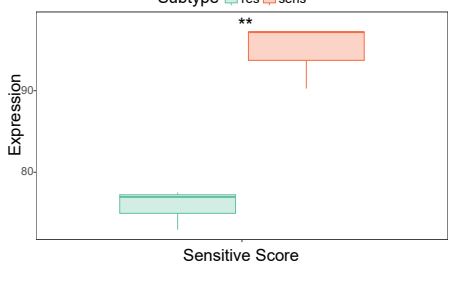
G

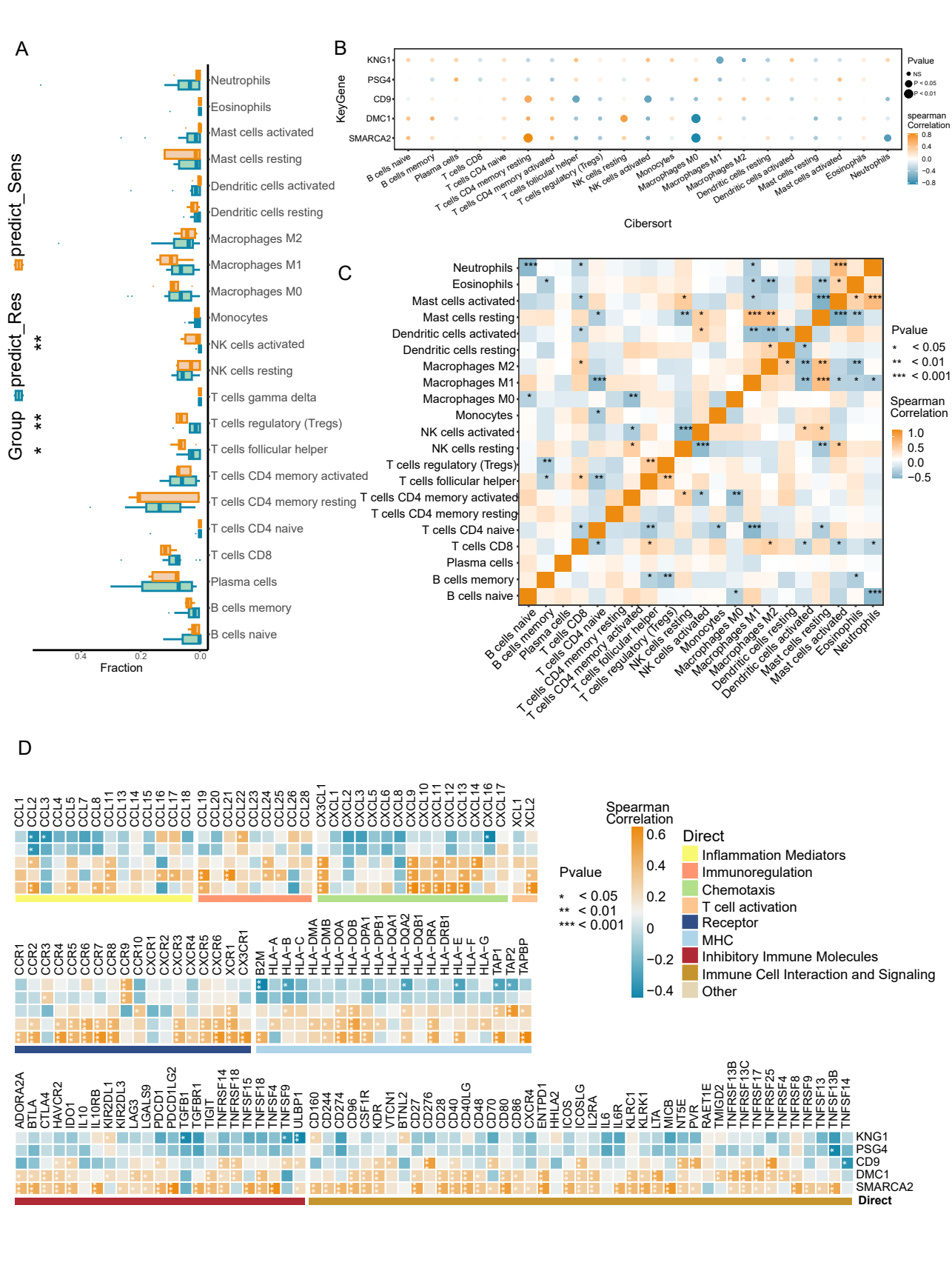


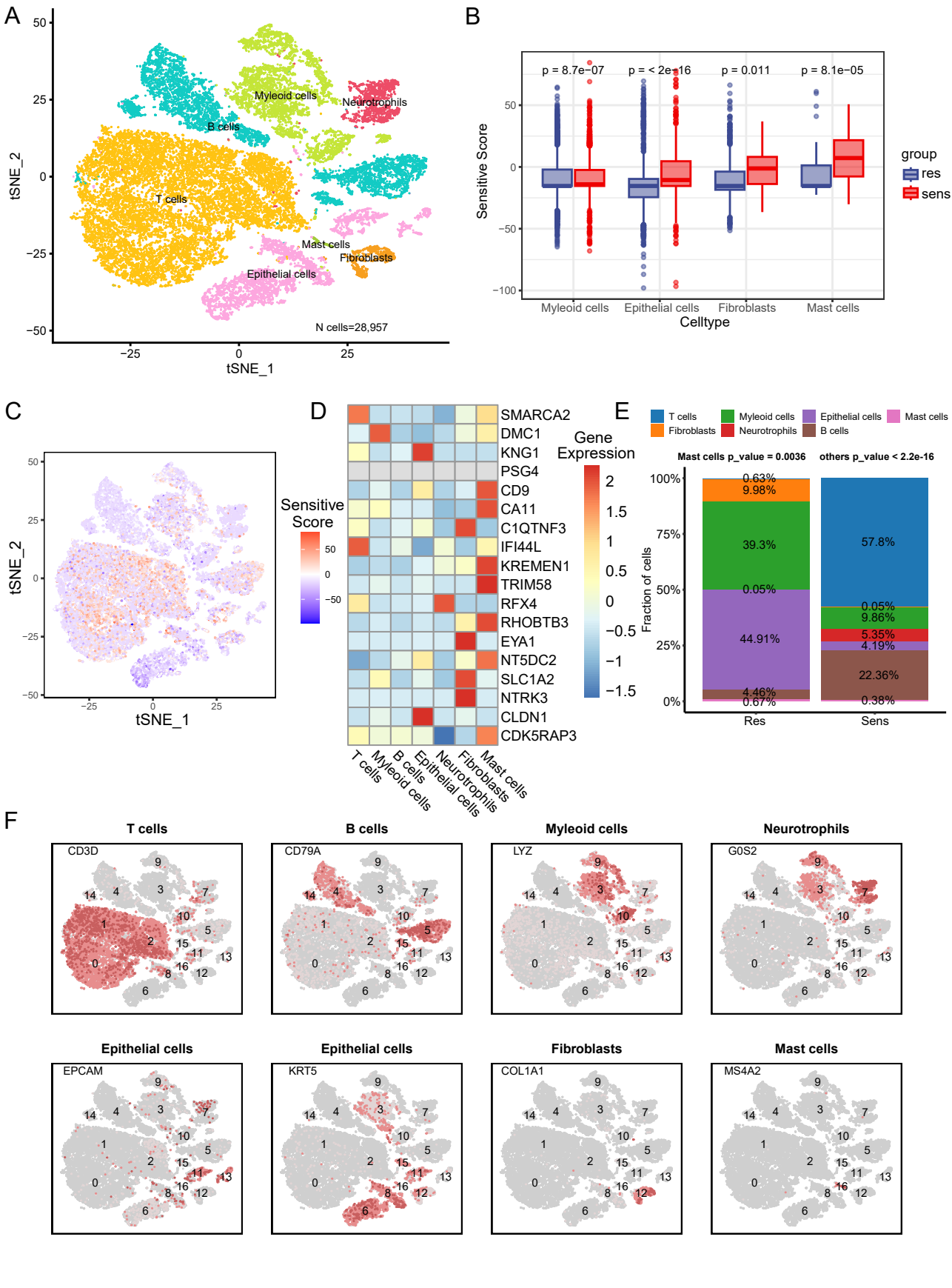
H

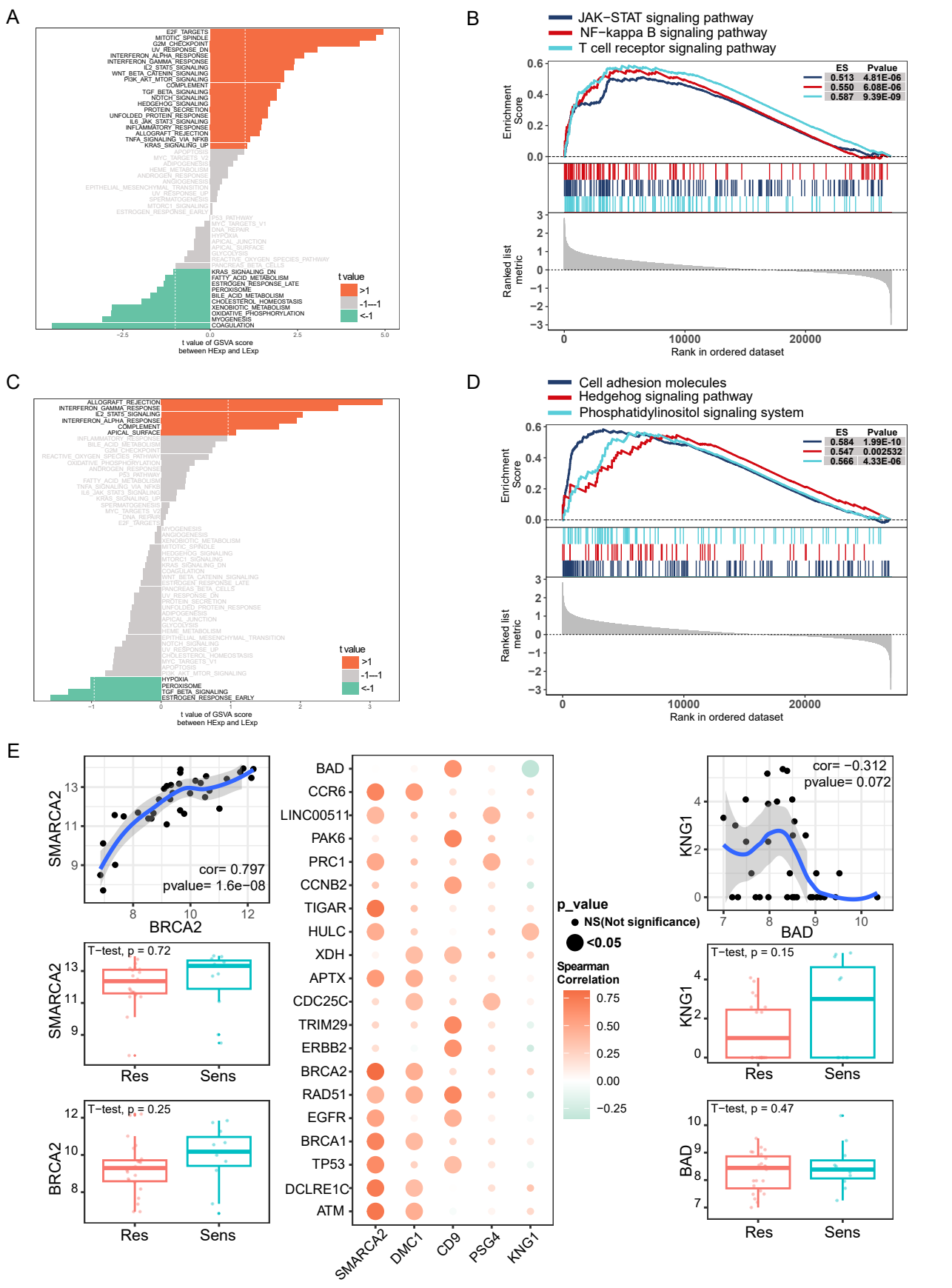


I




A
Group predict_Res predict_Sens
Neutrophils
Eosinophils
Mast cells activated
Mast cells resting
Dendritic cells activated
Dendritic cells resting
Macrophages M2
Macrophages M1
Macrophages M0
Monocytes
NK cells activated
NK cells resting
T cells gamma delta
T cells regulatory (Tregs)
T cells follicular helper
T cells CD4 memory activated
T cells CD4 memory resting
T cells CD4 naive
T cells CD8
Plasma cells
B cells memory
B cells naive
Fraction
B
KeyGene
KNG1
PSG4
CD9
DMC1
SMARCA2
Cibersort
Pvalue
NS
P < 0.05
P < 0.01
spearman Correlation
C
Pvalue
* < 0.05
** < 0.01
*** < 0.001
Spearman Correlation
D
Spearman Correlation
Pvalue
* < 0.05
** < 0.01
*** < 0.001
Direct
Inflammation Mediators
Immunoregulation
Chemotaxis
T cell activation
Receptor
MHC
Inhibitory Immune Molecules
Immune Cell Interaction and Signaling
Other
Direct


A
Myeloid cells
Neurotrophils
B cells
T cells
Mast cells
Fibroblasts
Epithelial cells
N cells=28,957
tSNE_1
tSNE_2
B
p = 8.7e−07
p = < 2e−16
p = 0.011
p = 8.1e−05
Sensitive Score
Celltype
Myeloid cells
Epithelial cells
Fibroblasts
Mast cells
group
res
sens
C
tSNE_1
tSNE_2
D
Sensitive Score
Gene Expression
SMARCA2
DMC1
KNG1
PSG4
CD9
CA11
C1QTNF3
IFI44L
KREMEN1
TRIM58
RFX4
RHOBTB3
EYA1
NT5DC2
SLC1A2
NTRK3
CLDN1
CDK5RAP3
T cells
Myeloid cells
B cells
Epithelial cells
Neurotrophils
Fibroblasts
Mast cells
E
T cells
Myeloid cells
Epithelial cells
Mast cells
Fibroblasts
Neurotrophils
B cells
Mast cells p_value = 0.0036
others p_value < 2.2e-16
Fraction of cells
Res
Sens
F
T cells
B cells
Myeloid cells
Neurotrophils
CD3D
CD79A
LYZ
G0S2
Epithelial cells
Epithelial cells
Fibroblasts
Mast cells
EPCAM
KRT5
COL1A1
MS4A2


A
E2F_TARGETS
MITOTIC_SPINDLE
G2M_CHECKPOINT
UV_RESPONSE_DN
INTERFERON_ALPHA_RESPONSE
INTERFERON_GAMMA_RESPONSE
E2_STAT5_SIGNALING
WNT_BETA_CATENIN_SIGNALING
PI3K_AKT_MTOR_SIGNALING
COMPLEMENT
TGF_BETA_SIGNALING
NOTCH_SIGNALING
HEDGEHOG_SIGNALING
PROTEIN_SECRETION
UNFOLDED_PROTEIN_RESPONSE
IL6_JAK_STAT3_SIGNALING
INFLAMMATORY_RESPONSE
ALLOGRAFT_REJECTION
TNFA_SIGNALING_VIA_NFKB
KRAS_SIGNALING_UP
APOPTOSIS
MYC_TARGETS_V2
ADIPOGENESIS
HEME_METABOLISM
ANDROGEN_RESPONSE
ANGIOGENESIS
EPITHELIAL_MESENCHYMAL_TRANSITION
UV_RESPONSE_UP
SPERMATOGENESIS
MTORC1_SIGNALING
ESTROGEN_RESPONSE_EARLY
P53_PATHWAY
MYC_TARGETS_V1
DNA_REPAIR
HYPOXIA
APICAL_JUNCTION
APICAL_SURFACE
GLYCOLYSIS
REACTIVE_OXYGEN_SPECIES_PATHWAY
PANCREAS_BETA_CELLS
KRAS_SIGNALING_DN
FATTY_ACID_METABOLISM
ESTROGEN_RESPONSE_LATE
PEROXISOME
BILE_ACID_METABOLISM
CHOLESTEROL_HOMEOSTASIS
XENOBIOTIC_METABOLISM
OXIDATIVE_PHOSPHORYLATION
MYOGENESIS
COAGULATION
t value
>1
-1~-1
<-1
−2.5
0.0
2.5
5.0
t value of GSVA score
between HExp and LExp
B
JAK−STAT signaling pathway
NF−kappa B signaling pathway
T cell receptor signaling pathway
ES Pvalue
0.513 4.81E-06
0.550 6.08E-06
0.587 9.39E-09
Enrichment Score
Ranked list metric
0
10000
20000
Rank in ordered dataset
C
ALLOGRAFT_REJECTION
INTERFERON_GAMMA_RESPONSE
E2_STAT5_SIGNALING
INTERFERON_ALPHA_RESPONSE
COMPLEMENT
APICAL_SURFACE
INFLAMMATORY_RESPONSE
BILE_ACID_METABOLISM
G2M_CHECKPOINT
REACTIVE_OXYGEN_SPECIES_PATHWAY
OXIDATIVE_PHOSPHORYLATION
ANDROGEN_RESPONSE
P53_PATHWAY
FATTY_ACID_METABOLISM
TNFA_SIGNALING_VIA_NFKB
IL6_JAK_STAT3_SIGNALING
KRAS_SIGNALING_UP
SPERMATOGENESIS
MYC_TARGETS_V2
DNA_REPAIR
E2F_TARGETS
MYOGENESIS
ANGIOGENESIS
XENOBIOTIC_METABOLISM
MITOTIC_SPINDLE
HEDGEHOG_SIGNALING
MTORC1_SIGNALING
KRAS_SIGNALING_DN
COAGULATION
WNT_BETA_CATENIN_SIGNALING
ESTROGEN_RESPONSE_LATE
PANCREAS_BETA_CELLS
UV_RESPONSE_DN
PROTEIN_SECRETION
UNFOLDED_PROTEIN_RESPONSE
ADIPOGENESIS
APICAL_JUNCTION
GLYCOLYSIS
HEME_METABOLISM
EPITHELIAL_MESENCHYMAL_TRANSITION
NOTCH_SIGNALING
UV_RESPONSE_UP
CHOLESTEROL_HOMEOSTASIS
MYC_TARGETS_V1
APOPTOSIS
PI3K_AKT_MTOR_SIGNALING
HYPOXIA
PEROXISOME
TGF_BETA_SIGNALING
ESTROGEN_RESPONSE_EARLY
t value
>1
-1~-1
<-1
−1
0
1
2
3
t value of GSVA score
between HExp and LExp
D
Cell adhesion molecules
Hedgehog signaling pathway
Phosphatidylinositol signaling system
ES Pvalue
0.584 1.99E-10
0.547 0.002532
0.566 4.33E-06
Enrichment Score
Ranked list metric
0
10000
20000
Rank in ordered dataset
E
SMARCA2
BRCA2
cor= 0.797
pvalue= 1.6e−08
T−test, p = 0.72
Res
Sens
T−test, p = 0.25
BAD
CCR6
LINC00511
PAK6
PRC1
CCNB2
TIGAR
HULC
XDH
APTX
CDC25C
TRIM29
ERBB2
BRCA2
RAD51
EGFR
BRCA1
TP53
DCLRE1C
ATM
SMARCA2
DMC1
CD9
PSG4
KNG1
p_value
NS(Not significance)
<0.05
Spearman Correlation
0.75
0.50
0.25
0.00
−0.25
KNG1
BAD
cor= −0.312
pvalue= 0.072
T−test, p = 0.15
T−test, p = 0.47