

Development of a Ferroptosis-Associated Diagnostic Model for Colorectal Cancer Liver Metastasis and Virtual Screening of Potential Therapeutic Compounds

Hui Chen

Xuan Wu Hospital of the Capital Medical University

Zhenshun Wang

Tongji Hospital

Zhiyi Fan

Xuan Wu Hospital of the Capital Medical University

Zhaobo Liu

Xuan Wu Hospital of the Capital Medical University

Dongdong Lin

1dd1231@126.com

Xuan Wu Hospital of the Capital Medical University

Dongbin Liu

Xuan Wu Hospital of the Capital Medical University

Research Article

Keywords: Colorectal cancer, Liver metastasis, Ferroptosis, Diagnostic biomarker, Immune infiltration

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11041829/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

In colorectal cancer (CRC), liver metastasis is the primary factor leading to death, yet effective biomarkers for liver metastatic colorectal cancer (LMCRC) remain scarce. Ferroptosis, a regulated type of cell death spurred by iron, has been demonstrated to have a significant role in tumor progression, metabolic adjustment, and immune regulation. This study centered on discovering ferroptosis-related molecular alterations in LMCRC and establishing a relevant diagnostic framework.

Method

Genes that exhibited differential expression between primary colorectal tumors and liver metastases were identified by utilizing the GSE6988 dataset and cross-checked with ferroptosis-related genes obtained from FerrDb V3 and a published ferroptosis gene set. Functional enrichment analysis was performed employing Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Gene Set Enrichment Analysis (GSEA), and Gene Set Variation Analysis (GSVA) methodologies. Key biomarkers were discerned employing an integrative analytical framework that incorporated logistic regression, Support Vector Machine-Recursive Feature Elimination (SVM-RFE), and Least Absolute Shrinkage and Selection Operator (LASSO) regression. The diagnostic efficacy of the model was subsequently validated in the external GSE81558 dataset employing Receiver Operating Characteristic (ROC) curve and decision curve analysis. Single-sample Gene Set Enrichment Analysis (ssGSEA) was employed to quantify the levels of immune cell infiltration. Finally, computational strategies, including virtual screening and molecular docking, were applied to screen for potential compounds targeting the protein of interest.

Results

In the GSE6988 dataset, investigators identified 272 genes that had notable expression alterations. Among them, 173 genes had elevated expression, and 99 genes had reduced expression. By cross-checking with ferroptosis-related gene sets (FRGs), 28 ferroptosis-associated genes with changed expression (FRDEGs) were found. These genes were primarily associated with processes like oxidative stress reaction, iron/copper ion regulation, ferroptosis, mineral absorption, fatty acid routes, and bile acid metabolism. The amalgamation of logistic regression, SVM-RFE, and LASSO regression identified four crucial genes: WNT5A, LCN2, TRPA1, and AKR1C2. A risk-scoring model developed from these genes displayed outstanding accuracy in the initial cohort (AUC > 0.90) and sustained strong performance in external validation sets (AUC falling within the range of 0.70 to 0.90). The ssGSEA analysis revealed statistically significant disparities in 14 immune cell populations between the LMCRC and CRC cohorts. WNT5A demonstrated the strongest positive correlation with mast cells ($r = 0.60$, $p < 0.05$), while AKR1C2 exhibited the most pronounced negative correlation with gamma delta T cells ($r = -0.433$, $p < 0.05$). Molecular docking investigations proposed promising predicted binding capabilities of Z9405216322 and D481-2213 for AKR1C2.

Conclusion

This study examined the molecular characteristics related to ferroptosis in LMCRC and established a diagnostic model built on four crucial genes that hold promising clinical implications. The findings suggest potential biomarkers and treatment approaches for further experimental and clinical validation.

Introduction

Colorectal cancer (CRC) is among the most frequently diagnosed malignancies globally, occupying the third position in terms of cancer incidence^[1]. The liver acts as the main location for hematogenous metastasis of CRC, and approximately half of patients develop liver metastases as the disease advances. These metastases constitute a significant contributor to mortality in colorectal cancer^[2], markedly diminishing the 5-year survival rate upon their occurrence. Consequently, timely and accurate detection of hepatic metastases from colorectal cancer is critical for optimizing patient outcomes.

The identification of liver metastases derived from colorectal cancer predominantly relies on serum biomarkers like carcinoembryonic antigen (CEA) and CA19-9, together with imaging methods such as ultrasound, CT scans, and MRI. Nevertheless, the conventional serum markers frequently possess restricted sensitivity and specificity, which may impede their efficacy in early-stage discovery. Moreover, diagnostic methods frequently face difficulties in efficiently detecting minute metastatic foci. Additionally, early-phase liver metastases may not present symptoms, possibly resulting in delayed identification among some people. As a result, the investigation of extremely accurate and sensitive molecular markers is crucial for enhancing the early detection of liver metastases originating from colorectal cancer.

Ferroptosis is a form of regulated cell death driven by iron-dependent lipid peroxidation and compromised antioxidant defense mechanisms^[3]. Unlike apoptosis, necroptosis, or pyroptosis, ferroptosis operates through distinct pathways and has garnered substantial interest in the field of cancer research^[4]. Ferroptosis is associated with tumor suppression, metastasis progression, treatment resistance, and immune system regulation^[5]. Mechanisms associated with ferroptosis have been recognized as crucial elements influencing tumor advancement and treatment results in colorectal cancer. However, their effect on liver metastases is still not well-defined.

This study comprehensively explored the changes in LMCRC driven by ferroptosis. It made use of multiple publicly available transcriptomic data sources, along with ferroptosis-related gene datasets, pathway enrichment analysis, biomarker discovery through machine-learning methods, immune microenvironment characterization, and molecular docking simulations. First, Pathway enrichment analyses were performed to elucidate the functional

mechanisms and biological pathways associated with ferroptosis in LMCRC.. Meanwhile, machine-learning approaches were used to find potential diagnostic biomarkers and to develop and validate a diagnostic system. Concurrently, Immunoprofiling was performed to examine the interaction between ferroptosis and the tumor immune microenvironment, thereby elucidating the immunological determinants modulating LMCRC progression.. Moreover, molecular docking evaluated the possible binding interactions between the proposed compounds and the identified protein target. Overall, these integrated methods aimed to identify diagnostic biomarkers, build and evaluate predictive models, and analyze the relationship between ferroptosis-related molecular patterns and immune infiltration in LMCRC.

Materials and Methods

2.1 Data acquisition and pretreatment of colorectal liver metastases

By making use of the GEOquery R library^[6] (v2.70.0), two datasets of colorectal liver metastasis (LMCRC), GSE6988^[7] (GPL4811, having 49 primary tumor specimens and 26 liver metastasis specimens) and GSE81558^[8] (GPL15207, with 23 primary tumor specimens and 19 liver metastasis specimens), were fetched from the GEO platform^[9] (<https://www.ncbi.nlm.nih.gov/geo/>). All samples came from human primary colorectal tumors or metastatic colorectal cancer lesions in the liver. The data sets were made consistent using the `normalizeBetweenArrays` function from the `limma` package^[10]. More details can be seen in Table 1.

Genes related to ferroptosis (FRGs) were obtained from the FerrDb V3 ferroptosis database^[11] (<https://www.zhounan.org/ferrdb/v3/>). The FerrDb V3 database provides meticulously curated data regarding regulatory elements and their associations with diseases across 22 types of regulated cell death. Four gene sets—`ferrdb_driver`, `ferrdb_suppressor`, `ferrdb_marker`, and `ferrdb_unclassified`—were obtained (accessed on June 1, 2026; Tables S1–S4). Only genes related to ferroptosis (FRGs) with the annotated RCD type being ferroptosis were retained. Moreover, another previously published gene collection associated with ferroptosis (FRGs, Table S5) was retrieved through a PubMed search (<https://pubmed.ncbi.nlm.nih.gov/>) using the keyword Ferroptosis^[12]. After merging and removing duplicate entries, a combined dataset of 1680 ferroptosis-related genes was compiled (Table S6, FRGs).

2.2 Screening of differentially expressed genes related to colorectal liver metastases and ferroptosis

Based on the grouping information obtained from the GSE6988 dataset, the samples were classified into the LMCRC and CRC groups.. Differential gene expression analysis was carried out using the `limma` package, with filters of $|\log FC| > 1$ and an adjusted p-value < 0.05 after Benjamini-Hochberg correction. The results were shown via a volcano graph generated using the `ggplot2` library. Genes related to ferroptosis were cross-checked with the identified differentially expressed genes to identify those associated with ferroptosis. Their chromosomal positions and expression patterns were then depicted with a heatmap (`pheatmap`, version 1.0.12) and a circos chart (`RCircos`, version 1.2.2).

2.3 Validation and correlation analysis of differential expression

Group contrast graphs were produced to assess FRDEG expression in LMCRC and CRC samples from the GSE6988 dataset, by making use of FRDEG expression profiles.

To investigate the associations among FRDEGs, Spearman correlation analysis was utilized to assess the correlations between FRDEG expression levels within the GSE6988 dataset. We made a heatmap to show the correlation matrix, using the `pheatmap` R package (version 1.0.12). Moreover, scatter plots presenting the four gene pairs with the strongest correlations were made with `ggplot2` (Version 3.4.4).

2.4 Gene Ontology (GO) and pathway (KEGG) enrichment analysis

GO analysis^[13] is a common method used for large-scale functional enrichment studies. It covers three main areas: Biological Process (BP), Cellular Component (CC), and Molecular Function (MF). KEGG^[14] is a well-known database that's used for pathway-focused research. We did enrichment analyses for both GO and KEGG on the FRDEGs using the `clusterProfiler` R package (Version 4.10.0)^[15]. We picked terms based on a significance threshold: the p-value had to be less than 0.05, and the FDR q-value below 0.25. Then, we filtered and sorted the results by p-value in ascending order. From that, we identified the top five most significant terms for BP, CC, MF, and KEGG pathways.

2.5 Gene Set Enrichment Analysis (GSEA)

This study employed Gene Set Enrichment Analysis (GSEA)^[16]. Initially, genes in the GSE6988 dataset were arranged according to their log-fold change ($\log FC$) values obtained via differential expression assessment. Subsequently, Gene set enrichment analysis was carried out for the entire set of genes in dataset GSE6988 by using the `clusterProfiler` package in R (version 4.10.0). The configuration for GSEA had a seed of 2022, enabling each gene set to contain between 10 and 500 genes. We got the gene sets from the Molecular Signatures Database (MSigDB), specifically from the `c2.all.v2026.1.Hs.symbols` collection. To find the notable gene sets, we used thresholds: an adjusted p-value less than 0.05 and an FDR (q-value) less than 0.05, with adjustments done via the Benjamini-Hochberg method.

2.6 Gene Set Variation Analysis (GSVA)

Gene Set Variation Analysis, or GSVA for short ^[17], is a method that doesn't rely on parameters or supervision. Basically, it lets us check how pathways are enriched across different samples. So, to check if there are any specific pathways that stand out as especially active, we grabbed the h.all.v2026.1.Hs.symbols.gmt gene set from the Molecular Signatures Database, which is often called MSigDB for short ^[18]. By using the GSVA R package (Version 1.50.0), all the genes from the GSE6988 dataset were put through GSVA to compare pathway enrichment levels between the LMCRC and CRC groups. For GSVA recognized pathways as significant when p-value was less than 0.05. It picked the top 10 pathways having the largest positive log-fold change (logFC) and the bottom 10 pathways with the smallest negative logFC for the analysis.

2.7 Construction of diagnostic model for colorectal liver metastases

To figure out the key genes related to LMCRC, we used three machine learning methods. They were Logistic Regression, Support Vector Machine Recursive Feature Elimination (SVM-RFE), and Least Absolute Shrinkage and Selection Operator (LASSO) regression.

In instances where the outcome variable was a binary one, precisely differentiating between LMCRC and CRC specimens, logistic regression was utilized to assess the association between predictors and the binary outcome. Among the 28 genes associated with ferroptosis that showed differential expression, a criterion of $p < 0.05$ was used to recognize genes related to ferroptosis, and a logistic regression model was set up.

The SVM-RFE approach makes use of support vector machines to carry out a procedure for feature selection. Through the iterative removal of characteristics that exert minimal influence on classification, the most important ones are recognized. Applying the logistic algorithm with a random seed set at 2026, 28 genes were obtained^[19] and used to construct the SVM-RFE model. We figured out which genes were key based on the number of genes that gave us the best accuracy with the fewest errors.

We used the glmnet package ^[20] in R to run LASSO regression with the four genes that the SVM-RFE approach picked. The process was executed with a random seed of 2026 (set.seed(2026)) and set up with the binomial family option. LASSO regression, which is a form of linear regression different from the basic one, incorporates an L1 regularization term ($\lambda \times \sum |\beta_i|$) to restrict the magnitude of the regression coefficients. This method aids in minimizing overfitting and enhances the models generalization ability. The outcomes of the LASSO regression analysis were presented via diagnostic plots illustrating the coefficient paths of the variables.

In summary, by making use of the Key Genes recognized via the diagnostic framework, a logistic regression method was created to calculate the RiskScore inside the system. The method of computation is detailed as follows:

$$\text{RiskScore} = \sum_i \text{Coefficient}(\text{gene}_i) * \text{mRNA Expression}(\text{gene}_i)$$

2.8 Internal/External validation of the diagnostic model for colorectal liver metastases

Using the pROC^[21] package in R (version 1.18.5), the ROC curve was constructed to assess the predictive performance of the sample risk score (RiskScore), which was derived from the logistic regression model for LMCRC based on the specified formula. So, the AUC—that's the area under the ROC curve—usually falls between 0.5 and 1. The diagnostic performance gets better as the AUC nears 1. The values of the AUC spanning from 0.5 to 0.7 signified low accuracy. Meanwhile, an AUC within the range of 0.7 and 0.9 presented moderate precision, and an AUC surpassing 0.9 showed high dependability. To illustrate the interrelationships among pivotal genes, a nomogram was constructed employing the R package rms (version 6.7-1). Calibration analysis made it possible to generate a calibration curve, which evaluated the diagnostic models exactness and discrimination regarding LMCRC, according to the model results. Decision Curve Assessment (DCA)^[22] constitutes a method utilized for the evaluation of models in clinical prognosis, diagnostic testing, and biomolecular marker analysis. The ggDCA package (version 1.1) was applied to generate DCA plots employing logistic regression predictive probabilities derived from the GSE6988 dataset.

The RiskScore of samples in the GSE81558 dataset was computed using the established logistic regression framework. Subsequently, the R package pROC (version 1.18.5) was utilized to create the ROC curve, evaluating the capacity of the RiskScore—derived through the logistic regression model—to forecast the occurrence of LMCRC. Moreover, DCA graphs were generated using the R package ggDCA (version 1.1), making use of the prediction probabilities of the model in the GSE81558 dataset.

The assessment of Gene Ontology (GO) annotations via semantic comparison provides a quantitative way to measure the comparative analysis of genomic homology, acting as an essential basis for diverse bioinformatics methods. The functional correlation of Key Genes was ascertained by utilizing the R package GOSemSim^[23], where functional likeness (Friends) was utilized to assess the relationships between Key Genes.

2.9 Differential expression verification and ROC curve analysis

To assess the variation of Key Gene expression between CRC and LMCRC in the datasets GSE6988 and GSE81558, comparison graphs were generated with the Key Gene expression data.

The ROC curve for the key genes was visualized via the R package pROC (Version 1.18.5), and AUC was calculated. The AUC was employed to evaluate the diagnostic efficacy of the expression levels of crucial genes in forecasting LMCRC.

2.10 Analysis of immune infiltration in disease control group

Single-specimen gene-set enrichment analysis (ssGSEA) serves as an approach for gauging the relative amounts of diverse immune cell infiltrates. We figured out how to group the immune cells based on clear gene signatures^[24]. These signatures cover different types, like activated CD8+ T lymphocytes, activated dendritic cells, gamma-delta T lymphocytes, natural killer cells, and regulatory T lymphocytes. The enrichment scores derived from ssGSEA provided a quantitative assessment of the degree of infiltration for each immune cell type within individual specimens. By adopting this method, the immune cell infiltration pattern of the GSE6988 dataset was built. Also, we made group comparison charts using the R package ggplot2 (v3.4.4) to show the differences in immune cell expression between CRC and LMCRC samples in the GSE6988 dataset. Then, we looked at the correlations between immune cells with the Spearman method, and we used a heatmap made with the R package pheatmap (v1.0.12) to display how these immune cells relate to each other. Next, we checked the association between key genes and immune cell groups using the Spearman approach. Finally, we created a correlation bubble graph with the ggplot2 package (Version 3.4.4) in R to show the connection between crucial genes and immune cells.

2.11 PPI network screening

This study employed the STRING platform, focusing on core genes and utilizing a minimum interaction threshold of 0.15 as the benchmark. So, we built a protein-protein interaction (PPI) network using these key genes. You see, in the network, there are clusters where nodes are closely linked. These clusters might stand for molecular complexes with specific biological functions. Then, we selected the genes that showed interactions with other genes for further analysis.

2.12 Virtual screening (DrugCLIP)

To recognize lead compounds having high affinity for essential genes and notable therapeutic potential, the DrugCLIP tool^[25] was utilized for virtual protein analysis. At first, the protein structure was obtained from the PDB repository and refined into single protein crystals. CB-Dock2 helped in the identification of the docking site. The largest pocket volume at the C1 site was selected as the target region, recording its central coordinates (Center x, y, z) and pocket radius parameters. The altered protein structure along with pocket details was then submitted to DrugCLIP for comprehensive virtual screening. The outcomes were arranged in descending sequence according to the DrugCLIP score, and the compounds with the highest scores were recognized as potential candidates. Vina scores higher than -5 kcal/mol were regarded as indicating minimal or unimportant binding predictions; scores ranging from -7 to -5 kcal/mol indicated moderate binding predictions; scores lower than -7 kcal/mol reflected strong binding predictions.

2.13 Statistical analysis

In this study, We used R software, version 4.3.3, to handle the data processing and analysis. Now, when comparing continuous variables between two groups, we checked for statistical significance with the independent Student's t-test, but only if the variables were normally distributed. Otherwise, we'd go with a different method, unless specified. For variables that were not normally distributed, comparisons were conducted using the Wilcoxon rank-sum test (Mann-Whitney U test). Correlations among various molecules were analyzed via Spearman correlation analysis. Unless otherwise specified, all reported p-values are two-tailed, with statistical significance defined as $p < 0.05$.

Results

3.1 Technology Roadmap

The schematic representation of the study workflow is illustrated in Fig. 1. The GSE6988 dataset served as the discovery cohort for identifying differentially expressed genes (DEGs) between primary colorectal cancer (CRC) and liver metastatic colorectal cancer (LMCRC) samples. The intersection of DEGs with ferroptosis-related genes (FRGs) yielded ferroptosis-related differentially expressed genes (FRDEGs). Functional enrichment analyses, including Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, Gene Set Enrichment Analysis (GSEA), and Gene Set Variation Analysis (GSVA), were subsequently conducted to characterize the biological functions and pathways associated with FRDEGs. Logistic regression, support vector machine-recursive feature elimination (SVM-RFE), and least absolute shrinkage and selection operator (LASSO) regression were then integrated to identify key genes and construct a diagnostic model. The diagnostic performance of the model was evaluated in the GSE6988 cohort and externally validated in the GSE81558 cohort. Furthermore, immune infiltration analysis, protein-protein interaction (PPI) network analysis, virtual screening, and molecular docking were performed to investigate the biological relevance and potential therapeutic implications of the identified key genes.

LMCRC, hepatic metastatic colorectal cancer; DEGs, genes showing significant expression disparities; FRGs, genes related to ferroptosis; FRDEGs, differentially expressed genes associated with ferroptosis; GSEA, gene set enrichment assessment; GSVA, gene set variation analysis; GO, gene ontology; KEGG, genomic pathways mapped by Kyoto Encyclopedia; PPI, protein-interaction network.

3.2 Screening of differentially expressed genes related to colorectal liver metastases and ferroptosis

At first, we carried out normalization for the datasets GSE6988 and GSE81558 of LMCRC. Subsequently, we assessed the expression levels of the datasets before and after normalization via box plot visualization (Figure S1). The plots showed that normalization led to a more uniform distribution of expression

values among the samples.

Specimens from the GSE6988 dataset were classified into LMCRC and CRC cohorts. Conducting differential expression analysis using limma identified 272 DEGs having $|\log FC| > 1$ and an adjusted p-value < 0.05 , which included 173 genes with elevated expression and 99 with reduced expression. The results are presented via a volcano plot (Fig.2A).

We wanted to find the genes that are differentially expressed and linked to ferroptosis, called FRDEGs. So, we checked the overlap between DEGs and the ferroptosis-related gene collections, or FRGs. We figured out what they had in common and showed it using a Venn diagram, which is Figure 2B.. Altogether, 28 ferroptosis-related genes showing significant differential expression (FRDEGs) were recognized. Based on the intersection analysis results, the alterations in FRDEG expression levels across distinct sample groups in dataset GSE6988 were examined. We used the R package pheatmap to make heatmaps that clearly show the analysis results for the top 10 up-regulated and down-regulated ferroptosis-related genes (FRDEGs), which you can see in Fig.2C. In the end, the RCircos R package enabled the exploration of the chromosomal dispersion of FRDEGs and the creation of a visualization chart showing their locations on human chromosomes (Fig.2D). The genetic mapping demonstrated the distribution of ferroptosis-related variably expressed genes on human chromosomes.

A. Volcano plot for differential gene expression analysis comparing LMCRC and CRC from dataset GSE6988. B. Venn diagram depicting the intersection between genes showing differential expression (DEGs) and those linked to ferroptosis (FRGs) in the GSE6988 dataset. C. Visual display of different gene expression profiles linked to ferroptosis-related DEGs in dataset GSE6988. D. Genetic mapping of genes with variable expression that are associated with ferroptosis (FRDEGs); DEGs refer to Genes that have Differential Expression; FRGs denote Genes that are Associated with Ferroptosis; FRDEGs represent Differentially Expressed Genes that are Linked to Ferroptosis. The LMCRC cohort is shown in yellow, whereas the CRC cohort is presented in purple. In the heat-map, sections marked in red suggest increased levels of expression, while zones colored in blue represent decreased expression.

3.3 Differential expression verification and correlation analysis

To explore the alterations in the expression of 28 FRDEGs between LMCRC and CRC in the GSE6988 dataset, we generated comparative visual representations (Fig.3A). MT1G expression exhibited a significant difference among the groups ($p < 0.01$). On the contrary, the other 27 FRDEGs showed much larger differences between the LMCRC and CRC groups (all $p < 0.001$).

Subsequently, we evaluated the interrelationships among 28 genes with differential expression associated with ferroptosis and presented the outcomes via a heat map to show pairwise correlations (Fig.3B). The results showed us that most of these FRDEGs were positively connected. Look, among them, MT1L and MT1X were really strongly linked (with a correlation $r = 0.968$). Then, we used the R package ggplot2 to make a picture showing how these genes correlate, and we highlighted the four pairs of genes that had the strongest connections (you can see them in Fig. 3C to 3F).

A. Comparative examination of ferroptosis-related genes showing differential expression (FRDEGs) between LMCRC and CRC in the GSE6988 dataset. B. Correlation heat-map of genes related to ferroptosis and differentially expressed (FRDEGs), with a comparison between LMCRC and CRC in dataset GSE6988. C-F. Visual representation of the connection between the four most significant associations among ferroptosis-related differentially expressed genes (FRDEGs) via a scatter plot. LMCRC represents Liver Metastases originating from Colorectal Cancer; FRDEGs stand for Differentially Expressed Genes linked to Ferroptosis. ** represents $p < 0.01$, which means a significant statistical importance; *** stands for $p < 0.001$, showing an even greater degree of statistical significance. An r-value (correlation coefficient) less than 0.3 implies a very small or almost non-existent relationship, whereas values between 0.3 and 0.5. A feeble relationship was lower than 0.5, a mediocre connection was within 0.5 to 0.8, and a strong association went beyond 0.8. In the comparison group chart, LMCRC is depicted in yellow, whereas CRC is presented in purple. In the association heat-diagram, red shows a positive correlation, while blue shows a negative one. The color's intensity reflects how strong the association is.

3.4 Gene ontology (GO) and pathway (KEGG) enrichment analysis

We did GO and KEGG pathway enrichment assessments to find out how the biological processes (BP), cellular components (CC), molecular functions (MF), and KEGG pathways of these 28 FRDEGs are connected to LMCRC. Then, these 28 FRDEGs were put through GO and KEGG pathway analysis. The full details of what we found are in Table 2. The GO enrichment findings suggested that these FRDEGs were related to BP categories like iron ion transportation, copper ion elimination of toxicity, and reactions to metal ion stress. CC categories included the endocytic vesicle cavity, collagen-containing extracellular matrix, specific granule cavity, endoplasmic reticulum cavity, and endocytic vesicles. MF categories included things like bile acid binding, iron ion binding, oxidoreductase activity, and protein serine/threonine kinase activator activity. Moreover, these genes were enriched in KEGG pathways such as mineral uptake, chemical carcinogenesis with DNA adducts, ferroptosis, fatty acid breakdown, fatty acid metabolism, and other biological pathways. We showed the GO and KEGG pathway analysis results using bubble plots, which you can see in Fig.4A. At the same time, based on those enrichment results, we also made an interaction map. This map covers the biological processes (BP), cellular components (CC), molecular functions (MF), and the KEGG pathways (Fig.4 B–E). In these figures, the relevant molecules are shown right next to their category labels. Also, the size of each dot (or node) tells you how many molecules are in that particular category.

A. A bubble graph showing results from gene ontology (GO) and pathway (KEGG) enrichment assessments of ferroptosis-related differentially expressed genes (FRDEGs): including biological processes (BP), cellular components (CC), molecular functions (MF), and KEGG pathway estimations. The horizontal axis shows GO categories and KEGG pathways. B-E. The network visual display of gene ontology (GO) and pathway (KEGG) enrichment assessment for ferroptosis-related differentially-expressed genes (FRDEGs) shows biological processes (BP, B), cellular components (CC, C), molecular functions (MF, D), and KEGG pathways (E). The tan-colored circles represent entries, and the connections show the associations among entries and molecules. The intensity of red within the molecular nodes represents higher LogFC values, showing up-regulation. In contrast, greater blue tinting indicates lower LogFC values, suggesting

down-regulation. FRDEGs stands for genes that show differential expression linked to ferroptosis. GO is Gene Ontology concepts, KEGG is Kyoto Encyclopedia of Genes and Genomes paths, BP is biological activities, CC is cellular constituents, and MF is molecular functions. In the bubble graph, the bubble's size tells you the gene number, and its color shows the p-value. So, a darker red means a smaller p-value, and a deeper blue means a larger p-value. For checking GO and KEGG enrichment, the standards were a p-value under 0.05 and a q-value below 0.25.

3.5 Gene Set Enrichment Analysis (GSEA)

To investigate pathway-specific changes associated with LMCRC, genes from GSE6988 were arranged according to their logFC values that indicate differential expression and examined via GSEA. So, in Fig. 5A, the mountain plot shows how affected cellular structures and molecular functions are linked, and you can find the specific results in Table 3. GSEA showed remarkable enrichment for REACTOME SYNTHESIS OF BILE ACIDS AND BILE SALTS VIA 7ALPHA HYDROXYCHOLESTEROL (Fig. 5B), WP ETHANOL EFFECTS ON HISTONE MODIFICATIONS (Fig. 5C), VILLANUEVA LIVER CANCER KRT19 DN (Fig. 5D), along with KEGG pathways associated with GLYCINE, SERINE, and THREONINE processing (Fig. 5E), together with extra genetic clusters.

A. GSEA mountain plot presenting four biological processes obtained from dataset GSE6988. B-E. The findings of GSEA indicated that the GSE6988 dataset showed significant enrichment in REACTOME SYNTHESIS OF BILE ACIDS AND BILE SALTS VIA 7ALPHA HYDROXYCHOLESTEROL (B), WP ETHANOL EFFECTS ON HISTONE MODIFICATIONS (C), VILLANUEVA LIVER CANCER KRT19 DN (D) and KEGG GLYCINE, SERINE, AND THREONINE PATHWAY (E); GSEA, Enrichment Analysis of Gene Sets. Throughout the mountain visualization, the color denotes the p-value. A darker red implies a smaller p-value, and a more intense blue suggests a larger p-value. The criteria for selecting gene sets in the gene set enrichment analysis (GSEA) were a p-value smaller than 0.05 and a false discovery rate (q-value) below 0.05. The Benjamini-Hochberg method was employed for p-value adjustment.

3.6 Gene Set Variation Analysis (GSVA)

GSVA using the h.all.v2026.1.Hs.symbols.gmt gene set was carried out to assess disparities in pathway enrichment profiles between LMCRC and CRC specimens in the GSE6988 dataset. The comprehensive results are presented in Table 4. Subsequently, a heat map was generated to visualize the outcomes (Fig. 6A). So, in the end, we found the top 10 pathways that were positively enriched, which met the criteria of $p < 0.05$ and showed logFC. We also identified the top 10 negatively enriched pathways. Then, we looked at how these 20 pathways were expressed differently between the LMCRC and CRC groups, and we showed this using a two-way bar chart (Fig. 6B).

A-B. Heatmap (A) and two-direction bar graph (B) showing gene set variation analysis (GSVA) results when comparing LMCRC with CRC from the GSE6988 dataset. LMCRC denotes Liver Metastases from Colorectal Cancer, and GSVA represents Gene Set Variation Evaluation. The LMCRC cohort is presented in yellow, whereas the CRC cohort is displayed in purple. The criteria for selecting GSVA involved a p-value under 0.05 and logFC, emphasizing the top ten pathways with the most positive enrichment and the top ten pathways with the most negative enrichment. In the thermal map, the color blue signifies minimal enrichment, while red indicates substantial enrichment.

3.7 Construction of diagnostic model for colorectal liver metastases

The genes that are differentially expressed in relation to ferroptosis (FRDEGs) were employed to construct the logistic regression model. The analysis showed that all 28 genes were statistically significant in the model ($p < 0.05$), so we picked these genes to look into further.

The 28 genes found via logistic regression analysis were further treated using SVM-RFE along with five-fold cross-validation. The rankings of genes were averaged, and the group of genes linked to the minimum error rate was ascertained (Fig. 7A) in conjunction with the highest accuracy ratio (Fig. 7B). The configuration of the model was attained. So, the results showed that the SVM-RFE method got the best precision when four genes were selected. These genes, recognized as ferroptosis-related differentially expressed genes (FRDEGs), were WNT5A, LCN2, TRPA1, and AKR1C2.

Within the LASSO framework, the lowest lambda value was identified as optimal for the four genes selected through the SVM-RFE approach. These four FRDEGs, which were flagged as crucial genes, consisted of WNT5A, LCN2, TRPA1, and AKR1C2. The relevant visualization was accomplished by generating a diagram depicting the LASSO regression model (Fig. 7C) in conjunction with the LASSO variable path visualization (Fig. 7D).

A-B The quantity of genes having minimal error rates (A) and those attaining optimal accuracy (B) recognized by the SVM approach are shown. Graphical presentation of C-D diagnostic model (C) and variable path chart (D) for the LASSO regression analysis. LMCRC stands for Liver Metastatic Colorectal Cancer; FRDEGs refer to Ferroptosis-Associated Differentially Expressed Genes; LASSO represents Least Absolute Shrinkage and Selection Technique; SVM-RFE means Recursive Feature Elimination using Support Vector Machine.

3.8 Validation of diagnostic model for colorectal liver metastases

Subsequently, the Receiver Operating Characteristic (ROC) curve was constructed utilizing the R package pROC, based on the predicted probabilities obtained from the logistic regression model applied to the GSE6988 dataset. The way of calculating to determine RiskScore is described as follows:

$$\text{RiskScore} = -37.2107 \cdot \text{WNT5A} - 7.8651 \cdot \text{LCN2} - 19.6854 \cdot \text{TRPA1} + 15.222 \cdot \text{AKR1C2}$$

The ROC analysis showed that the logistic regression model presented robust differentiation between LMCRC and primary CRC specimens in GSE6988 (AUC > 0.90; Fig.8A). Subsequently, To check how accurate and good the diagnostic framework is at telling LMCRC apart, we made a calibration curve using calibration analysis. We checked how well the model predicts by looking at how closely the actual chances match the predicted ones in different situations, as shown in Figure 8B. So, the calibration curve for the LMCRC diagnostic model showed that the dotted line was really close to the diagonal line of the best model. Then, we used DCA to check how useful the model is in practice, using the logistic regression prediction probabilities from the GSE6988 dataset (you can see it in Fig.8C). The DCA plot demonstrated that the model exhibited a greater net benefit relative to both the treat-all and treat-none strategies across varying threshold probabilities. To further verify the effectiveness of the diagnostic model for LMCRC, a Nomogram was built to depict the association of key genes in the GSE6988 dataset (Fig.8D). The research findings indicated that the expression of WNT5A in the Key Genes showed significantly higher efficacy than other factors in the diagnostic system for LMCRC. On the contrary, the expression of LCN2 displayed markedly diminished diagnostic value compared to other variables in the model.

Afterward, the analysis scores of functional similarity (Friends) were employed to pinpoint crucial genes associated with the biological mechanisms of LMCRC (Fig.8E). The results showed that WNT5A was a key player in how LMCRC progresses..

The R package pROC to generate ROC curves, based on the predicted probabilities from logistic regression analysis in the GSE81558 dataset.

The depicted ROC curve (Fig.8F) indicated that the logistic regression algorithm showed a moderately precise diagnostic capacity (AUC ranging from 0.7 to 0.9) when differentiating LMCRC samples from control (CRC) samples. The forecasting ability of the model within the GSE81558 dataset was evaluated, with emphasis on its clinical significance in the diagnosis of LMCRC. DCA was employed to appraise its practical use, and the results were presented (Fig.8G). The results showed that the curve of the model constantly went beyond those of the completely positive and completely negative groups in a certain range, which illustrated the models better net benefits and enhanced performance.

A. ROC curves were obtained from the logistic regression analysis of LMCRC specimens and CRC specimens. Calibration Curve plot (B) and Decision Curve Evaluation (DCA) diagram (C) showing the predictive model for LMCRC. D. Diagram depicting the key genes from dataset GSE6988 within the predictive model for detecting LMCRC. E. Visual depiction of functional similarity (Friends) analysis results for notable genes. F. Receiver operating characteristic curves for the logistic regression model within samples of LMCRC and CRC. G. DCA chart showing the predictive probability for diagnosing LMCRC via the diagnostic framework obtained from the model in dataset GSE81558. In a graph of DCA, the vertical axis shows the net benefit, while the horizontal axis stands for the threshold probability or probability threshold. DCA, Decision Curve Evaluation; ROC, Receiver Operating Features; AUC, Area Measurement of the Curve.

3.9 Differential expression validation and ROC curve analysis

To investigate the differential expression levels of the four pivotal genes between LMCRC and CRC cases within the GSE6988 dataset, comparative analyses were conducted (Fig.9A). The results demonstrate that the expression levels of AKR1C2, LCN2, TRPA1, and WNT5A were markedly elevated, achieving high statistical significance ($p < 0.001$).

The pROC library in the R programming language was utilized to generate the ROC curve based on the expression profiles of the key genes within the GSE6988 dataset. The ROC curve (Fig.9C), demonstrated that the expression levels of AKR1C2 and LCN2 exhibited moderate diagnostic efficacy for distinguishing between LMCRC and CRC, with AUC values ranging between 0.7 and 0.9. On the contrary, TRPA1 and WNT5A showed higher diagnostic accuracy (AUC > 0.9) while differentiating between LMCRC and primary CRC lesions.

In a comparable way, to explore the different expression levels of four key genes between LMCRC and CRC by using dataset GSE81558, we display the results via a comparative-group plot (Fig.9B), and the results suggest the following: AKR1C2 showed a notably significant expression ($p < 0.05$), whereas WNT5A presented an extremely significant expression level ($p < 0.001$).

The ROC curve was built by utilizing the pROC library within R, making use of the expression data of the significantly changed key genes discovered in the GSE81558 dataset (Fig.9D) showed that the levels of AKR1C2 and WNT5A displayed remarkable accuracy in differentiating LMCRC from CRC ($0.7 < \text{AUC} < 0.9$).

A. Examination of notable gene disparities between the LMCRC group and CRC group in dataset GSE6988. B. Comparison of key genes between the LMCRC cohort and the CRC cohort in dataset GSE81558. C. Receiver-operating-characteristic plots were generated for key genes (AKR1C2, LCN2, TRPA1, WNT5A) that showed notable differential expression among groups in dataset GSE6988. D. Receiver operating characteristic curves for the critical genes AKR1C2 and WNT5A display remarkable disparities in expression amounts among different groups in dataset GSE81558. Metastatic liver tumors originating from colorectal cancer (MLTCC). ns suggests that p is greater than or equal to 0.05, which means there is no statistical significance; * means p is less than 0.05, showing that there is statistical validity; ** emphasizes that p is less than 0.01, indicating a strong statistical connection; *** represents p is less than 0.001, reflecting extremely high statistical validity. In the comparison-group graph, yellow stands for LMCRC and purple represents CRC.

3.10 Analysis of immune infiltration in disease control group

Employing the GSE6988 dataset, the ssGSEA algorithm was employed to quantify the infiltration levels of 28 distinct immune cell subtypes. So, first off, we used comparison plots to see how immune cell infiltration levels differed between the LMCRC group and other subgroups in the CRC population. The group comparison graph (Fig.10A) showed statistically significant variances ($p < 0.05$) among all 14 types of immune cells. Then, we used a heatmap to show how the infiltration levels of these 14 immune cell types are linked in the immune infiltration analysis of samples from the GSE6988 dataset (Fig. 10B). We found that most of the immune cells were strongly positively linked. In particular, Regulatory T cells and Type 1 T helper cells had the strongest positive link ($r =$

0.677, $p < 0.05$). Ultimately, a bubble chart was employed to illustrate the association between core genes and the degree of immune cell infiltration (Fig. 10C). The correlation bubble plot analysis demonstrated that the majority of genes exhibited significant associations with immune cells. Among these, WNT5A presented the most prominent positive link with Mast cells ($r = 0.6$, $p < 0.05$). On the contrary, AKR1C2 showed the most robust negative correlation with Gamma delta T cells ($r = -0.433$, $p < 0.05$).

A. Examination of immune cell distribution in dataset GSE6988 between the LMCRC group and the CRC group (A). B. Visualization of immune cell correlation matrix from dataset GSE6988. C. Scatterplot shows the association between immune cell infiltration degrees and key gene expression in dataset GSE6988. ns stands for $p \geq 0.05$, which means there is no statistical significance; * indicates $p < 0.05$, showing statistical relevance; ** represents $p < 0.01$, meaning significant statistical importance; *** emphasizes $p < 0.001$, indicating extraordinary statistical relevance. The absolute value (r) of the correlation coefficient being below 0.3 implies little to no correlation. When it lies between 0.3 and 0.5, it shows a slight relationship. An r -value from 0.5 to 0.8 represents a moderate association, and when the values surpass 0.8, it means a strong connection. A red hue indicates a positive connection, while blue represents a negative association. The strength of the hue shows the size of the relationship.

3.11 Virtual Screening (DrugCLIP)

At the beginning, an examination of the interactions among proteins encoded by Key Genes was carried out. By making use of the STRING database, a network graph showing these PPI network for the Key Genes was created (Fig. 11A). The PPI network examination showed the linkages among the four genes, and the specific data are presented in Table S7.

To pinpoint potential lead compounds that possess high binding capacity for AKR1C2 and have therapeutic value, the DrugCLIP software was utilized to carry out virtual screening focused on the AKR1C2 protein (PDB ID: 4XO6, Fig. 11B). DrugCLIP identified the top 100 molecules that showed high affinity potential for the AKR1C2 gene (Table S8). The rankings were arranged according to DrugCLIP scores, which reflect how well a small-molecule can interact with its target protein, where higher scores signify better binding ability. The detailed data of the binding pocket for the target is available in Table S9. The top candidate molecules for AKR1C2 determined by Vina Score and DrugCLIP score are Z9405216322 and D481-2213 respectively.

The interaction examination of AKR1C2 and Z9405216322 (Fig. 11C) showed considerable affinity (Vina Score = -10.38 kcal/mol). Moreover, AKR1C2 demonstrated remarkable binding to D481-2213 (Fig. 11D) (Vina Score = -8.99 kcal/mol).

A. Interaction network of crucial genes established through protein link-ups analyzed by means of the STRING platform. B. Visual display of the AKR1C2 protein structure. C-D. Presentation of the docking results between AKR1C2 and Z9405216322 (C) together with D481-2213 (D). From left-hand side to right-hand side, the overall docking graph is followed by the interaction force illustration. Faint blue broken lines stand for hydrogen bonding; light blue-dotted lines point to weak hydrogen bonds; grey dashed lines display hydrophobic interactions; turquoise dashed lines mark halogen bonds; gold dashed lines indicate ionic connections; peach dashed lines depict cation- π interactions; lime-green dashed lines denote π - π stacking.

Discussion

Despite advancements in surgical methods, pharmaceutical treatments, and novel therapies such as immunotherapy and targeted strategies, the prognosis for patients with LMCRC remains unfavorable. This is mainly due to the complex characteristics and substantial variability of metastatic neoplasms. Therefore, a thorough examination of the molecular features associated with LMCRC may assist in uncovering biomarkers and formulating new diagnostic and treatment regimens.

This study integrated transcriptomic analysis, annotation of genes linked to ferroptosis, enrichment analyses, feature selection using machine learning, assessment of immune infiltration, examination of PPI networks, and molecular docking to explore the molecular traits related to ferroptosis in LMCRC. From the dataset GSE6988, 28 differentially expressed genes associated with ferroptosis (FRDEGs) were identified, suggesting distinct transcriptional alterations related to ferroptosis in LMCRC when compared to primary CRC specimens. The detected dysregulation of FRDEGs in LMCRC might imply changes in iron regulation, oxidative stress handling, and lipid metabolism within the liver's metastatic milieu.

Functional enrichment research indicated that the recognized FRDEGs primarily take part in activities like iron ion conveyance, copper ion detoxifying, reactions to metal ion stressors, bile acid interactions, oxidoreductase functions, and metabolic routes related to ferroptosis. These findings are in line with biological anticipations, since the liver is a highly specialized organ that actively participates in bile acid generation, regulation of iron amounts, and lipid processing. Consequently, tumor cells adapting to the liver's micro-environment encounter distinctive pressures that might benefit clones able to withstand ferroptosis or display metabolic adaptability. Pathways associated with mineral uptake, chemical-induced carcinogenesis (DNA adduct formation), fatty acid degradation, and lipid metabolism were enriched, mirroring potential disruptions in metal ion equilibrium, oxidative stress mechanisms, and metabolic trends in LMCRC. Additionally, GSEA and GSVA emphasized different pathway enrichments associated with bile acid metabolism, oxidative stress reactions, hypoxia, inflammatory signaling, and PI3K-AKT-mTOR activity, suggesting possible links between metabolic and stress-related mechanisms and the LMCRC phenotype.

Among the four core genes, WNT5A appears to connect ferroptosis-related changes with cancer progression and microenvironmental modification. In our dataset, WNT5A gradually became a key marker, and it was linked to immune cell infiltration. This suggests that when WNT5A is expressed in an abnormal way, it might be connected to both tumor-specific processes and what happens in the microenvironment. As an important ligand in the non-canonical Wnt signaling pathway, WNT5A has been related to cell movement, tissue invasion, and stromal interplays in different malignancies^[26, 27]. Regarding colorectal cancer, research has demonstrated that WNT5A promotes a more aggressive state, consistent with features of metastatic development^[28]. Overall, our

findings suggest that WNT5A may affect the metastatic traits of LMCRC and also play a part in immune regulation within the metastatic area. However, considering the correlative aspect of this analysis, additional experimental studies are required to determine whether WNT5A directly affects ferroptosis sensitivity or mainly acts as an indicator of invasive phenotypic changes.

LCN2 appeared as a crucial gene in the ultimate framework, possibly suggesting disturbances in iron regulation and stress tolerance within LMCRC. Results from this study highlighted processes related to iron ion transport, metal-associated stress reactions, and oxidoreductase functions, which are consistent with the biological functions of LCN2. Also known as neutrophil gelatinase-associated lipocalin, this gene is capable of interacting with siderophores to regulate iron metabolism in cells, influencing oxidative stress resistance and adaptability to difficult environments^[29, 30]. LCN2 has been related to tumor growth, the course of epithelial-mesenchymal transition, and inflammation-induced remodeling, especially in instances of colorectal cancer^[31, 32]. The outcomes of our study suggest that alterations in LCN2 levels may be associated with adaptations in iron metabolism, which are conducive to the growth of metastases and strengthen the defense against oxidative harm. Considering that ferroptosis is associated with iron-mediated lipid peroxidation, it is scientifically justifiable to include LCN2 in our signature, which is in line with the potential for dysregulated iron control in LMCRC.

TRPA1 may act as a distinct axis for sensing stress in LMCRC. This ion channel, classified as a polymodal transient receptor potential type, responds to electrophiles, reactive oxygen substances, and pro-inflammatory stimuli^[33, 34]. Our results recognized TRPA1 as a vital predictive gene, implying possible connections between oxidative stress detection and calcium-related signaling within LMCRC. In support of this idea, our enrichment analysis showed participation in ferroptosis, mineral absorption, and pathways related to redox reactions. TRPA1 has mostly been investigated in the fields of sensory and inflammatory biology instead of cancer studies; nevertheless, its known function in redox-sensitive pathways implies that it might have a part in handling cellular responses to oxidative harm^[34]. Thus, TRPA1 may have a function in the way LMCRC cells detect and react to stress in their nearby environment. At present, the data suggest a predictive and correlational function rather than a direct causal link, and additional mechanistic verification is needed to elucidate TRPA1's role in ferroptosis regulation.

AKR1C2 is notably prominent. It did not merely act as a crucial part of the diagnostic structure but also emerged as a potential treatment-oriented target in our molecular docking assessment. AKR1C2 belongs to the aldo-keto reductase family. It participates in steroid metabolism, detoxifies reactive carbonyl substances, and controls redox balance^[35, 36]. These procedures are tightly linked to ferroptosis, which entails lipid peroxidation and the accumulation of oxidative stress. In our dataset, AKR1C2 demonstrated notable variation in its expression and stayed essential to the classification system, suggesting a potential link between AKR1C2-related redox regulation and the LMCRC profile. Earlier studies have correlated members of the AKR family with cancer progression, therapy resistance, and metabolic alterations in different kinds of tumors^[35, 36]. Moreover, molecular docking simulations demonstrated favorable interactions between AKR1C2 and the compounds under test, corroborating the viability of AKR1C2 as a drug target worthy of more in-depth research.

The combined assessment of WNT5A, LCN2, TRPA1, and AKR1C2 implies that a ferroptosis-related multi-gene profile might offer higher diagnostic significance for the identification of LMCRC when compared to depending on a solitary marker. This corresponds to the complex characteristics of metastatic colorectal cancer, in which pathways related to signaling, metabolic activities, oxidative stress control, and immune system mechanisms are closely interlinked. Multi-gene constructs often outperform single biomarkers in terms of effectiveness since they cover varied but complementary aspects of tumor biology^[37, 38]. In our study, external verification provided initial proof of the models replicability; nevertheless, larger-scale multicenter cohorts and additional specimen kinds are required for prospective validation before clinical application. Overall, these four genes seem to stand for unique ferroptosis-related molecular characteristics relevant to differentiating LMCRC from primary CRC. Our exploration of immune infiltration demonstrates that the ferroptosis-associated signature is closely associated with the tumor immune micro-environment. The genes WNT5A and AKR1C2 showed associations with significant changes in immune cell numbers, suggesting that ferroptosis-related transcriptional variations could regulate immune recruitment or indicate immune-state regulation within metastatic sites. It is being more and more recognized as a type of cell death that has a notable immunological effect, affecting antigen spread, cytokine routes, and the activity of the immune system^[39, 40]. In the situation of colorectal cancer, immune escape and immune suppression have crucial functions in tumor advancement and the response to therapy^[41]. Therefore, our findings indicate that genes associated with ferroptosis are essential parts of a complex regulatory web that connects oxidative harm, immune regulation, and metastatic growth. This also implies that controlling ferroptosis-related pathways might potentially enhance the efficacy of immunotherapy in LMCRC, although this idea needs experimental verification.

Moreover, network examination of PPI and computational selection gave initial understandings to promoting drug exploration. The PPI research uncovered potential interaction connections among the four crucial genes by utilizing data from the STRING database. Virtual screening using DrugCLIP singled out AKR1C2 as a viable target, where molecular docking outcomes indicated strong binding capabilities between AKR1C2 and compounds Z9405216322 and D481-2213^[42]. Although these results are in the preliminary stage, they present promising lead molecules for further experimental assessment. Nevertheless, the findings from molecular docking ought to be interpreted carefully, since biochemical binding and functional impacts must be validated through cellular and in-vivo investigations.

This study reveals several limitations. Primarily, the assessment was based on available transcriptomic data sets, where the external validation cohort had a limited quantity of samples. Secondly, the findings were mainly derived from computational approaches, lacking experimental verification at the transcript, protein, and functional levels. Thirdly, differences among data sets might have influenced the dependability of the results. Fourthly, the analyses regarding immune infiltration and molecular interactions are hypothesis-led and cannot replace experimental studies. Future research ought to strive to confirm these findings using larger multi-center groups and forward-looking clinical specimens. Single-cell and spatially resolved transcriptomic investigations, along with experimental verification, can offer more profound understandings of the functional implications and medical importance of these ferroptosis-related genes in LMCRC.

Conclusion

This study explored the molecular characteristics related to ferroptosis in LMCRC and identified 28 FRDEGs. By using machine learning methods, we constructed a diagnostic model relying on four genes: WNT5A, LCN2, TRPA1, and AKR1C2. The framework demonstrated robust diagnostic abilities in both the discovery stage and external validation cohorts, highlighting its potential use in clinical environments. Examination of enrichment and immune cell invasion showed that transcriptional changes related to LMCRC are associated with pathways related to metal ion control, ferroptosis and lipid mechanisms, bile acid kinetics, and altered immune invasion. Additionally, computational screening and molecular docking identified AKR1C2 as a prospective target that warrants further experimental exploration. In general, These findings provide preliminary computational perspectives on ferroptosis-related molecular alterations in LMCRC and emphasize potential diagnostic biomarkers as well as a prospective therapeutic target that needs further experimental and clinical verification.

Abbreviations

AKR, aldo-keto reductase
AKR1C2, aldo-keto reductase family 1 member C2
AUC, area under the receiver operating characteristic curve
BP, biological process
CC, cellular component
CEA, carcinoembryonic antigen
CRC, colorectal cancer
CT, computed tomography
DCA, decision curve analysis
DEGs, differentially expressed genes
FDR, false discovery rate
FRDEGs, ferroptosis-related differentially expressed genes
FRGs, ferroptosis-related genes
GEO, Gene Expression Omnibus
GO, Gene Ontology
GSEA, Gene Set Enrichment Analysis
GSVA, Gene Set Variation Analysis
KEGG, Kyoto Encyclopedia of Genes and Genomes
LASSO, least absolute shrinkage and selection operator
LCN2, lipocalin 2
LMCRC, liver metastatic colorectal cancer
logFC, log fold change
MF, molecular function
MRI, magnetic resonance imaging
MSigDB, Molecular Signatures Database
PDB, Protein Data Bank
PPI, protein-protein interaction
RCD, regulated cell death
ROC, receiver operating characteristic
SVM-RFE, support vector machine-recursive feature elimination
ssGSEA, single-sample Gene Set Enrichment Analysis

TRPA1, transient receptor potential cation channel subfamily A member 1
WNT5A, Wnt family member 5A

Declarations

Declaration of competing interest

The authors declare that they have no competing interests.

Ethics approval and consent to participate

This study used publicly available, de-identified datasets and did not involve the collection of new human or animal specimens. Therefore, ethics approval and consent to participate were not required.

Consent for publication

The study is based on data from public databases and therefore does not require informed consent from patients.

Funding

This work was supported by the S&T Program of Xiongan New Area (XA202401102001K) and the S&T Program of Hebei (244X7704D).

Author Contribution

H.C.and ZSW: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft; Z.Y.F:Data curation, Validation, Visualization; Z.B.L: Resources, Software, Writing - review & editing; D.D.L and D.B.L: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - review & editing.

Acknowledgements

The authors sincerely thank all team members for their valuable contributions to this study. **Hui Chen** and **Zhenshun Wang** contributed to the study conceptualization, data curation, formal analysis, investigation, methodology development, and preparation of the original draft. **Zhiyi Fan** contributed to data curation, validation, and visualization. **Zhaobo Liu** provided resources and software support and contributed to manuscript review and editing. **Dongdong Lin** and **Dongbin Liu** contributed to the conceptualization of the study, funding acquisition, project administration, supervision, and manuscript review and editin

Availability of data and materials

The datasets analyzed in this study were obtained from the Gene Expression Omnibus (GEO) database under accession numbers [GSE6988 and GSE81558]. The datasets are publicly available at <https://www.ncbi.nlm.nih.gov/geo/>. Genes related to ferroptosis (FRGs) were obtained from the FerrDb V3 ferroptosis database (<https://www.zhounan.org/ferrdb/v3/>). Additional data generated or analyzed during this study are included in this published article and its supplementary files. The analysis code is available from the corresponding author upon reasonable request.

Tables

Table 1
GEO Microarray Chip Information

Platform	GSE6988	GSE81558
	GPL4811	GPL15207
Species	Homo sapiens	Homo sapiens
Tissue	Initial tumors of the colon; secondary lesions in the liver	Primary colorectal cancerous tumors; liver metastatic neoplasms
Samples in LMCRC group	26	19
Samples in CRC group	49	23
Reference	PMID: 17640062	PMID:27662660
GEO, Gene Expression Data Repository; LMCRC, Colorectal Carcinoma with Liver Metastases.		

Table 2
Results of GO and KEGG Enrichment Analysis

ONTOLOGY	ID	Description	GeneRatio	BgRatio	p-value	q-value
BP	GO:0006826	iron ion transport	4/27	58/18986	1.30553407623486e-06	0.000383659881007415
BP	GO:0010273	detoxification of copper ions	3/27	16/18986	1.41865900130156e-06	0.000383659881007415
BP	GO:1990169	stress response to copper ions	3/27	16/18986	1.41865900130156e-06	0.000383659881007415
BP	GO:0009636	response to toxic substances	6/27	270/18986	1.80211068952803e-06	0.000383659881007415
BP	GO:0097501	stress response to metal ions	3/27	21/18986	3.35337627260306e-06	0.000485767439250655
CC	GO:0071682	endocytic vesicle lumen	2/27	23/19960	0.000438070597543757	0.0207507125152306
CC	GO:0062023	collagen-containing extracellular matrix	4/27	430/19960	0.00251678629398874	0.0500461724053391
CC	GO:0035580	specific granule lumen	2/27	62/19960	0.00316959091900481	0.0500461724053391
CC	GO:0005788	endoplasmic reticulum lumen	3/27	314/19960	0.0085287150590389	0.100997941488619
CC	GO:0030139	endocytic vesicle	3/27	356/19960	0.0119841339224071	0.113533900317541
MF	GO:0032052	bile acid binding	2/27	10/18737	8.93472725222703e-05	0.00921687653387631
MF	GO:0016709	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, NAD(P) H as one donor, and incorporation of one atom of oxygen	2/27	35/18737	0.00115541126732189	0.0465020405865121
MF	GO:0005506	iron ion binding	3/27	153/18737	0.00135235526195469	0.0465020405865121
MF	GO:0016627	oxidoreductase activity, acting on the CH-CH group of donors	2/27	59/18737	0.00325254216551759	0.0767367765499688
MF	GO:0043539	protein serine/threonine kinase activator activity	2/27	69/18737	0.00442031228907053	0.0767367765499688
KEGG	hsa04978	Mineral absorption	3/23	61/9386	0.000421671903190484	0.0156385877428788
KEGG	hsa05204	Chemical carcinogenesis - DNA adducts	3/23	70/9386	0.000632198227903612	0.0156385877428788
KEGG	hsa04216	Ferroptosis	2/23	42/9386	0.00465969344933384	0.0603546353810303
KEGG	hsa00071	Fatty acid degradation	2/23	43/9386	0.00487973647761521	0.0603546353810303
KEGG	hsa01212	Fatty acid metabolism	2/23	57/9386	0.00844716680597712	0.0835824926065104
GO stands for Gene Ontology; BP stands for Biological Process; CC stands for Cellular Structure; MF stands for Functional Molecules; KEGG stands for Kyoto Pathway Database.						

Table 3
Results of GSEA

ID	setSize	enrichmentScore	NES
WP_ETHANOL_EFFECTS_ON_HISTONE_MODIFICATIONS	25	0.789484795607219	2.005986660858
VILLANUEVA_LIVER_CANCER_KRT19_DN	33	0.75654084236482	1.996270267232
REACTOME_SYNTHESIS_OF_BILE_ACIDS_AND_BILE_SALTS_VIA_7ALPHA_HYDROXYCHOLESTEROL	14	0.881976461380247	1.983282178803
KEGG_GLYCINE_SERINE_AND_THREONINE_METABOLISM	21	0.801376086797948	1.957798404265
Genomic Context-based Enrichment Evaluation (GCEA).			

Table 4
Results of GSVA

ID	logFC	AveExpr	t	P. Value
HALLMARK_XENOBIOTIC_METABOLISM	0.43486692921451	-0.059820104	9.05424797151304	1.0265260692782e-14
HALLMARK_BILE_ACID_METABOLISM	0.404204937284542	-0.054991729	8.11805374817446	1.14990472112536e-13
HALLMARK_HEDGEHOG_SIGNALING	0.242522219729093	0.0145632706662126	3.83290189753291	0.00021954240144380
HALLMARK_COAGULATION	0.190519433092447	-0.013564464	3.06759477148148	0.00276530667140999
HALLMARK_REACTIVE_OXYGEN_SPECIES_PATHWAY	0.184992513740077	-0.009719976	2.92106527451365	0.00429736361706866
HALLMARK_HYPOXIA	0.163460861848784	-0.025530259	2.92246556611723	0.00427959895873938
HALLMARK_COMPLEMENT	0.148650931464293	-0.002918883	2.49807118319098	0.014089573519927
HALLMARK_FATTY_ACID_METABOLISM	0.145977657734686	-0.042309493	2.26602774921103	0.0255676596916284
HALLMARK_PEROXISOME	0.143645732112098	-0.028818047	2.4535911424447	0.0158434822428091
HALLMARK_CHOLESTEROL_HOMEOSTASIS	0.127877731420369	0.0173268231745791	2.04936495133881	0.0430000646323911
HALLMARK_G2M_CHECKPOINT	-0.155521473	0.0143829411501276	-2.051317082	0.0428059393647468
HALLMARK_MYC_TARGETS_V1	-0.164419851	0.0104685449033655	-2.100300281	0.0381753133719155
HALLMARK_IL6_JAK_STAT3_SIGNALING	-0.164434318	0.0106100675534064	-2.564767563	0.0117845229946724
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	-0.169878423	-0.006744839	-2.719247677	0.00769498719182181
HALLMARK_ESTROGEN_RESPONSE_LATE	-0.178840143	0.00794079340878144	-3.523717967	0.000639125066226294
HALLMARK_PI3K_AKT_MTOR_SIGNALING	-0.181044857	0.0149093402576423	-3.234937465	0.00164170596193699
HALLMARK_PANCREAS_BETA_CELLS	-0.183141051	0.0290540704515446	-2.805860357	0.00601399309148632
HALLMARK_INFLAMMATORY_RESPONSE	-0.195396637	0.00643374252443648	-3.478390788	0.000743778661443173
HALLMARK_MYC_TARGETS_V2	-0.196093954	0.00506114849368251	-2.516177508	0.0134268910595047
HALLMARK_NOTCH_SIGNALING	-0.322673722	0.00120240239231436	-5.194287467	1.05985857047729e-06
Gene Set Variation Assessment (GSVA).				

References

1. H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and their co-workers. Global Cancer Statistics 2020: GLOBOCANs evaluation of incidence and mortality rates for 36 cancer kinds in 185 countries. *CA: A Journal for Medical Specialists*. 2021;71(3):209–249.
2. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Cancer of the colon and rectum. *Lancet*. 2019;394(10207):1467–1480.
3. SJ Dixon, KM Lemberg, MR Lamprecht, R Skouta, EM Zaitsev, CE Gleason, along with their co-workers. Ferroptosis: a particular form of cell death propelled by iron and different from apoptosis. *Cell*. 2012;149(5):1060–1072.
4. Wang W, Green M, Choi JE, along with their co-workers. CD8 + T lymphocytes have a function in modulating tumor ferroptosis in the context of cancer immunotherapy. *Nature*. 2019;569(7755):270–274.
5. Xiang Jiang, Brent R. Stockwell, as well as Marcus Conrad. Ferroptosis: mechanisms, biological importance, and relevance in diseases. *Overview of the Molecular Cellular Biology Journal*. 2021;22(4):266–282.
6. D. S. Individual selected to work in the legislative body. GEOquery: Combining Gene Expression Omnibus (GEO) datasets using the BioConductor platform. *Bioinformatics*, as of July 15 in 2007; Volume 23, Number 14: pages 1846–1847.
7. Ki DH, Jeung HC, Park CH, Kang SH, Lee GY, and Lee WS worked together with their team. Comprehensive genomic analysis pinpoints crucial gene indicators linked to liver metastases in colorectal cancer. *International Journal of Cancer*, Volume 121, Number 9, Pages 2005–2012, Year 2007.
8. Sayagues JM, Corchete LA, Gutierrez ML, Sarasquete ME, M. Abad del Mar, Bengoechea O, along with their co-workers. Genetic analysis of liver metastatic tumors in patients with colorectal cancer. *Oncotarget*, 2016; Volume 7, Number 45: pages 72908–72922.
9. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013 Vol 41 Iss Database issue pp D991-5 2013.
10. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W. and Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015 Vol 43 Iss 7 pp e47 2015.

11. N. Zhou, L. Peng, Q. Luo, T. Yin, H. Sun, Y. Zhang, and other participants. FerrDb Version 3: Improving the curated database for regulatory elements and disease associations, expanding its range from ferroptosis to regulated cell death. *Nucleic Acids Research* in 2026; Volume 54, Issue D1: Pages D572–D582.
12. Zhu J, Wu Y, Ge X, Chen X, and Mei Q. Recognition and Validation of Genes Associated with Ferroptosis in Ulcerative Colitis. *Journal of Inflammation Research* 2024; Volume 17: Page 44.67–4482.
13. Mi H, Muruganujan A, Ebert D, Huang X, and Thomas PD. PANTHER 14th version: more extensive genomic coverage, revised PANTHER GO-slim classes, and enhanced enrichment analysis functions. *Nucleic Acids Research*, Volume 47, Issue D1, Pages D419–D426, Published in 2019.
14. Kanehisa MaG, S., the KEGG: Kyoto source for gene and genome data. *Nucleic Acids Research*, Volume 28, Number 1, pages 27 to 30, released in 2000.
15. G. Yu and L. Wang. G., Han, Y. collaborated with He, Q. Y. clusterProfiler: an R software crafted for examining and contrasting biological functions among gene sets. *Omics*, Vol. 16, Iss. 5, pp. 284–287, Published in the year 2012.
16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005. 2005;102:15545–50.
17. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinf* 2013. 2013;14:7.
18. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov. J. P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011 Vol 27 Iss 12 pp 1739-40 2011.
19. Sanz H, Valim, Clarissa, Vegas, Esteban, Oller, Josep M, Reverter. Ferran. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. *BMC bioinformatics* 2018 Vol 19 Iss 1 pp 432 2018.
20. Engebretsen SaB, J. Utilizing glmnet for statistical prediction. *Clin Epigenetics*, Vol. 11, Iss. 1, Pg. 123, 2019.
21. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M. pROC: an open-source package for R and S + to analyze and compare ROC curves. *BMC Bioinf* 2011. 2011;12:77.
22. Van Calster B, Wynants L, Verbeek JFM, Verbakel JY, Christodoulou E, Vickers AJ, Roobol MJ, Steyerberg EW. Reporting and Interpreting Decision Curve Analysis: A Guide for Investigators. *Eur Urol* 2018. 2018;74:796–804.
23. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinf* 2010. 2010;26:976–8.
24. P. Charoentong, F. Finotello, M. Angelova, C. Mayer, M. Efremova, D. Rieder, H. Hackl, together with Z. Trajanoski. Comprehensive immunogenomic investigations spanning various cancer kinds reveal associations between genetic alterations and immune characteristics, along with factors influencing the efficacy of checkpoint inhibitors. *Cell Reports*, Volume 18, Issue 1, Pages 248–262, Released in 2017.
25. Zhang J. Contrastive Learning-Powered Ultrafast Genome-Wide Virtual Screening: The Rise of DrugCLIP. *Drug Dev Res*. 2026;87(5):e70351.
26. Anastas JN along with Moon RT. WNT signaling pathways as possible cancer therapy approaches. *Natures Perspectives on Oncology*. 2013;13(1):11–26.
27. Kikuchi A, Yamamoto H, Sato A, Matsumoto S. Wnt5a: the pathway, functions, and connection to diverse disorders. *Physiological Actions*. 2012;204(1):17–33.
28. Weeraratna AT, Jiang Y, Hostetter G, Rosenblatt K, Duray P, Bittner M, along with their co-workers. The Wnt5a signaling pathway impacts the motility and invasiveness of metastatic melanoma cells. *Malignant tumor cell*. 2002;1(3):279–288.
29. DH Goetz, MA Holmes, N Borregaard, ME Bluhm, KN Raymond, RK Strong. Neutrophil lipocalin NGAL acts as an antibacterial protein that interferes with siderophore-related iron uptake processes. *Mol Cell*. 2002;10(5):1033–1043.
30. Devireddy L. R., Teodoro J. G., Richard F. A., Green M. R. A lipocalin 24p3 membrane receptor particularly promotes cell death and iron uptake. *Cell*. 2001;104(1):45–56.
31. X. Leng, T. Ding, H. Lin, Y. Wang, L. Hu, J. Hu, along with their co-researchers. Inhibiting lipocalin 2 hinders breast cancer growth and metastasis. *Journal of Studies in Biochemistry*. 2009;284(5):3301–3308.
32. Candido S, Abrams SL, Steelman LS, Lattanzio R, Curcuro G, Montalto G, together with multiple collaborators. Comprehensive examination of colon cancer gene expression data sets uncovers a prognostic biomarker profile and implies a potential participation of lipocalin–2. *Oncotarget*. 2016;7(14):19595–19617.
33. DM Bautista, SE Jordt, T Nikai, PR Tsuruda, AJ Read, J Poblete, and their co-workers. TRPA1 has a crucial function in triggering inflammation brought about by environmental irritants and substances that induce pain. *Cell*. 2006;124(6):1269–1282.
34. Nilius B, Szallasi A. Exploring transient receptor potential channels: Linking basic research and clinical advancement. *Examination and Anticipations in the Pharmaceutical Domain*. 2014;66(3):676–814.
35. TextMaster Composition Assistant. A synopsis of aldo-keto reductase (AKR) catalysts. *Interactions among Molecules in Biological Chemistrys Realm*. 2015;234:236–246.
36. Penning TM, Byrns MC. Steroid hormone transforming aldo-keto reductases and cancer. *Ann N Y Acad Sci*. 2009;1155:33–42.
37. Chen X, Liu L, Mims J, Pate L, Shang H, Shi X, and co-workers. Multiple-gene markers in colorectal carcinoma: significance for diagnosis and prognosis. *Oncology Treatment Alternatives*. 2021;10(14):4860–4872.
38. Chen Y, Li H. Prognostic and Predictive Models for Left- and Right- Colorectal Cancer Patients: A Bioinformatics Analysis Based on Ferroptosis-Related Genes. *Front Oncol*. 2022;12:833834.
39. Wang W, Green M, Choi JE, Gijon M, Kennedy PD, Johnson JK, et al. CD8 + T cells regulate tumour ferroptosis during cancer immunotherapy. *Nature*. 2019;569(7755):270–4.

40. Friedmann Angeli JP, Krysko DV, Conrad M. Ferroptosis at the crossroads of cancer-acquired drug resistance and immune evasion. Nat Rev Cancer. 2019;19(7):405–14.

41. Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, et al. Lineage tracking reveals a subset of tumor-derived dendritic cells in colorectal cancer. Nat Med. 2018;24(4):508–16.

42. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. 2021;49(D1):D605–12.

Figures

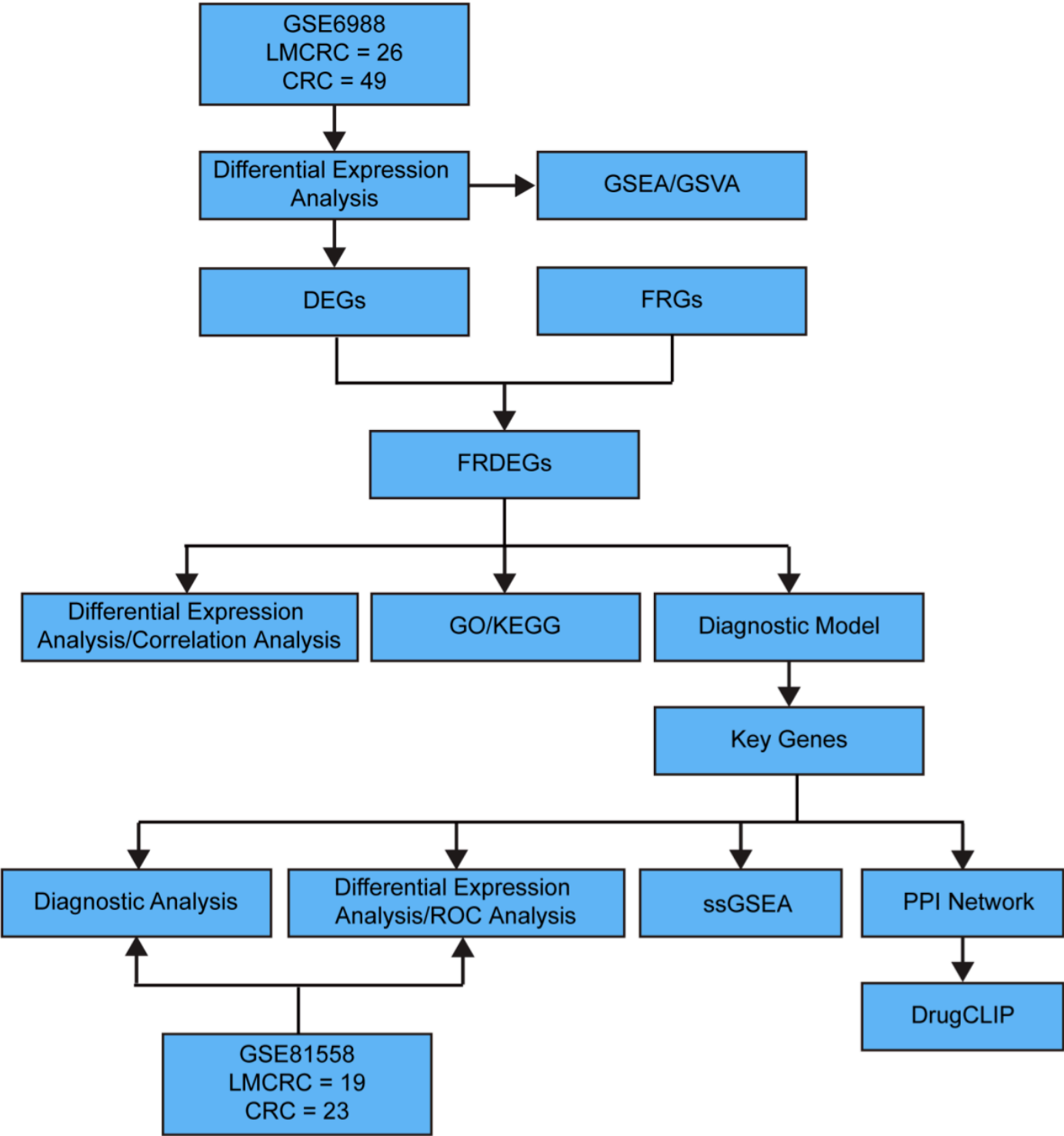


Figure 1

Flow Chart for the Comprehensive Analysis

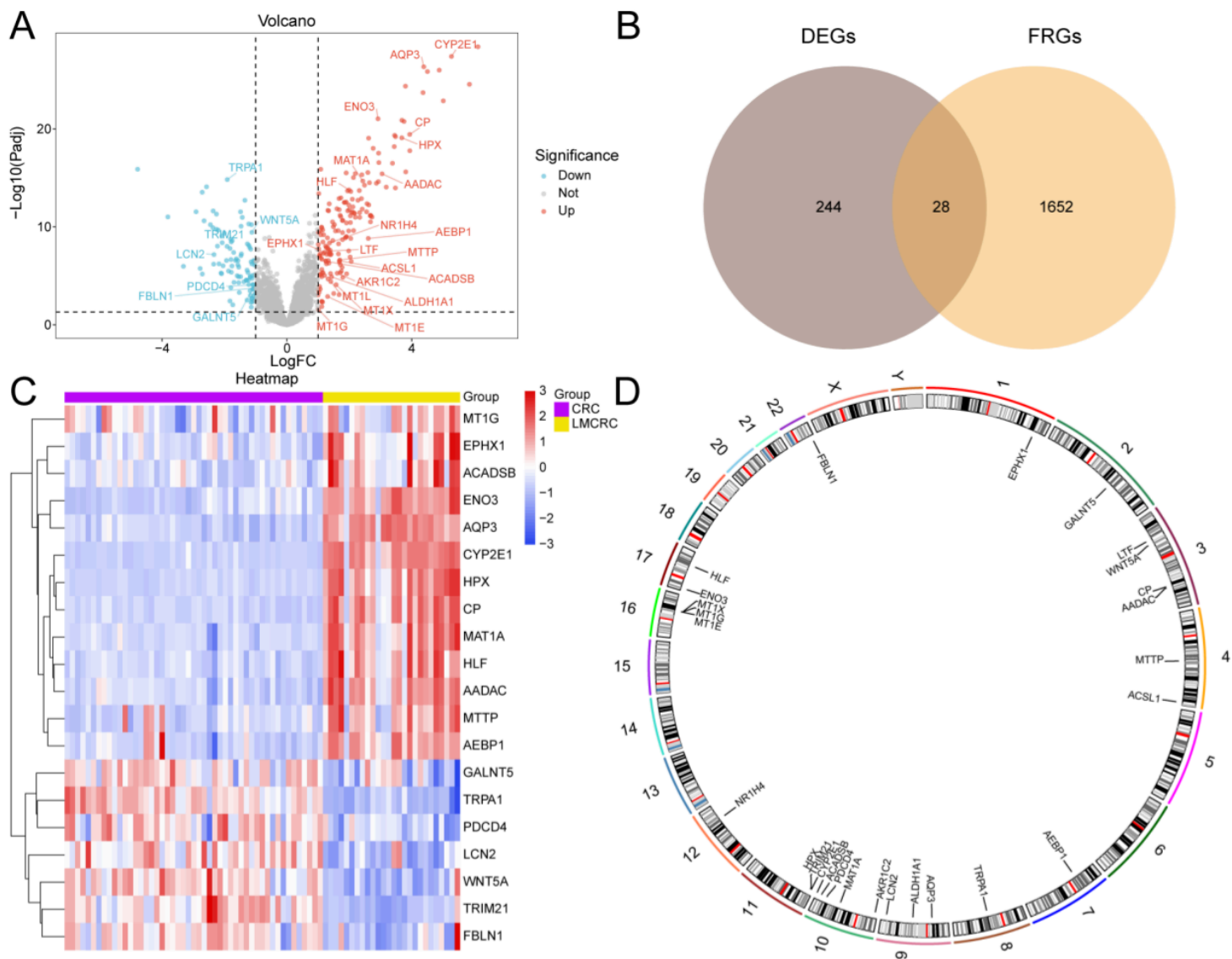


Figure 2

Differential Gene Expression Analysis

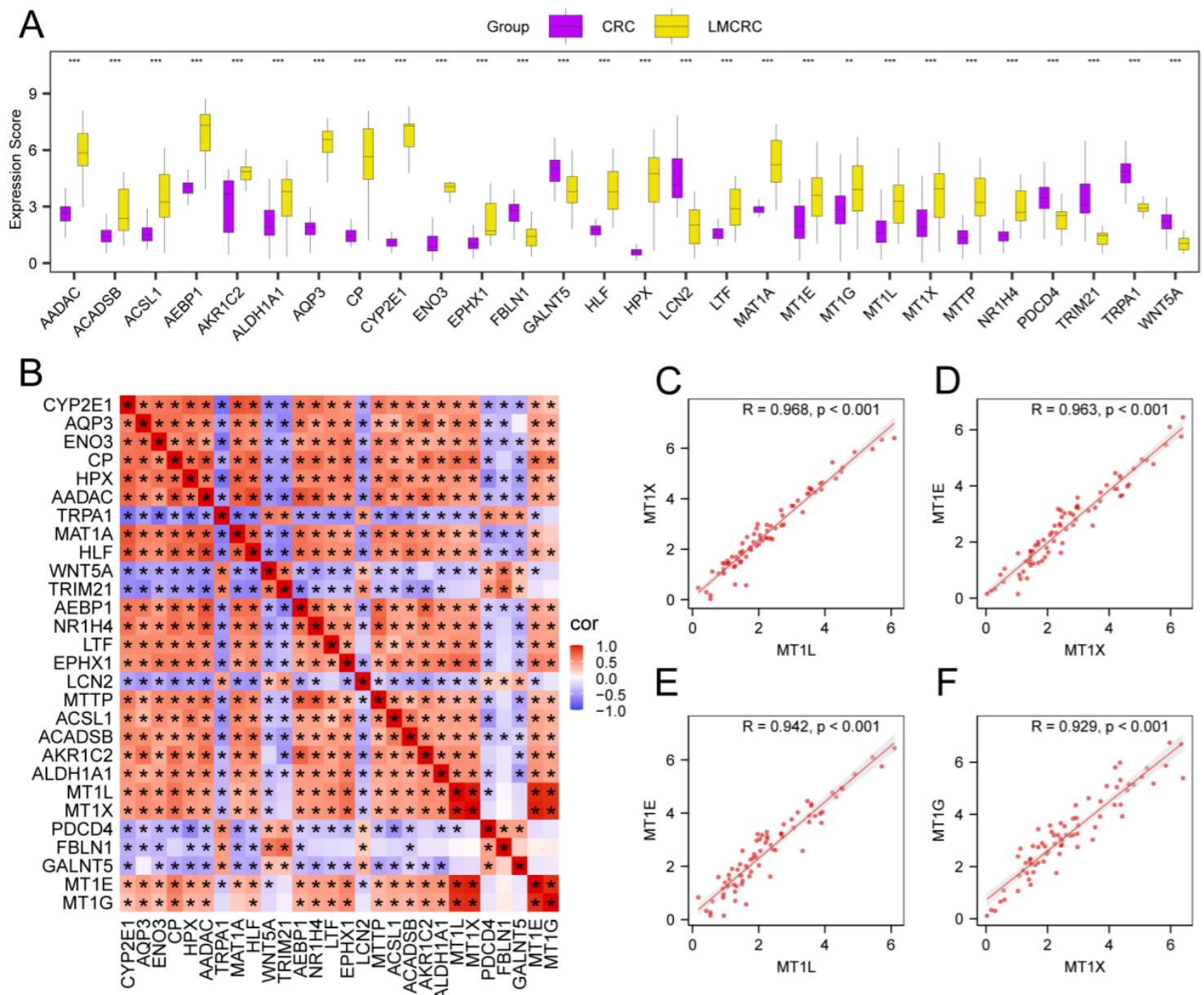


Figure 3

Differential Expression and Correlation Analysis

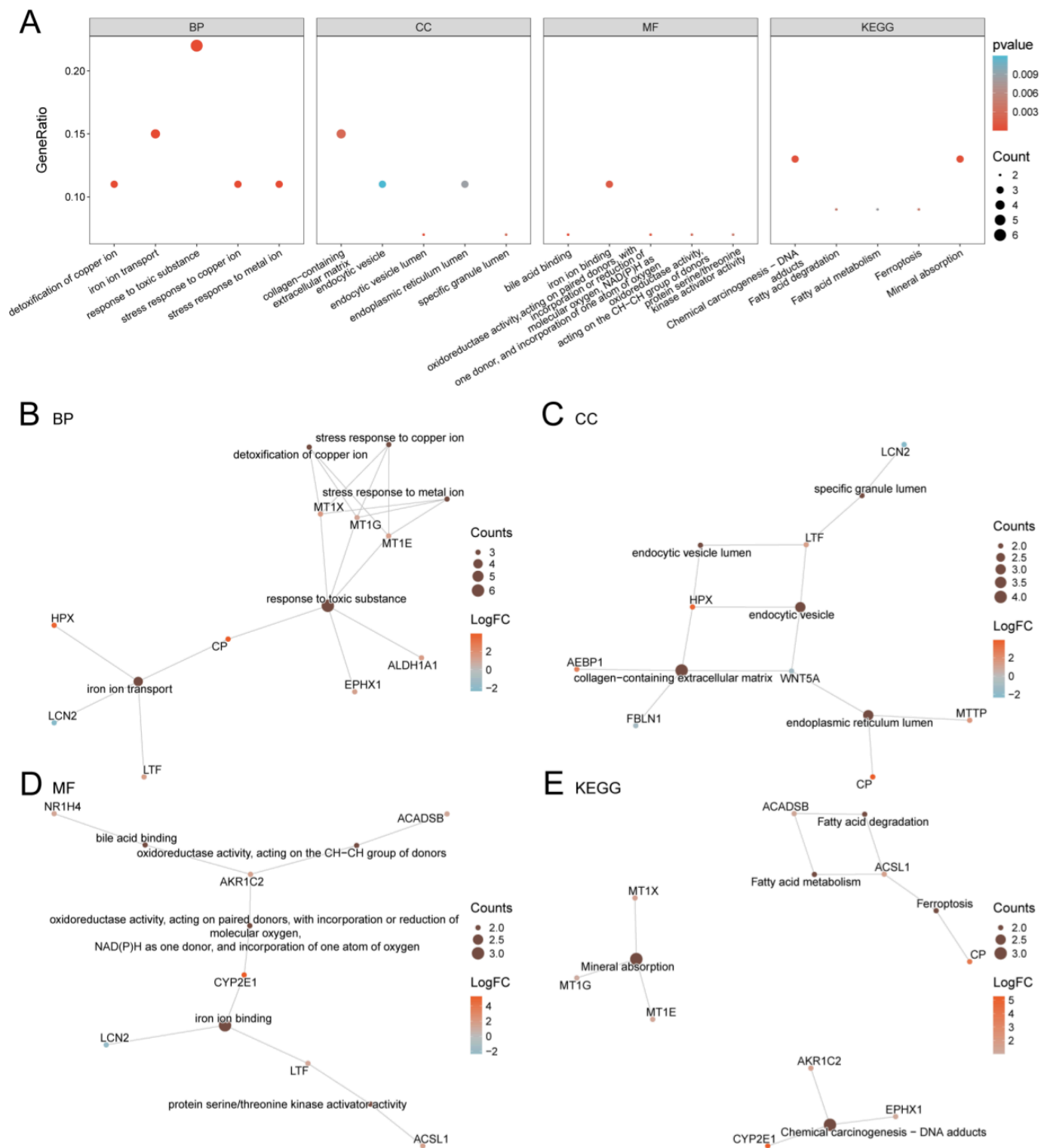


Figure 4

GO and KEGG enrichment analysis for FRDEGs

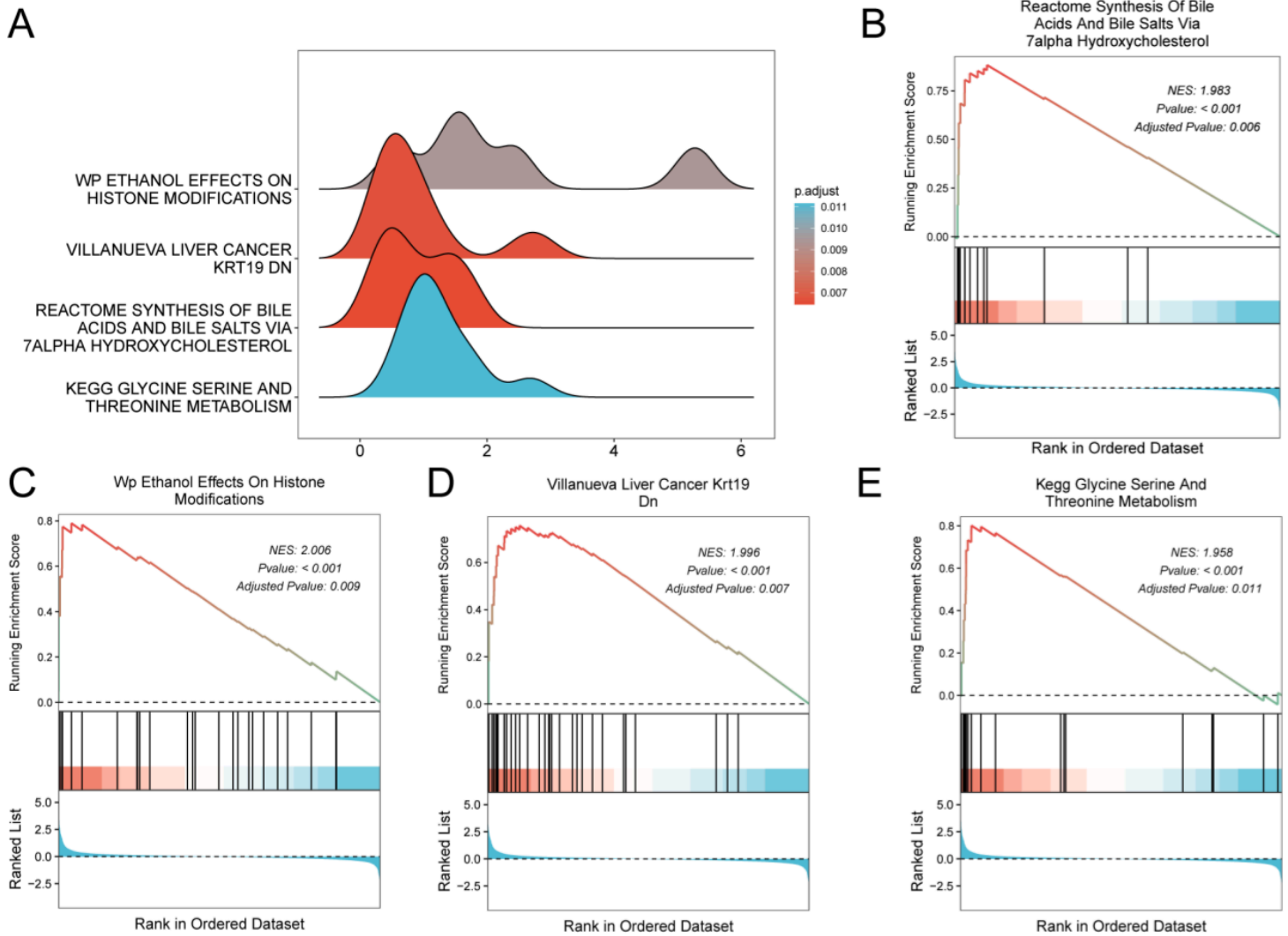


Figure 5

GSEA Analysis

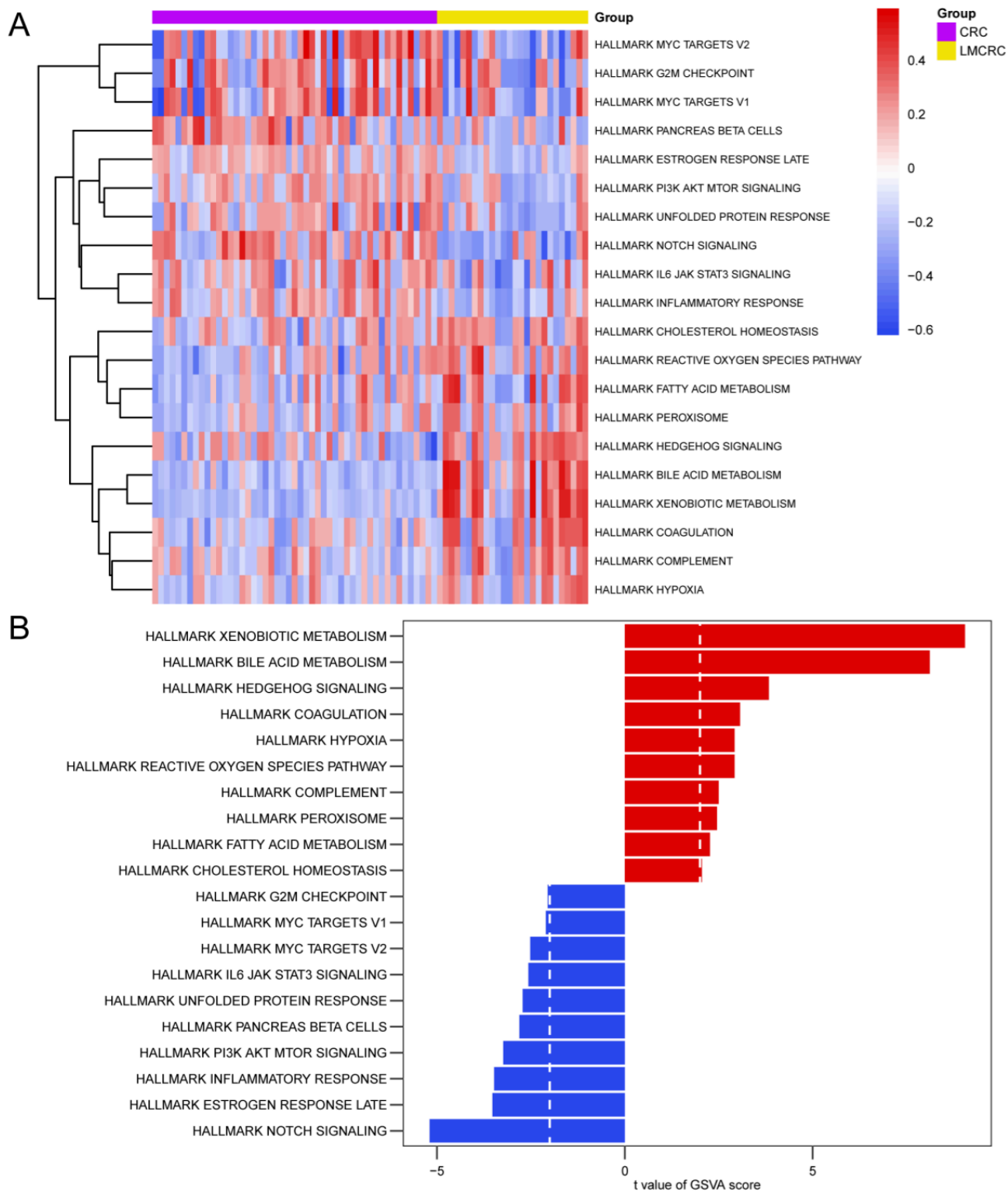


Figure 6

GSEA Analysis

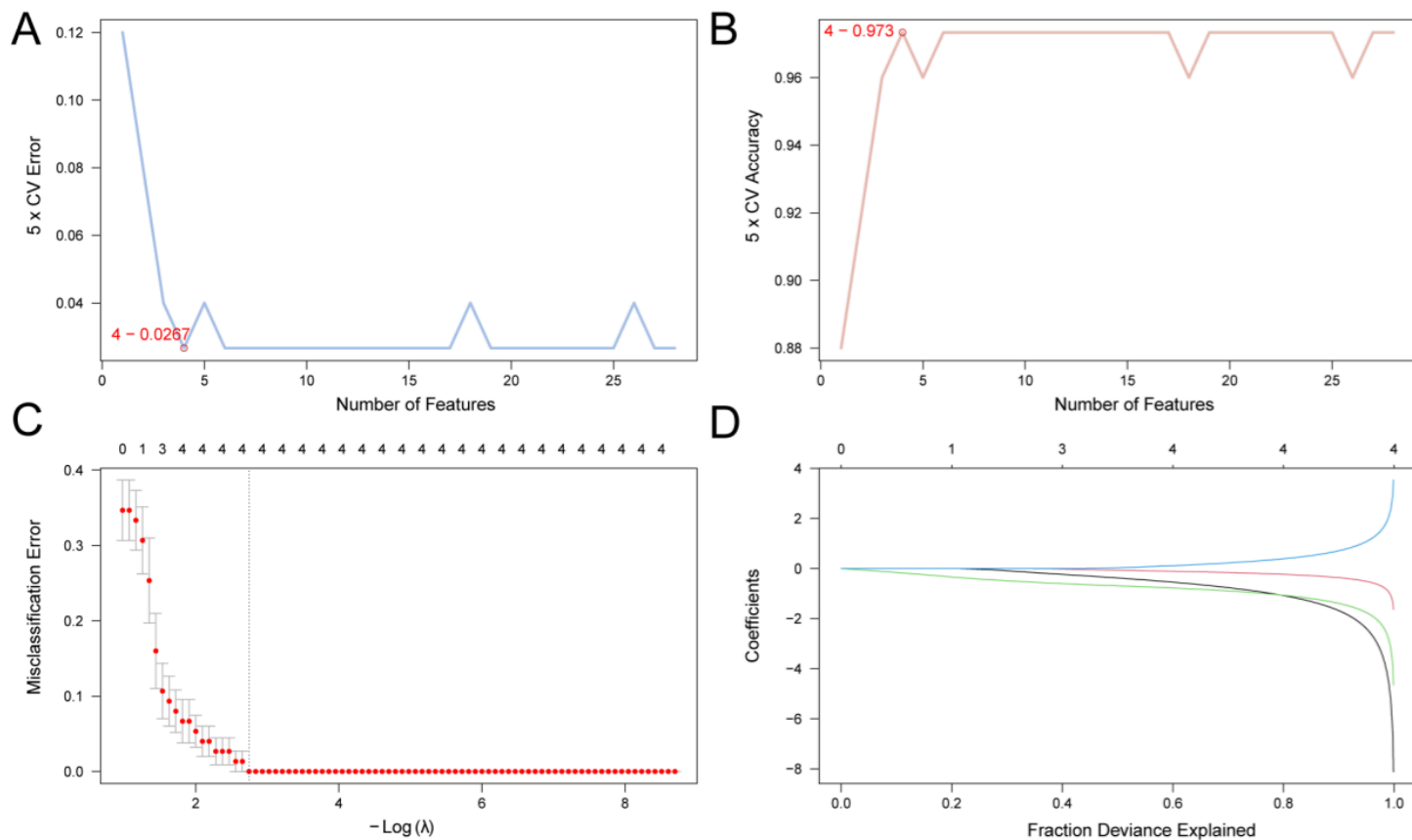


Figure 7

Machine Learning

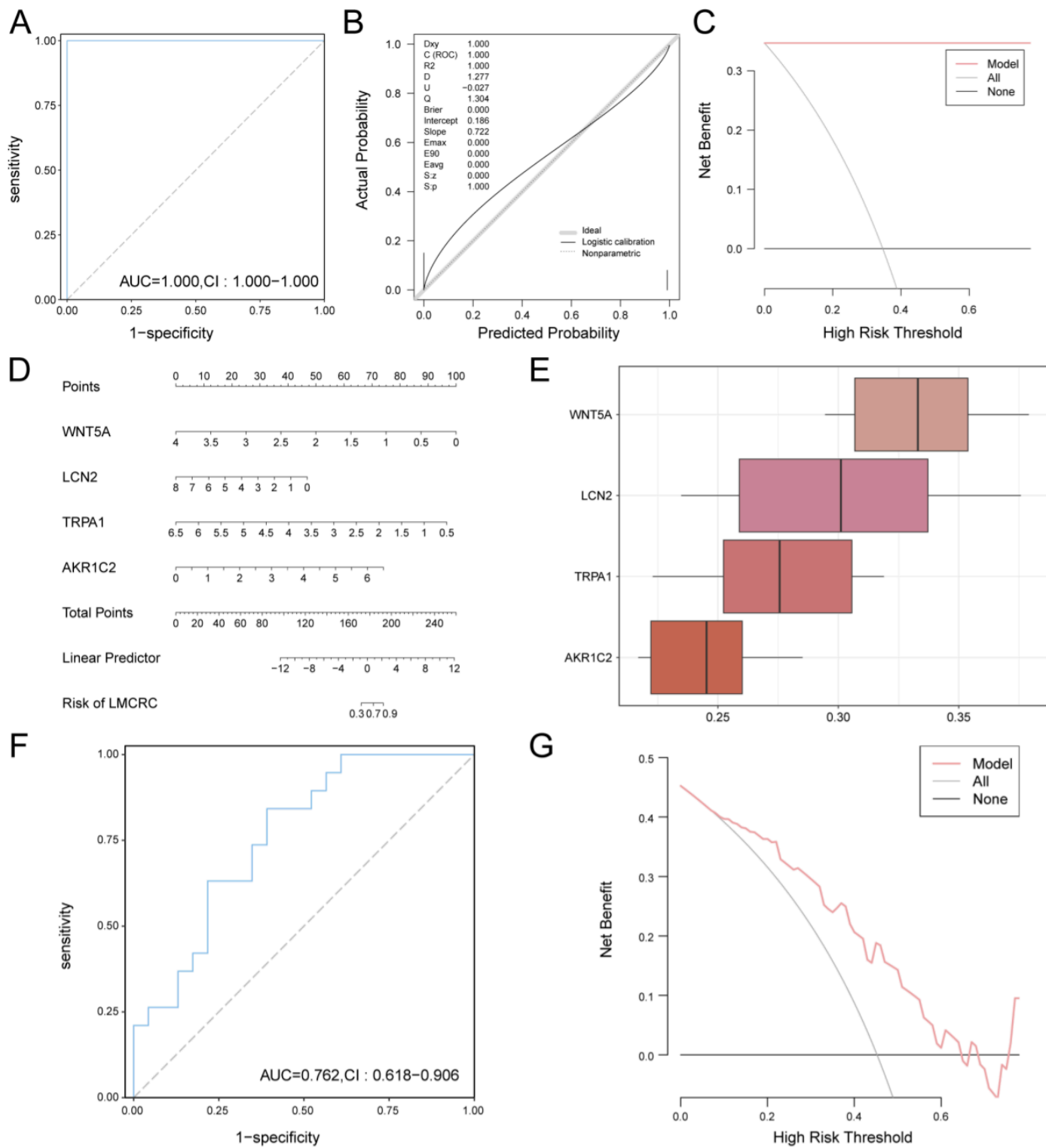


Figure 8

Diagnostic and Validation Analysis of LMCRC and Friends Analysis

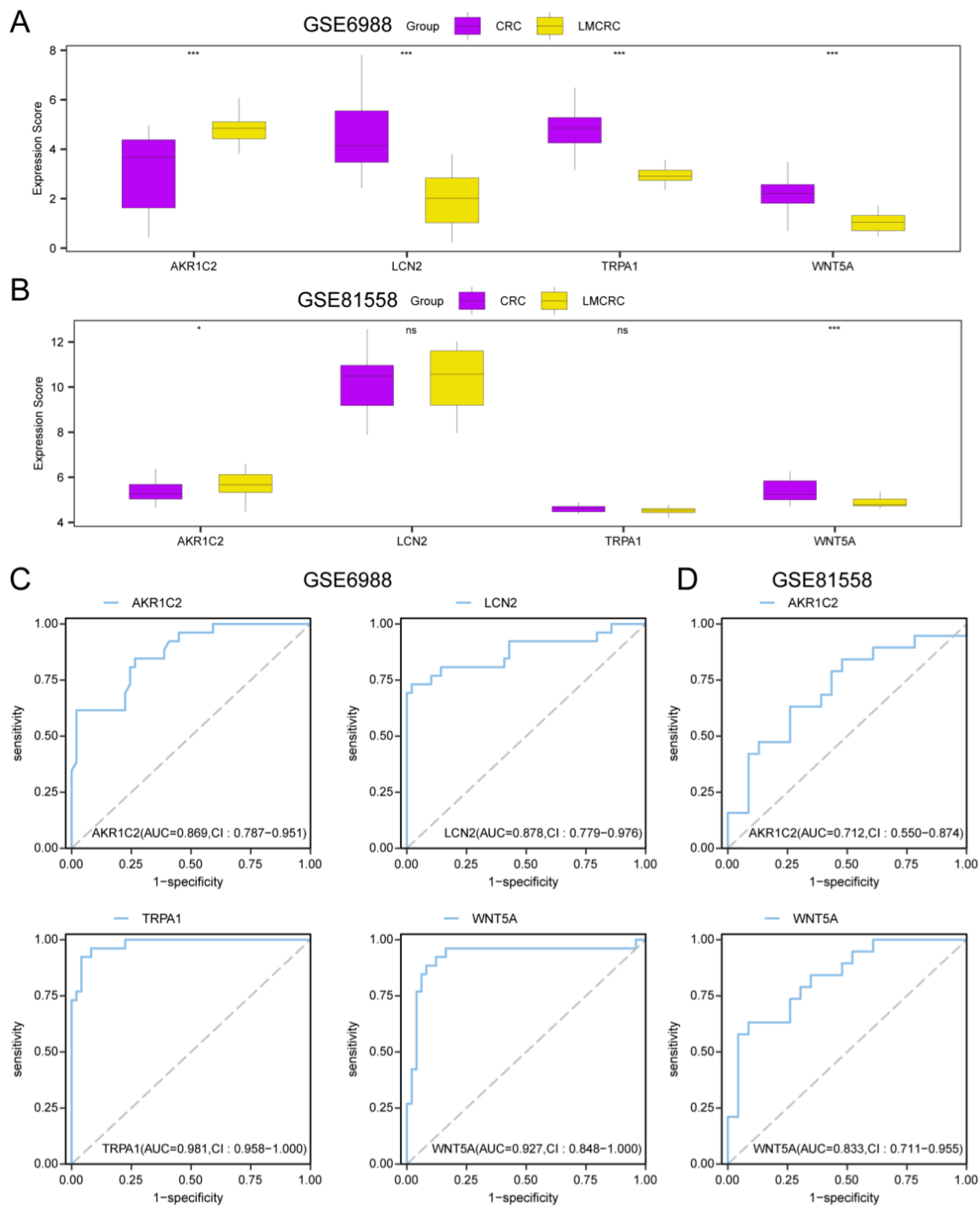


Figure 9

Differential Expression and ROC Curve

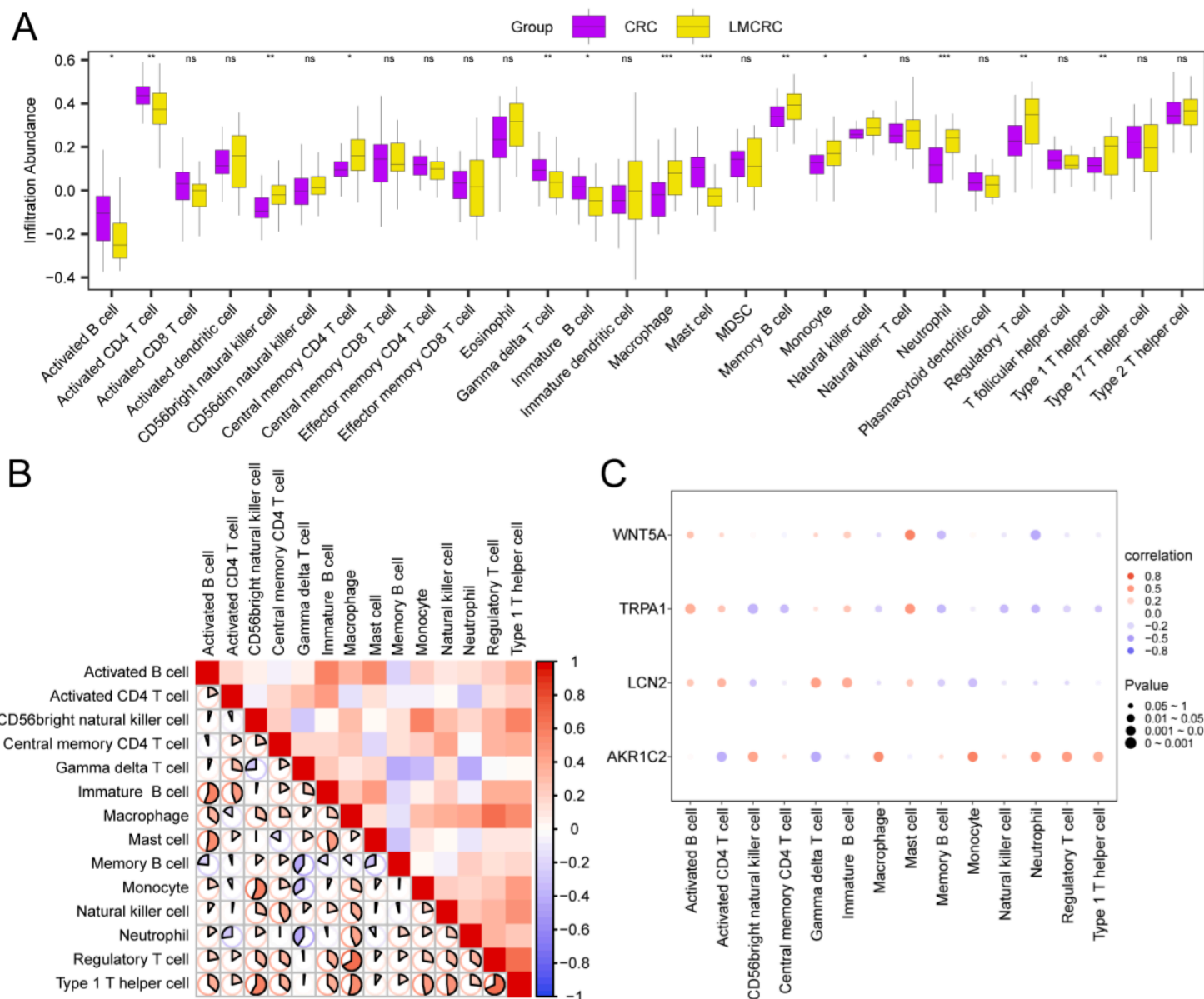


Figure 10

Immune Infiltration Analysis by ssGSEA Algorithm

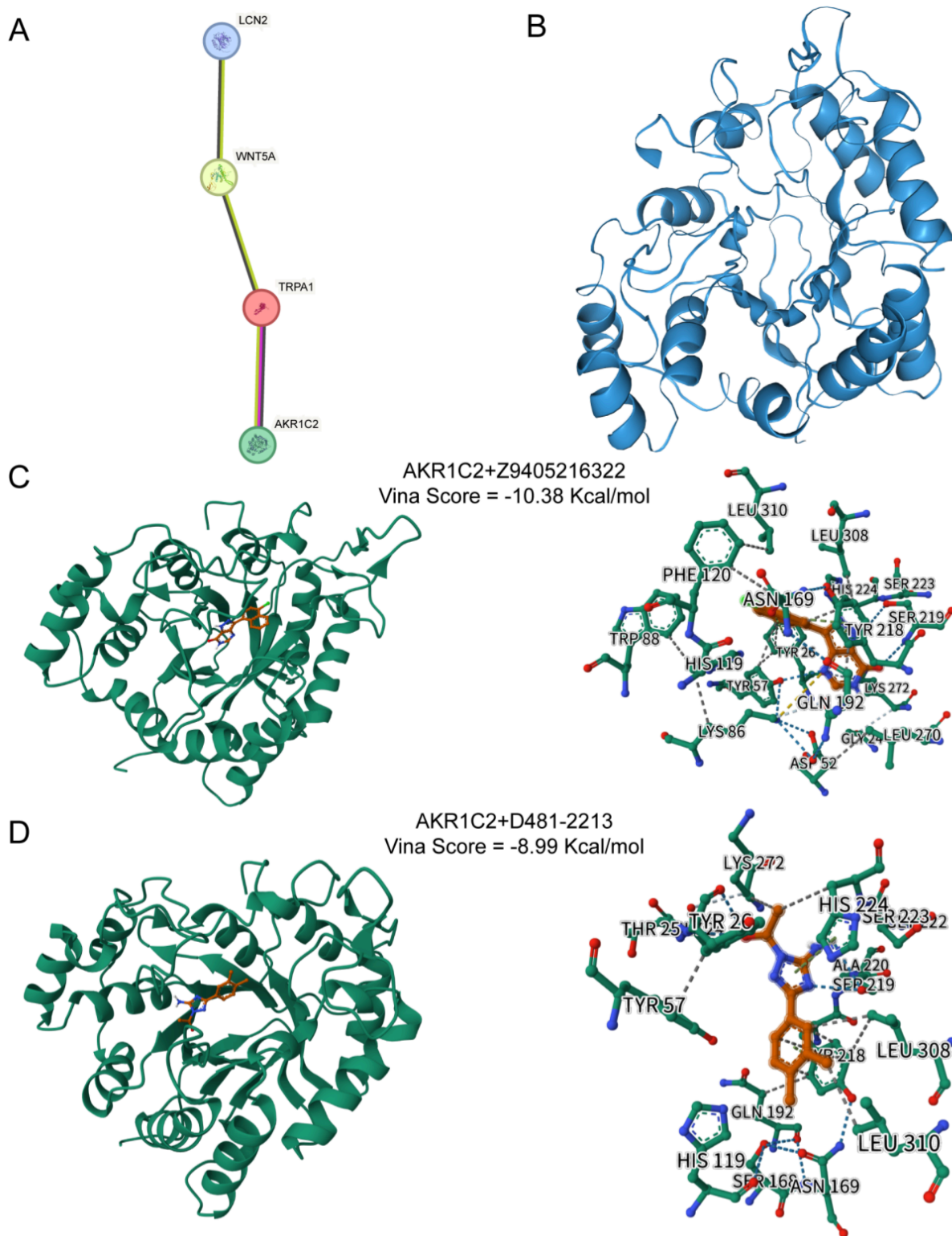


Figure 11

Molecular Docking