

Immune Escape and Metabolic Reprogramming in Colon Cancer: Insights from Endocytosis-Related Genes

Yu Qin¹, Yang Mao¹, Yuxin Han¹, Kai Cheng², Jiangcheng Shi^{1*}

¹School of Life Sciences, Tiangong University, Tianjin, China

²School of Computer Science and Technology, Tiangong University, Tianjin, China

Abstract. Colon cancer (COAD) is a common malignancy, yet its etiology is not fully elucidated. This study gathered endocytosis-related genes, using gene expression profiles from TCGA databases to categorize molecular subtypes of COAD into Cluster1 and Cluster2 based on genes related to endocytosis, and further explored the connection between the two molecular subtypes and prognostic characteristics. Differential expression analysis of the two subtypes revealed 3412 differentially expressed genes (DEGs), whose functions were remarkably enriched in the cellular metastasis and oxidative phosphorylation in Cluster1 by fGSEA. Additionally, Cluster1 exhibited higher expression levels of DEGs associated with immune infiltration and metabolism in comparison to Cluster2 by GSVA, and TIDE scores indicated patients with Cluster2 may benefit more from immunotherapy. Based on the DEGs, we utilized univariate Cox regression to identify 759 prognostic genes, which were then screened by three machine learning models (Lasso, RF, SVM-RFE) simultaneously, resulting in four feature genes: NEK4, MED13, OXSR1, and SLAIN2. Moreover, in Cluster1, these feature genes displayed consistent positive or negative correlations with immune escape-related and metabolic reprogramming-related pathways and genes in Pearson heatmap. These results suggest that there are significant differences in immune escape and metabolic reprogramming between colon cancer subtypes Cluster1 and Cluster2 as determined by genes according to endocytosis.

1 Introduction

Colon cancer (COAD) is a common gastrointestinal malignancy that occurs in the colon region and is the most prevalent molecular subtype of colorectal cancer (CRC), accounting for the third most common gastrointestinal tumor [1]. It is a major cause of cancer-related death, ranking second in the world in terms of mortality and third in terms of incidence, accounting for approximately 10% of cancer incidence and cancer mortality, posing a significant threat to human health. It is projected that by 2030, the global worldwide burden posed by colon cancer will rise by 60%, resulting in 2.2 million new cases and 1.1 million fatalities [2]. Recent studies have also shown a progressive increase in the incidence of colon cancer in adults under the age of 50, with an increase of 1.6% from 2000 to 2013, and a sustained annual growth of 1.2% in the colon cancer mortality rate in this population [3]. Furthermore, Vieira et al. have demonstrated that colon cancer may be caused by smoking, family genetics, poor diet, physical inactivity, overweight, and obesity [4]. Since colon carcinogenesis is a complex and long-term process, patients typically receive treatment only after developing advanced symptoms such as perforation and obstruction. Colon cancer often has an insidious onset and a poor

prognosis [5]. Surgical resection is currently the most common treatment for early-stage colon cancer. Additionally, chemotherapy, targeted therapy, and immunotherapy have provided new options for colon cancer patients in different situations. However, the heterogeneity of tumors, drug resistance, and local recurrence and metastasis during treatment still have an unfavourable impact on the prognosis and recovery of COAD patients [6].

Endocytosis is the process of transporting extracellular substances into the intracellular compartment through the deformational movement of the plasma membrane. Research has shown that invasive pseudopods in phagocytosis drive the infiltration and metastatic processes in colon cancer, as well as in all malignant tumors [7]. Colon cancer cells attach to the extracellular matrix through invasive pseudopods and produce cytolytic substances to break through the matrix barrier and invade the periphery. Endocytosis harnesses its function in this process [8]. Colorectal cancer cells metastasize and proliferate in humans by evading the immune system. Tumor immune escape is a crucial characteristic of tumor formation. Colon cancer cells modify themselves and the tumor microenvironment to evade immune surveillance [9]. Gao et al. discovered that Small Heterodimer Partner 2 (SHP2) in macrophages can promote tumor immune escape by

* Corresponding author: shijiangcheng@tiangong.edu.cn

inhibiting the STING-TBK1-IRF3 pathway, resulting in the downregulation of type I interferon signaling in colon cancer patients [10]. Activated SHP2 disrupts the phosphorylation homeostasis of adhesion patch-associated proteins at the phagocytic cups of macrophages, leading to inactivation of $\beta 2$ integrins and cytoskeletal remodeling, which inhibits macrophage phagocytosis. Enriching mimetic inactivated SHP2 can enhance the phagocytosis of tumor cells by macrophages both in vivo and in vitro. Research has demonstrated that SHP2 plays a crucial role in modulating macrophage phagocytosis [11]. Metabolic reprogramming refers to the alteration of major pathways of intracellular energy metabolism, resulting in the accumulation of lactate as an intermediate product. This accumulation is derived from the tumor acidic microenvironment, which promotes tumor immune escape and the degradation of extracellular matrix, leading to local tumor invasion [12]. Endocytosis had been shown to fulfil its role in tumor development by regulating epithelial-mesenchymal plasticity (EMP) during metastasis. Castosa et al. found that high levels of Hakai expression were linked to increased tumor lymph node metastasis (TNM) staging in human colon cancer. In a mouse xenograft model, Hakai expression induced tumor growth, invasion, and metastasis [13]. Metabolic reprogramming enhances cancer cell invasiveness and alters metabolic patterns in metastatic cancer cells, contributing to colon cancer progression. [14]. Therefore, immune escape and metabolic reprogramming, which occur during the progression of colon cancer, form a vicious circle and are also causative of each other.

However, it is important to note that colon cancer patients may have different pathomechanisms, and even the same type of colon cancer can exhibit significant phenotypic differences and high heterogeneity in both histopathology and molecular biology. This heterogeneity can lead to substantial variations in disease progression, clinical efficacy, sensitivity to radiotherapy, and prognosis among different patients [15]. Cancer heterogeneity arises primarily due to genomic instability, which can occur at the level of single nucleotides or in a larger scope [16]. Transcriptomics analysis techniques have enabled researchers to acquire large-scale genomic data and are widely used in the study of cancer subtypes in tissues and organs such as breast, lung, and liver [17]. In summary, in order to dig deeper into the physiological and pathological mechanisms of colon cancer cell endocytosis-related genes in their immune escape and metabolic reprogramming, this paper is based on transcriptomics data, combined with supervised and unsupervised machine learning algorithms, analyzed at the level of colon cancer subtypes, to provide an important insight and a fresh perspective for the discovery of molecular mechanism biomarkers of colon cancer development.

2 Materials and Methods

2.1 Data collection

The Cancer Genome Atlas Database (TCGA) is a reference database for tumor research that collects and organizes cancer-related genomic data. It is widely accepted as an effective resource to support cancer research, as the number of samples it provides has been validated by several studies. We employed a R package named "TCGAbiolinks" to call the COAD API to remotely download all RNA-Seq data and clinical information from colon cancer samples [18]. The data were then preprocessed using the GDCprepare function and normalized to a TPM sample expression profile matrix with 483 cancer samples and 59,427 genes.

The Molecular Signatures Database (MSigDB) is a well-annotated repository of genes, available at <https://www.gsea-msigdb.org/gsea/msigdb> [19]. To prepare cancer tissue samples for typing, it is necessary to download a list of genes associated with endocytosis, immune checkpoints, and metabolic reprogramming in colon cancer cells from the MSigDB public database. This list will be used for subsequent differential analysis, enrichment analysis, and visualization of associated gene expression.

The UCSC Xena database (<https://xenabrowser.net>) collected and organized clinical and survival information for all TCGA colon cancer samples in this study. This study followed 483 patients, including 252 males and 229 females, with ages ranging from 30 to 90 years old. The racial distribution was 234 whites, 63 African Americans, and 11 Asians. However, racial information was missing for 173 samples. There are four definitions of survival time criteria for colon cancer: overall survival (OS), disease-specific survival (DSS), disease-free interval (DFI), and progression-free interval (PFI).

2.2 K-Means clustering algorithm

An efficient unsupervised learning technique for handling large datasets is the K-means clustering algorithm. It takes a collection of n samples and clusters them into k groups based on minimum variance. This ensures that cohorts in the same cluster are similar, while those in different clusters are not analogous [20]. This study utilized the K-Means clustering algorithm to cluster and categorize colon cancer tissue samples into two subtypes. The process was carried out using R packages including "mlbench", "factoextra", "cluster", and "fpc".

2.3 DEGs identification and enrichment analysis

The combined expression profile matrix and clustering and typing results of colon cancer tissue samples underwent limma analysis to identify differentially expressed genes (DEGs) between two colon cancer subtypes through the "limma" R package. [21]. In this analysis, we set $p < 0.05$ and $|\log_2 FC| > \log_2(1.5)$ as screening criteria. We then fitted a linear model using

lmFit and evaluated the differential expression values using the eBayes function.

SangerBox (<http://sangerbox.com>) is a web-based platform that offers various bioinformatic analysis methods and tools. The provided online tool enables the utilization of GO enrichment analysis to probe the biological meaningfulness of DEGs, such as Biological Process (BP), Cellular Composition (CC), and Molecular Function (MF). [22]. The R package "clusterProfiler" was utilized to cluster up- and down-regulated DEGs into KEGG pathways [23]. Additionally, to enhance the reference abundance of the pathway enrichment, we conducted a rapid analysis using the fast GSEA (fGSEA) method based on the Hallmark gene set of MSigDB, with a cut-off criterion of $p < 0.05$.

2.4 Differential analysis of prognosis, immunity and metabolism of colon cancer subtypes

For further survival analysis, we utilized two R packages, "survival" and "survminer", and employed Kaplan-Meier analysis to illustrate the survival time distribution for the two molecular colon cancer subtypes.

To evaluate the metabolic differences between the two colon cancer subtypes, A differential analysis was performed to determine the enrichment of genes associated with colon cancer subtypes on metabolic pathways. To evaluate the gene set enrichment results of microarrays and transcriptomes, we applied GSVA, a non-parametric, unsupervised analysis approach. This was done using the R packages "GSEABase" and "GSVA" [24]. The "limma" R package was used to analyze the difference in GSVA scores of metabolic pathways in the clustered colon cancer samples. The significance of the difference in pathway enrichment was measured by $t > 2$. The expression of genes related to metabolic reprogramming was visualized in the molecular subtypes of colon cancer samples at the gene level.

The IOBR R package (<https://github.com/IOBR/IOBR>) integrates eight previously published tumor microenvironment (TME) quantification methods, comprising CIBERSORT, ESITMATE, quanTIseq, xCell, TIMER, EPIC, MCPcounter, and IPS [25]. The results of eight algorithms were combined using the "limma" R package, and the significance of the difference in the enrichment of immune pathways was determined using the same criterion as before ($t > 2$). Since the present study targeted immune escape, the genes in the immunogenetic analysis were derived from the files of three types of immune checkpoint genes (Inhibitory, Stimulatory, Immune), and the expression of the genes of each type of immune checkpoint was extracted and visualized separately [26].

TIDE (<http://tide.dfci.harvard.edu/>) is an algorithm that predicts the efficacy of immunotherapy response in cancer patients by evaluating the presence of two different immune escape mechanisms in tumors. We used this algorithm to evaluate the likelihood of immune escape in molecular subtypes of colon cancer and to guide relevant therapeutic studies, the gene expression

profile of 483 COAD patients is regarded as the input of TIDE [27].

2.5 Feature genes screening

A univariate Cox regression model was utilized to analyze a batch of 3671 DEGs across four indicators: OS, DSS, DFI, and PFI, using univariate Cox regression model that relies on the coxph function of the "survival" package [28]. The Hazard Ratio (HR) was defined as the exponential value of Cox's regression coefficient for each DEG, and the upper and lower bounds of the HR were set to 95% confidence intervals. The calculation of HR for all DEGs across the four indicators was completed through 3671×4 loop iterations to screen for significant prognostic DEGs. These were then categorized based on the size of p-value into $p < 0.05$, $p < 0.01$, and $p < 0.001$.

LASSO (least absolute shrinkage and selection operator) is a type of regularized regression model that is well-suited for large-scale data analysis. In this study, we utilized the "glmnet" R package to analyze the expression values of 759 prognostic genes as features. We then condensed these prognostic genes into more important feature genes employing the root mean square error as the cost function [29].

RF (random forest) is a supervised learning method that integrates decision trees as base classifiers. Each tree randomly selects samples and features for classifier construction, and the final result is determined by voting to obtain representative features, i.e., the feature genes that are needed. The feature screening metrics of Random Forest (RF) are categorized into two types: MeanDecreaseAccuracy (MDA) and MeanDecreaseGini (MDG). MDA refers to the degree of reduction in the predictive accuracy of random forests by changing the value of a variable to a random number. MDG calculates the effect of each variable on the heterogeneity of the observations at each node of the classification tree. Thus, the comparison of variables involves calculating the effect of each variable on the heterogeneity of observations at each node of the classification tree. This comparison determines the importance of the variables, and the values of the two metrics are proportional to the importance of the corresponding genes. The R package "randomForest" [30] was used for this purpose. Finally, we selected the top 20 genes based on their importance for each of the two metrics as feature genes.

The SVM-RFE algorithm combines Support Vector Machine (SVM) with Recursive Feature Filtering (RFE) to achieve dimensionality reduction. It belongs to the backward search algorithm and selectively eliminates unnecessary features to ultimately retain important genes [31]. The SVM-RFE algorithm, which utilizes the R package "e1071", is based on the AvgRank. The higher the AvgRank, the more important the corresponding gene is. Finally, we consider the genes with AvgRank values in the top 100 as the feature genes [32]. Additionally, the exotic R script 'msvmRFE.R' (<https://gitee.com/lzh23/GEO1/blob/master/msvmRFE.R>) was applied throughout the analysis.

Finally, we obtain the feature genes by taking the intersection of the key genes obtained from the three algorithms and selecting only those genes that are common to all three methods.

2.6 Correlation between feature genes and immunity and metabolism

After screening the DEGs using univariate Cox regression and three machine learning algorithms, we need to analyze the correlation between the feature genes and metabolic and immune functions, metabolic reprogramming-related genes, and immune escape (immune checkpoint)-related genes. We will use the Pearson correlation coefficient method and present the results in heat maps. Correlation analysis was performed to consider the interconnection between the metabolism and immunity of colon cancer subtypes.

3 Results

3.1 Identification of molecular subtypes and DEGs involved in colon cancer

The tissue samples were initially classified into Cluster1 and Cluster2, utilizing K-Means algorithm based on colon cancer histopathology (Fig. 1A). Subsequently, we analyzed the DEGs in the two subtypes, adhering strictly to the screening criteria of $p < 0.05$ and $|\log_2FC| > \log^2(1.5)$. In total, 3671 DEGs were recognized, in which 3412 DEGs were remarkably upregulated while 259 DEGs were significantly downregulated (Fig. 1B). Figure 1C displays volcano and heat maps that illustrate the distribution of DEGs in two colon cancer subtypes. The top 25 upregulated DEGs and the top 25 downregulated DEGs, ranked by p-value, were selected for visualization. When combined with prognostic data, no significant difference in prognosis was found between Cluster1 and Cluster2 in OS, DSS, DFI, and PFI (Fig. 1D-Fig. 1G).

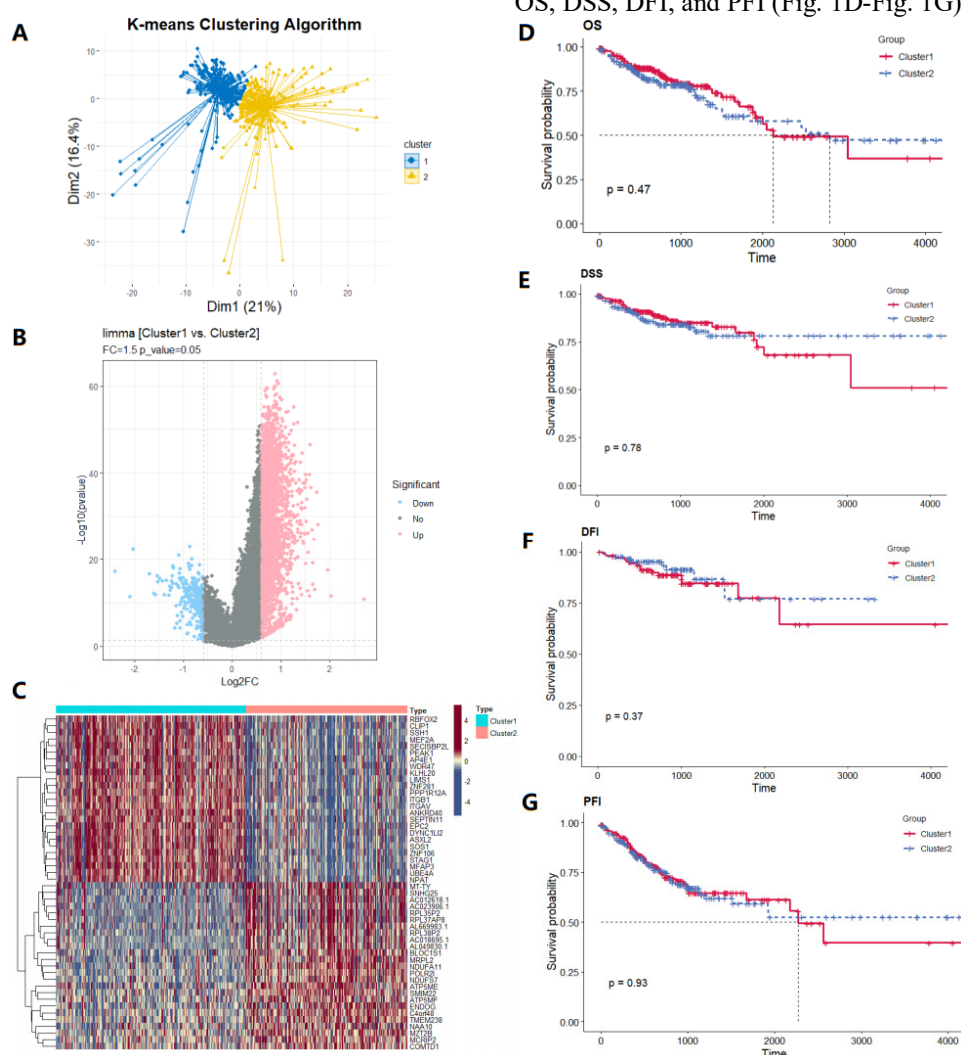


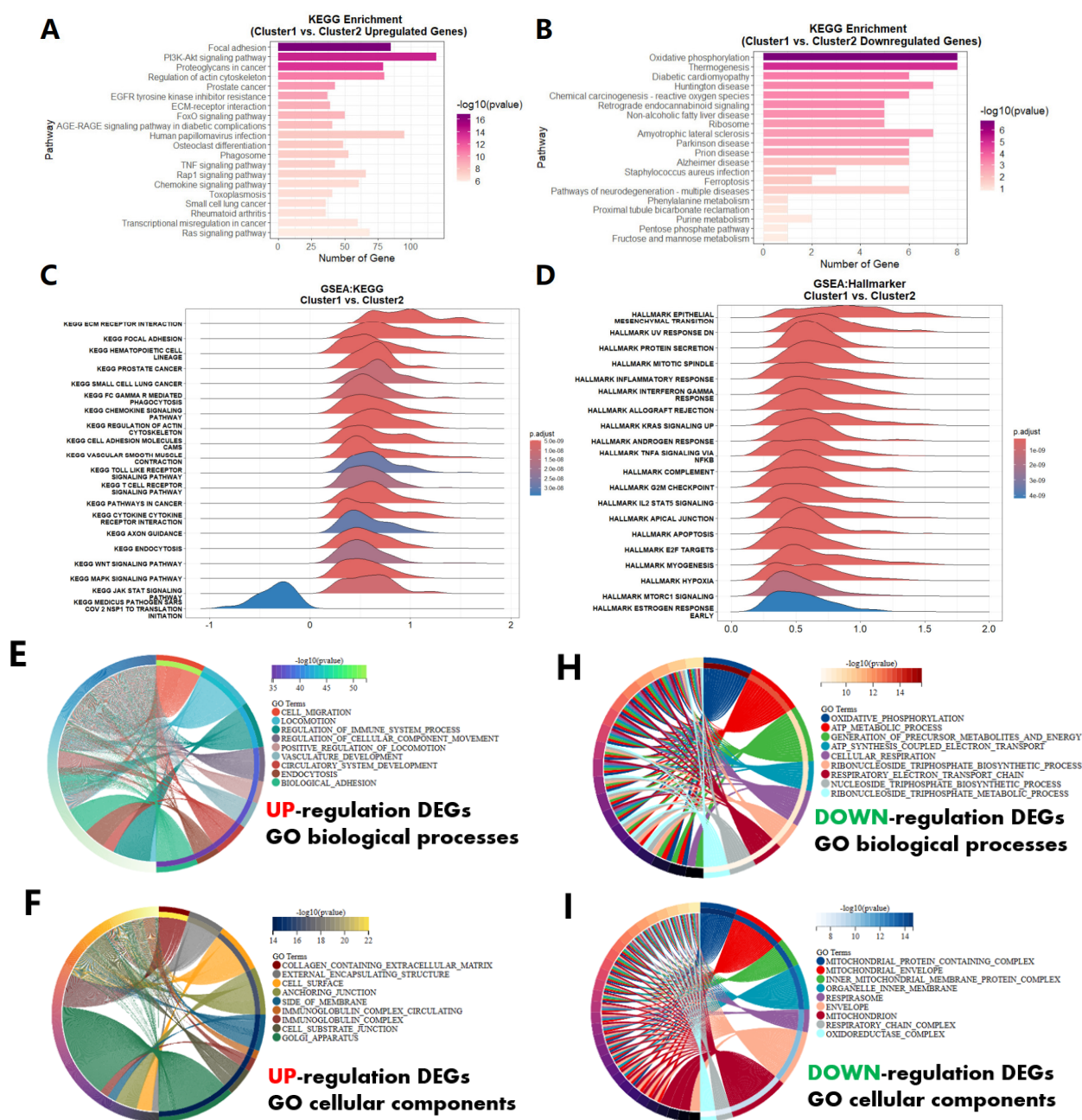
Fig. 1. Results of DEGs analysis of data for two colon cancer subtypes with survival prognosis for two subtypes of colon cancer. (A) K-Means clustering PCA downscaling plot for colon cancer. (B) Volcano plot of DEGs using FC values and adjusted p-values; pink scatters represent up-regulation of differential genes, light blue scatters represent down-regulation of DEGs, and gray scatters indicate non-significant DEGs. (C) Heat map of DEGs. Different colors represent the expression trends of genes in samples with different colon cancer subtypes (top25 for up-regulation and top25 for down-regulation, sorted according to p-value). (D) KM curve under OS indicator. (E) KM curves under DSS indicator. (F) KM curve under the DFI indicator. (G) KM curve under the PFI indicator.

3.2 Functional analysis of DEGs between two colon cancer subtypes

To further evaluate the function of DEGs among colon cancer subtypes, we conducted KEGG and GO analyses on up- and down-regulated DEGs, respectively. It revealed that the up-regulated DEGs were markedly enriched in biological processes (BP) related to colon cancer cell migration. The cellular components (CC) were mainly located in the extracellular matrix containing collagen, while the molecular functions (MF) were most significantly enriched in extracellular matrix structural composition (Fig. 2E-Fig. 2G). In contrast, the KEGG analysis of up-regulated DEGs showed significant enrichment in the PI3K-AKT signaling pathway (Fig. 2A). The down-regulated DEGs indicated significant enrichment in BP related to oxidative

phosphorylation, CC related to mitochondrial protein complexes and mitochondrial envelope, and MF related to electron transfer activities in the GO functions (Fig. 2H-Fig. 2J). On the other hand, the KEGG of down-regulates DEGs suggested significant enrichment in oxidative phosphorylation and thermogenesis (Fig. 2B).

To determine the effects of subtle expression changes of differential genes on signaling pathways, we analyzed them using the fGSEA method. The results indicate that GSEA-KEGG was most significantly enriched in the extracellular matrix (ECM) receptor interaction pathway, and more interestingly, endocytosis was also remarkably enriched (Fig. 2C). The results of fGSEA based on Hallmark gene sets demonstrated that DEGs were most significantly enriched in the protein secretion and apoptosis pathways (Fig. 2D). This is coincident with the results showed in Figure 1B.



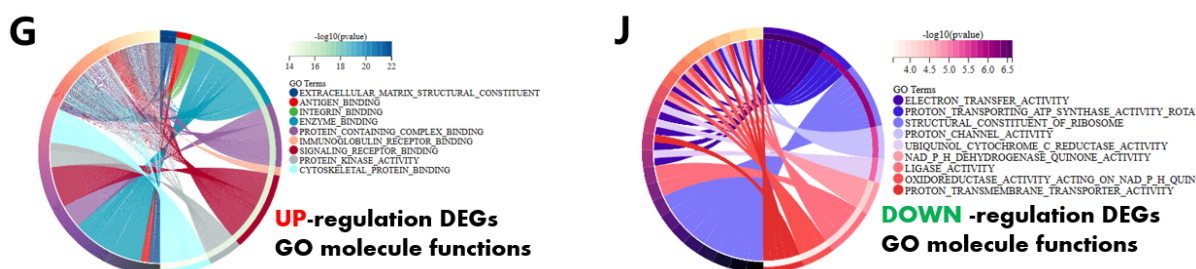


Fig. 2. KEGG and GO enrichment analysis of DEGs for colon cancer subtypes. (A) KEGG enrichment results of up-regulated DEGs. (B) KEGG enrichment results of down-regulated DEGs. (C) The fGSEA-KEGG enrichment ridgeline map of DEGs. (D) GSEA-Hallmark enrichment ridgeline of DEGs. (E) Up-regulated GOBP enrichment results of DEGs. (F) GOCC enrichment results of up-regulated DEGs. (G) Up-regulated GOMF enrichment results of DEGs. (H) Down-regulation of GOBP enrichment results of DEGs. (I) Down-regulation of GOCC enrichment results of DEGs. (J) Down-regulation of GOMF enrichment results of DEGs.

3.3 Differences between immune infiltration and metabolic reactions among colon cancer subtypes

We selected 43 metabolic pathways from the KEGG pathway gene set. Using GSVA enrichment analysis and limma difference analysis, we identified 28 significantly enriched metabolic pathways ($|t| > 2$) (Fig. 3B). These pathways include myo-inositol phosphate metabolism, aldosterone regulation of sodium reabsorption, metabolic effects of cytochrome P450 on xenobiotics, and retinol metabolism. The degree of metabolism was more pronounced in the colon cancer subtype Cluster1 than in the subtype Cluster2, as evidenced by the magnitude of GSVA score t-value of the first two pathways exceeding 10 and the number of pathways with positive t-value being greater than the number with negative t-value. In Figure 3B, at the gene level, most of the metabolic reprogramming-related gene expressions in Cluster1 were clearly higher than those in Cluster2 in Figure 3F, such as ACLY, DLST, SUCLG2, etc., but there were also two counterexamples, TALDO1, GAPDH.

Meanwhile, the IOBR package identified 138 immune indicators for colon cancer subtypes using eight immune infiltration algorithms. Of these, 97 were significantly enriched ($|t| > 2$) according to limma difference analysis (Fig. 3A), including endothelial cells (in MCPcounter) and macrophages (in TIMER). The

immunological factor scores associated with the four types of immunophenotypes (MHC, EC, SC, and CP) and immune factor scores (in IPS) were also considered. Additionally, other uncharacterized cells (in quantiseq) were analyzed. The analysis showed that Cluster1 had a significantly higher number of indicators with positive t-values compared to negative ones. Additionally, there were more indicators with t-values greater than 10 in GSVA scores for Cluster1, suggesting that at the level of immune indicators, the activation of immune function in Cluster1 was significantly greater than that in Cluster2. At the gene level, Figures 3C, 3D, and 3E demonstrate significantly higher expression of the three classes of immune checkpoint genes in Cluster1 compared to Cluster2. Specifically, HMGB1 and CXCL10 in the Stimulatory class, IDO1 and TGFB1 in the Inhibitory class, and HLA-DPB1 in the Immune class showed significant differences. The only exception was TNFRSF14 in the Stimulatory class, which had opposite results in Cluster2.

Figure 3G showed that the TIDE scores of Cluster1 were lower than those of Cluster2 when the TIDE algorithm was applied. This suggested that patients with Cluster1 responded poorly to immunotherapy for colon cancer compared to those with Cluster2 ($p < 0.001$). We hypothesized that the former might be resistant to immunotherapy.

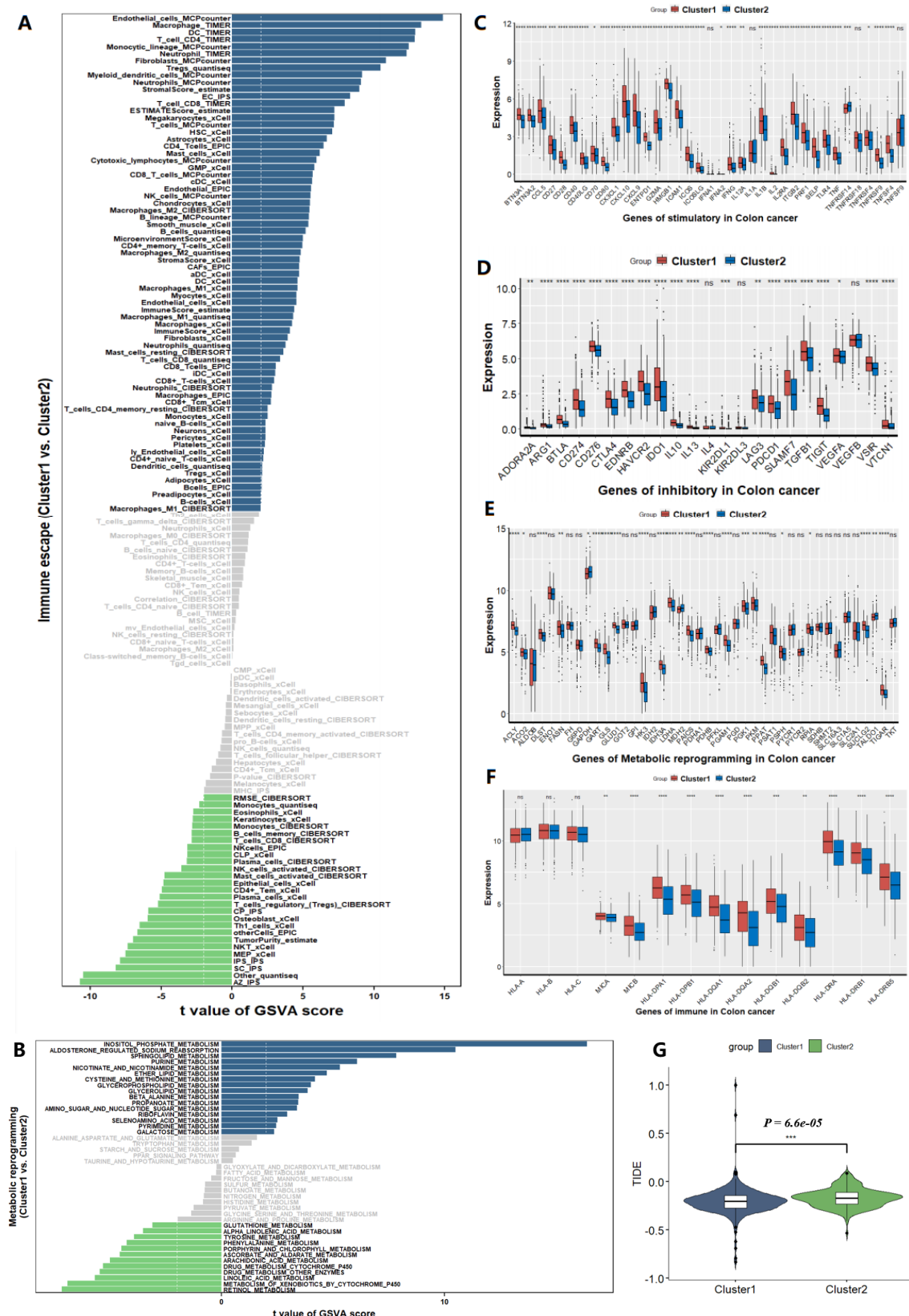


Fig. 3. Metabolic analysis and immune infiltration analysis of colon cancer subtypes. (ns, $p > 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$) (A) Histogram of 97 significantly enriched immune indicators. (B) Histogram of 28 significantly enriched metabolic pathways. (DC: Dendritic Cells; AZ: All relevant immune factors score, the higher the score, the higher the expression of relevant factors; MHC: Antigen Processing relevant immune factors score; EC: Effector Cells relevant immune factors score; SC: Suppressor Cells relevant immune factors score; CP: Checkpoints or Immunomodulators related immune factor score (C) Box plots of the expression of Stimulatory class genes in two subtypes of colon cancer. (D) Box plots of the expression of Inhibitory class genes in two subtypes of colon cancer. (E) Box plots of expression of Immune class genes in two subtypes of colon cancer. (F) Box plots of Expression of metabolic reprogramming-related genes in two subtypes of colon cancer. (G) Comparison of TIDE levels for Cluster1 and Cluster2 (***, $p < 0.001$)

3.4 Obtaining prognostic genes by Cox regression

In the TCGA samples, univariate Cox regression analysis screening (Fig. 5A-Fig. 5D) revealed that more than half of the prognostic DEGs for the three indicators, other than the DSS indicator, were risk factors for the

DEGs between the two types of colon cancer subtypes. To comprehensively analyze the prognostic genes, we combined all the prognostic DEGs involved in the four indicators. This resulted in 759 prognostic genes, which were used as input data for the machine learning feature screening after taking the concatenated set (Fig. 4E).

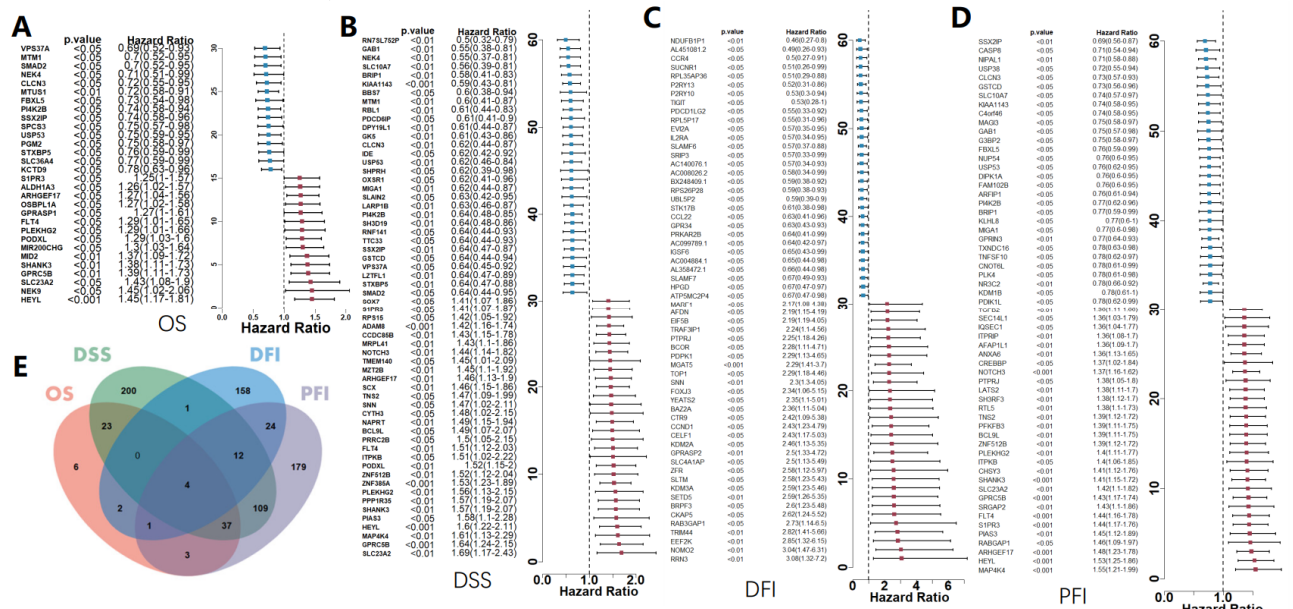


Fig. 4. univariate Cox regression analysis of differential genes. (A) 76 prognostic DEGs (29 protective factors, 47 risk factors, only the top 15 of each protective and risk factor were displayed) under the OS indicator. (B) 386 prognostic DEGs under the DSS indicator (216 protective factors, 170 risk factors, only the top 30 of each protective and risk factor were displayed). (C) 202 prognostic DEGs under the DFI indicator (45 protective factors, 157 risk factors, only the top 30 of each protective and risk factor were displayed). (D) 369 prognostic DEGs under the PFI indicator (54 protective factors, 305 risk factors, only the top 30 of each protective and risk factor were displayed). (E) Venn diagram of prognostic genes under the four survival indicators.

3.5 NEK4, MED13, OXSR1, SLAIN2 are identified as feature genes by machine learning

The 759 genes with prognostic value underwent screening using three different supervised machine learning algorithms. The LASSO model was found to be the best fit when genes' number reached 80, as the binomial deviation was smallest at this point (Fig. 5A, Fig. 5B). These 80 genes are associated with colon cancer subtypes and are considered the feature genes. The RF was utilized to compute and rank prognostic

gene importance scores. Model error was stabilized at approximately 200 trees (Fig. 6C). The top 20 genes, based on both MDA and MDG indicators, were selected as feature genes, totaling 26 (Fig. 5D). Furthermore, the SVM-RFE sorted each prognostic gene based on its calculated AvgRank and ultimately selected 39 feature genes with 100 as the cutoff criterion (Fig. 5E).

Finally, we identified the four most representative feature genes by intersecting the feature genes screened by the three supervised machine learning models: NEK4, MED13, OXSR1, and SLAIN2.

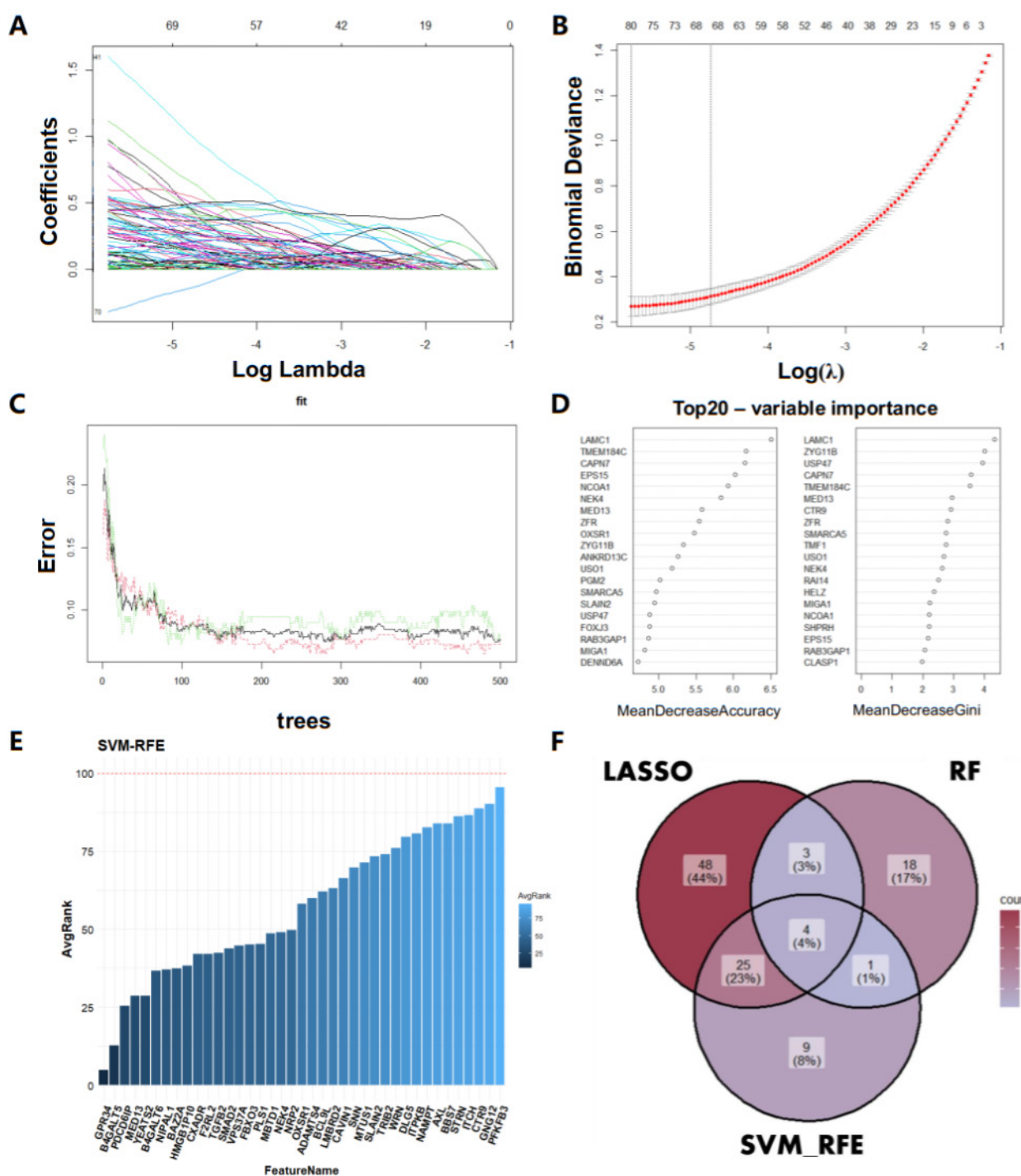


Fig. 5. Supervised machine learning algorithm screening feature genes. (A,B) LASSO regression algorithm screens 80 feature genes. (C,D) RF algorithm screens 40 feature genes. (E) SVM-RFE algorithm screens 39 feature genes. (F) Venn diagram of feature genes screened by machine learning algorithm.

3.6 Expression of feature genes shows relevance in immune escape and metabolic reprogramming

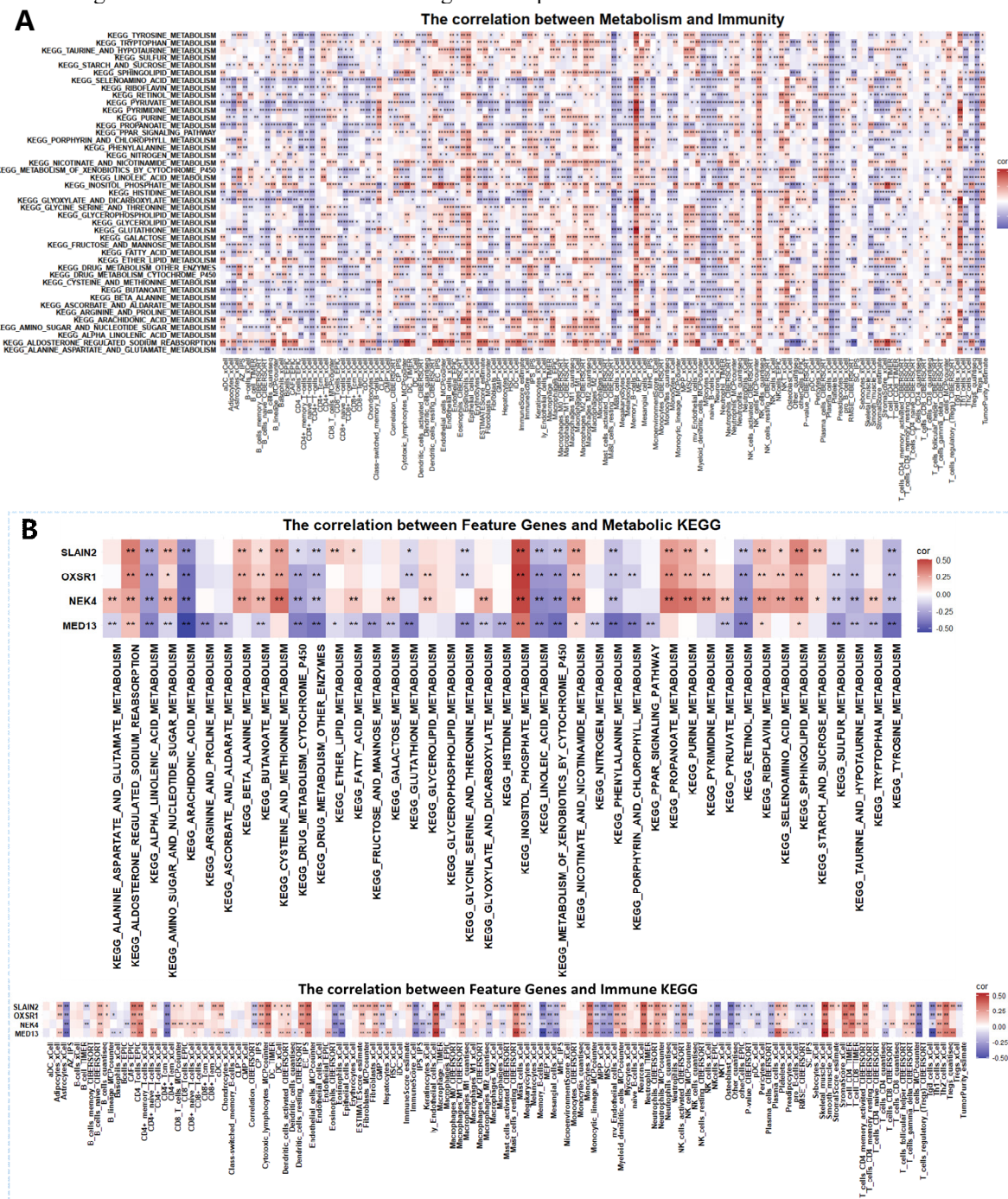
Based on Fig. 6A, it's clear that the immune infiltration-related functions and pathways of metabolic activities in colon cancer showed multiple consistencies of positive or negative correlation, which indicated that immune infiltration and metabolic responses were closely related, intertwined and influenced each other in an intricate

relationship. At the pathway level, the relationship between the four feature genes, NEK4, MED13, OXSR1 and SLAIN2, and the metabolic and immune activities of the colon cancer subtype samples is shown in Fig. 6A and Fig. 6B, and there were consistencies of the four feature genes in the metabolic and immune pathways. For instance, the four feature genes exhibited significant positive correlations with metabolic pathways, such as phosphatidylinositol metabolism and regulation of sodium reabsorption by aldosterone, as well as immune indicators, such as macrophages (in TIMER) and mast cells (in xCell) (Fig. 1). In Fig. 6A, there were no

significant correlations found. However, in Fig. 6C, highly significant negative correlations were observed with metabolic pathways such as linoleate metabolism, arachidonic acid metabolism, and regulatory T cells (Treg) (in CIBERSORT). Additionally, the epithelial cells (in xCell) and other immune indicators also showed highly significant negative correlations.

At the gene level, the four feature genes exhibited a strong positive correlation with EDNRB and ARG1 genes in the Inhibitory class, HMGB1 and ENTPD1 genes in the Stimulatory class, and the HLA-DQA1 gene in the Immune class. The analyzed data revealed highly significant negative correlations between the gene

VEGFB and the Inhibitory class, as well as between the genes TNFRSF14 and TNFRSF4 and the Immune class gene HLA-A. However, the co-expression of the feature genes with Immune class genes was weak (Fig. 6C). Furthermore, the four feature genes showed a significant positive correlation with metabolic reprogramming-related genes, including GLS, GART, PPAT, and SUCLG2. Notably, there was no negatively correlated co-expression observed (Fig. 6D). Overall, the correlations between the feature genes (NEK4, MED13, OXSR1, and SLAIN2) and the metabolic and immune aspects of colon cancer subtypes were predominantly positive.



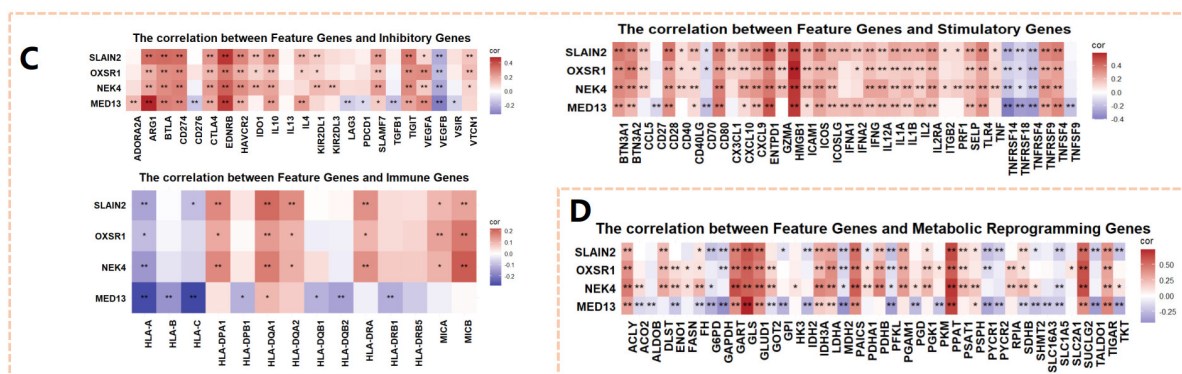


Fig. 6. Heat map of Pearson correlation analysis. (*, $p < 0.05$; **, $p < 0.01$) (A) Heat map of the relationship between immunity and metabolism. (B) Heat map of the relationship between the 4 feature genes and metabolic pathways/immunological indicators. (C) Heat map of the relationship between the 4 feature genes and 3 classes of genes of immune checkpoint. (D) Heat map of the relationship between the 4 feature genes and genes related to metabolic reprogramming.

4 Discussion

Colon cancer is a prevalent gastrointestinal tumor with a high degree of invisibility and poor prognosis. It has the highest number of new cases and deaths among gastrointestinal tumors, and its incidence is much higher than that of other gastrointestinal tumors. Being diagnosed with colon cancer represents a substantial physical and emotional impact for patients, as well as on the economy. [33]. This study analyzed the differences in immune escape and metabolic reprogramming in colon cancer subtypes according to endocytosis-related genes. The aim is to fill gaps in research mechanisms related to the occurrence and development of colon cancer subtypes and provide an effective reference for the formation of new clinical treatment strategies.

The study's analysis revealed 3671 DEGs between Cluster1 and Cluster2, with more up-regulated genes than down-regulated genes (Fig. 1B). Subsequent GSEA functional enrichment and immune infiltration of the DEGs confirmed this result. The majority of significantly enriched pathways in the ridgeline plot were in subtype Cluster1 (Fig. 2C-Fig. 2D). Cluster1 showed significant enrichment in most of the immune function indicators in the bar graph. Additionally, the expression of immune checkpoint-related genes and metabolic reprogramming-related genes were significantly higher in Cluster1 than in Cluster2 (Fig. 3C-Fig. 3F). Meanwhile, the survival analysis results showed no significant difference between the two molecular subtypes of colon cancer in terms of prognosis (Fig. 1D-Fig. 1G). However, this did not impact the study of the significance of the differences between the two subtypes of colon cancer in terms of immune escape and metabolic reprogramming. After functional enrichment of DEGs, we identified four feature genes that were finally screened using univariate Cox regression and three supervised machine learning algorithms: NEK4, MED13, OXSR1, and SLAIN2, and the expression of these genes showed consistency with a number of immune escape-associated functional and metabolic activity pathways collectively in the final Pearson correlation analysis (Fig. 6B).

Endocytosis is a complex process that involves the packaging, sorting, and internalization of cell surface proteins, lipids, and fluids from the extracellular environment into the cell [34]. Recent research has shown that endocytosis plays a crucial role in regulating immune surveillance and response, as well as tumor immune escape, tumor metastasis, and drug delivery [35]. Hui et al. discovered that inhibiting endocytosis with Prochlorperazine (PCZ) can effectively enhance the therapeutic effect of monoclonal antibodies on tumors. After the administration of PCZ, the binding of epidermal growth factor (EGF) on tumor tissues was significant, which further improved the tumor-killing ability of mononuclear macrophages (PBCMs). This treatment has promising clinical applications in cancer treatment [36]. In cellular metabolic activities, membrane compartments contain varying levels of phosphatidylinositol (PIP). PI(4,5)P₂ is involved in endocytosis, cyclic transport, and changes in the PIP level on the endosomal membrane are critical for the directional flow of cargoes. For instance, cyclic transport of clathrin-independent endocytosis (CIE) cargo membrane proteins is closely linked to cancer development [37]. Additionally, Bang et al. confirmed that LRP1 promotes the endocytosis of retinol complexes. This finding has significant implications for vitamin A-dependent intestinal immunity. Intestinal myeloid cells, including macrophages and dendritic cells (DCs), can provide antigenic guidance to B and T cells, promoting further functional immune differentiation and ultimately inducing the production of immunoglobulin A with metabolites of retinol [38]. It is possible that this process is related to immune activities associated with colon carcinogenesis. These findings are consistent with our immune and metabolic analyses of colon cancer (Fig. 3A-Fig. 3B, Fig. 6A).

NEK4 is a gene that encodes a serine/threonine kinase participates in cellular activities, such as maintenance of cilia stability, cycle regulation, and DNA damage response [39]. Research has demonstrated that NEK4 regulates positive regulators of Epithelial-Mesenchymal Transition (EMT) in lung adenocarcinoma cells by enhancing the transcription factors Zeb1 and Smad. In our study, NEK was identified as a feature gene that may have a similar mechanism of influence in

EMT regulation in colon cancer cells [40]. HMGB1 is a gene that is highly co-expressed with NEK4 (Fig. 6C). It can mediate lipopolysaccharide (LPS)-induced inflammation in colon cancer cells, which promotes an increase in colon cancer incidence and metastasis, leading to poor prognosis. NEK4 may indirectly regulate immune escape in colon cancer, but the exact mechanism is still unclear [41].

MED13 is a gene that encodes proteins. Disruption of the Mediator complex's structure can lead to dysfunction in transcriptional regulation. This dysfunction is strongly linked to the evolution of malignancies like COAD [42]. Research results indicate that the deletion of MED13 has a stronger specific effect on the expression of acquired super enhancer genes in COAD than BRD4. Dual targeting of the two leads to a drastic reduction in the expression of super-enhancer-related genes, resulting in a decline in cellular activity accompanied by a weakening of proliferation, and a cumulative effect on evolution of COAD [43]. Regarding metabolism, it was discovered that overexpression of Med13 specifically in the heart led to a significant increase in oxygen consumption and carbon dioxide in laboratory mice, and subsequent studies have shown that these factors based on metabolism are also manifested in adipose and liver tissues [44]. It was hypothesized that MED13's overexpression in COAD cells results in an upregulation in terms of lipid uptake, β -oxidation, and mitochondrial content. This, in turn, promotes the formation of an acidic environment in TME [45].

OXSRI is a gene encoding threonine-protein kinase OSR1, and its high expression can be attributed to mutations in the p53 protein, which is a positive regulator of PD-L1 expression. The level of OXSRI expression is associated with the level of PD-L1 and TILs positively, and PD-L1, as an important immune checkpoint molecule that acts as a key in the tumor escape in cancer, but further research is necessary to explore the molecular mechanisms involved in OXSRI during p53-mediated regulation of PD-L1 expression in COAD cells [46]. Metabolic reprogramming is a critical in the formulation and development of functional immune cells. Cao et al. showed that the metabolic enzyme encoded by the SUCLG2 gene was involved in maintaining the immunoregulatory function of diffDCs, which is essential for maintaining immune homeostasis, and there was a significant positive correlation between OXSRI and SUCLG2 (Fig. 6D), suggesting that OXSRI may be indirectly involved in the immune and metabolic activities of DCs in colon cancer by affecting SUCLG2 expression through some pathway [47].

SLAIN2 is a gene that codes for a protein and is co-expressed with the immune checkpoint-associated ENTPD1 gene. ENTPD1 encodes the dominant exonuclease CD39, which is expressed by endothelial cells (ECs) and Treg. According to Feng et al.'s study, the expression of Cd39 on Tregs inhibits NK cell-mediated anti-tumor response and enhances angiogenesis, thereby promoting the development and metastasis of MCA38 colon cancer tumors in mice model, which is coincident with the consequence of our correlation analysis (Figure 6B, Figure 6C) [48].

Although the results of the analysis on the differences in immune escape and metabolic reprogramming between the two types of colon cancer subtypes based on endocytosis-related genes were obtained, this study has some shortcomings. There are two potential areas for improvement in this study. Firstly, the colon cancer subtypes used do not show significant prognostic differences (Fig. 1D-Fig. 1G). It may be necessary to adjust the number of subtypes or utilize better subtyping algorithms to improve the results. Secondly, this paper only utilizes information from colon cancer samples in the TCGA database. The diversity of difference analysis can be further enriched by combining with colon cancer cohorts from the GEO datasets. In conclusion, the continuous development of biomedical big data allows us to utilize emerging bioinformatics technologies, such as spatial transcriptomics and imaging genomics, to deepen our understanding of this study from multiple perspectives. Further experimental studies can be planned to elucidate the molecular mechanisms involved.

5 Highlights

- Colon cancer based on endocytosis-related genes was categorized into two subtypes according to the K-means algorithm.
- There are significant differences between the two subtypes of COAD in terms of immune escape and metabolic reprogramming.
- Through the utilization of both machine learning techniques and univariate Cox regression analysis, four feature genes closely associated with immune escape and metabolic reprogramming in two subtypes of colon cancer were identified as being of particular interest: NEK4, MED13, OXSRI, and SLAIN2.

Funding

This study was supported by National Natural Science Foundation of China (62202330).

References

1. Yang, Z., Wu, H., Dai, D., Yuan, Y., & Shao, X. ZNF692 Promotes the Progression of Colon Adenocarcinoma by Regulating HSF4 Expression. *Iran J Public Health*, **52**(12), 2601-2610 (2023)
2. Bray, F., Parkin, D. M., & African Cancer Registry Network. Cancer in sub-Saharan Africa in 2020: a review of current estimates of the national burden, data gaps, and future needs. *The Lancet. Oncology*, **23**(6), 719-728 (2022)
3. Rosenberg, P. S., & Miranda-Filho, A. Advances in statistical methods for cancer surveillance research: an age-period-cohort perspective. *Frontiers in oncology*, **13**, 1332429 (2024)

4. Vieira, A. R., Abar, L., Greenwood, D., & Norat, T. Foods and beverages and colorectal cancer risk: a systematic review and meta-analysis of cohort studies, an update of the evidence of the WCRF-AICR Continuous Update Project. *Annals of oncology*, **28**(8), 1788–1802 (2017)
5. Nakagawa, K., Ishibe, Goto, K., & Endo, I. Effects of neoadjuvant chemotherapy for patients with obstructive colon cancer: A multicenter propensity score-matched analysis (YCOG2101). *Annals of gastroenterological surgery*, **8**(2), 262–272 (2023)
6. Berlin, C., Mauerer, B., H., Fichtner-Feigl, S., ... Kesselring, R. Single-cell deconvolution reveals high lineage- and location-dependent heterogeneity in mesenchymal multivisceral stage 4 colorectal cancer. *The Journal of clinical investigation*, **134**(5), e169576 (2023)
7. Harada, A., Matsumoto, S., Eguchi, H., & Kikuchi, A. Localization of KRAS downstream target ARL4C to invasive pseudopods accelerates pancreatic cancer cell invasion. *eLife*, **10**, e66721 (2021)
8. Karlsson, S., & Nyström, H. The extracellular matrix in colorectal cancer and its metastatic settling - Alterations and biological implications. *Critical reviews in oncology/hematology*, **175**, 103712 (2022)
9. Starzer, A. M., Preusser, M., & Berghoff, A. S. Immune escape mechanisms and therapeutic approaches in cancer: the cancer-immunity cycle. *Therapeutic advances in medical oncology*, **14**, 17588359221096219 (2022)
10. Gao, J., Wu, Z., Zhao, Xu, Q., Chen, H., Chen, D., & Sun, Y. Allosteric inhibition reveals SHP2-mediated tumor immunosuppression in colon cancer by single-cell transcriptomics. *Acta pharmaceutica Sinica. B*, **12**(1), 149–166 (2022)
11. Li, Y., Zhou, H., Liu, P., Lv, D., H., Zhang, X., & Ke, Y. SHP2 deneddylation mediates tumor immunosuppression in colon cancer via the CD47/SIRPα axis. *The Journal of clinical investigation*, **133**(4), e162870 (2023)
12. Estrella, V., Chen, T., Gatenby, R. A., & Gillies, R. J. Acidity generated by the tumor microenvironment drives local invasion. *Cancer research*, **73**(5), 1524–1535 (2013)
13. Castosa, R., Martinez-Iglesias, O., Roca-Lema, M., Concha, Á., & Figueroa, A. Hakai overexpression effectively induces tumour progression and metastasis in vivo. *Scientific reports*, **8**(1), 3466 (2018)
14. Nenkov, M., Ma, Y., Gaßler, N., & Chen, Y. Metabolic Reprogramming of Colorectal Cancer Cells and the Microenvironment: Implication for Therapy. *International journal of molecular sciences*, **22**(12), 6262 (2021)
15. Zhao, L., Lee, V. H. F., Ng, M. K., H., & Bijlsma, M. F. Molecular subtyping of cancer: current status and moving toward clinical applications. *Briefings in bioinformatics*, **20**(2), 572–584 (2019)
16. Chen, M., Linstra, R., & van Vugt, M. A. T. M. Genomic instability, inflammatory signaling and response to cancer immunotherapy. *Biochimica et biophysica acta. Reviews on cancer*, **1877**(1), 188661 (2022)
17. Nakazawa, M. A., M., Higashihara, K., & Okuno, Y. . Novel cancer subtyping method based on patient-specific gene regulatory network. *Scientific reports*, **11**(1), 23653 (2021)
18. Colaprico, A., Silva, T. C., Olsen, Bontempi, G., & Noushmehr, H. TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, **44**(8), e71 (2016)
19. Liberzon, A., Birger, C., Mesirov, J. P., & Tamayo, P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems*, **1**(6), 417–425 (2015)
20. Karim, M.R., Beyan, O.D., Cochez, M., & Decker, S. Deep learning-based clustering approaches for bioinformatics. *Briefings in Bioinformatics*, **22**, 393–415 (2020)
21. Ritchie, M. E., Phipson, B., C. W., Shi, W., & Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47 (2015)
22. Shen, W., Song, Z., Zhong, X., He, X., Wang, T., Li, S., & Song, X. Sangerbox: A comprehensive, interaction-friendly clinical bioinformatics analysis platform. *iMeta*, **1**(3), e36 (2022)
23. Wu, T., Hu, E., Xu, S., Chen, M., Fu, X., Liu, S., Bo, X., & Yu, G. . clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation*, **2**(3), 100141. (2021)
24. Hänzelmann, S., Castelo, R., & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, **14**, 7 (2013)
25. Zeng, D., Ye, Z., Shen, R., Yu, G., Shi, M., & Liao, W. IOBR: Multi-Omics Immuno-Oncology Biological Research to Decode Tumor Microenvironment and Signatures. *Frontiers in immunology*, **12**, 687975 (2021)
26. Thorsson, V., Gibbs, D. L., E., Gao, G. F., Mose, L. E., ... Shmulevich, I. The Immune Landscape of Cancer. *Immunity*, **48**(4), 812–830.e14 (2018)
27. Jiang, P., Gu, S., A., Wucherpfennig, K. W., & Liu, X. S. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nature medicine*, **24**(10), 1550–1558 (2018)
28. Pentiuk, N., & Motsiuk, V. Evaluation of nutritional status in patients with liver cirrhosis. Validity and prognostic value of the Patient-Generated Subjective Global Assessment. *Przegląd gastroenterologiczny*, **18**(3), 327–333 (2023)
29. Lam, C. S. N., Bharwani, A. A., Lo, C. M., & Cheung, T. T. A machine learning model for colorectal liver metastasis post-hepatectomy prognostications. *Hepatobiliary surgery and nutrition*, **12**(4), 495–506 (2023)

30. Shoeibi, A., Ghassemi, N., Khadem, A., & Rajendra Acharya, U. Automatic diagnosis of schizophrenia and attention deficit hyperactivity disorder in rs-fMRI modality using convolutional autoencoder model and interval type-2 fuzzy regression. *Cognitive neurodynamics*, **17**(6), 1501–1523 (2023)
31. Yang, Y. Y., Gao, Z. X., Liu, Z. S., & Wu, P. Identification of ULK1 as a novel mitophagy-related gene in diabetic nephropathy. *Frontiers in endocrinology*, **13**, 1079465 (2023)
32. Mao, Y., Hou, X., Fu, S., & Luan, J. Transcriptomic and machine learning analyses identify hub genes of metabolism and host immune response that are associated with the progression of breast capsular contracture. *Genes & diseases*, **11**(3), 101087 (2023)
33. Cheng, V., Oveisi, N., J. M., Murphy, R. A., & De Vera, M. A. Colorectal Cancer and Onset of Anxiety and Depression: A Systematic Review and Meta-Analysis. *Current oncology*, **29**(11), 8751–8766 (2022)
34. Mellman, I., & Yarden, Y. Endocytosis and cancer. *Cold Spring Harbor perspectives in biology*, **5**(12), a016949 (2013)
35. Banushi, B., Joseph, S. R., Lum, B., Lee, J. J., & Simpson, F. Endocytosis in cancer and cancer therapy. *Nature reviews. Cancer*, **23**(7), 450–473 (2023)
36. Chew, H. Y., De Lima, P. O., Gonzalez Cruz, F., Foote, M., ... Simpson, F. Endocytosis Inhibition in Humans to Improve Responses to ADCC-Mediating Antibodies. *Cell*, **180**(5), 895–914.e27 (2020)
37. Wills, R. C., & Hammond, G. R. V. PI(4,5)P2: signaling the plasma membrane. *The Biochemical journal*, **479**(21), 2311–2325 (2022)
38. Bang, Y. J., Hu, Z., Li, Y., Raj, P., Herz, J., & Hooper, L. V. Serum amyloid A delivers retinol to intestinal myeloid cells to promote adaptive immunity. *Science*, **373**(6561), eabf9232 (2021).
39. Doles, J., & Hemann, M. T. Nek4 status differentially alters sensitivity to distinct microtubule poisons. *Cancer research*, **70**(3), 1033–1041 (2010)
40. Ding, N. H., Zhang, L., Xiao, Z., W. H., & Sun, L. Q. NEK4 kinase regulates EMT to promote lung cancer metastasis. *Journal of cellular and molecular medicine*, **22**(12), 5877–5887 (2018)
41. Basei, F. L., de Castro Ferezin, Zorzano, A., & Kobarg, J. Nek4 regulates mitochondrial respiration and morphology. *The FEBS journal*, **289**(11), 3262–3279 (2022)
42. Roliński, M., Montaldo, N. P., Sætrom, P., & van Loon, B. Loss of Mediator complex subunit 13 (MED13) promotes resistance to alkylation through cyclin D1 upregulation. *Nucleic acids research*, **49**(3), 1470–1484 (2021)
43. KuKuuluvainen, E., Domènech-Moreno, E., Niemelä, E. H., & Mäkelä, T. P. Depletion of Mediator Kinase Module Subunits Represses Superenhancer-Associated Genes in Colon Cancer Cells. *Molecular and cellular biology*, **38**(11), e00573-17 (2018)
44. Zhang, N., Song, Y., Xu, Y., & Yang, M. MED13L integrates Mediator-regulated epigenetic control into lung cancer radiosensitivity. *Theranostics*, **10**(20), 9378–9394 (2020)
45. Fant, C. B., & Taatjes, D. J. Regulatory functions of the Mediator kinases CDK8 and CDK19. *Transcription*, **10**(2), 76–90 (2019)
46. Glorieux, C., Xia, X., & Huang, P. The Role of Oncogenes and Redox Signaling in the Regulation of PD-L1 in Cancer. *Cancers*, **13**(17), 4426 (2021)
47. Zhang, X., Liu, J., Cheng, Y., Zhu, & Cao, X. Metabolic enzyme Suc1g2 maintains tolerogenicity of regulatory dendritic cells diffDCs by suppressing Lactb succinylation. *Journal of autoimmunity*, **138**, 103048 (2023)
48. Feng, L., Sun, X., Csizmadia, S., Murakami, T., & Wu, Y. Vascular CD39/ENTPD1 directly promotes tumor cell growth by scavenging extracellular adenosine triphosphate. *Neoplasia*, **13**(3), 206–216 (2011)