Machine Learning Models For Predicting Lymph Node And Distant Metastases In Colorectal Cancer

Xiaoyu Dai
Ningbo No 2 Hospital
Siqi Dai
Zhejiang University School of Medicine Second Affiliated Hospital
Xi Yang
Huzhou Central Hospital
Jing Zhuang
Huzhou University
Jin Liu
Huzhou Central Hospital
Shuwen Han (shuwenhan985@163.com)
Huzhou Central Hospital https://orcid.org/0000-0002-1119-4010

Research article

Keywords: Colorectal cancer, lymph node metastases, distant metastases, machine learning models

DOI: https://doi.org/10.21203/rs.3.rs-45372/v1

Abstract

Background: Colorectal cancer (CRC) is the third most common malignancy in the world and metastasis is responsible for a major proportion of the cancer-related deaths in CRC patients.

Aims: To construct machine learning models for predicting lymph node and distant metastases in colorectal cancer and analyze biological functions features of metastasis-related genes.

Methods: RNA-seq and miRNA-seq data as well as corresponding clinical data from colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) were obtained from The Cancer Genome Atlas (TCGA) database. The differentially expressed RNAs (DE-RNAs) in non-LNM (N0) and LNM (N1/N2) as well as non-distant metastases (M0) and distant metastases (M1) were analyzed. Six machine learning models including logistic regression (LR), random forest (RF), support vector machine (SVM), Catboost, gradient boosting decision tree (GBDT), and artificial neural network (NN) were constructed to predict cancer metastasis and the feature genes of the optimal model were further analyzed by functional enrichment, protein-protein interaction (PPI) network, and drug-target analyses.

Results: Differential RNA expression profiles of LNM and non-LNM as well as M0 vs. M1 were observed in both COAD and READ samples. NN model was determined to be the optimal model for predicting distant metastases, while Catboost and LR models were the optimal models for predicting LNM in COAD and READ samples, respectively. PPI analysis indicated that KIR2DL4, chemokine-related genes CXCL9/10/11/13 and CCL25, and gamma-aminobutyric acid (GABA) receptor genes (GABRR1, GABRB2 and GABRA3) were key genes in metastasis. In addition, atorvastatin and eszopiclone were identified as potential therapeutic agents as they target these genes.

Conclusions: We constructed six machine learning models for predicting colorectal cancer metastases and identify the optimal model. We analyzed biological functions features of metastasis-related RNAs in colorectal cancer.

Highlights

1. We analyzed the differential RNA expression related to profiles metastasis-related molecules in colorectal cancer.
2. We constructed six machine learning models for predicting lymph node and distant metastases in colorectal cancer and identify the optimal model.
3. We analyzed biological functions features of metastasis-related RNAs in colorectal cancer.

Introduction

Colorectal cancer (CRC) is the third most common malignancy in the world. In the United States, clinicians observed a decreased CRC incidence rate in adults older than 50 years of age between 2000 and 2014, especially in distal tumors of individuals aged ≥ 65 years, while an increased incidence rate

was observed in people less than 50 years of age during the same period [1]. The incidence and mortality of CRC vary between human development levels, with both increasing rapidly in low/middle-income countries while highly developed countries seem to be experiencing a decrease in these parameters. It is estimated that the global burden of CRC will increase by 60% by the year 2030 with over 2.2 million new cases and 1.1 million deaths per year [2]. Tumor cells from primary tumors can migrate to regional lymph nodes and distant organs. Metastasis is responsible for most cancer-related deaths in patients with CRC [3]. Although most local CRC patients can be cured using surgical resection, only around 70% of CRC patients with regional lymph node metastasis (LNM) can be cured by surgery coupled with adjuvant chemotherapy, and advanced metastatic CRCs are still mostly incurable, in spite of the improvements in medical treatment [4, 5]. Thus, the accurate prediction of metastasis in CRC is crucial for its treatment.

Recently, aberrant expression of genes, and several kinds of non-coding RNAs have been demonstrated to function in CRC metastasis and could serve as predictive biomarkers for this condition [6–8]. For example, Schmid *et al*. showed that the high expression of Spondin 2, a transcriptional target of the metastasis gene *MACC1*, was associated with specific prognostic outcomes, and that overexpression of Spondin 2 could promote CRC metastasis both *in vitro* and *in vivo* [8]. Hur *et al*. showed that patients with elevated serum miR-203 levels were at higher risk for developing LNM and distant organs metastasis, and serum miR-203 levels obviously increased in CRC metastasis mouse models when compared to those of the control [9]. Yoshida *et al*. suggested that small nucleolar RNA 21 (SNORA21) was significantly up-regulated in CRC using both RNA profiling datasets and clinical validations, and proliferation and invasion of CRC cells were reduced following inhibition of SNORA21 [10]. High expression of long non-coding RNA (lncRNA) XIST has been reported in CRC cells and tissues, and XIST has been shown to regulate tumor growth and metastasis via the miR-200b-3p-ZEB1 axis.

Machine learning has been shown to be extremely accurate and precise when used to predict medical outcomes outstripping standard statistics and human judgment [11, 12]. A previous review showed that machine learning models displayed excellent performance in predicting the outcomes of various neurosurgical conditions, outstripping established prognostic indicators and clinical experts [13]. Andrés *et al*. reported that machine learning showed increased sensitivity and specificity when predicting nodal metastasis compared to the tumor depth invasion model developed using early oral squamous cell carcinoma, and the forest algorithm was shown to be the most efficient model for this data [14]. Zhi *et al*. revealed that the genes identified by the support vector machine (SVM) classifier algorithm could accurately distinguish metastatic and non-metastatic CRC samples [15]. Artificial neural network (NN) can be used to realize effective analysis of non-linear datasets and have been successfully used to assist clinical decision-making in neurosurgery [16]. In addition, it has been reported that NN is better at predicting relapse in breast cancer than the logistic regression (LR) model [17]. However, there are no studies on the applications of different machine learning models to CRC metastasis.

In this study, RNA expression data and clinical phenotype data from The Cancer Genome Atlas (TCGA) database, was analyzed using six machine learning models, including LR, random forest (RF), SVM, Catboost, gradient boosting decision tree (GBDT), and NN to predict CRC metastasis. Among the six

models, the genes identified by the optimal models were further analyzed using functional enrichment analysis, protein-protein interaction (PPI) network and drug-target analysis (Supplemental Fig. 1 showing the workflow). This study could provide valuable insight for the treatment of CRC by providing biomarkers for the prediction of CRC metastasis.

Materials And Methods

Data acquisition and preprocessing

The sequencing data from RNA (RNA-seq) and microRNA (miRNA-seq) experiments as well as the corresponding clinical data of TCGA-colon adenocarcinoma (TCGA-COAD) and TCGA-rectum adenocarcinoma (TCGA-READ) were downloaded from the university of California Santa Cruz (UCSC, http://xena.ucsc.edu/) Genome Browser database. Based on the annotation file (gencode.v22.annotation.gene) available from the GENCODE database, the RNA-seq and miRNA-seq data were annotated to identify mRNA, lncRNAs, snoRNAs, and miRNA transcripts.

Differential expression analysis

For both COAD and READ data sets, the differential expression analysis of non-LNM (N0) and LNM (N1/N2), and non-distant (M0) and distant metastases (M1) were performed using the limma package. The differentially expressed mRNA (DE- mRNA), lncRNAs (DE- lncRNAs), snoRNAs (DE- snoRNAs), and miRNAs (DE-miRNAs) with a P value < 0.05 and |log fold change (FC)| > 0.585 or 1 (the number of feature genes used in the construction of the data models must be less than the number of samples, therefore the thresholds were different when identifying the candidate feature genes) were selected. The ggpubr package was used to visualize the volcano plot.

Training of the optimal model and performance evaluation

The count value of the candidate feature genes was standardized to log2 (x+1) data and binary labels were added to each sample: metastases was 1 while non-metastases was 0. After that, samples were divided into a training (80% samples) and a test (20% samples) dataset using the train_test_split method from the scikit learn package in python. The sklearn.linear_model, sklearn.ensemble, sklearn.svm and sklearn.neural_network methods were used to construct LR, RF, GBDT, SVM and NN machine learning models, and the Catboost package was used to construct the Catboost machine learning model. The expression value of the feature genes in the samples were used as the feature value to classify and discriminate the samples. The Recursive Feature Elimination (RFE) algorithm was implemented in the sklearn.feature_selection method. After cross-validation, the optimal feature genes from each model were identified based on the area under the receiver operating characteristic (ROC) curve (AUC). The LNM and distant metastases in COAD and READ were predicted using the six models, and the model with the highest AUC was selected as the optimal model. The formula or code for all six models used in this study are shown in Supplemental files 1-4. The feature genes of the optimal model were used in the following analyses.

Functional enrichment analysis

To investigate the biological function of the feature genes, the Gene Ontology (GO) annotation terms and KEGG pathways were enriched using the clusterProfiler tool with a cut-off value of $P < 0.05$ and count ≥ 2.

Construction of the PPI network

The interactions between the feature genes were retrieved using the STRING database (Version: 11.0, http://www.string-db.org/) with the parameters set as follows: (1) 0.4 (medium confidence) PPI score; (2) species: Homo sapiens; and (3) disable structure previews inside network bubbles, hide disconnected nodes in the network. The PPI network was visualized using Cytoscape software.

Prediction of drug-target pairs

The drug-target pairs for the feature genes were predicted using the DGIdb 3.0 database. The drug-target pairs with FDA approval or those reported in published studies were selected and visualized using Cytoscape software.

Results

RNAs were differentially expressed in LNM and non-LNM samples

A total of 438 COAD samples with RNA-seq, miRNA-seq and clinical data were included, of which 255 were N0 and 183 were N1/N2 samples. In total, 352 RNAs were differentially expressed between LNM and non-LNM samples, including 263 unique DE-mRNAs, 60 DE-lncRNAs, 10 DE-miRNAs and 19 DE-undefined genes (Figure 1A and Table 1).

For READ, a total of 160 samples with RNA-seq, miRNA-seq and clinical data were included, of which 80 were N0 samples and 76 were N1/N2 samples. A total of 474 RNAs were differentially expressed in LNM vs. non-LNM samples, including 244 unique DE-mRNAs, 87 DE-lncRNAs, 26 DE-miRNAs and 117 DE-undefined genes (Figure 1B and Table 1).

RNAs were differentially expressed in distant metastases and non-distant metastases samples

Among the 438 COAD samples, there were 316 M0 and 63 M1 samples. In all, 129 RNAs were differentially expressed in M0 and M1 samples, including 81 unique DE-mRNAs, 22 DE-lncRNAs, 5 DE-miRNAs, and 21 DE-undefined genes (Figure 1C and Table 1).

Among the 160 READ samples, there were 121 M0 and 22 M1 samples. A total of 134 RNAs were differentially expressed in the M0 and M1 samples, including 90 unique DE-mRNAs, 34 DE-lncRNAs, and 10 DE-undefined genes (Figure 1D and Table 1).

Identifying the optimal model to predict LNM in COAD and READ samples

For LNM prediction in COAD we evaluated the AUC for each of the six machine learning models. They were as follows: 0.7671 for the LR model, 0.7722 for the NN model, 0.7532 for the SVM model, 0.7610 for the RF model, 0.7634 for the GBDT model, and 0.8040 for the Catboost model (Figure 2A). Therefore, Catboost was identified as the optimal model for the prediction of LNM in COAD samples. There were 236 feature genes identified by the Catboost model (Supplemental Table 1), of which *C6orf15* and *CXCL11* were shown to make the largest contribution (Supplemental Figure 2).

For LNM prediction in READ samples, the LR model showed the largest AUC at 0.9254 and thus was identified as the optimal model (Figure 2B). A total of 292 feature genes were identified by this model (Supplemental Table 2), and the significance of each of these genes is shown in Supplemental Figure 3.

Identification of the optimal model to predict distant metastases in COAD and READ samples

For the prediction of distant metastases in COAD, the AUC for all six models were evaluated. The values were as follows: 0.6914 for the LR model, 0.8047 for the NN model, 0.6432 for the SVM model, 0.6992 for the RF model, 0.6406 for the GBDT model, and 0.6953 for the Catboost model (Figure 2C). The NN model was thus identified as the optimal model to predict distant metastases in COAD. There were 129 feature genes identified by the NN model.

For the prediction of distant metastases in READ samples, the NN model was also identified as offering the best AUC value (0.8600) making it the optimal model (Figure 2D). A total of 134 feature genes were identified by the NN model.

Biological functions of the feature mRNAs identified by the optimal models

The feature mRNAs identified by the optimal models are listed in Table 2. In order to investigate the biological functions of these mRNAs, GO annotation and KEGG pathway enrichment was analyzed. For the feature mRNAs distinguished by the Catboost model for LNM prediction in COAD, we observed a significant enrichment in calcium ion homeostasis, calcium ion transport into the cytosol, the calcium signaling pathway amongst others (Figure 3A).

The mRNAs identified by the LR model for LNM prediction in READ were implicated in retinoid/diterpenoid/ metabolic process, cholesterol metabolism, and remodeling related biological processes, including, protein–containing complex remodeling, protein–lipid complex remodeling, plasma lipoprotein particle remodeling (Figure 3B).

The mRNAs identified by the NN model for distant metastases prediction in COAD were found to participate in terms that included regulation of transmembrane transport/ion transport, T cell chemotaxis, chemokine signaling pathway amongst others (Figure 4A). For the mRNAs identified by the NN model for distant metastases prediction in READ, the enriched results mainly contained killing of cells of other organisms, disruption of cells of other organisms, triglyceride–rich lipoprotein particle remodeling, and cholesterol/thiamine metabolism (Figure 4B).

PPI network from the feature mRNAs identified by the optimal models

PPI analysis was performed on the feature mRNAs identified by the optimal models, and the mRNAs were resolved using the STRING database were listed in Supplemental Table 3. Isthmin 2 (ISM2) and killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 4 (KIR2DL4) were the only overlapping mRNAs between all four groups. Figure 5A shows the PPI network of the mRNAs identified in the Catboost prediction model for LNM in READ. Among the mRNAs in figure 5A, *C6orf15*, an uncharacterized protein located on chromosome 6 open reading frame 15, had the greatest log FC value (positively correlated with node size) in the differential expression analyses. While GNG4, G protein subunit gamma 4, interacted with the most proteins. The PPI network of the mRNAs identified by the LR prediction model for LNM in READ is shown in Figure 5B and identifies KIR2DL4 as the most significant component.

The PPI network of the mRNAs identified by the NN model for distant metastases prediction in COAD is shown in Figure 6A, and was shown to include several chemokine genes, including C-X-C motif chemokine ligand 9 (CXCL9), CXCL10, CXCL11, CXCL13 and C-C motif chemokine ligand 25 (CCL25). The PPI network of the mRNAs identified by the NN model for distant metastases prediction in READ is shown in Figure 6B. Regenerating family member 3 alpha (REG3A), defensin alpha 5 (DEFA5), and DEFA6 were shown to have the greatest log FC (positively correlated with node size) values in the differential expression analysis.

Drug-target pairs for feature mRNAs

For LNM prediction in COAD, 264 drug-gene pairs were predicted to target 29 feature mRNAs (Figure 7A and Supplemental table 4). Gamma-aminobutyric acid type A receptor alpha3 subunit (GABRA3), cholinergic receptor muscarinic 2 (CHRM2), Fc fragment of IgG receptor IIIb (FCGR3B), and dopamine receptor D2 (DRD2) were targeted by the most drugs. A total of 9 drugs were found to target CXCL10. From the LNM prediction in READ, 113 drug-gene pairs were predicted to target 12 feature mRNAs (Figure 7B and Supplemental table 5). GABRB2 was targeted by 11 drugs, and all 11 drugs were potentiators of GABRB2. Cyclosporine and deferoxamine were found to target CXCL2.

For the distant metastases prediction in COAD, 83 drug-gene pairs were predicted to target 17 feature mRNAs (Figure 8A and Supplemental table 6). For the distant metastases prediction in READ, 119 drug-gene pairs were predicted to target 17 feature mRNAs (Figure 8B and Supplemental table 7). A total of 9 drugs were found to target CXCL10. GABRA3, 5-hydroxytryptamine receptor 1D (HTR1D), HTR2C, and solute carrier family 6 member 4 (SLC6A4) were targeted by the most drugs, and the interaction types were all known.

Discussion

Machine learning has recently been shown to be an accurate and precise method for predicting medical outcomes outstripping standard statistics and human judgment. In this study, the NN model was

identified as the optimal model for predicting distant metastases, while Catboost and LR models were shown to be optimal models for predicting LNM in COAD and READ samples, respectively. Consistent with these observations Biglarian *et al*. showed that the ROC for NN and LR models predicting distant metastasis of CRC were 0.82 and 0.77, respectively, suggesting that the NN model was more suitable for the prediction of distant metastasis in CRC [18]. The feature genes from the optimal models were significantly enriched in calcium ion homeostasis, transmembrane transport/ion transport, T cell chemotaxis, chemokine signaling pathway amongst others. While the PPI analysis indicated that KIR2DL4, chemokine-related genes (CXCL9/10/11/13 and CCL25), and gamma-aminobutyric acid (GABA) receptor genes (GABRR1, GABRB2 and GABRA3) are all key genes in metastasis. In addition, atorvastatin and eszopiclone were shown to target those genes, suggesting they may have some therapeutic value.

KIR2DL4, also known as CD158D, is a member of the killer cell immunoglobulin-like receptor (KIR) family expressed by natural killer (NK) cells [19]. KIRs can recognize histocompatibility complex (MHC) ligands and mediate the function of NK cells. KIR2DL4 functions as an inhibitory receptor that releases inhibitory signals to NK cells [19, 20]. HLA-G, a non-classical class I human leukocyte antigen, is the only known ligand for KIR2DL4 [19]. Notably, expression of HLA-G has been reported to be a predisposing factor for metastasis [21]. Studies have revealed the associations between HLA-G expression and lymph node metastasis in cervical cancer [22], papillary thyroid cancer [23], and CRC-associated liver metastases [24]. Reportedly, HLA-G upregulates the expression of matrix metalloproteinases (MMPs) and other tumor promoting factors, so that tumor cells have a higher invasion and metastasis potential [25]. Ueshima *et al*. suggested that KIR2DL4 + tissue mast cells promote LNM and lymph-vascular invasion in HLA-G + breast cancer cells [26]. Thus, we hypothesize that KIR2DL4 plays a crucial role in CRC metastases probably via interactions with its ligand HLA-G.

Studies have shown that chemokines play important roles in clinical outcome prediction and invasion/metastasis of various cancers [27, 28]. For example, CXCL10 elevates the expression of MMP9, and triggers cell migration and invasion in metastatic CRC cells rather than primary cancer cells [29]. Tokunaga *et al*. revealed that the chemokine CXCL9/CXCL10/CXCL11/CXCR3 axis mediates differentiation and migration of immune cells via the paracrine axis and promotes cancer metastasis via the autocrine axis, suggesting that this axis is a potential target for anti-cancer therapies [30]. This is consistent with the results of our study where we found that aberrant expression of CXCL9/CXCL10/CXCL11/CXCL13/CCL25 were identified as features in the NN model used to predict distant metastases and were enriched in T cell chemotaxis and chemokine signaling pathways. We suggest that this implicates these chemokines in the regulation of CRC metastases. Notably, several drugs were found to target CXCL10, for example, atorvastatin. Atorvastatin has been reported to decrease plasma CXCL10 levels in the treatment of Crohn's disease [31], and decrease plasma CXCL9 levels in the treatment of systemic lupus erythematosus [32]. In addition, combined treatment with atorvastatin and phloretin induces apoptosis and G2/M arrest in colon cancer cells [33]. Thus we speculate that atorvastatin may be a promising drug for targeting chemokines during the treatment of CRC.

Gamma-aminobutyric acid (GABA) is an inhibitory neurotransmitter which interacts with two types of receptors, including GABA A and GABAB. GABAA receptors are ionotropic receptors consisting of various subunits and functioning as chloride channels [34, 35]. GABA has been implicated in cancer metastasis [36, 37]. GABRR1, GABRB2 and GABRA3 are all subunits of the GABAA receptor. Liu *et al*. suggested that GABRA3 could promote LNM in lung adenocarcinoma by inducing MMP-2 and MMP-9 expression [38]. In this study, GABRR1, GABRB2 and GABRA3 were all aberrantly expressed in CRC, and were among the feature genes identified by the optimal machine learning models for the prediction of CRC metastasis. They were enriched in ion transport and transporter activity related terms. Reportedly, ion channels/transporters are important operators of various cell-cell signaling pathways as they sense and respond to changes in the environment. Ion channels/transporters participate in each step of the cascade in cancer metastasis, suggesting that they could be potential therapeutic targets in cancer therapy and metastasis prevention [39–41]. In this study, various drugs were predicted to target GABRR1, GABRB2 and GABRA3. These included eszopiclone, which is an agonist/positive allosteric modulator of GABRA3. Studies have shown the interactions between eszopiclone and GABA receptors [42, 43]. However, despite these novel findings, this study did have some limitations. (1) We analyzed the expression of different kinds of RNAs in CRC, and the role of differentially expressed mRNAs in the prediction of CRC metastasis. These analyses should be extended to the effects of the differential expression of the various miRNAs, lncRNAs and snoRNAs identified in previous CRC metastasis studies. (2) The effect of the key mRNAs identified in the machine models should be validated by experimental and clinical data. (3) The predicted drug-gene interactions should be confirmed to provide insight into their application in CRC treatment and the prevention of CRC metastasis.

Conclusions

We constructed multiple machine learning models for colorectal cancer metastasis, identified the optimal model, and applied the optimal model to analyze the valuable RNAs related to colorectal cancer metastasis. The machine learning models could contribute to the prediction of LNM and distant metastases in CRC. This study might provide a novel routine for screening the promising targets for CRC treatment and the prevention of CRC metastases.

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing Interests

The authors declare that no conflicts of interest exist.

Funding

This work was supported by the Major Science and Technology Projects for Medical and Health Care of Zhejiang Province (No. WKJ-ZJ-2013), Huzhou Public Welfare Application Research Project (No.2019GZ39).

Authors' Contributions

All authors participated in the conception and design of the study;

Conceived the manuscript: Han Shuwen and Yang Xi;

Wrote the paper: Han Shuwen, Yang Xi and Zhuang Jing;

Processed the data: Dai Xiaoyu and Dai Siqi;

Drew figures: Zhuang Jing and Liu Jin;

All authors read and approved the paper.

Acknowledgements

The authors gratefully acknowledge the database available to us for this study. **Availability of data and materials**

The datasets generated during the current study are not publicly available but obtained from corresponding authors on reasonable request.

References

1. R.L. Siegel, K.D. Miller, S.A. Fedewa, D.J. Ahnen, R.G.S. Meester, A. Barzi, A. Jemal, Colorectal cancer statistics, 2017, CA: A Cancer Journal for Clinicians, 67 (2017) 177-193.
2. M. Arnold, M.S. Sierra, M. Laversanne, I. Soerjomataram, A. Jemal, F. Bray, Global patterns and trends in colorectal cancer incidence and mortality, Gut, 66 (2017) 683-691.
3. E. Van Cutsem, A. Cervantes, B. Nordlinger, D. Arnold, Metastatic colorectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up, Annals of oncology, 25 (2014) iii1-iii9.
4. M.G. Fakih, Metastatic Colorectal Cancer: Current State and Future Directions, Journal of Clinical Oncology, 33 (2015) 1809-1824.
5. H. Ling, K. Pickard, C. Ivan, C. Isella, M. Ikuo, R. Mitter, R. Spizzo, M.D. Bullock, C. Braicu, V. Pileczki, The clinical and biological significance of MIR-224 expression in colorectal cancer metastasis, Gut, 65 (2016) 977-989.

6. E. Saus, A. Brunet-Vega, S. Iraola-Guzmán, C. Pegueroles, T. Gabaldón, C. Pericay, Long non-coding RNAs as potential novel prognostic biomarkers in colorectal cancer, Frontiers in genetics, 7 (2016) 54.
7. I. Garajová, M. Ferracin, E. Porcellini, A. Palloni, F. Abbati, G. Biasco, G. Brandi, Non-coding RNAs as predictive biomarkers to current treatment in metastatic colorectal cancer, International journal of molecular sciences, 18 (2017) 1547.
8. F. Schmid, Q. Wang, M. Huska, M. Andrade-Navarro, M. Lemm, I. Fichtner, M. Dahlmann, D. Kobelt, W. Walther, J. Smith, SPON2, a newly identified target gene of MACC1, drives colorectal cancer metastasis in mice and is prognostic for colorectal cancer patient survival, Oncogene, 35 (2016) 5942-5952.
9. K. Hur, Y. Toiyama, Y. Okugawa, S. Ide, H. Imaoka, C.R. Boland, A. Goel, Circulating microRNA-203 predicts prognosis and metastasis in human colorectal cancer, Gut, 66 (2017) 654-665.
10. K. Yoshida, S. Toden, W. Weng, K. Shigeyasu, J. Miyoshi, J. Turner, T. Nagasaka, Y. Ma, T. Takayama, T. Fujiwara, A. Goel, SNORA21 – An Oncogenic Small Nucleolar RNA, with a Prognostic Biomarker Potential in Human Colorectal Cancer, EBioMedicine, 22 (2017) 68-77.
11. Z. Obermeyer, E.J. Emanuel, Predicting the future—big data, machine learning, and clinical medicine, The New England journal of medicine, 375 (2016) 1216.
12. A.M. Bur, M. Shew, J. New, Artificial intelligence for the otolaryngologist: a state of the art review, Otolaryngology–Head and Neck Surgery, 160 (2019) 603-611.
13. J.T. Senders, P.C. Staples, A.V. Karhade, M.M. Zaki, W.B. Gormley, M.L.D. Broekman, T.R. Smith, O. Arnaout, Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review, World Neurosurgery, 109 (2018) 476-486.e471.
14. A.M. Bur, A. Holcomb, S. Goodwin, J. Woodroof, O. Karadaghy, Y. Shnayder, K. Kakarala, J. Brant, M. Shew, Machine learning to predict occult nodal metastasis in early oral squamous cell carcinoma, Oral Oncology, 92 (2019) 20-25.
15. J. Zhi, J. Sun, Z. Wang, W. Ding, Support vector machine classifier for prediction of the metastasis of colorectal cancer, International journal of molecular medicine, 41 (2018) 1419-1426.
16. P. Azimi, H.R. Mohammadi, E.C. Benzel, S. Shahzadi, S. Azhari, A. Montazeri, Artificial neural networks in neurosurgery, Journal of Neurology, Neurosurgery & Psychiatry, 86 (2015) 251.
17. J. Faradmal, A.R. Soltanian, G. Roshanaei, R. Khodabakhshi, A. Kasaeian, Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse, Asian Pac J Cancer Prev, 15 (2014) 5883-5888.
18. A. Biglarian, E. Bakhshi, M.R. Gohari, R. Khodabakhshi, Artificial neural network for prediction of distant metastasis in colorectal cancer, Asian Pacific Journal of Cancer Prevention, 13 (2012) 927-930.
19. S. Rajagopalan, E.O. Long, KIR2DL4 (CD158d): an activation receptor for HLA-G, Frontiers in immunology, 3 (2012) 258.

20. T.R. Kataoka, C. Ueshima, M. Hirata, S. Minamiguchi, H. Haga, Killer Immunoglobulin-Like Receptor 2DL4 (CD158d) Regulates Human Mast Cells both Positively and Negatively: Possible Roles in Pregnancy and Cancer Metastasis, International Journal of Molecular Sciences, 21 (2020) 954.
21. I. Zidi, N.B. Amor, HLA-G as predisposing for metastasis, Medical hypotheses, 77 (2011) 134-139.
22. D.M. Ferns, A.M. Heeren, S. Samuels, M.C. Bleeker, T.D. de Gruijl, G.G. Kenter, E.S. Jordanova, Classical and non-classical HLA class I aberrations in primary cervical squamous-and adenocarcinomas and paired lymph node metastases, Journal for immunotherapy of cancer, 4 (2016) 78.
23. L.M. Nunes, F.M. Ayres, I.C.M. Francescantonio, V.A. Saddi, M.A.G. Avelino, R.d.C.G. Alencar, R.C.d. Silva, A.J. Meneghini, I.J. Wastowski, Association between the HLA-G molecule and lymph node metastasis in papillary thyroid cancer, Human Immunology, 74 (2013) 447-451.
24. M. Swets, M.H. König, A. Zaalberg, N.G. Dekker-Ensink, H. Gelderblom, C.J.H. van de Velde, P.J. van den Elsen, P.J.K. Kuppen, HLA-G and classical HLA class I expression in primary colorectal cancer and associated liver metastases, Human Immunology, 77 (2016) 773-779.
25. A. Lin, W.-H. Yan, Human leukocyte antigen-G (HLA-G) expression in cancers: roles in immune evasion, metastasis and target for therapy, Molecular Medicine, 21 (2015) 782-791.
26. C. Ueshima, T.R. Kataoka, M. Hirata, A. Furuhata, E. Suzuki, M. Toi, T. Tsuruyama, Y. Okayama, H. Haga, The killer cell Ig-like receptor 2DL4 expression in human mast cells and its potential role in breast cancer invasion, Cancer immunology research, 3 (2015) 871-880.
27. A. Mitchell, S.L. Hasanali, D.S. Morera, R. Baskar, X. Wang, R. Khan, A. Talukder, C.S. Li, M. Manoharan, A.R. Jordan, A chemokine/chemokine receptor signature potentially predicts clinical outcome in colorectal cancer patients, Cancer Biomarkers, (2019) 1-11.
28. Y. Itatani, K. Kawada, S. Inamoto, T. Yamamoto, R. Ogawa, M.M. Taketo, Y. Sakai, The role of chemokines in promoting colorectal cancer invasion/metastasis, International journal of molecular sciences, 17 (2016) 643.
29. A. Zipin-Roitman, T. Meshel, O. Sagi-Assif, B. Shalmon, C. Avivi, R.M. Pfeffer, I.P. Witz, A. Ben-Baruch, CXCL10 Promotes Invasion-Related Properties in Human Colorectal Carcinoma Cells, Cancer Research, 67 (2007) 3396.
30. R. Tokunaga, W. Zhang, M. Naseem, A. Puccini, M.D. Berger, S. Soni, M. McSkane, H. Baba, H.-J. Lenz, CXCL9, CXCL10, CXCL11/CXCR3 axis for immune activation – A target for novel cancer therapy, Cancer Treatment Reviews, 63 (2018) 40-47.
31. O. Grip, S. Janciauskiene, Atorvastatin reduces plasma levels of chemokine (CXCL10) in patients with Crohn's disease, PLoS One, 4 (2009) e5263-e5263.
32. G. Ferreira, A. Teixeira, E. Sato, Atorvastatin therapy reduces interferon-regulated chemokine CXCL9 plasma levels in patients with systemic lupus erythematosus, Lupus, 19 (2010) 927-934.
33. M. Zhou, J. Zheng, J. Bi, X. Wu, J. Lyu, K. Gao, Synergistic inhibition of colon cancer cell growth by a combination of atorvastatin and phloretin, Oncology letters, 15 (2018) 1985-1992.

34. K. Gumireddy, A. Li, A.V. Kossenkov, M. Sakurai, J. Yan, Y. Li, H. Xu, J. Wang, P.J. Zhang, L. Zhang, L.C. Showe, K. Nishikura, Q. Huang, The mRNA-edited form of GABRA3 suppresses GABRA3-mediated Akt activation and breast cancer metastasis, Nature Communications, 7 (2016) 10715.
35. M. Vithlani, M. Terunuma, S.J. Moss, The dynamic modulation of GABAA receptor trafficking and its role in regulating the plasticity of inhibitory synapses, Physiological reviews, 91 (2011) 1009-1022.
36. H. Azuma, T. Inamoto, T. Sakamoto, S. Kiyama, T. Ubai, Y. Shinohara, K. Maemura, M. Tsuji, N. Segawa, H. Masuda, γ-Aminobutyric acid as a promoting factor of cancer metastasis; induction of matrix metalloproteinase production is potentially its underlying mechanism, Cancer research, 63 (2003) 8090-8096.
37. P.H. Thaker, K. Yokoi, N.B. Jennings, Y. Li, R.B. Rebhun, D.L. Rousseau Jr, D. Fan, A.K. Sood, Inhibition of experimental colon cancer metastasis by the GABA-receptor agonist nembutal, Cancer biology & therapy, 4 (2005) 753-758.
38. L. Liu, C. Yang, J. Shen, L. Huang, W. Lin, H. Tang, W. Liang, W. Shao, H. Zhang, J. He, GABRA3 promotes lymphatic metastasis in lung adenocarcinoma by mediating upregulation of matrix metalloproteinases, Oncotarget, 7 (2016) 32341.
39. C. Stock, A. Schwab, Ion channels and transporters in metastasis, Biochimica et Biophysica Acta (BBA) - Biomembranes, 1848 (2015) 2638-2646.
40. S. Martial, Involvement of ion channels and transporters in carcinoma angiogenesis and metastasis, American Journal of Physiology-Cell Physiology, 310 (2016) C710-C727.
41. M.B. Djamgoz, R.C. Coombes, A. Schwab, Ion transport and cancer: from initiation to metastasis, in, The Royal Society, 2014.
42. C.L. Dixon, N.L. Harrison, J.W. Lynch, A. Keramidas, Zolpidem and eszopiclone prime α1β2γ2 GABAA receptors for longer duration of activity, British journal of pharmacology, 172 (2015) 3522-3536.
43. S.V. Fox, A.L. Gotter, S.J. Tye, S.L. Garson, A.T. Savitz, J.M. Uslaner, J.I. Brunner, P.L. Tannenbaum, T.P. McDonald, R. Hodgson, Quantitative electroencephalography within sleep/wake states differentiates GABA A modulators eszopiclone and zolpidem from dual orexin receptor antagonists in rats, Neuropsychopharmacology, 38 (2013) 2401-2408.

Tables

Table 1
Statistics of differentially expressed genes.

	COAD (Lymph nodes metastasis)	READ (Lymph nodes metastasis)	COAD (Distant metastases)	READ (Distant metastases)
Log FC	0.585	0.585	1	1
Up probes	310	328	88	112
Down probes	42	146	41	22
Total probes	352	474	129	134
Not matched probes	62012	61890	62235	62230
mRNA (up/down)	263(230/33)	244(191/53)	81(58/23)	90(74/16)
lncRNA (up/down)	60(54/6)	87(45/42)	22(20/2)	34(29/5)
snoRNA (up/down)	0	0	0	0
miRNA (up/down)	10(10/0)	26(26/)	5(5/0)	0
Undefined (up/down)	19(16/3)	117(66/51)	21(5/16)	10(9/1)
Unique mRNA	263	244	81	90

The differentially expressed mRNAs, lncRNAs, snoRNAs, miRNAs and undefined genes between N0 vs. N1/N2 in COAD, N0 vs. N1/N2 in READ, M0 vs. M1 in COAD, and M0 vs. M1 in READ. Up, up-regulated genes; Down, down-regulated genes; COAD, colon adenocarcinoma; READ, rectum adenocarcinoma.

Table 2
Statistics of feature mRNAs distinguished by optimal models.

	COAD (Lymph nodes metastasis)	READ (Lymph nodes metastasis)	COAD (Distant metastases)	READ (Distant metastases)
Optimal model	Catboost	LR	NN	NN
Feature mRNA	178	149	81	90
Trans mRNA	177	149	80	90
No dup	177	149	80	90

Trans mRNA, the mRNAs converted by clusterprofier package and org.hs.eg.db package; No dup, the amount of mRNAs after removal of duplications; COAD, colon adenocarcinoma; READ, rectum adenocarcinoma.

Figures

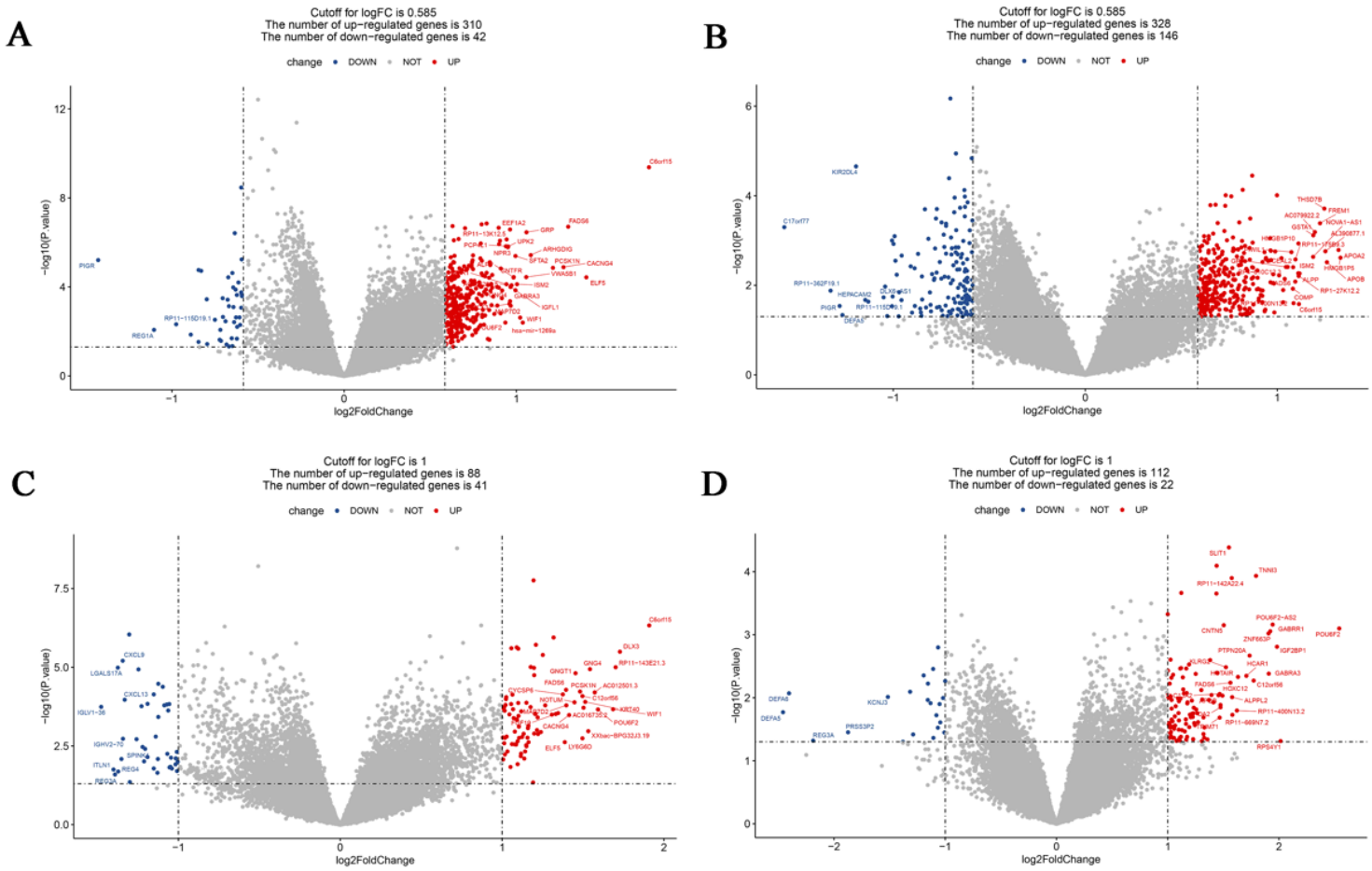


Figure 1

The volcano plot of the differentially expressed genes. Volcano plot of the differentially expressed genes (including mRNAs, lncRNAs, snoRNAs, miRNAs and undefined genes) between N0 and. N1/N2 in COAD samples, N0 and. N1/N2 in READ samples, M0 and M1 in COAD samples, and M0 and M1 in READ samples. Red dots represent the up-regulated genes, and blue dots represent the down-regulated genes. The dots with genes name were in the top 30 most differentially expressed genes.

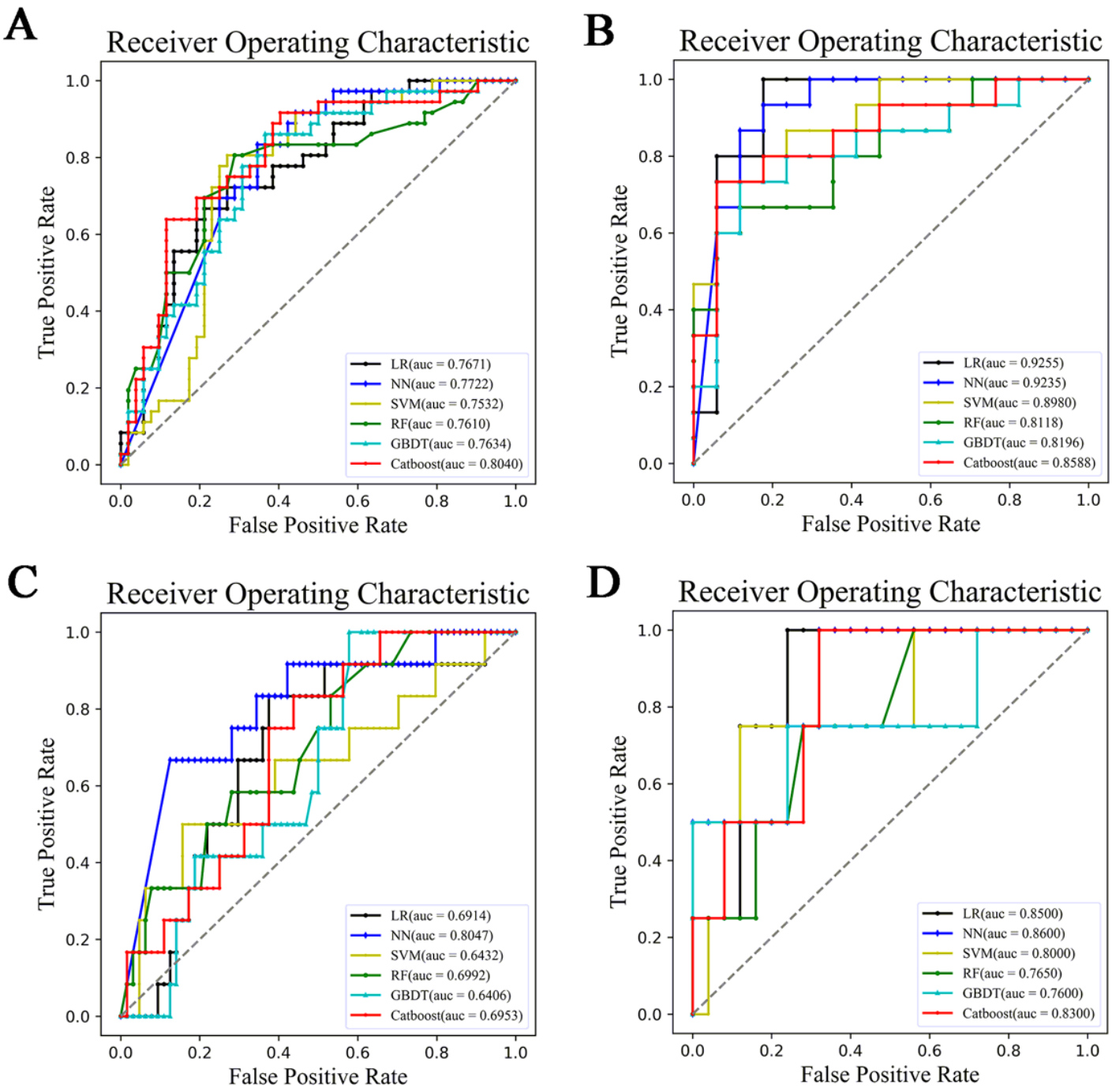
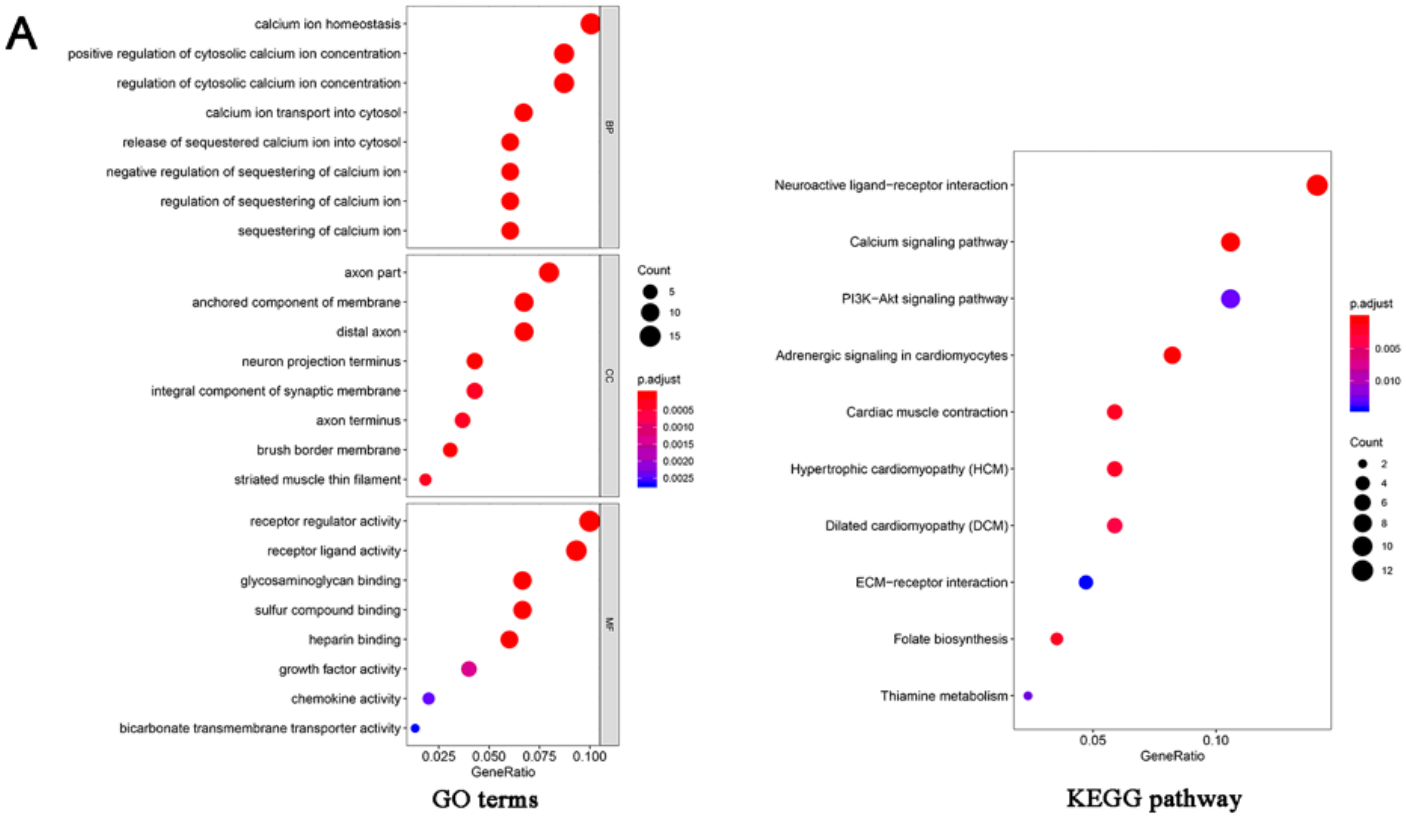


Figure 2

Receiver operating characteristic (ROC) curve for each of the six prediction models. Receiver operating characteristic (ROC) curve analyses for the six lymph node metastasis prediction models in COAD (A) and READ (B) samples showing the prediction performance of each model; ROC curves for the six distant metastases prediction models in COAD (C) and READ (D) samples showing the prediction performance of each of these models.


A
calcium ion homeostasis
positive regulation of cytosolic calcium ion concentration
regulation of cytosolic calcium ion concentration
calcium ion transport into cytosol
release of sequestered calcium ion into cytosol
negative regulation of sequestering of calcium ion
regulation of sequestering of calcium ion
sequestering of calcium ion
BP
axon part
anchored component of membrane
distal axon
neuron projection terminus
integral component of synaptic membrane
axon terminus
brush border membrane
striated muscle thin filament
CC
receptor regulator activity
receptor ligand activity
glycosaminoglycan binding
sulfur compound binding
heparin binding
growth factor activity
chemokine activity
bicarbonate transmembrane transporter activity
MF
0.025 0.050 0.075 0.100
GeneRatio
GO terms
Count
5
10
15
p.adjust
0.0005
0.0010
0.0015
0.0020
0.0025
Neuroactive ligand−receptor interaction
Calcium signaling pathway
PI3K−Akt signaling pathway
Adrenergic signaling in cardiomyocytes
Cardiac muscle contraction
Hypertrophic cardiomyopathy (HCM)
Dilated cardiomyopathy (DCM)
ECM−receptor interaction
Folate biosynthesis
Thiamine metabolism
0.05 0.10
GeneRatio
KEGG pathway
p.adjust
0.005
0.010
Count
2
4
6
8
10
12

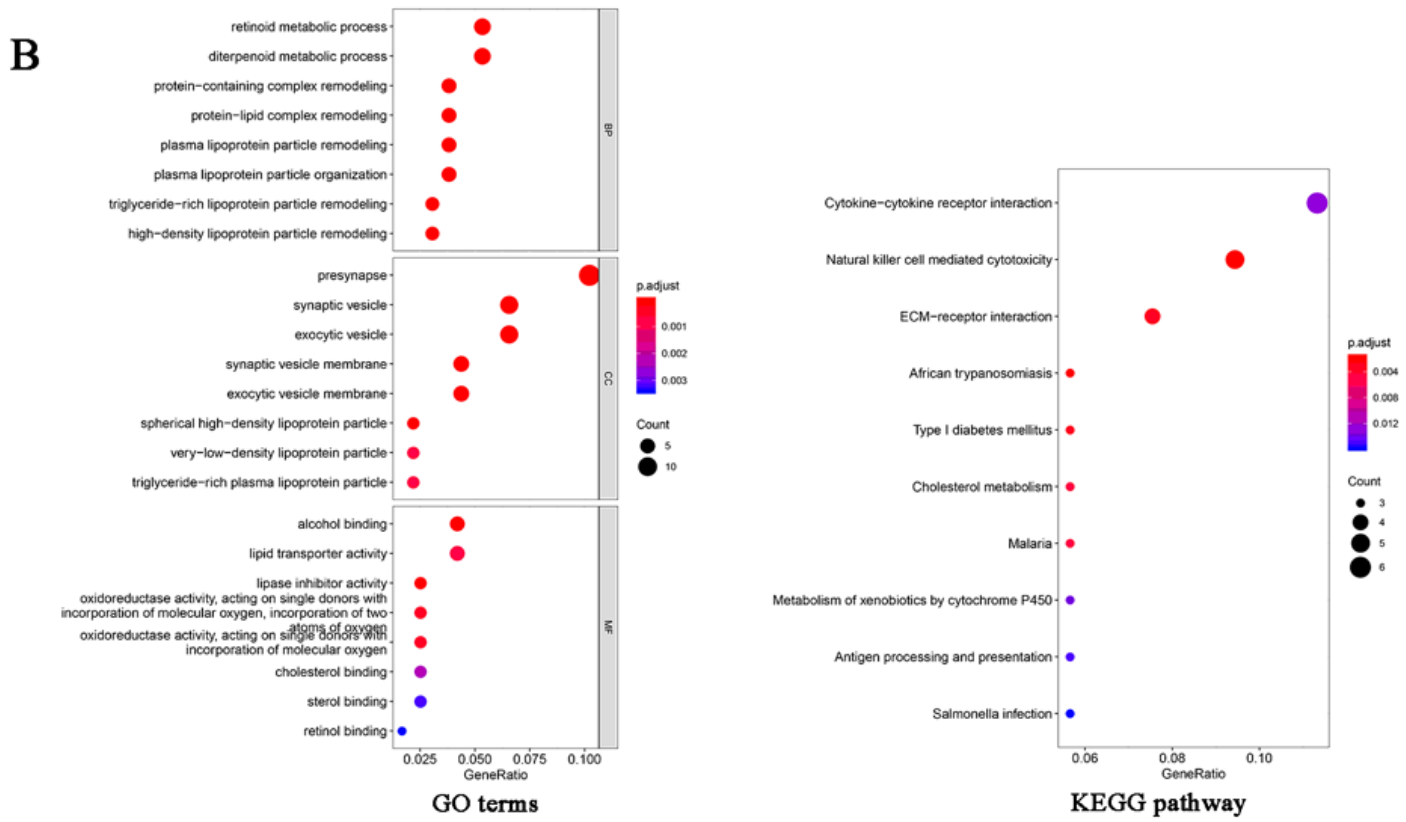

B
retinoid metabolic process
diterpenoid metabolic process
protein−containing complex remodeling
protein−lipid complex remodeling
plasma lipoprotein particle remodeling
plasma lipoprotein particle organization
triglyceride−rich lipoprotein particle remodeling
high−density lipoprotein particle remodeling
BP
presynapse
synaptic vesicle
exocytic vesicle
synaptic vesicle membrane
exocytic vesicle membrane
spherical high−density lipoprotein particle
very−low−density lipoprotein particle
triglyceride−rich plasma lipoprotein particle
CC
alcohol binding
lipid transporter activity
lipase inhibitor activity
oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen
oxidoreductase activity, acting on single donors with incorporation of molecular oxygen
cholesterol binding
sterol binding
retinol binding
MF
0.025 0.050 0.075 0.100
GeneRatio
GO terms
p.adjust
0.001
0.002
0.003
Count
5
10
Cytokine−cytokine receptor interaction
Natural killer cell mediated cytotoxicity
ECM−receptor interaction
African trypanosomiasis
Type I diabetes mellitus
Cholesterol metabolism
Malaria
Metabolism of xenobiotics by cytochrome P450
Antigen processing and presentation
Salmonella infection
0.06 0.08 0.10
GeneRatio
KEGG pathway
p.adjust
0.004
0.008
0.012
Count
3
4
5
6

Figure 3

Functional enrichment analysis for the genes identified in the optimal LNM prediction models. Bubble chart of GO terms and KEGG pathways enriched for the genes identified in the optimal lymph node metastasis prediction models for COAD (A) and READ (B) data showing the functions of those genes; bubble size represents the count of these enriched genes; the color (red to blue) represents decreasing P value (from high to low).

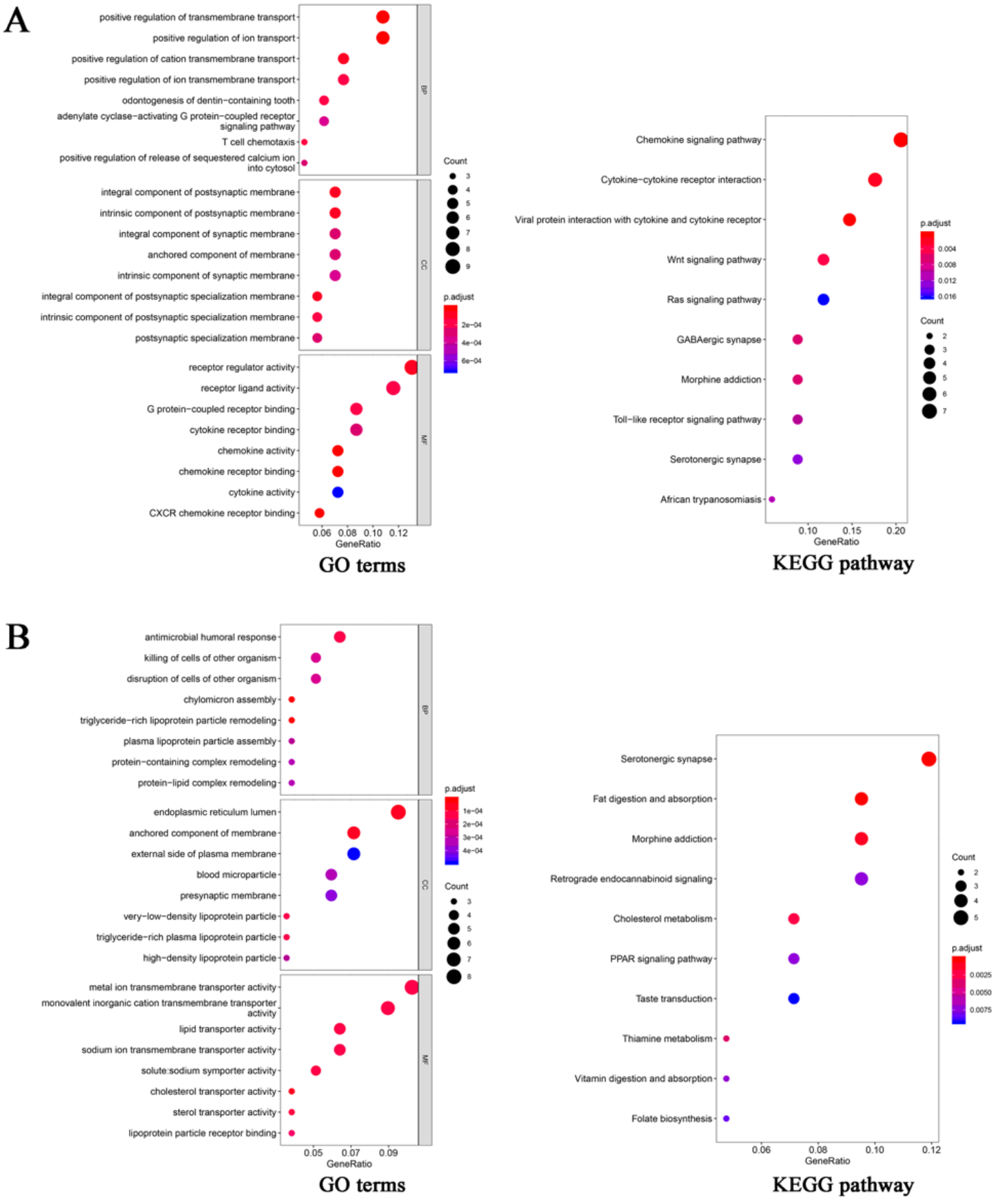

A
positive regulation of transmembrane transport
positive regulation of ion transport
positive regulation of cation transmembrane transport
positive regulation of ion transmembrane transport
odontogenesis of dentin-containing tooth
adenylate cyclase-activating G protein-coupled receptor signaling pathway
T cell chemotaxis
positive regulation of release of sequestered calcium ion into cytosol
integral component of postsynaptic membrane
intrinsic component of postsynaptic membrane
integral component of synaptic membrane
anchored component of membrane
intrinsic component of synaptic membrane
integral component of postsynaptic specialization membrane
intrinsic component of postsynaptic specialization membrane
postsynaptic specialization membrane
receptor regulator activity
receptor ligand activity
G protein-coupled receptor binding
cytokine receptor binding
chemokine activity
chemokine receptor binding
cytokine activity
CXCR chemokine receptor binding
BP
CC
MF
0.06 0.08 0.10 0.12
GeneRatio
Count
p.adjust
2e-04
4e-04
6e-04
GO terms
Chemokine signaling pathway
Cytokine-cytokine receptor interaction
Viral protein interaction with cytokine and cytokine receptor
Wnt signaling pathway
Ras signaling pathway
GABAergic synapse
Morphine addiction
Toll-like receptor signaling pathway
Serotonergic synapse
African trypanosomiasis
0.10 0.15 0.20
GeneRatio
p.adjust
0.004
0.008
0.012
0.016
Count
KEGG pathway
B
antimicrobial humoral response
killing of cells of other organism
disruption of cells of other organism
chylomicron assembly
triglyceride-rich lipoprotein particle remodeling
plasma lipoprotein particle assembly
protein-containing complex remodeling
protein-lipid complex remodeling
endoplasmic reticulum lumen
anchored component of membrane
external side of plasma membrane
blood microparticle
presynaptic membrane
very-low-density lipoprotein particle
triglyceride-rich plasma lipoprotein particle
high-density lipoprotein particle
metal ion transmembrane transporter activity
monovalent inorganic cation transmembrane transporter activity
lipid transporter activity
sodium ion transmembrane transporter activity
solute:sodium symporter activity
cholesterol transporter activity
sterol transporter activity
lipoprotein particle receptor binding
BP
CC
MF
0.05 0.07 0.09
GeneRatio
p.adjust
1e-04
2e-04
3e-04
4e-04
Count
GO terms
Serotonergic synapse
Fat digestion and absorption
Morphine addiction
Retrograde endocannabinoid signaling
Cholesterol metabolism
PPAR signaling pathway
Taste transduction
Thiamine metabolism
Vitamin digestion and absorption
Folate biosynthesis
0.06 0.08 0.10 0.12
GeneRatio
Count
p.adjust
0.0025
0.0050
0.0075
KEGG pathway

Figure 4

Functional enrichment analysis for the genes identified in the optimal distant metastases' prediction models. Bubble chart of GO terms and KEGG pathways enriched for the genes identified in the optimal distant metastases prediction models from COAD (A) and READ (B) data showing the functions of these genes; bubble size represents the count for each of these enriched genes; the color (red to blue) represents changes in the P value (high to low).

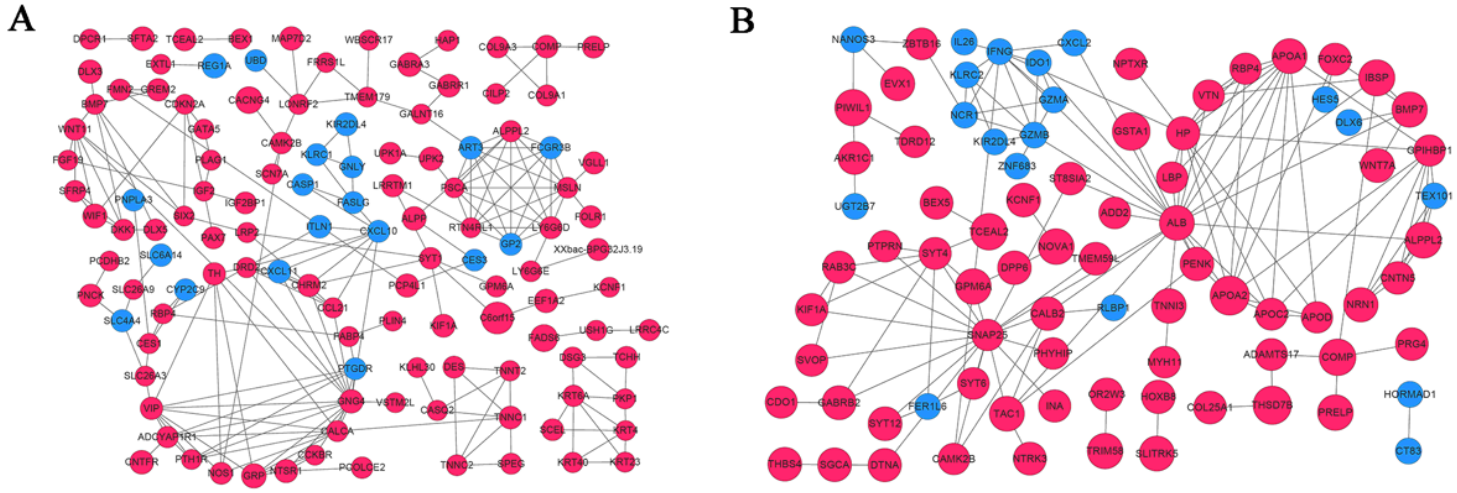


Figure 5

PPI network for the genes identified by the optimal LNM prediction models. PPI network for the genes identified by the optimal lymph node metastasis prediction models in COAD (A) and READ (B) demonstrating interactions among these genes; red nodes represent up-regulated genes; blue nodes represent down-regulated genes; lines represent interactions between two nodes.

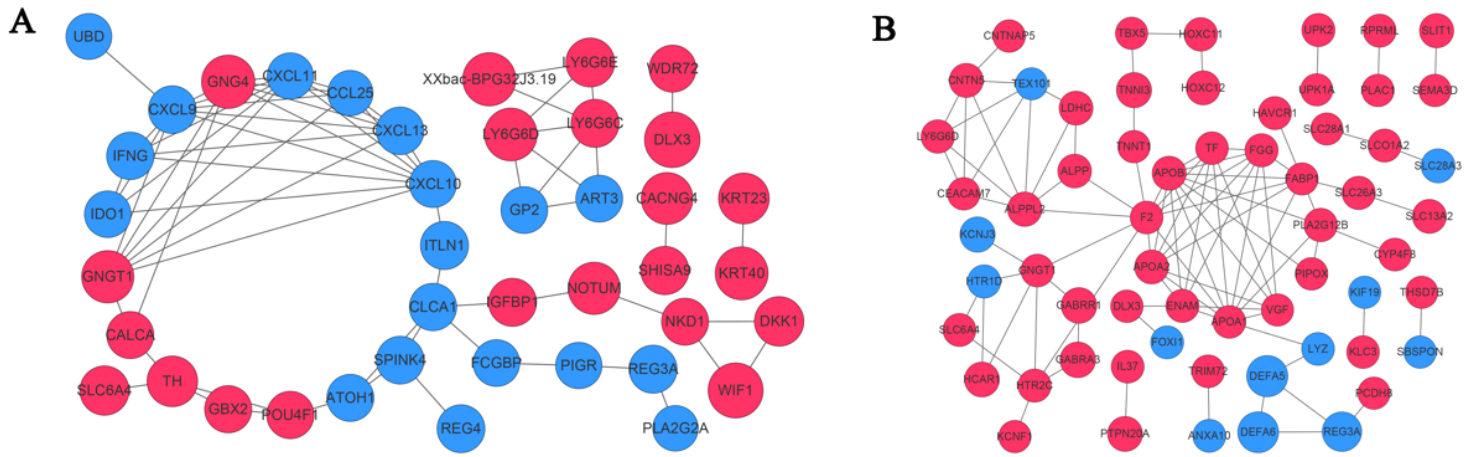


Figure 6

PPI network for the genes identified by the optimal distant metastases prediction models. PPI network for the genes identified by the optimal distant metastases prediction models in COAD (A) and READ (B) demonstrating the interactions between these genes; red nodes represent up-regulated genes; blue nodes represent down-regulated genes; lines represent interactions between two nodes.

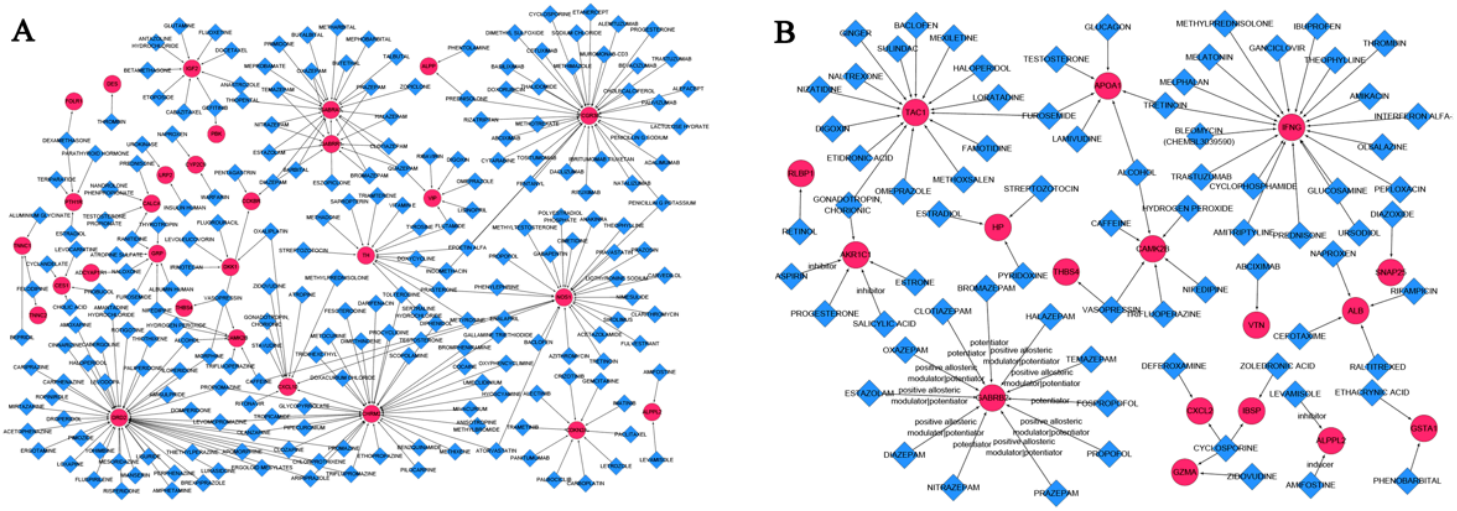


Figure 7

Drug-gene network for the genes identified by the optimal LNM prediction models. Drug-gene network for the genes identified by the optimal lymph node metastasis prediction models in COAD (A) and READ (B) showing the drugs that target these genes; red nodes represent genes; blue nodes represent drugs; lines represent interactions between gene and drug, and the phrase around the lines represent the interaction type.

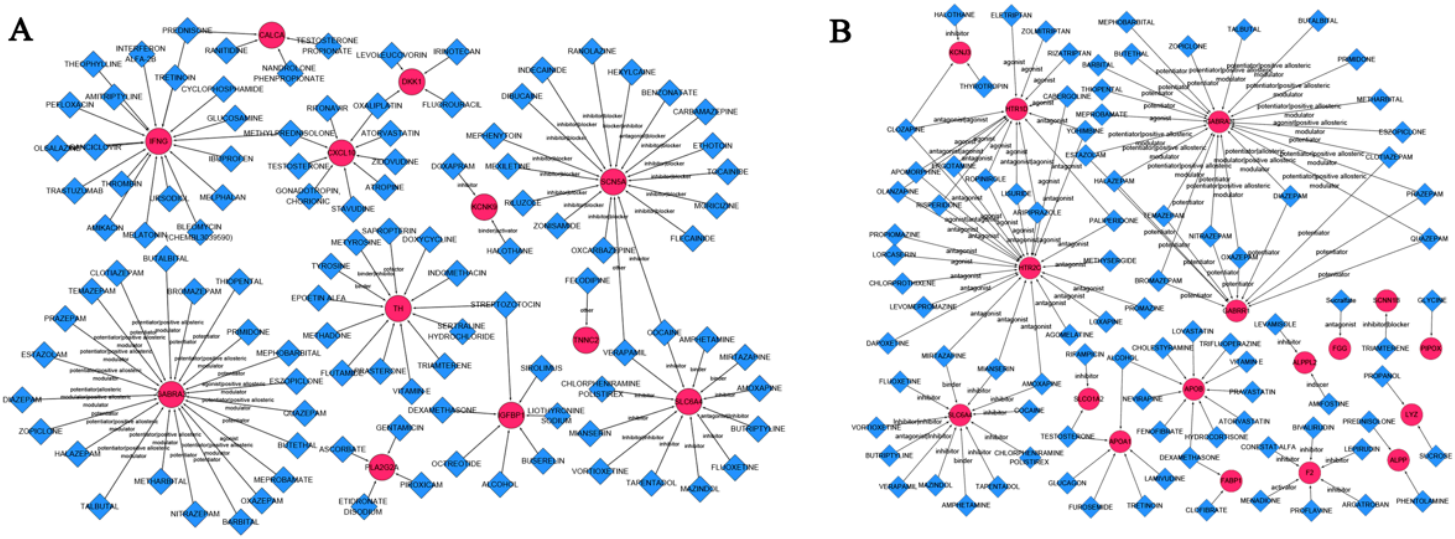


Figure 8

Drug-gene network for the genes identified by the optimal distant metastases prediction models. Drug-gene network for the genes identified by the optimal distant metastases prediction models in COAD (A) and READ (B) showing the drugs that target these genes; red nodes represent genes; blue nodes represent drugs; lines represent interactions between gene and drug, and the phrase around the lines represent the interaction type.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementalfigure1.Png
- Supplementalfigure3.Png
- Supplementalfigure2.Png
- Supplementalfile4.ipynb
- Supplementalfile3.ipynb
- Supplementalfile2.ipynb
- Supplementalfile1.ipynb
- SupplementalTable5.xls
- SupplementalTable4.xls
- SupplementalTable6.xls
- SupplementalTable7.xls
- SupplementalTable1.xls
- SupplementalTable2.xls
- SupplementalTable3.xls