Enhancing Chemotherapy Response Prediction via Matched Colorectal Tumor-Organoid Gene Expression Analysis and Network-Based Biomarker Selection

Wei Zhang[1], Chao Wu[2], Hanchen Huang[1], Paulina Bleu[2], Wini Zambare[2], Janet Alvarez[2], Lily Wang[1,3,4,5], Philip B. Paty[2], Paul B. Romesser[6], J. Joshua Smith[2*], X. Steven Chen[1,3*]

[1] Department of Public Health Sciences, University of Miami Miller School of Medicine, Miami, FL, 33136, USA;

[2]Colorectal Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA;

[3] Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, 33136, USA;

[4] Dr. John T Macdonald Foundation Department of Human Genetics, University of Miami, Miller School of Medicine, Miami, FL, 33136, USA;

[5] John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, 33136, USA;

[6]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA;

Correspondence and requests for materials should be addressed to J.J.S. (email: smithj5@mskcc.org), X.S.C. (email: steven.chen@miami.edu)

Abstract

Colorectal cancer (CRC) poses significant challenges in chemotherapy response prediction due to its molecular heterogeneity. This study introduces an innovative methodology that leverages gene expression data generated from matched colorectal tumor and organoid samples to enhance prediction accuracy. By applying Consensus Weighted Gene Co-expression Network Analysis (WGCNA) across multiple datasets, we identify critical gene modules and hub genes that correlate with patient responses, particularly to 5-fluorouracil (5-FU). This integrative approach advances precision medicine by refining chemotherapy regimen selection based on individual tumor profiles. Our predictive model demonstrates superior accuracy over traditional methods on independent datasets, illustrating significant potential in addressing the complexities of high-dimensional genomic data for cancer biomarker research.

Introduction

Colon and rectal cancer remains a major global health challenge [1], ranking as the third most prevalent cancer diagnosis and the third leading cause of cancer-related mortality for both men and women in the United States. It is estimated that more than 150,000 new cases will be diagnosed in the coming year, with >50,000 of those being rectal cancer (RC). It is expected that 52,000 deaths are expected to occur relative to both colon and RC this year [2]. Chemotherapy remains central to the treatment of both colon and rectal cancers and centers around drugs such as 5-fluorouracil (5-FU), Oxaliplatin, and Irinotecan. However, the effectiveness of these treatments varies considerably due to the molecular heterogeneity of colon and rectal tumors, with response rates ranging widely. For instance, the use of 5-fluorouracil (5-FU) in combination with leucovorin as a first-line treatment for metastatic disease shows an overall response rate of approximately 20-30%. Moreover, the introduction of irinotecan or oxaliplatin to this regimen increases the response rate to about 40-50%, demonstrating significant variability in treatment efficacy [3]. Further, this variability is specifically seen in RC where the response to triplet chemotherapy (e.g., FOLFIRI (5-FU, Oxaliplatin, and Irinotecan)) is associated with better prognosis than standard therapy [4, 5]. In addition, given the fact that some RC patients may only require upfront chemotherapy for cure, the proper selection of neoadjuvant chemotherapy will become even more critical [6]. Variability in treatment response highlights the critical need for precise predictive models to better forecast individual chemotherapy responses. Accurate prediction models are essential not just for enhancing treatment effectiveness, but also for avoiding unnecessary side effects in non-responsive patients, ultimately leading to more personalized and effective cancer care.

Current commercially available colorectal cancer gene signature panels, like OncotypeDX, ColoPrint, and others, primarily serve prognostic purposes, but their effectiveness in predicting neoadjuvant therapy response is not well-established [7-12]. Moreover, the likelihood of individual cancer biomarkers reaching clinical significance remains notably low, a reality shaped by multiple factors. One notable challenge is that prediction models derived from cell lines often fail in human tumors [13, 14]. In addition, intratumoral heterogeneity (ITH) via stromal cells in the tumor microenvironment may hide subtle gene expression alterations associated with genetic

diversity and heterogeneity within the tumor epithelium [15]. Finally, the complexity of genomic features (i.e., Curse of dimensionality) presents a key challenge to prediction tasks involving the drug response [16-18]. This means the top genomic features selected from primary tumors may not be predictive in different patient cohorts partially due to the false positive signals from high-dimensional molecular data, which was shown in the colorectal cancer biomarker discovery [7-10].

Tumor organoids, three-dimensional cell culture systems, offer a revolutionary approach in cancer research. They closely mimic the structural and functional attributes of original tissues or organs, providing a more accurate representation of human tumors. These organoids preserve the genetic, phenotypic, and behavioral characteristics of their source tumors, making them highly relevant for drug discovery, treatment response tracking, and personalized medicine. Their ability to maintain the heterogeneity of the source tumors enhances the study of tumorigenesis, drug screening, and precision medicine, presenting a significant advantage over traditional two-dimensional cell cultures. Previously, our group has shown that RC organoids correspond to the patient-specific outcomes observed including chemotherapy response and clinical outcomes [19]. This advancement facilitates more reliable investigations into tumor pathogenesis, offering a promising platform for cancer research and clinical applications [20, 21].

Recent advancements in colon and RC organoid research demonstrate effective replication of the intricate cellular diversity and molecular heterogeneity seen in patient tumors, thus providing an essential tool for exploring disease development, progression, and treatment responses [19, 22-25]. The transcriptome data from organoids has been used to predict anti-cancer drug efficacy [26]. Moreover, harnessing the power of colon and RC organoid models and integrating molecular data from matched primary tumors and organoids allows an opportunity to identify novel biomarkers that predict treatment response. Directly selecting markers from tumor samples can result in a high rate of false-positive signals due to the vast number of features relative to the number of observations, which can obscure true biological signals. To mitigate this issue, organoids serve as an amplified biological system that retains the complexity of the original tumor but in a more controlled environment that allows for clearer observation of treatment responses. By comparing the molecular profiles of organoids with those of the corresponding tumor samples, we can more

effectively filter out noise and retain robust signals that reflect intrinsic tumor biology. A previous study showed that a cancer-cell intrinsic gene expression signature has excellent predictive performance by minimizing intratumoral heterogeneity bias in colon and RC prognostic/predictive classification [27]. Another recent study found that tumor intrinsic immune signatures developed via a matched organoid–primary tumor system are effective tissue biomarkers of prognosis in colon and RC [28], indicating organoids reflect intrinsic properties of the tumor and its microenvironment [19, 29-33].

Here, we developed new strategies that leverage matched colorectal tumor and organoid transcriptome data and the consensus gene network approach to identify key gene expression biomarkers predictive of 5-FU-based chemotherapy response. Our results indicate that these tumor-based biomarkers are strongly correlated with patient survival outcomes when treated with specific chemotherapeutic agents. This innovative approach marks a notable shift towards precision medicine, offering the potential to customize therapeutic strategies based on individual colorectal patient profiles to enhance treatment efficacy.

Results

An Overview of the Integrative Analysis of Matched Colorectal Tumor and Organoid Data for Chemotherapy Response Prediction

This study employs a unique integrative analysis approach, as depicted in **Figure 1**, focusing on matched colorectal cancer (CRC) tumor tissues and patient-derived organoids. The analysis begins by leveraging gene expression data from both CRC tumors and corresponding organoids, drawn primarily from datasets GSE171680 and GSE171681, which include a substantial number of matched samples (87 samples)[28]. Given the lack of direct drug response data in these matched datasets, we next integrated an additional colorectal cancer organoid dataset, GSE64392, which contained IC50 values. The detailed information of three training datasets is listed in **Table 1A**.

Consensus Weighted Gene Co-expression Network Analysis (WGCNA) is a powerful method that identifies clusters of genes (modules) with similar expression patterns across multiple datasets [34, 35]. This approach is particularly effective in integrating data from different sources [36], like tumor and organoid gene expression profiles in our study. By analyzing these patterns

collectively on three datasets - colorectal tumors (GSE171680), colorectal tumor-matched organoids (GSE171681), and an independent colorectal organoid data (GSE64392), consensus WGCNA provides a more robust and comprehensive understanding of the underlying biological relationships. It helps in pinpointing key gene modules that are consistently associated across datasets, which are crucial for understanding complex traits such as drug response in cancer. The study then centers on hub genes within these modules, vital for developing a predictive model for chemotherapy response. This workflow combines matched tumor-organoid gene expression with additional drug response data, aiming to develop a robust predictive model for disease-relevant chemotherapy response.

To build and validate our drug-response prediction models, we incorporated a detailed methodological approach as depicted in Figure 1. Hub genes identified from the organoid WGCNA were employed to construct predictive models using ridge regression, random forest, and an ensemble of these methods. The models were trained and optimized through cross-validation to minimize prediction error, selecting the best-performing model based on its predictive accuracy in terms of area under the curve (AUC) on independent validation sets. The selected ensemble model demonstrated superior performance, indicating its effectiveness in predicting chemotherapy response. Further validation involved calculating patient-specific drug-resistance scores using the optimal model, allowing for a thorough assessment of individual response to chemotherapy.

Identification of coherent gene modules through consensus WGCNA

Given the robust nature of the weighted correlation network to the choice of soft-thresholding power, we selected $\beta = 12$ for the signed network. This ensured a scale-free topology model fit above 0.75, in which the network conforms to a scale-free topology, a characteristic of biological networks where few nodes (genes) are highly connected [34]. (**Figure 2A**). The consensus WGCNA identified 16 modules, including a grey module. Grey modules typically contain genes that do not correlate well with any others and thus are not grouped into specific functional modules. In contrast, non-grey modules such as the turquoise module (n=214), blue module (n=190), brown module (n=157), yellow module (n=144), and green module (n=141), which contain the most genes, show more homogenous and potentially functionally relevant expression patterns. Although the grey module comprised 1910 genes, it displayed heterogeneous expression patterns and was

not assigned to any particular function. **Figure 2B** shows the dendrogram of modules clustered by hierarchical clustering based on consensus Topological Overlap Matrix (TOM).

Prognostic relevance and correlation of gene modules with clinical outcomes

Among the modules identified, the tan, salmon, and magenta modules showed statistically significant associations with Overall Survival (OS) and Recurrence-Free Survival (RFS) in Cox regression models in independent datasets (**Figure 2C**). These modules demonstrated protective effects, with hazard ratios less than one. Notably, the tan and salmon modules had significant Spearman correlations between their eigengenes in matched organoid (GSE171681) and tumor samples (GSE171682, suggesting robust biomarker potential. **Figure 3A** shows the scatterplot of Spearman correlations between the eigengenes of these two modules. The tan module, in particular, showed the highest correlation ($R^2_{tan} = 0.7, p = 2.2 \times 10^{-16}$), and the salmon module also achieves a moderate correlation coefficient ($R^2_{salmon} = 0.5, p = 1.4 \times 10^{-6}$). **Figure 3B** shows the consensus network of the tan and salmon modules with the names of the hub genes colored in brown. The tan module has more connections than the salmon module. In each pair among the three datasets, the individual module membership Spearman correlations are all significant (**Figure 3C**). This indicates a high consistency of gene findings in the tan and salmon modules across the three datasets and confirms the reliability of these gene modules in predicting clinical outcomes. We identified and selected 35 hub genes with an absolute consensus module membership (MM) of 0.5 or higher for subsequent prediction modeling.

Building organoid drug-response models and selecting drug response-related genes

The organoid 5-FU drug-response models were built using the 35 hub genes selected from WGCNA. To further identify the biomarkers following 5FU treatment in colorectal cancer, we developed an ensemble model that combined two machine learning methods: random forest and ridge regression (see "Methods"). We compared the performance of our ensemble method to that of using only ridge or only random forest methods through cross-validation in the training data GSE64392. We also validated these models on the GSE171680 dataset using the Area Under the Curve (AUC) metric (see **Supplementary Figure 1**). The results suggested that the biomarkers identified by the ensemble model demonstrated higher predictive performance across all models.

Using this model, seven genes were selected as the final biomarkers for the 5-FU drug response based on training data GSE64392 (**Table 2**).

Validation of organoid prediction model with six independent datasets

The prediction performance of our organoid model was then validated in six independent GSE datasets: GSE39582, GSE17538, GSE106584, GSE72970, and GSE87211. The details of the survival and transcriptomic data for these six datasets are described in **Table 1B**. The patient-specific drug-response scores of each validation dataset were calculated (see "Methods") and the statistical difference in overall survival (OS) between the drug-sensitive and drug-resistant groups was assessed by the Kaplan-Meier survival curves and log-rank tests. As shown in **Figure 4** and **Table 3**, the drug-sensitive group had a significantly longer OS than the drug-resistant group for all six datasets: the p-values of log-rank tests are 1.32×10^{-04} (GSE39582), 7.09×10^{-04} (GSE17538), 4.95×10^{-02} (TCGA-COAD), 2.08×10^{-02} (GSE106584), 4.60×10^{-03} (GSE72970), and 1.25×10^{-03} (GSE87211). We examined the pattern of the seven selected drug-related genes.

Interestingly, we found that all the genes with lower IC50 values show higher expression in GSE64932 (**Figure 5A**). Furthermore, to determine if the 35 hub genes identified in the consensus WGCNA have similar validation outcomes as the seven drug-related genes, we computed patient-specific drug-resistant scores using the ridge regression coefficients of the 35 hub genes. The prognosis test results are displayed in **Supplementary Figure 2**, which shows significant p-values were achieved in all datasets except TCGA-COAD, where the significance was marginally achieved. The selection of 35 hub genes from the WGCNA, and the further selection of 7 genes from these 35, provide a reliable predictor of drug response. This can be showed by the significant log-rank tests in 5 out of 6 validation datasets, which were based on survival outcomes of the 35 hub genes. Furthermore, all 6 validation datasets showed significant results in log-rank tests using the 7-gene organoid model.

Functional enrichment analysis of hub genes

We next performed enrichment analysis for the 35 selected hub genes in the KEGG, REACTOME, and GO pathway databases. **Figure 5B** presents the significant pathways with adjusted p-values less than 0.05. It can be observed that certain pathways are prominently associated with colon and

rectal cancer chemotherapy response. Top pathways included those related to DNA repair mechanisms, cell cycle regulation, apoptosis, and drug metabolism. For instance, pathways involved in DNA damage response are crucial, as chemotherapy often targets rapidly dividing cancer cells by inducing DNA damage. Similarly, pathways regulating apoptosis is also significant, as the effectiveness of chemotherapy is partly determined by the ability of cancer cells to undergo programmed cell death.

Prediction performance of alternative gene selection processes

To demonstrate the consistency and utility of our approach in predicting the survival of colorectal cancer patients, we evaluated the prediction performance on the validation datasets using two alternative gene selection methods: one based on three additional WGCNA strategies, and another based on two gene association tests (Table 3). We found that the genes chosen by Model 2 (See Methods), which were selected based on consensus WGCNA applied to two datasets, yielded significant results in the log-rank tests for four datasets (Supp Figure 4). This was followed by Model 1, where genes were selected based on WGCNA applied to tissue data GSE171680; it produced significant results in the log-rank tests for three datasets (Supp Figure 3). Model 3, in which genes were selected based on WGCNA applied to organoid data GSE64932, did not consistently yield significant results in log-rank tests (Supp Figure 5). This suggests that applying more data to construct consensus WGCNA yields more robust results in predicting survival. Next, we examined the gene selection process based on the two different criteria of gene filtering from the association tests. We found that the small sample size of the organoid and tissue data could potentially affect the reliability of test results, thereby impacting the validation results in survival prediction (Supp Figure 6-7).

Discussion

In this study, we introduced a novel methodology for predicting chemotherapy responses in colon and rectal cancer, utilizing matched tumor-organoid gene expression data. Our approach incorporated Consensus Weighted Gene Co-expression Network Analysis (WGCNA) across diverse datasets, including colorectal tumors, corresponding organoids, and an independent CRC organoid dataset with drug response data (IC50 values). This method effectively identified key

gene modules and hub genes associated with colorectal chemotherapy response, marking a significant advancement in personalized medicine approaches for colorectal treatment.

The prediction results of our study highlight the substantial advantages of using matched organoid and tumor samples for biomarker selection in colorectal cancer treatment. By integrating gene expression data from these matched samples, our analysis was able to identify more precise and relevant biomarkers for chemotherapy response prediction. This approach led to the discovery of specific gene modules and hub genes that are critically involved in response to chemotherapy in colorectal cancer, particularly 5-fluorouracil (5-FU). Our predictive model, built on these findings, demonstrated a notable improvement in accuracy and reliability compared to traditional methods as results showed in **Table 3**.

The superior performance of our proposed model underscores the effectiveness of the two strategies we employed. First, matched tumor-organoid data in capturing the intrinsic signature of chemo-response complex molecular dynamics of colon and rectal cancers, thus yielding a more robust clinical outcome. Second, the WGCNA network-based biomarker selection offered notable advantages in understanding colorectal cancer chemotherapy response. The ability of WGCNA to identify modules of co-expressed genes allowed us to discern complex gene interaction networks relevant to response to treatment in colon and rectal cancer. This network-based approach facilitated the identification of not just individual genes but also clusters of genes (modules) that collectively contribute to drug responsiveness. This method, by capturing the systemic relationships and dependencies among genes, provided a more holistic view of the molecular mechanisms underlying the response to chemotherapy in colorectal cancer, thereby enhancing the accuracy and relevance of the selected biomarkers for clinical application.
A limitation of the current work is the lack of direct drug response data from the matched colorectal cancer organoid dataset (GSE171681), necessitating reliance on similar in vitro organoid experiments (GSE64392). Future research could benefit from matched tumor-organoid datasets inclusive of patient and organoid drug response outcomes. Additionally, network modeling approaches could be refined to integrate more extensive biological information [37, 38], potentially offering more profound insights into CRC chemotherapy response mechanisms. The strength of this work highlights a new pathway in biomarker discovery for colon and rectal

cancer chemotherapy response prediction, addressing typical high-dimensional genomic data challenges in cancer research like intratumoral heterogeneity and the curse of dimensionality. This method could extend to more advanced organoid cultures, including those that incorporate integrating the potential of organoid and the patient-tumor microenvironment [39], offering the potential to identify biomarkers for immunotherapy responses.

Discussion

In this study, we introduced a novel methodology for predicting chemotherapy responses in colon and rectal cancer, utilizing matched tumor-organoid gene expression data. Our approach incorporated Consensus Weighted Gene Co-expression Network Analysis (WGCNA) across diverse datasets, including colorectal tumors, corresponding organoids, and an independent CRC organoid dataset with drug response data (IC50 values). This method effectively identified key gene modules and hub genes associated with colorectal chemotherapy response, marking a significant advancement in personalized medicine approaches for colorectal treatment.

The prediction results of our study highlight the substantial advantages of using matched organoid and tumor samples for biomarker selection in colorectal cancer treatment. By integrating gene expression data from these matched samples, our analysis was able to identify more precise and relevant biomarkers for chemotherapy response prediction. This approach led to the discovery of specific gene modules and hub genes that are critically involved in response to chemotherapy in colorectal cancer, particularly 5-fluorouracil (5-FU). Our predictive model, built on these findings, demonstrated a notable improvement in accuracy and reliability compared to traditional methods as results showed in **Table 3**.

The superior performance of our proposed model underscores the effectiveness of the two strategies we employed. First, matched tumor-organoid data in capturing the intrinsic signature of chemo-response complex molecular dynamics of colon and rectal cancers, thus yielding a more robust clinical outcome. Second, the WGCNA network-based biomarker selection offered notable advantages in understanding colorectal cancer chemotherapy response. The ability of WGCNA to identify modules of co-expressed genes allowed us to discern complex gene interaction networks relevant to response to treatment in colon and rectal cancer. This network-

based approach facilitated the identification of not just individual genes but also clusters of genes (modules) that collectively contribute to drug responsiveness. This method, by capturing the systemic relationships and dependencies among genes, provided a more holistic view of the molecular mechanisms underlying response to chemotherapy in colorectal cancer, thereby enhancing the accuracy and relevance of the selected biomarkers for clinical application.

A limitation of the current work is the lack of direct drug response data from the matched colorectal cancer organoid dataset (GSE171681), necessitating reliance on similar in vitro organoid experiments (GSE64392). Future research could benefit from matched tumor-organoid datasets inclusive of patient and organoid drug response outcomes. Additionally, network modeling approaches could be refined to integrate more extensive biological information [37, 38], potentially offering more profound insights into CRC chemotherapy response mechanisms.

The strength of this work highlights a new pathway in biomarker discovery for colon and rectal cancer chemotherapy response prediction, addressing typical high-dimensional genomic data challenges in cancer research like intratumoral heterogeneity and the curse of dimensionality. This method could extend to more advanced organoid cultures, including those that incorporate integrating the potential of organoid and the patient-tumor microenvironment [39], offering the potential to identify biomarkers for immunotherapy responses.

Method

Study cohorts

We selected three colorectal cancer datasets from research carried out by van de Wetering et al.[22] and Cho et al. [28] for use in the consensus WGCNA algorithm and development of organoid drug-response model. The datasets from van de Wetering et al. contains 22 organoid samples of microarray data and drug-response for colorectal cancer. The microarray data can be retrieved from the Gene Expression Omnibus (GEO) with the study accession GSE64932 [22]. In our study, we used IC_{50} values as drug sensitivity measurements and selected 19 samples tested with 5-Fu. The two datasets from Cho et al. include paired expression data from 87 organoid (GEO accession: GSE171681) and patient tissue (GEO accession: GSE171680) samples with colorectal cancer. The detailed information of the three training datasets can be found in Table 1A. Furthermore, we validated the prognosis predictive value of our organoid drug-response model to the overall survival (OS) outcomes of five GEO datasets for colorectal and colon cancer (GSE39582 [40], GSE17538 [41], GSE106584 [42], GSE72970 [43], and GSE87211 [44]). We also included one TCGA data (https://www.cancer.gov/tcga) for colorectal cancer (TCGA-COAD). All six validation datasets contain OS outcomes and the samples with 5-FU based treatment were selected. The detailed information of validation data can be found in Table 1B.

Data preprocessing

For the expression datasets from GEO, we used the robust multichip average (RMA) normalized expression data [45]. The genes were represented by the probes with the largest interquartile range (IQR) statistics using *findLargest* function in genefilter R/Bioconductor package. The gene symbols were annotated using the AnnotationDbi R/Bioconductor package. The TCGA-COAD patient data were downloaded from the TCGA data portal using the TCGAbiolinks Bioconductor/R package [46]. For expression analysis of TCGA-COAD, we used the FPKM-UQ dataset and performed a log2 transformation.

Consensus WGCNA

We used consensus WGCNA to study the relationships among the three expression profiles (GSE64932, GSE171680, and GSE171681) [34]. To perform this analysis, we first filtered the three expression profiles, selecting only the 3637 common genes that were among the top 50% most variable genes across all datasets. For each dataset, a signed correlation weight, $s_{ij} = \frac{1+cor(x_i, x_j)}{2}$, is assigned to each gene pair x_i and x_j via a positive soft thresholding parameter β. The signed network weighted adjacency matrix is defined as:

$$a_{ij} = |s_{ij}|^{\beta}$$

Here, β is the raised power of similarity measures, which emphasize more on the strong associations [47]. The three adjacency matrices were then transformed into the topological overlap matrices (TOM) that provide a robust measure of connections between gene pairs [48]. The individual TOMs are calibrated using the full quantile normalization such that all quantiles equal each other. To obtain the consensus TOM, we calculated the component-wise mean of individual TOMs for each set. The genes were further clustered by average linkage hierarchical clustering using the dissimilarity measure of the consensus TOM (1 – consensus TOM). The dendrogram cut height for module detection was set to 0.999 and the minimum module size of each module was set to 30 genes. The consensus network analysis was conducted using the *blockwiseConsensusModules* R function in WGCNA package.

Significant modules and hub genes selection

To identify the significant modules, we first performed Cox proportional hazards regression models to test association between patient OS and recurrent free survival (RFS) outcomes with the module eigengenes (MEs). This approach was adopted as the survival outcomes are significantly related to drug response. The module eigengenes were obtained as the 1_{st} principal component of the patient tissue expression. Three modules were selected with coefficient *P-value* < 0.05 in either OS or RFS cox regression model. Furthermore, to determine the concordance of modules between organoid and tissue expression, we estimated Spearman correlations between the eigengenes of each module in the two paired organoid and tissue expressions (GSE171681 and GSE171682). Only two modules, tan and salmon, among the three survival outcomes significant modules, were highly significant correlated ($R^2_{tan} = 0.7, p = 2.2 \times 10^{-16}$ and $R^2_{salmon} = 0.5, p = 1.4 \times 10^{-6}$) between their eigengenes. These two modules were then selected as the significant modules for further model training. To assess the module membership (MM) of genes in each module, we calculated the correlation between the gene expressions and MEs of each module. This is denoted as kME. To evaluate the MM across all the expressions, we used the consensus kME that obtained by average aggregation of the kMEs for each expression set. The consensus kME was implemented in the function *consensusKME*. The selection of hub genes varies as each dataset has different clinically related information. For example, GSE64932 only contains drug response data, while GSE171681 only includes survival outcome data. We selected the hub genes of the significant modules using only the criterion of consensus $|\mathrm{MM}| \geq 0.5$.

Organoid drug-response model training

To build the drug-response models, we used the hub genes selected from the WGCNA of the organoid expression profile against the median IC_{50} of 5-FU as drug response. We selected ridge, random forest, and an ensemble method of random forest and ridge as the training models. Elastic net model was performed using the glmnet R package. The optimal α and λ parameters were

selected based on the lowest mean squared error between the drug response and predicted values in the validation sets, using 3-fold cross validation (CV). This was implemented in the *cv.glmnet* function. The Random Forest (RF) was constructed using the *rfsrc* function from the randomForestSRC R package. The final RF model was built using genes with a permutation importance greater than 0. The ensemble method was created by choosing genes based on their permutation importance obtained from random forest model. These selected genes were then fitted into a ridge regression model. Following the selection of drug-response related genes, we conducted a 3-fold CV on each model to decide the optimal model as our final organoid model. Furthermore, to ensure the optimal model is selected, we applied the three models to patient data GSE171680, which was used to conduct the consensus WGCNA. The Area Under the Curve (AUC) was calculated based on the predicted values of each model, using the OS outcome as binary. The ensemble model was selected as the optimal model as it achieved both lowest 3-fold CV values and highest AUC among all three model. To stabilize the CV errors and assess the model powers, we repeated running the 3-fold CV for each model with 100 times. Seven genes were selected by the optimal model as the drug-response related genes.

Patient specific drug-resistant score

To validate our optimized organoid drug-response model, we calculated the patient specific drug-resistance score for each patient using the corresponding expression data. Specifically, the score from the optimized organoid drug-response is calculated as follow:

$$Score_{patient} = \sum_{i \in G} \beta_i \cdot Exp_{patient,i}$$

where i is gene from the 7 drug-response related genes G, $Exp_{patient,i}$ represents the expression level of gene i of the patient, and β_i is the ridge regression coefficient of gene i from the optimized organoid model. In the random forest model, the patient-specific score is derived from the predicted values. These values are estimated by the model using the expression data of the selected genes. For each validation dataset, the drug-resistant scores were separated into two groups: the drug-resistant group (with score $\geq$ cut point) and the drug-sensitive group (with score $<$ cut point). The maximum rank statistic, which implemented in the MaxStat R package, was used to determine the cut point for each validation dataset. The Kaplan-Meier survival analysis and the log-rank test were used to visualize and evaluate the statistical differences in overall survival (OS) between the two groups.

Functional enrichment analysis of hub genes

The pathway analysis was performed for the selected 35 hub genes using over representation analysis, which was implemented in clusterProfiler R package [49]. Briefly, this method determined whether biological processes that were over-represented in the gene list of interest using p-values calculated by hypergeometric distribution and adjusted by Benjamini-Hochberg

(BH) method to calculate false discover rate (FDR). For the selected hub genes, we analyzed pathways from KEGG, Reactome and GO pathways from the Molecular Signatures Database (MSigDB), which can be accessed by the msigdbr R/Bioconductor package. The minimum and maximum sizes of gene sets used for analysis were set to 10 and 500 respectively.

Gene selection process based on other WGCNA approaches

Additionally, we performed three other WGCNA models to select candidate genes: Model 1, WGCNA based on only the tissue expression (GSE171680); Model 2, consensus WGCNA based on the matched organoid and tissue expressions (GSE171681 and GSE171680); and Model 3, WGCNA based on only the organoid expression with drug response (GSE64932). For Model 1 and 2, the modules were identified as significant based on the Cox regression results regarding the OS outcome. We filtered the hub genes for further model training using $|MM| \geqslant 0.5$. As there was no public drug response data available for this study, we used the OS outcome as response to train three models: ridge, random forest, and the ensemble model. On the other hand, the significance of the modules of Model 3 was determined by the Spearman correlation between the IC_{50} and the eigengene of each module. The same criterion was conducted to select the hub genes and followed by training the organoid models.

Gene selection process based on gene association tests

We compared gene selection methods by conducting gene filtering based on the results of three association tests: 1) Gene-OS Cox regression test, 2) Gene-drug response Spearman correlation test, and 3) Gene-paired Spearman correlation test. Specifically, for the gene-OS Cox regression test, we fit the model to the OS outcome of GSE171680, with each gene as the dependent variable. Each model was adjusted for age and sex to account for confounding variables. The gene-drug Spearman correlation test was conducted between the IC_{50} drug response and each gene of GSE64932. The gene-paired Spearman correlation test was conducted between the gene expression levels of the paired organoid and tissue data. The candidate genes of training the organoid models were selected based on their significance in gene-paired correlations, gene-OS association tests and gene-drug correlation tests. We employed two gene filtering criteria and built organoid models using the genes chosen based on these criteria respectively. The first criterion selects genes where all test result p-values are smaller than 0.05. The second criterion selects genes where all test result p-values are smaller than 0.05 and the sign of the gene-OS and gene-drug coefficients are in agreement.

References:

1. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB: **Colorectal cancer**. *Lancet* 2019, **394**(10207):1467-1480.
2. Siegel RL, Wagle NS, Cercek A, Smith RA, Jemal A: **Colorectal cancer statistics, 2023**. *CA Cancer J Clin* 2023, **73**(3):233-254.
3. Gustavsson B, Carlsson G, Machover D, Petrelli N, Roth A, Schmoll HJ, Tveit KM, Gibson F: **A review of the evolution of systemic chemotherapy in the management of colorectal cancer**. *Clin Colorectal Cancer* 2015, **14**(1):1-10.
4. Conroy T, Bosset JF, Etienne PL, Rio E, Francois E, Mesgouez-Nebout N, Vendrely V, Artignan X, Bouche O, Gargot D *et al*: **Neoadjuvant chemotherapy with FOLFIRINOX and preoperative chemoradiotherapy for patients with locally advanced rectal cancer (UNICANCER-PRODIGE 23): a multicentre, randomised, open-label, phase 3 trial**. *Lancet Oncol* 2021, **22**(5):702-715.
5. Bascoul-Mollevi C, Gourgou S, Borg C, Etienne PL, Rio E, Rullier E, Juzyna B, Castan F, Conroy T: **Neoadjuvant chemotherapy with FOLFIRINOX and preoperative chemoradiotherapy for patients with locally advanced rectal cancer (UNICANCER PRODIGE 23): Health-related quality of life longitudinal analysis**. *Eur J Cancer* 2023, **186**:151-165.
6. Schrag D, Shi Q, Weiser MR, Gollub MJ, Saltz LB, Musher BL, Goldberg J, Al Baghdadi T, Goodman KA, McWilliams RR *et al*: **Preoperative Treatment of Locally Advanced Rectal Cancer**. *N Engl J Med* 2023, **389**(4):322-334.
7. Salazar R, Roepman P, Capella G, Moreno V, Simon I, Dreezen C, Lopez-Doriga A, Santos C, Marijnen C, Westerga J *et al*: **Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer**. *J Clin Oncol* 2011, **29**(1):17-24.
8. O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, Lopatin M, Watson D, Baehner FL, Shak S, Baker J *et al*: **Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin**. *J Clin Oncol* 2010, **28**(25):3937-3944.
9. Lenehan PF, Boardman LA, Riegert-Johnson D, De Petris G, Fry DW, Ohrnberger J, Heyman ER, Gerard B, Almal AA, Worzel WP: **Generation and external validation of a tumor-derived 5-gene prognostic signature for recurrence of lymph node-negative, invasive colorectal carcinoma**. *Cancer* 2012, **118**(21):5234-5244.
10. Kennedy RD, Bylesjo M, Kerr P, Davison T, Black JM, Kay EW, Holt RJ, Proutski V, Ahdesmaki M, Farztdinov V *et al*: **Development and independent validation of a prognostic assay for stage II colon cancer using formalin-fixed paraffin-embedded tissue**. *J Clin Oncol* 2011, **29**(35):4620-4626.
11. Park YY, Lee SS, Lim JY, Kim SC, Kim SB, Sohn BH, Chu IS, Oh SC, Park ES, Jeong W *et al*: **Comparison of prognostic genomic predictors in colorectal cancer**. *PLoS One* 2013, **8**(4):e60778.
12. Di Narzo AF, Tejpar S, Rossi S, Yan P, Popovici V, Wirapati P, Budinska E, Xie T, Estrella H, Pavlicek A *et al*: **Test of four colon cancer risk-scores in formalin fixed paraffin embedded microarray gene expression data**. *J Natl Cancer Inst* 2014, **106**(10).

13. Gillet JP, Varma S, Gottesman MM: **The clinical relevance of cancer cell lines**. *J Natl Cancer Inst* 2013, **105**(7):452-458.
14. Borst P, Wessels L: **Do predictive signatures really predict response to cancer chemotherapy?** *Cell Cycle* 2010, **9**(24):4836-4840.
15. Burrell RA, McGranahan N, Bartek J, Swanton C: **The causes and consequences of genetic heterogeneity in cancer evolution**. *Nature* 2013, **501**(7467):338-345.
16. Crawford J, Greene CS: **Incorporating biological structure into machine learning models in biomedicine**. *Curr Opin Biotechnol* 2020, **63**:126-134.
17. Berisha V, Krantsevich C, Hahn PR, Hahn S, Dasarathy G, Turaga P, Liss J: **Digital medicine and the curse of dimensionality**. *NPJ Digit Med* 2021, **4**(1):153.
18. Barbour DL: **Precision medicine and the cursed dimensions**. *NPJ Digit Med* 2019, **2**:4.
19. Ganesh K, Wu C, O'Rourke KP, Szeglin BC, Zheng Y, Sauve CG, Adileh M, Wasserman I, Marco MR, Kim AS *et al*: **A rectal cancer organoid platform to study individual responses to chemoradiation**. *Nat Med* 2019, **25**(10):1607-1614.
20. Veninga V, Voest EE: **Tumor organoids: Opportunities and challenges to guide precision medicine**. *Cancer Cell* 2021, **39**(9):1190-1201.
21. Drost J, Clevers H: **Organoids in cancer research**. *Nat Rev Cancer* 2018, **18**(7):407-418.
22. van de Wetering M, Francies HE, Francis JM, Bounova G, Iorio F, Pronk A, van Houdt W, van Gorp J, Taylor-Weiner A, Kester L *et al*: **Prospective derivation of a living organoid biobank of colorectal cancer patients**. *Cell* 2015, **161**(4):933-945.
23. Ooft SN, Weeber F, Dijkstra KK, McLean CM, Kaing S, van Werkhoven E, Schipper L, Hoes L, Vis DJ, van de Haar J *et al*: **Patient-derived organoids can predict response to chemotherapy in metastatic colorectal cancer patients**. *Sci Transl Med* 2019, **11**(513).
24. Vlachogiannis G, Hedayat S, Vatsiou A, Jamin Y, Fernandez-Mateos J, Khan K, Lampis A, Eason K, Huntingford I, Burke R *et al*: **Patient-derived organoids model treatment response of metastatic gastrointestinal cancers**. *Science* 2018, **359**(6378):920-926.
25. Weeber F, van de Wetering M, Hoogstraat M, Dijkstra KK, Krijgsman O, Kuilman T, Gadellaa-van Hooijdonk CG, van der Velden DL, Peeper DS, Cuppen EP *et al*: **Preserved genetic diversity in organoids cultured from biopsies of human colorectal cancer metastases**. *Proc Natl Acad Sci U S A* 2015, **112**(43):13308-13311.
26. Kong J, Lee H, Kim D, Han SK, Ha D, Shin K, Kim S: **Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients**. *Nat Commun* 2020, **11**(1):5485.
27. Dunne PD, Alderdice M, O'Reilly PG, Roddy AC, McCorry AMB, Richman S, Maughan T, McDade SS, Johnston PG, Longley DB *et al*: **Cancer-cell intrinsic gene expression signatures overcome intratumoural heterogeneity bias in colorectal cancer patient classification**. *Nat Commun* 2017, **8**:15657.
28. Cho EJ, Kim M, Jo D, Kim J, Oh JH, Chung HC, Lee SH, Kim D, Chun SM, Kim J *et al*: **Immuno-genomic classification of colorectal cancer organoids reveals cancer cells with intrinsic immunogenic properties associated with patient survival**. *J Exp Clin Cancer Res* 2021, **40**(1):230.
29. Ooft SN, Weeber F, Schipper L, Dijkstra KK, McLean CM, Kaing S, van de Haar J, Prevoo W, van Werkhoven E, Snaebjornsson P *et al*: **Prospective experimental treatment of**

colorectal cancer patients based on organoid drug responses. *ESMO Open* 2021, **6**(3):100103.
30. Hsu KS, Adileh M, Martin ML, Makarov V, Chen J, Wu C, Bodo S, Klingler S, Sauve CG, Szeglin BC *et al*: **Colorectal Cancer Develops Inherent Radiosensitivity That Can Be Predicted Using Patient-Derived Organoids**. *Cancer Res* 2022, **82**(12):2298-2312.
31. Yokota E, Iwai M, Yukawa T, Yoshida M, Naomoto Y, Haisa M, Monobe Y, Takigawa N, Guo M, Maeda Y *et al*: **Clinical application of a lung cancer organoid (tumoroid) culture system**. *NPJ Precis Oncol* 2021, **5**(1):29.
32. de Witte CJ, Espejo Valle-Inclan J, Hami N, Lohmussaar K, Kopper O, Vreuls CPH, Jonges GN, van Diest P, Nguyen L, Clevers H *et al*: **Patient-Derived Ovarian Cancer Organoids Mimic Clinical Response and Exhibit Heterogeneous Inter- and Intrapatient Drug Responses**. *Cell Rep* 2020, **31**(11):107762.
33. Yao Y, Xu X, Yang L, Zhu J, Wan J, Shen L, Xia F, Fu G, Deng Y, Pan M *et al*: **Patient-Derived Organoids Predict Chemoradiation Responses of Locally Advanced Rectal Cancer**. *Cell Stem Cell* 2020, **26**(1):17-26 e16.
34. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.
35. Langfelder P, Horvath S: **Eigengene networks for studying the relationships between co-expression modules**. *BMC Syst Biol* 2007, **1**:54.
36. Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks MP, van Eijk K, van den Berg LH, Ophoff RA: **Aging effects on DNA methylation modules in human brain and blood tissue**. *Genome Biol* 2012, **13**(10):R97.
37. Alvarez MJ, Shen Y, Giorgi FM, Lachmann A, Ding BB, Ye BH, Califano A: **Functional characterization of somatic mutations in cancer using network-based inference of protein activity**. *Nat Genet* 2016, **48**(8):838-847.
38. Colaprico A, Olsen C, Bailey MH, Odom GJ, Terkelsen T, Silva TC, Olsen AV, Cantini L, Zinovyev A, Barillot E *et al*: **Interpreting pathways to discover cancer driver genes with Moonlight**. *Nat Commun* 2020, **11**(1):69.
39. Neal JT, Li X, Zhu J, Giangarra V, Grzeskowiak CL, Ju J, Liu IH, Chiou SH, Salahudeen AA, Smith AR *et al*: **Organoid Modeling of the Tumor Immune Microenvironment**. *Cell* 2018, **175**(7):1972-1988 e1916.
40. Marisa L, de Reynies A, Duval A, Selves J, Gaub MP, Vescovo L, Etienne-Grimaldi MC, Schiappa R, Guenot D, Ayadi M *et al*: **Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value**. *PLoS Med* 2013, **10**(5):e1001453.
41. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE *et al*: **Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer**. *Gastroenterology* 2010, **138**(3):958-968.
42. Zhu J, Deane NG, Lewis KB, Padmanabhan C, Washington MK, Ciombor KK, Timmers C, Goldberg RM, Beauchamp RD, Chen X: **Evaluation of frozen tissue-derived prognostic gene expression signatures in FFPE colorectal cancer samples**. *Sci Rep* 2016, **6**:33273.
43. Del Rio M, Mollevi C, Bibeau F, Vie N, Selves J, Emile JF, Roger P, Gongora C, Robert J, Tubiana-Mathieu N *et al*: **Molecular subtypes of metastatic colorectal cancer are**

associated with patient response to irinotecan-based therapies. *Eur J Cancer* 2017, **76**:68-75.

44. Hu Y, Gaedcke J, Emons G, Beissbarth T, Grade M, Jo P, Yeager M, Chanock SJ, Wolff H, Camps J *et al*: **Colorectal cancer susceptibility loci as predictive markers of rectal cancer prognosis after surgery**. *Genes Chromosomes Cancer* 2018, **57**(3):140-149.
45. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. *Biostatistics* 2003, **4**(2):249-264.
46. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I *et al*: **TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data**. *Nucleic Acids Res* 2016, **44**(8):e71.
47. Zhang B, Horvath S: **A general framework for weighted gene co-expression network analysis**. *Stat Appl Genet Mol Biol* 2005, **4**:Article17.
48. Yip AM, Horvath S: **Gene network interconnectedness and the generalized topological overlap measure**. *BMC Bioinformatics* 2007, **8**:22.
49. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L *et al*: **clusterProfiler 4.0: A universal enrichment tool for interpreting omics data**. *Innovation (Camb)* 2021, **2**(3):100141.

Figure 1 - Workflow for the identification of gene biomarkers with WGCNA and development of organoid model. The process initiates with the selection of three colorectal cancer datasets for a consensus WGCNA. Potential gene biomarkers are then chosen to train the organoid models with GSE64392 and its drug response 5-FU. Subsequently, genes related to drug response are identified and utilized for prognosis testing on six independent colorectal cancer patient expression datasets.

Figure 2 – Construction of consensus WGCNA . (A) shows the scale-free topology model fit (y-axis) under different power values (x-axis); (B) shows the average connectivity (y-axis) under different power values (x-axis); (C) shows the dendrogram of the consensus module clustering based on the dissimilarity measure (1- consensus TOM); (D) the heatmap on the left plots the relationship of consensus module eigengenes and prognosis results of GSE171680; the heatmap on the right indicates the Spearman correlation between consensus module eigengenes of GSE171680 and GSE171681.

Figure 3 - Plots of significant modules: tan and salmon. (A) shows the scatterplots between eigengenes of GSE171680 (x-axis) and GSE171681 (y-axis) of tan (top) and salmon (bottom), with Spearman correlations and P-values labeled; (B) shows the consensus network of the tan and salmon modules (hub genes colored in brown); (C) shows the scatterplots of the module membership (MM) among all three datasets of tan (top) and salmon (bottom) modules.

Figure 4 - Drug-response predictions for 5-FU-based treated samples of six independent datasets. The predicted drug-resistant scores were divided into drug-sensitive and drug-resistant group and tested on the overall survival results from six independent datasets. Statistical significance was measured using Kaplan–Meier survival curves and log-rank tests. P-values <0.05 were considered significant.

Figure 5 - Heatmap of 7 drug-related biomarkers and functional enrichment plots of 35 hub genes selected from WGCNA. (A) displays a heatmap of seven drug-related biomarkers, accompanied by a hierarchical clustering dendrogram at the top. The coefficients of the biomarkers are depicted on the left. (B) presents dot plots for the functional enrichment analysis of 35 hub genes, selected from WGCNA. All significant pathways are included in the plot.

Supplementary Figure 1 - Showcases the cross-validation (CV) errors and ROC curves of the Ridge, RF, and ensemble models. On the left are boxplots of 100 repeated CV errors. The t-test results between Ridge vs ensemble and RF vs ensemble are displayed at the top of the boxplots, where "**" represents P-values < 0.01 and "****" signifies P-values < 0.0001. On the right are the ROC curves from testing on the binary label of the OS results of GSE171680 across the three models. The area under the curve (AUC) values are labeled for each model.

Supplementary Figure 2 - Drug-response predictions for 5-FU-based treated samples of six independent datasets with 35 hub genes selected from the consensus WGCNA

Supplementary Figure 3 - Drug-response predictions for 5-FU-based treated samples of six independent datasets with candidate genes selected from WGCNA Model 1

Supplementary Figure 4 - Drug-response predictions for 5-FU-based treated samples of six independent datasets with candidate genes selected from WGCNA Model 2

Supplementary Figure 5 - Drug-response predictions for 5-FU-based treated samples of six independent datasets with candidate genes selected from WGCNA Model 3

Supplementary Figure 6 - Drug-response predictions for 5-FU-based treated samples of six independent datasets with candidate genes selected from filtering criterion 1 of gene association tests

Supplementary Figure 7 - Drug-response predictions for 5-FU-based treated samples of six independent datasets with candidate genes selected from filtering criterion 2 of gene association tests

DECLARATIONS

Ethics approval and consent to participate
Not Applicable

Availability of data and materials
All datasets analyzed in this study are publicly available and listed in Table 1. The analysis code is available at https://github.com/TransBioInfoLab/Organoid-Prediction.

Competing Interest
The authors declare that they have no conflicts of interest.

Funding
This research was supported by US National Cancer Institute grant R37CA248289 (J.J.S), and Sylvester Comprehensive Cancer Center Intramural program SCCC-NIH-2022-11 (X.S.C).

Table 1A Gene expression training datasets

Dataset	Data type	Array type	Number of genes	Number of samples	Have drug response	Number of OS events	Median OS follow-up time	Number of RFS events	Median RFS follow-up time
GSE171680	Tissue	RNAseq	20501	87	No	15	38.27	19	29.90
GSE171681	Organoid	RNAseq	20501	87	No	15	38.27	19	29.90
GSE64392	Organoid	Microarray	25988	19	Yes	NA	NA	NA	NA

Table 1B Gene expression testing datasets

Datasets	Array type	Number of genes	Number of samples	Number of samples with 5-FU based chemotherapy	Number of OS events	Median follow-up time
GSE39582	Microarray	20824	167	102	30	41.00
GSE17538	Microarray	20824	232	88	40	47.83
TCGA-COAD	RNA-Seq	19462	456	122	18	25.15
GSE106584	Microarray	23145	156	83	37	81.27
GSE72970	Microarray	20824	124	124	92	22.80
GSE87211	Microarray	20816	203	203	28	62.25

Table 2 The seven genes selected by the ensemble organoid model

Gene	Coefficient	Module
CELP	-0.0798	tan
CPN1	-0.0282	tan
NEURL2	-0.0162	tan
PIPOX	-0.0392	tan
SLC19A3	-0.0568	tan
VAV3	0.0012	tan
HOXB13	-0.0325	salmon

Table 3 – Prediction results of the main models and all other gene selection processes.

Validation datasets	WGCNA based process					Three association tests based process	
	Main model		Other approaches			Criterion 1: Significant genes in three tests	Criterion 2: Concordant genes only
	Drug-related biomarkers	Hub genes	Model 1	Model 2	Model 3		
	Log-rank test P-value						
GSE39582	1.32E-04	2.00E-04	2.13E-02	1.06E-01	2.13E-01	1.36E-01	9.16E-02
GSE17538	7.09E-04	8.70E-04	2.56E-01	2.39E-02	1.53E-01	1.48E-01	1.36E-01
TCGA-COAD	4.95E-02	7.90E-02	1.32E-01	7.52E-03	7.26E-02	1.88E-01	6.88E-02
GSE106584	2.08E-02	3.70E-02	1.36E-01	4.84E-02	7.11E-02	2.19E-01	4.59E-02
GSE72970	4.60E-03	5.40E-04	2.25E-02	5.49E-02	4.42E-02	2.42E-02	3.04E-01
GSE87211	1.25E-03	1.80E-02	5.44E-03	1.63E-02	5.65E-02	2.87E-01	5.09E-02
Organoid model	Ensemble	Ridge	Ensemble	RF	Ensemble	Ensemble	Ensemble
Number of modules selected	2	2	3	1	7		
Number of genes selected	7	35	73	8	17	27	21

Fig.1

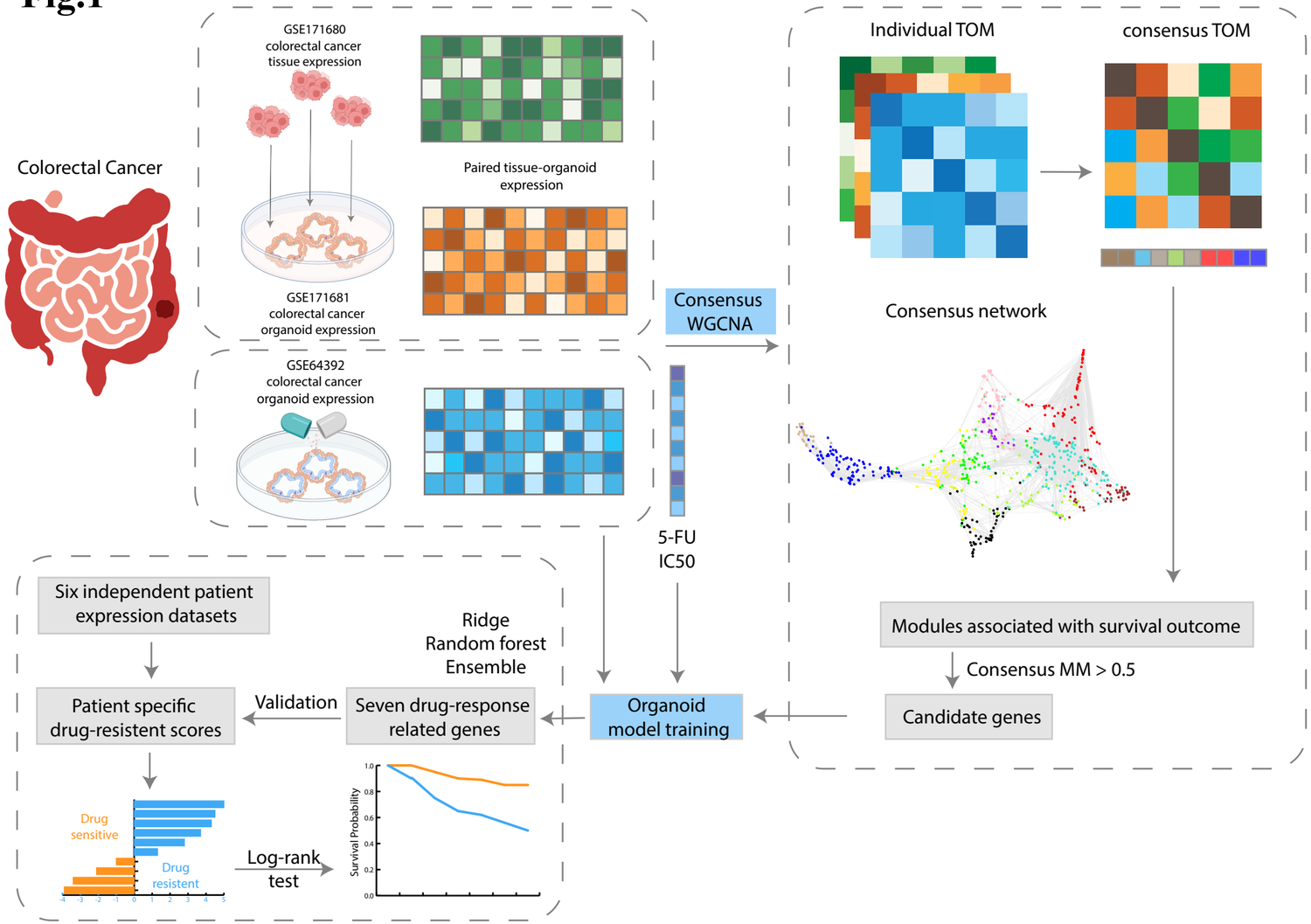
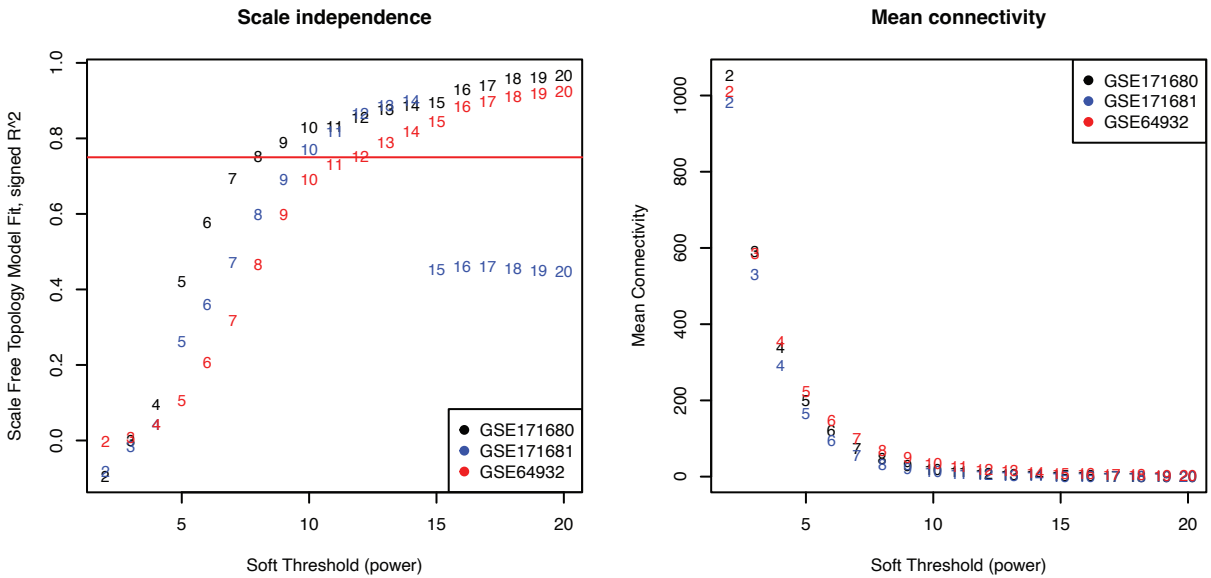
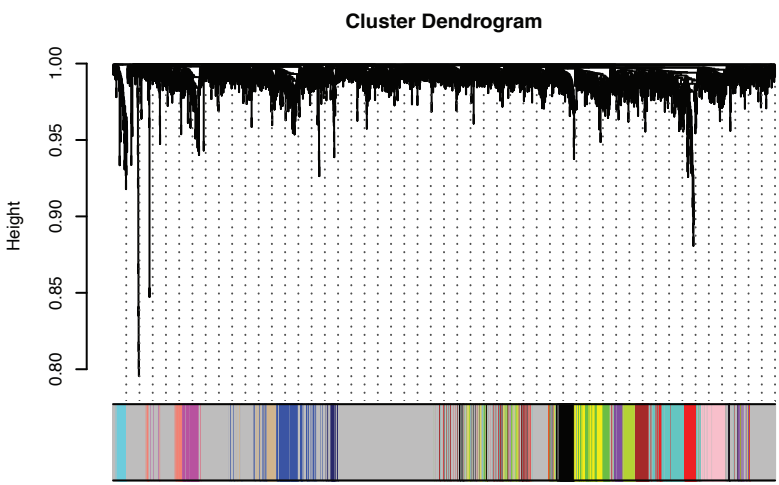

Colorectal Cancer
GSE171680
colorectal cancer
tissue expression
Paired tissue-organoid
expression
GSE171681
colorectal cancer
organoid expression
GSE64392
colorectal cancer
organoid expression
Consensus
WGCNA
5-FU
IC50
Individual TOM
consensus TOM
Consensus network
Modules associated with survival outcome
Consensus MM > 0.5
Candidate genes
Organoid
model training
Ridge
Random forest
Ensemble
Seven drug-response
related genes
Validation
Patient specific
drug-resistent scores
Six independent patient
expression datasets
Drug
sensitive
Drug
resistent
Log-rank
test
Survival Probability

Fig. 2

A


Scale independence
Scale Free Topology Model Fit, signed R^2
Soft Threshold (power)
GSE171680
GSE171681
GSE64932
Mean connectivity
Mean Connectivity
Soft Threshold (power)
GSE171680
GSE171681
GSE64932

B


Cluster Dendrogram
Height

C

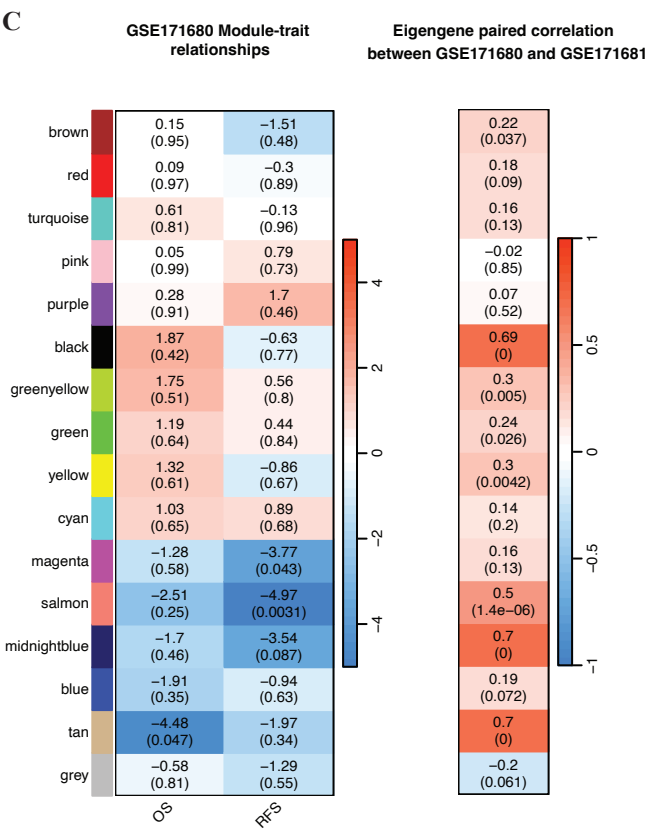

GSE171680 Module-trait relationships
Eigengene paired correlation between GSE171680 and GSE171681
brown 0.15 (0.95) −1.51 (0.48) 0.22 (0.037)
red 0.09 (0.97) −0.3 (0.89) 0.18 (0.09)
turquoise 0.61 (0.81) −0.13 (0.96) 0.16 (0.13)
pink 0.05 (0.99) 0.79 (0.73) −0.02 (0.85)
purple 0.28 (0.91) 1.7 (0.46) 0.07 (0.52)
black 1.87 (0.42) −0.63 (0.77) 0.69 (0)
greenyellow 1.75 (0.51) 0.56 (0.8) 0.3 (0.005)
green 1.19 (0.64) 0.44 (0.84) 0.24 (0.026)
yellow 1.32 (0.61) −0.86 (0.67) 0.3 (0.0042)
cyan 1.03 (0.65) 0.89 (0.68) 0.14 (0.2)
magenta −1.28 (0.58) −3.77 (0.043) 0.16 (0.13)
salmon −2.51 (0.25) −4.97 (0.0031) 0.5 (1.4e−06)
midnightblue −1.7 (0.46) −3.54 (0.087) 0.7 (0)
blue −1.91 (0.35) −0.94 (0.63) 0.19 (0.072)
tan −4.48 (0.047) −1.97 (0.34) 0.7 (0)
grey −0.58 (0.81) −1.29 (0.55) −0.2 (0.061)
OS
RFS

Fig. 3

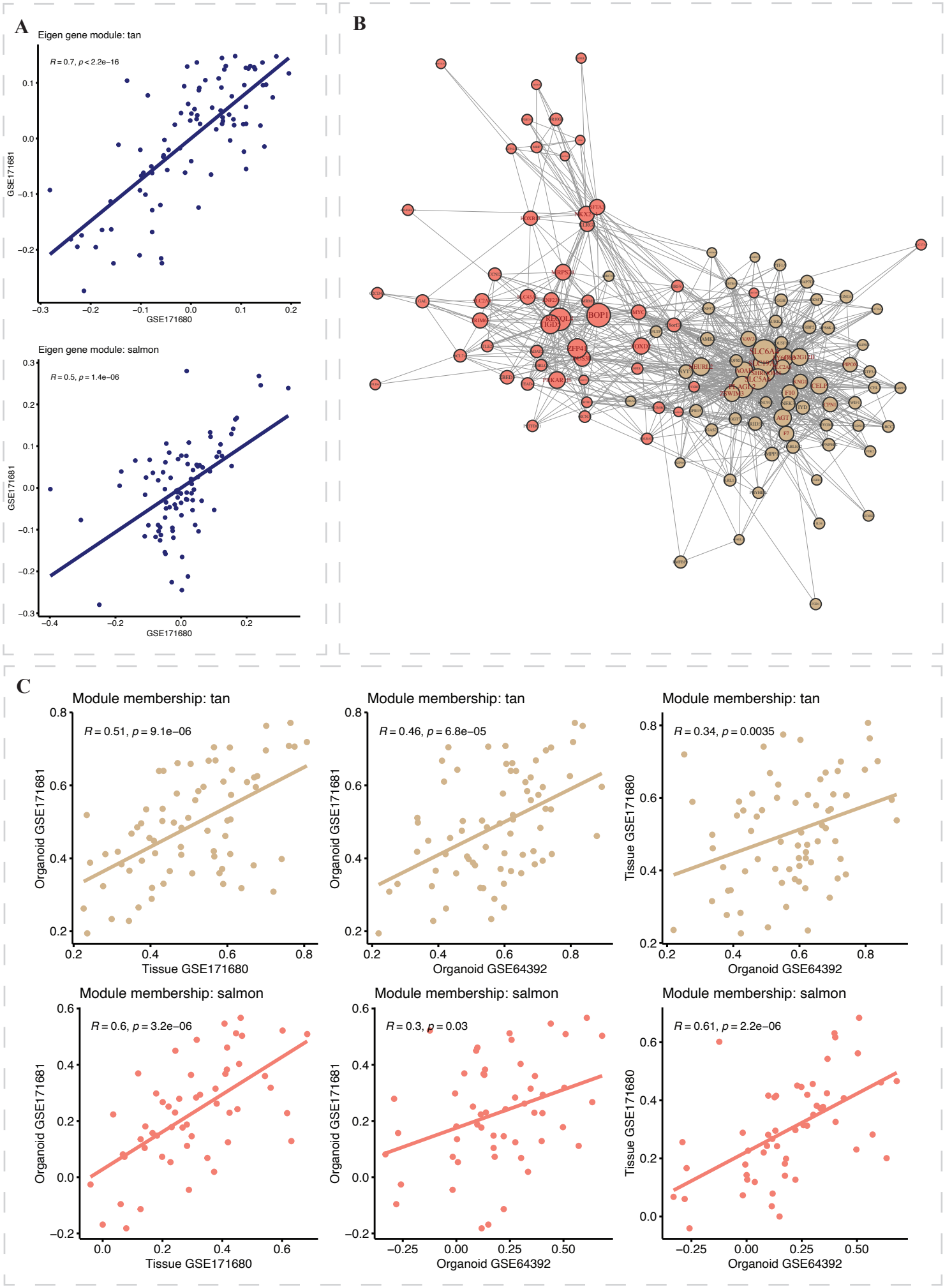
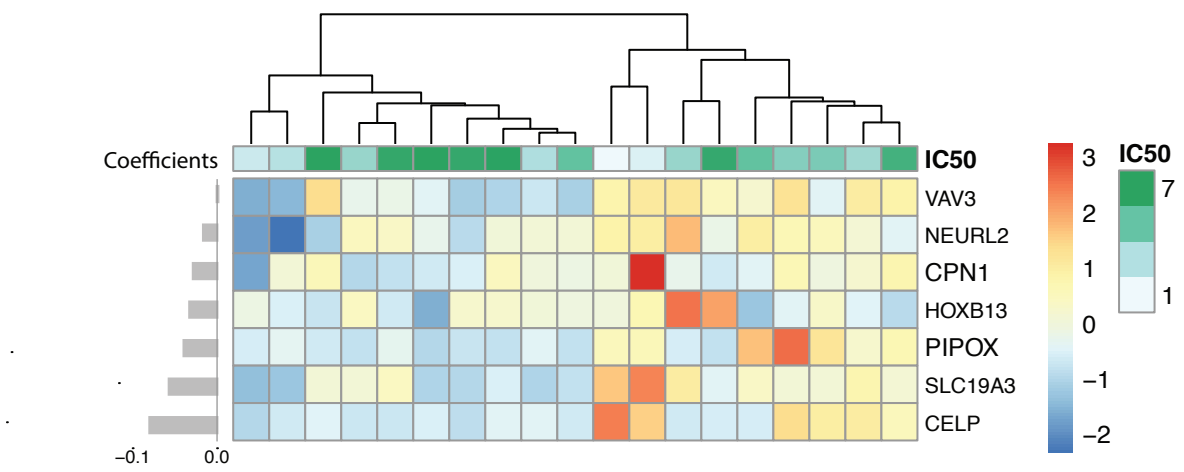

A
Eigen gene module: tan
R = 0.7, p < 2.2e−16
GSE171681
GSE171680
Eigen gene module: salmon
R = 0.5, p = 1.4e−06
GSE171681
GSE171680
B
C
Module membership: tan
R = 0.51, p = 9.1e−06
Organoid GSE171681
Tissue GSE171680
Module membership: tan
R = 0.46, p = 6.8e−05
Organoid GSE171681
Organoid GSE64392
Module membership: tan
R = 0.34, p = 0.0035
Tissue GSE171680
Organoid GSE64392
Module membership: salmon
R = 0.6, p = 3.2e−06
Organoid GSE171681
Tissue GSE171680
Module membership: salmon
R = 0.3, p = 0.03
Organoid GSE171681
Organoid GSE64392
Module membership: salmon
R = 0.61, p = 2.2e−06
Tissue GSE171680
Organoid GSE64392

Fig.4

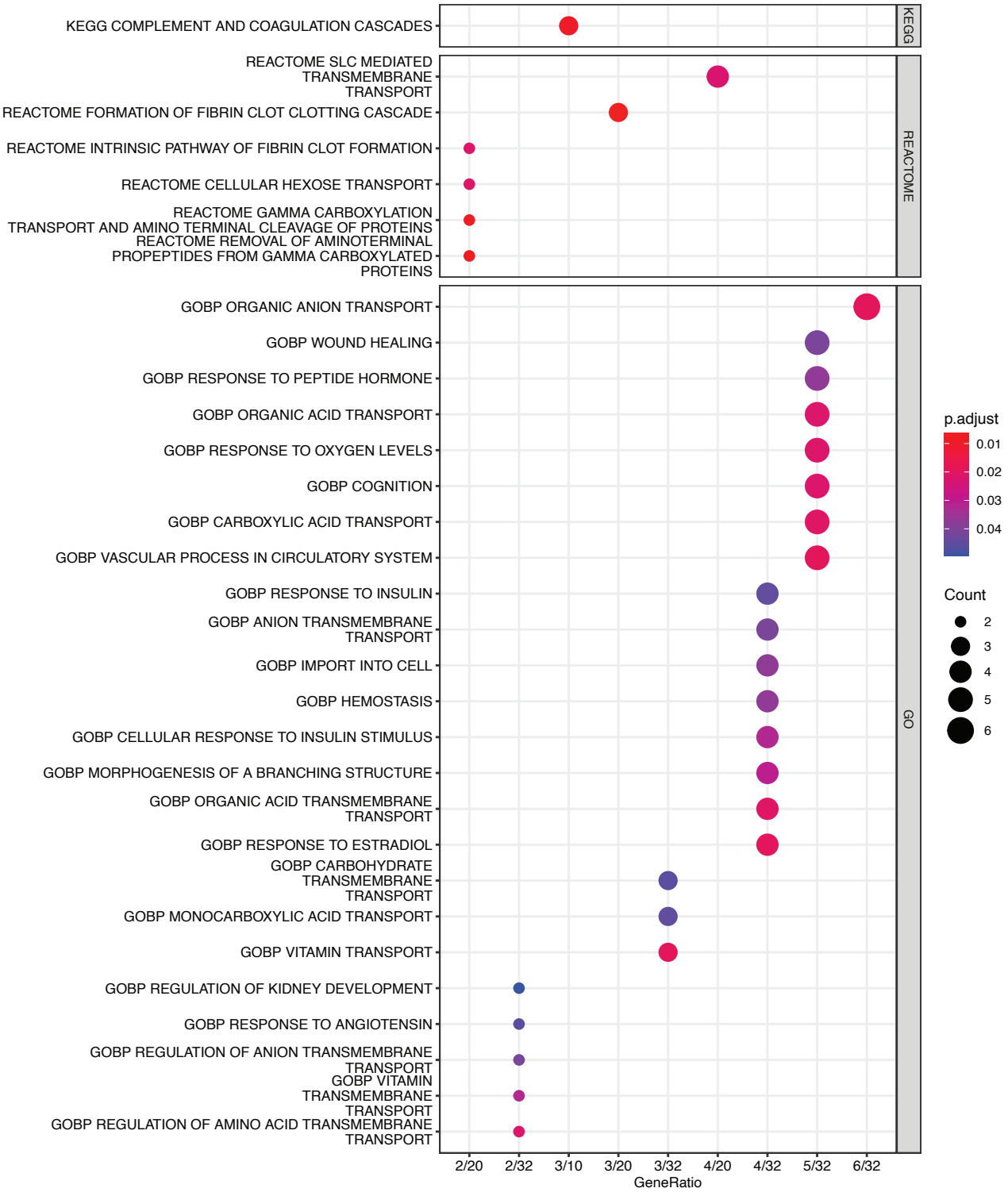
GSE39582
Strata group=Drug−resistant group=Drug−sensitive
Survival probability
p = 0.00013
Time
Number at risk
Strata
43 16 3 1 0
59 36 17 3 1
0 40 80 120 160
Time
GSE106584
Strata group=Drug−resistant group=Drug−sensitive
Survival probability
p = 0.021
Time
Number at risk
Strata
64 43 22 7 3 0
19 14 10 7 3 0
0 50 100 150 200 250
Time
GSE17538
Strata group=Drug−resistant group=Drug−sensitive
Survival probability
p = 0.00071
Time
Number at risk
Strata
15 6 3 1 0
73 53 23 10 1
0 30 60 90 120
Time
GSE72970
Strata group=Drug−resistant group=Drug−sensitive
Survival probability
p = 0.0046
Time
Number at risk
Strata
100 48 16 4 0
24 16 12 4 0
0 20 40 60 80
Time
TCGA−COAD
Strata group=Drug−resistant group=Drug−sensitive
Survival probability
p = 0.05
Time
Number at risk
Strata
24 1 0 0
87 10 1 0
0 50 100 150
Time
GSE87211
Strata group=Drug−resistant group=Drug−sensitive
Survival probability
p = 0.0012
Time
Number at risk
Strata
54 29 2 0
142 79 18 2
0 50 100 150
Time

Fig.5

A


Coefficients
IC50
VAV3
NEURL2
CPN1
HOXB13
PIPOX
SLC19A3
CELP
−0.1
0.0
3
2
1
0
−1
−2
IC50
7
1

B


KEGG COMPLEMENT AND COAGULATION CASCADES
KEGG
REACTOME SLC MEDIATED TRANSMEMBRANE TRANSPORT
REACTOME FORMATION OF FIBRIN CLOT CLOTTING CASCADE
REACTOME INTRINSIC PATHWAY OF FIBRIN CLOT FORMATION
REACTOME CELLULAR HEXOSE TRANSPORT
REACTOME GAMMA CARBOXYLATION TRANSPORT AND AMINO TERMINAL CLEAVAGE OF PROTEINS
REACTOME REMOVAL OF AMINOTERMINAL PROPEPTIDES FROM GAMMA CARBOXYLATED PROTEINS
REACTOME
GOBP ORGANIC ANION TRANSPORT
GOBP WOUND HEALING
GOBP RESPONSE TO PEPTIDE HORMONE
GOBP ORGANIC ACID TRANSPORT
GOBP RESPONSE TO OXYGEN LEVELS
GOBP COGNITION
GOBP CARBOXYLIC ACID TRANSPORT
GOBP VASCULAR PROCESS IN CIRCULATORY SYSTEM
GOBP RESPONSE TO INSULIN
GOBP ANION TRANSMEMBRANE TRANSPORT
GOBP IMPORT INTO CELL
GOBP HEMOSTASIS
GOBP CELLULAR RESPONSE TO INSULIN STIMULUS
GOBP MORPHOGENESIS OF A BRANCHING STRUCTURE
GOBP ORGANIC ACID TRANSMEMBRANE TRANSPORT
GOBP RESPONSE TO ESTRADIOL
GOBP CARBOHYDRATE TRANSMEMBRANE TRANSPORT
GOBP MONOCARBOXYLIC ACID TRANSPORT
GOBP VITAMIN TRANSPORT
GOBP REGULATION OF KIDNEY DEVELOPMENT
GOBP RESPONSE TO ANGIOTENSIN
GOBP REGULATION OF ANION TRANSMEMBRANE TRANSPORT
GOBP VITAMIN TRANSMEMBRANE TRANSPORT
GOBP REGULATION OF AMINO ACID TRANSMEMBRANE TRANSPORT
GO
2/20
2/32
3/10
3/20
3/32
4/20
4/32
5/32
6/32
GeneRatio
p.adjust
0.01
0.02
0.03
0.04
Count
2
3
4
5
6