Supplementary information:

Development of a novel chemokine signaling-based multigene signature to predict prognosis and therapeutic response in colorectal cancer

Xin QI, Donghui YAN, Jiachen ZUO, RuiWANG, Jiajia CHEN

School of Chemistry and Life Sciences, Suzhou University of Science and Technology, Suzhou 215011, China

Materials and methods

Data collection

Gene expression data from human CRC tissues and corresponding clinical data used in this study were retrieved from the public Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo) database. Given the critical role of distal metastasis in influencing patient survival, GSE131418 dataset with 332 primary and 184 metastatic human CRC samples in Moffitt Cancer Center (MCC) (Kamal et al., 2019) was firstly downloaded to identify differentially expressed genes (DEGs) involved in CRC metastasis. Next, using the GSE39582 (n=582) dataset (Marisa et al., 2013) as the discovery set, the prognostic gene signature associated with chemokine signaling pathway was screened out from the identified DEGs. Then, the prognostic performance of the gene signature was verified in the GSE17536 (n=145) dataset (Smith et al., 2010).

Identification of differentially expressed genes

The "limma" R package was used to identify differentially expressed mRNAs between primary and metastatic CRC samples, and the raw p-values were adjusted by the Benjamini-Hochberg false discovery rate method (Ritchie et al., 2015). The genes were considered as DEGs if they meet the screening criteria: $|\log_2(\text{fold change})| >= 1$ and adjusted p-value < 0.05.

Collection of genes in the chemokine signaling pathway

KEGG (http://www.genome.jp/kegg/) is an integrated database resource for metabolic pathways and gene signaling networks (Kanehisa and Goto, 2000). 192 genes in the chemokine signaling pathway were obtained from the KEGG database with the entry number hsa04062.

Establishment of the chemokine signaling-based multigene signature (CSbMgSig)

To construct the CSbMgSig for survival prediction, chemokine signaling-related genes (CSRGs) that differentially expressed between primary and metastatic CRC samples were firstly identified from the GSE131418 dataset. Since the GSE131418 dataset lacks survival information of CRC patients, we then employed the GSE39582 dataset in which samples with OS no more than 1 month or with incomplete clinical data were removed, to build the risk model according to the following steps: Firstly, the univariate Cox analysis was performed using the "survival" R package (Therneau, 2021) to screen differentially expressed CSRGs that were significantly related to OS of CRC

patients; Secondly, to further minimize over-fitting risk, the least absolute shrinkage and selection operator (LASSO)-penalized Cox regression analysis was conducted to establish the CSbMgSig utilizing the "glmnet" R package (Qi et al., 2021). According to the coefficient and normalized expression value of each gene constituting the prognostic signature, the risk score of each patient was computed by the formula: $Risk\ score = \sum_{i=1}^{n} Coef_i \times \mathrm{Exp}_i$,

Where n is the number of gene that constitutes the signature, Coef represents the regression coefficient of each prognostic gene in the signature, Exp equals the normalized expression level of each prognostic gene in the signature.

In addition, the constructed risk score formula was applied to the GSE17536 dataset to validate the prognostic performance of CSbMgSig. Furthermore, to examine whether the CSbMgSig was an independent prognostic factor for OS of patients with CRC, multivariate Cox regression analysis was carried out with the risk score and clinical indicators as variables.

Prognostic analysis of the CSbMgSig

To assess the prognostic performance of the established CSbMgSig, the CRC patients in each dataset were stratified into high- and low-risk groups based on the median risk score. Kaplan-Meier survival analysis and the log-rank test were performed to determine OS differences between the two groups. p value less than 0.05 was considered statistically significant. Besides, time-dependent ROC curve analysis was conducted to measure the predictive power of CSbMgSig with "timeROC" (Blanche et al., 2013) and "survival" (Therneau, 2021) R packages and the areas under the curve (AUC) was computed.

Functional enrichment analysis

Firstly, differential expression analysis was performed to screen DEGs between the high- and low- risk groups in the GSE39582 dataset with the criteria: fold change > 1.5 or < 0.67 and adjusted p-value < 0.05. Next, the "clusterProfiler" R package (v3.14.3) was employed to perform Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses of those DEGs associated with the CSbMgSig (Yu, 2018a). Then, gene set enrichment analysis (GSEA) was conducted to identify the significant enriched pathways in the high- or low-risk group by utilizing the MSigDB gene set "h.all.v7.2.entrez.gmt" in the Molecular Signatures Database (http://software.broadinstitute.org/gsea/msigdb/index.jsp), and the enriched GSEA pathways were exhibited using the "enrichplot" R package (Yu, 2018b).

Estimation of Tumor Infiltrating Immune Cells

To investigate the relationship between the chemokine signaling-based prognostic signature and immune status, single-sample gene set enrichment analysis (ssGSEA) was conducted with the "GSVA" R package (Hänzelmann et al., 2013) to determine infiltration scores of multiple immune cell types and immune-related functions based on the expressed level of the immune-related marker gene reported by Liang *et al* (Liang et al., 2020). Wilcoxon test (pvalue < 0.05) was used to assess the statistical significance between high-risk and low-risk groups. In addition, according to the expression pattern of genes in the prognostic signature, immune score and stromal score were calculated using the ESTIMATE algorithm (Meng et al., 2020), and the relationship between risk score and stromal/immune score was estimated by Pearson correlation analysis (p value < 0.05).

Chemotherapeutic sensitivity analysis

The response to chemotherapeutic drugs of CRC patients was analyzed based on the Genomics of Drug Sensitivity in Cancer (GDSC; https://www.cancerrxgene.org) (Yang et al., 2012), which is the largest public available pharmacogenomics database. The half-maximal inhibitory concentration (IC50) was calculated by the "pRRophetic" R package (Geeleher et al., 2014), and the Wilcoxon test (p value < 0.05) was used to assess the statistical significance between high- and low-risk groups.

References

Blanche P, Dartigues JF, Jacqmin-Gadda H, 2013. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*, 32(30):5381-5397. https://doi.org/10.1002/sim.5958

Geeleher P, Cox N, Huang RS, 2014. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PloS one*, 9(9):e107468. https://doi.org/10.1371/journal.pone.0107468

Hänzelmann S, Castelo R, Guinney J, 2013. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics*, 14(1):1-15. https://doi.org/10.1186/1471-2105-14-7

Kamal Y, Schmit SL, Hoehn HJ, et al., 2019. Transcriptomic differences between primary colorectal adenocarcinomas and distant metastases reveal metastatic colorectal cancer subtypes. *Cancer Res*, 79(16):4227-4241. https://doi.org/10.1158/0008-5472.CAN-18-3945

Kanehisa M, Goto S, 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27-30. https://doi.org/10.1093/nar/28.1.27

Liang JY, Wang DS, Lin HC, et al., 2020. A novel ferroptosis-related gene signature for overall survival prediction in patients with hepatocellular carcinoma. *Int J Biol Sci*, 16(13):2430-2441. https://doi.org/10.7150/ijbs.45050

Marisa L, De Reynies A, Duval A, et al., 2013. Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PLoS Med*, 10(5):e1001453. https://doi.org/10.1371/journal.pmed.1001453

Meng Z, Ren D, Zhang K, et al., 2020. Using ESTIMATE algorithm to establish an 8-mRNA signature prognosis prediction system and identify immunocyte infiltration-related genes in Pancreatic adenocarcinoma. *Aging (Albany NY)*, 12(6):5048. https://doi.org/10.18632/aging.102931

Qi X, Wang R, Lin YX, et al., 2021. A ferroptosis-related gene signature identified as a novel prognostic biomarker for colon cancer. *Front Genet*, 12:692426. https://doi.org/10.3389/fgene.2021.692426

Ritchie ME, Phipson B, Wu D, et al., 2015. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Res*, 43(7):e47. https://doi.org/10.1093/nar/gkv007

Smith JJ, Deane NG, Wu F, et al., 2010. Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. *Gastroenterology*, 138(3):958-968.

https://doi.org/10.1053/j.gastro.2009.11.005

Therneau TM, 2021. A package for survival analysis in r. 2020. R package version, 3

Yang W, Soares J, Greninger P, et al., 2012. Genomics of drug sensitivity in cancer (GDSC): A resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*, 41(D1):D955-D961. https://doi.org/10.1093/nar/gks1111

Yu G, 2018a. Clusterprofiler: Universal enrichment tool for functional and comparative study. *BioRxiv*:256784. https://doi.org/10.1101/256784

Yu G, 2018b. Enrichplot: Visualization of functional enrichment result. R package version, 1(2)

Supplemetary figures

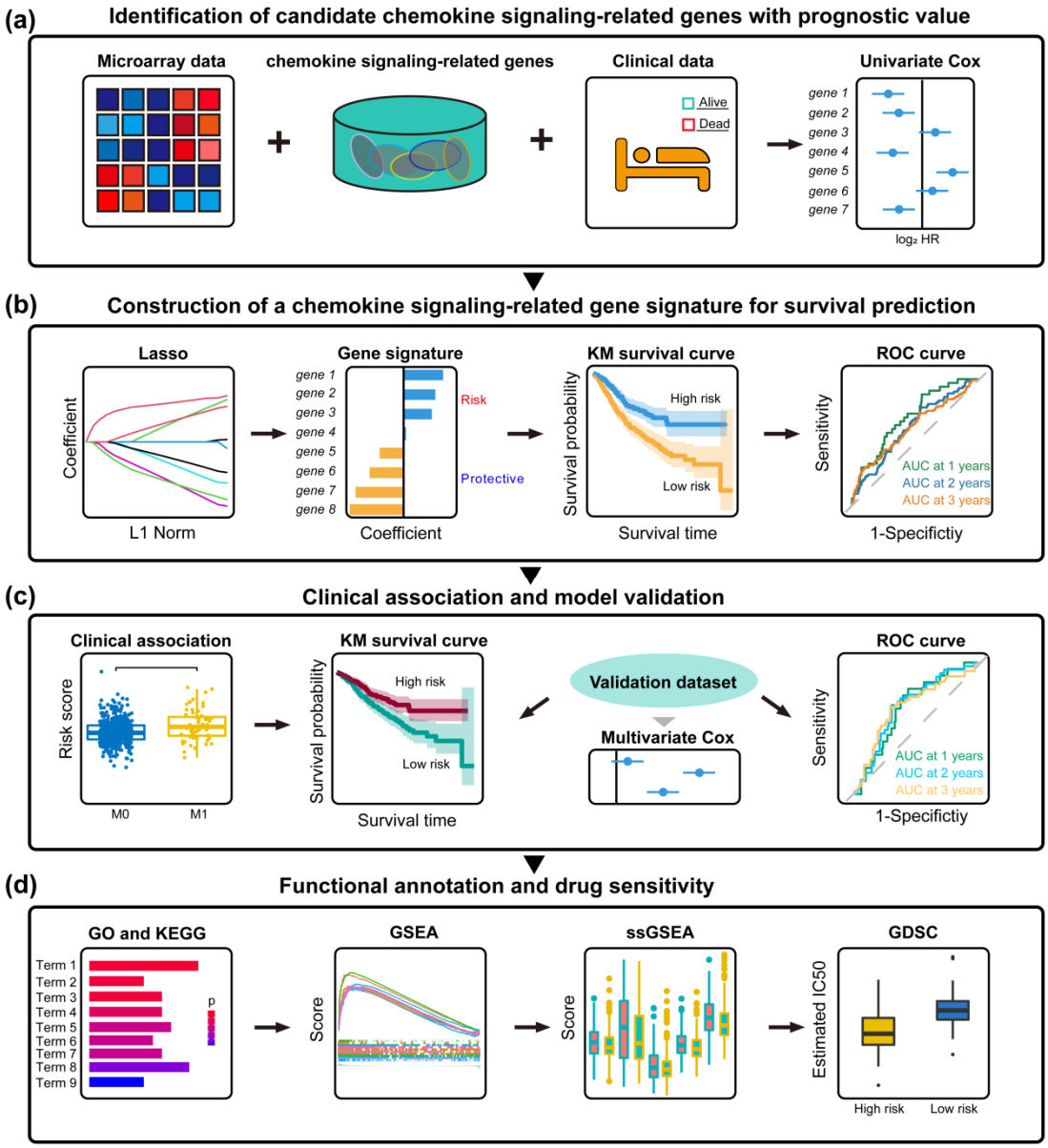

Fig. S1 Schematic diagram of the CSbMgSig construction and characterization.

(a) The prognostic CSRGs associated with CRC metastasis were identified through univariate Cox regression analysis. (b) The CSbMgSig for OS prediction of CRC patients was established by utilizing LASSO Cox regression analysis. (c) Clinical association analysis and independent validation of the CSbMgSig. (d) Functional implications of the CSbMgSig.

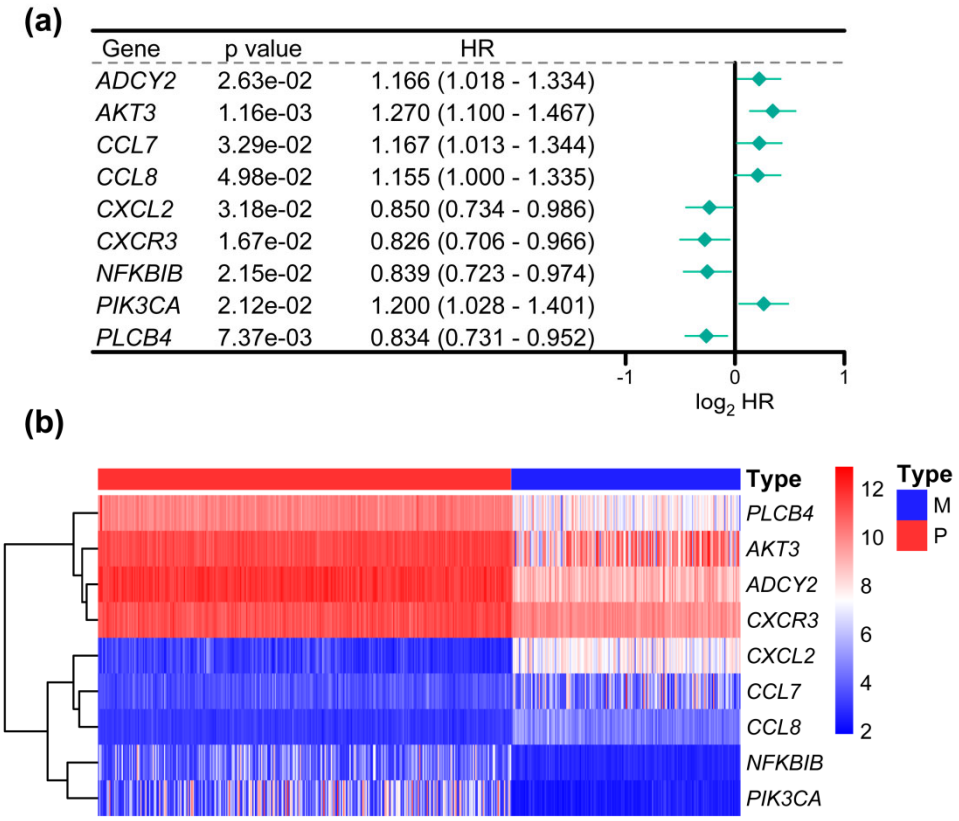

Fig. S2 Identification of prognostic CSRGs in CRC patients.

(a) Prognostic CSRGs that differentially expressed between primary and metastatic CRC samples were screened out by univariate Cox regression analysis. (b) The heatmap shows the changes in the prognostic CSRG expression level between the primary and metastatic CRC samples in the GSE131418 dataset.

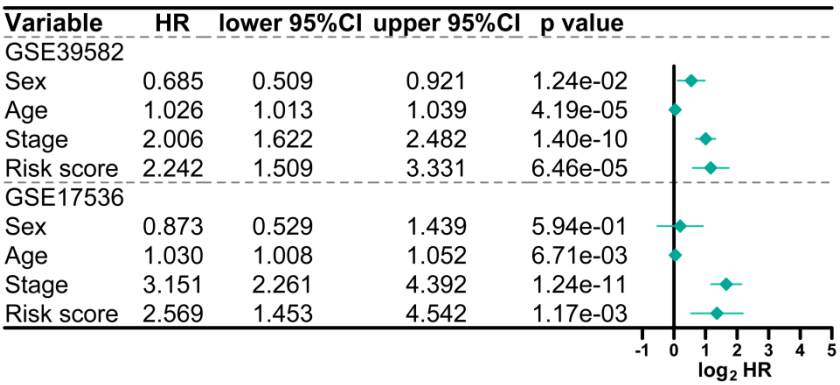

Fig. S3 Multivariable analysis of CSbMgSig and clinical factors in the training and validation cohort.

The prognostic independence of CSbMgSig in predicting OS was evaluated in the GSE39582 and GSE17536 datasets through multivariate Cox regression analysis.

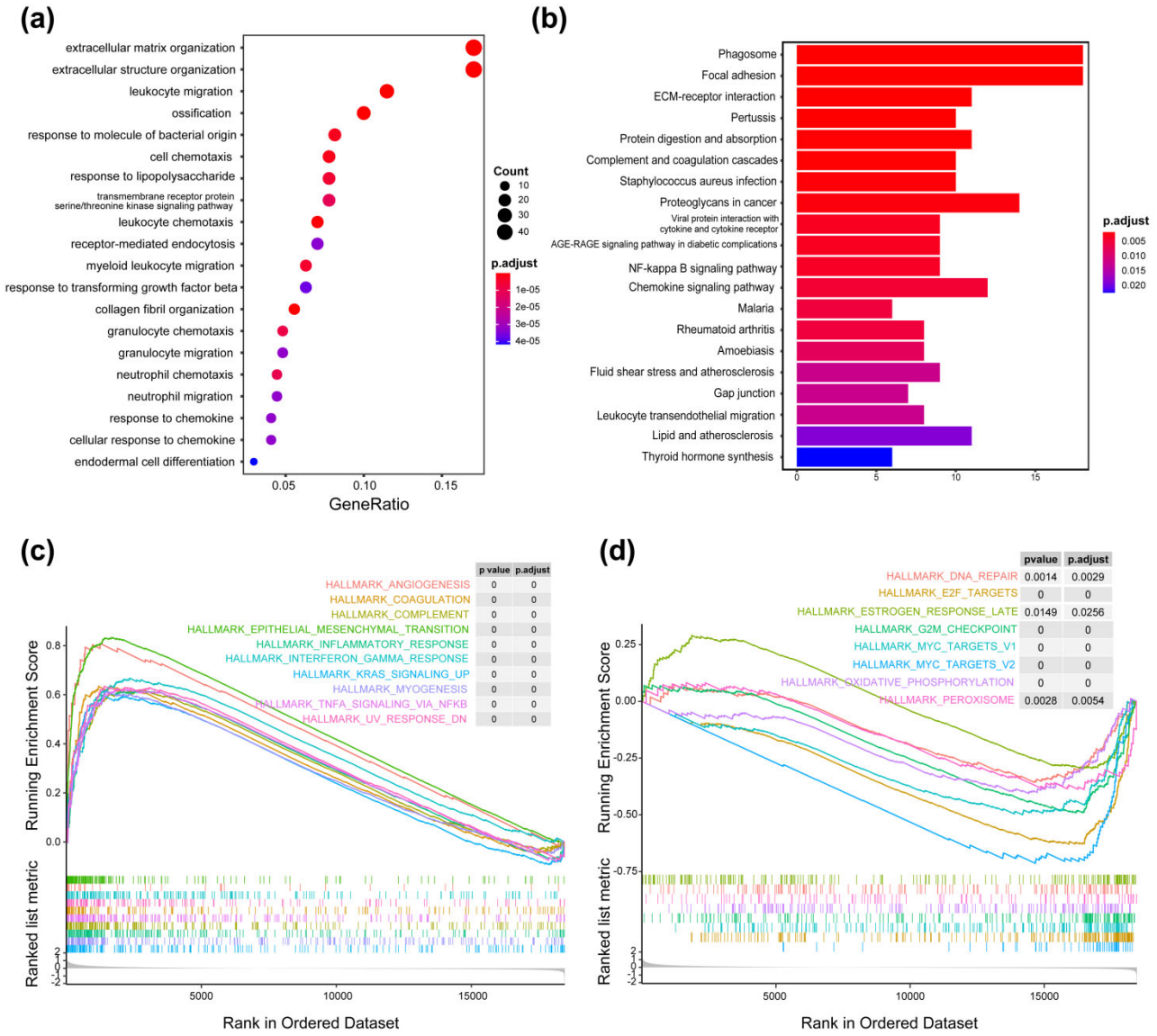

Fig. S4 Functional implications of the prognostic CSbMgSig in the GSE39582 dataset.

(a) Bubble diagram shows the top 20 enriched GO terms for the DEGs identified between high- and low-risk groups. (b) Barplot shows the top 20 enriched KEGG pathways for the DEGs identified between high- and low-risk groups. (c) and (d). Significant enriched hallmarks in the high-risk group (c) or low-risk group (d) identified by GSEA method.

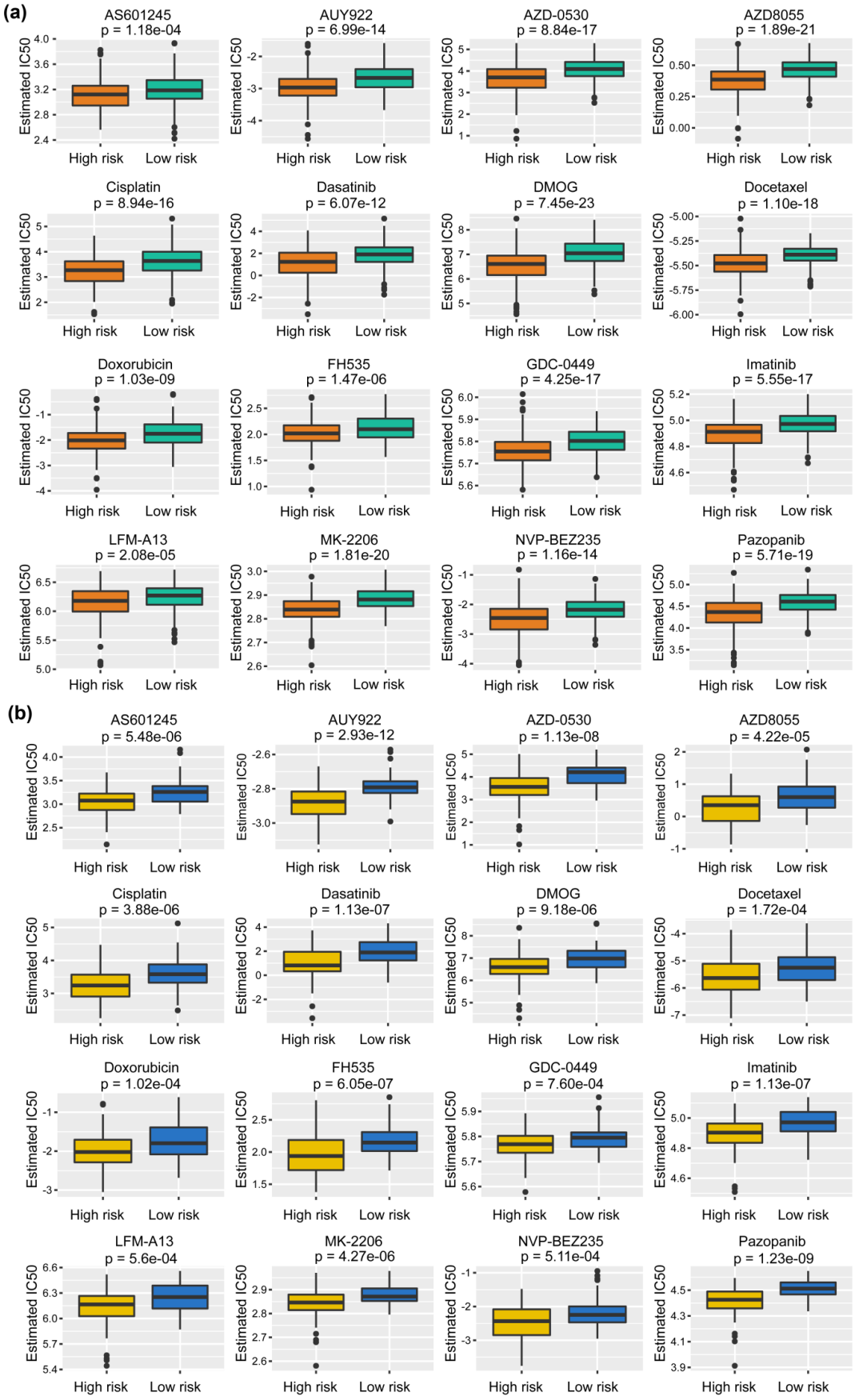

Fig. S5 Comparison of the chemotherapeutic responses in high- and low-risk patients.

Estimated IC_{50} values of CRC chemotherapeutic drugs were compared for high- and low-risk patients in the GSE39582 dataset (a) and GSE17536 dataset (b). *P* values were calculated by the Wilcoxon test.