Pipeline for Characterizing Alternative Mechanisms (PCAM) based on bi-clustering to study colorectal cancer heterogeneity

Sha Cao[1,2*], Wennan Chang[1,4], Changlin Wan[1,4], Yong Zang[1,2], Yijie Wang[5], Qin Ma[6*], Chi Zhang[1,3*]

[1]Center for Computational Biology and Bioinformatics, [2]Department of Biostatistics, [3]Department of Medical and Molecular Genetics, Indiana University, School of Medicine, Indianapolis, IN,46202, USA.

[4]Department of Electronic Computer Engineering, Purdue University, West Lafayette, IN 47907

[5]Department of Computer Science, Indiana University, Bloomington, IN, 43210

[6]Department of Biomedical Informatics, College of Medicine, the Ohio State University, Columbus, OH, 43210

*To whom correspondence should be addressed. +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Sha Cao: shcao@iu.edu and Qin Ma: maqin2001@gmail.com.

ABSTRACT

The cells of colorectal cancer (CRC) in their microenvironment experience constant stress, leading to dysregulated activity in the tumor niche. As a result, cancer cells acquire alternative pathways in response to the changing microenvironment, posing significant challenges for the design of effective cancer treatment strategies. While computational studies on high-throughput omics data have advanced our understanding of colorectal cancer subtypes, characterizing the heterogeneity of this disease remains remarkably complex. Here, we present a novel computational Pipeline for Characterizing Alternative Mechanisms (PCAM) based on biclustering to gain a more detailed understanding of cancer heterogeneity. Our application of PCAM to large-scale CRC transcriptomics datasets suggests that PCAM can generate a wealth of information leading to new biological understanding and predictive markers of alternative mechanisms. Our key findings include: 1) A comprehensive collection of alternative pathways in CRC, associated with biological and clinical factors. 2) Full annotation of detected alternative mechanisms, including their enrichment in known pathways and associations with various clinical outcomes. 3) A mechanistic relationship between known clinical subtypes and outcomes on a consensus map, visualized by the presence of alternative mechanisms. 4) Several potential novel alternative drug resistance mechanisms for Oxaliplatin, 5-Fluorouracil, and FOLFOX, some of which were validated on independent datasets. We believe that gaining a deeper understanding of alternative mechanisms is a critical step towards characterizing the heterogeneity of colorectal cancer. The hypotheses generated by PCAM, along with the comprehensive collection of biologically and clinically associated alternative pathways in CRC, could provide valuable insights into the underlying mechanisms driving cancer progression and drug resistance, which could aid in the development of more effective cancer therapies and guide experimental design towards more targeted and personalized treatment strategies.

Introduction

Colorectal cancer (CRC) is the third most frequent cancer type in the United States, which accounts for an estimated 8% of adult cancer incidence and more than 8% cancer deaths in 2023 (1). Epidemiology data suggests the average five-year survival rate of CRC is 64.9%, while more than 80% of patients die from the disease in five years in the case of metastasis (2,3). The tumor heterogeneity of CRC presents significant challenges in designing effective treatment strategies (4-6). The varying levels of sensitivity that different patient populations exhibit when receiving cytotoxic drugs can make it difficult to achieve successful therapies in heterogenic tumors (7).

A few molecular subtyping methods have been developed for CRC, aiming to facilitate personalized treatment (8-16). Among these, the Consensus Molecular Subtype (CMS) has been accepted as a standard CRC stratification (8,9). CMS was derived from a cohort of 18 independent gene expression data sets with 4,151 CRC samples, and it has stratified CRC patients into four classes with distinct molecular features and prognoses (8). Such patient stratifications that may predict treatment response or prognosis may not be widely applicable, as the genetic differences within and between CRC tumors are far more complicated, and further research into more comprehensive descriptions of CRC heterogeneity is still in progress (8). Based on CMS classification, deeper investigation into the molecular and phenotypic distinctions within each subtype has also been carried out. Though these works have largely contributed to the characterization of CRC heterogeneity, they are however under powered: the statistical power of detecting subtype biomarkers drops remarkably with more refined patients stratifications, where the samples attributed to each subtype becomes smaller. On the other hand, since all the subtypes are defined based on known clinical and/or biological characteristics, it will inevitably limit our power in identifying alternative mechanisms that could lead to novel clinical implications. These largely undermine the practicality of CMS classification in clinical translation. It is thus imperative to develop a computational framework to comprehensively detect alternative mechanisms in CRC in light of the inter-tumor heterogeneity, which is not restricted to existing molecular classifications, such as CMS.

Clearly, tumor heterogeneity has substantially hurdled the computational capability in mining gene expression data for studying the disease complexities. This is because the gene regulatory pathways are interwoven to ensure the robustness of the spatial and temporal regulation of the cell functions, resulting in multi-pathways from one stimulus to a single target (17,18). Thus, the set of genes used to execute a biological or clinical response may very likely exist in more than one alternative forms, and cells under different circumstances in different patient populations may likely choose to select any of them.

In light of the above challenges, we developed a novel computational **P**ipeline for **C**haracterizing **A**lternative **M**echanisms (**PCAM**) to study CRC heterogeneity. PCAM models alternative forms of biological/clinical response in a large gene expression matrix as submatrices, wherein the genes in the row subset correspond to the genes used to execute the response, and the samples in the column subset correspond to the patient circumstances that such an alternative form is activated and used. And PCAM replies on bi-clustering to identify such submatrices.

Bi-clustering analysis is a technique to identify gene co-expression structures specific to certain and sometimes to-be-identified subsets of samples (19,20). The algorithm outputs data blocks, each containing subset of samples and features in a sub-matrix format, called bi-clusters (BC). Our recently released bi-clustering tool QUBIC-R enables identification of BCs in whole-genome transcriptomics data set and has shown competitive performance (21-23). Bi-clustering algorithms have previously been used to study cancer expression datasets (24-26), to find clusters of patient samples, where in each cluster, the co-expressed genes may differ. However, these studies underestimate the complexities of cancer. Consider the complexity in whole genome transcriptional regulatory programs and the patient heterogeneity, there should exist many ways that the samples can be clustered. In other words, the level of similarity for two samples could vary drastically when looking at different biological pathways. The core algorithm of PCAM, QUBIC-R, could comprehensively identify all the significant submatrices in a large gene expression matrix, from tens to thousands of rows/columns. Therefore, in PCAM, one sample could fall into multiple BCs, allowing one sample to be involved in multiple activated response pathways.

Application of PCAM on a large collection of CRC gene expression datasets produced a wealth of information. PCAM fully recognizes the large heterogeneity within CRC patients, some of which may be strongly associated with existing CRC sub-classes defined by various clinical and genomic features, while the rest will provide novel alternative ways for us to better understand the disease. We believe PCAM is suitable for in-depth discovery of alternative biological mechanisms, by systematic interrogation of the disease in different clinical settings without compromising the analysis power.

Materials and Methods

Data collection

We have collected transcriptomics data of 1,440 colorectal cancer tissue samples including one RNA-Seq data from TCGA (The Cancer Genome Atlas) and seven microarray data sets from GEO (Gene Expression Omnibus) database. The micro-array datasets are selected with the following criteria: (1) data are collected by the top 10 most frequently utilized human microarray platforms in GEO database; (2) dataset has more than 50 samples; and (3) dataset provides certain prognostic or clinical outcome information. We use RPKM normalized expression value for RNA-Seq data and RMA normalized expression for microarray data. Detailed data information is provided in Table 1. In this study, DFS (disease free survival) refers to the duration between the primary treatment for cancer and the absence of any cancer-related symptoms, and OS (overall survival) represents the time elapsed from either the date of cancer diagnosis or the initiation of treatment until the patient's survival. Expression of each gene with multiple probes is assessed by expression of the probe with highest mean expression value in each data set. Genes of mean expressions at bottom 30% quantile in each microarray data set, and genes with 0 expression in more than 85% samples in the RNA-Seq data set are removed from the analysis, in order to control the noise of non- or lowly- expressed genes.

PCAM-step 1, Discretization: modeling the regulatory states of gene expressions via data discretization

To capture the regulatory states of a gene, we re-format the continuous expression data matrix into a larger binary matrix. Specifically, for a gene expression data $G_{\mathrm{m}\times\mathrm{n}}$ with m genes and n samples, we first find the $\mathrm{K}+1$ quantiles of each gene, and then generate a $\mathrm{K}\times\mathrm{n}$ binary matrix $\mathrm{D_g}$ for each gene g: $\mathrm{D_g}[i,j]=1$ if and only if expression of gene g in sample j is in the interval of $\left(\mathrm{Q}^{\mathrm{g}}_{\frac{\mathrm{i}-1}{\mathrm{K}}}, \mathrm{Q}^{\mathrm{g}}_{\frac{\mathrm{i}}{\mathrm{K}}}\right), i=1,\ldots,K$. Here $\mathrm{Q}^{\mathrm{g}}_{\alpha}$ represents the α quantile of the expression vector of gene g; and K is a hyper-parameter that controls the granularity of the discretization, with larger K capturing more potential transcriptional states of the gene. Obviously, each row of $\mathrm{D_g}$ indicates the samples with same expression patterns of g, and hence the same transcriptional regulatory states. Then we concatenate all the $\mathrm{D_g}$ by row to form a $\mathrm{Km}\times\mathrm{n}$ binary matrix $D_{\mathrm{Km}\times\mathrm{n}}$ and apply our in-house bi-clustering software QUBIC-R to identify the bi-clusters enriched by 1s in $D_{\mathrm{Km}\times\mathrm{n}}$. The rationality of this formulation is that each of the bi-cluster identified here corresponds to a group of genes, whose expression patterns are highly consistent over a subset of samples, hence representing a gene co-expression module specific to the subset of samples. It is worth noting that samples in one bi-cluster are highly likely to share similar transcriptional regulatory signals controlling the relevant genes. More discussion about the connection between bi-clusters and gene expression control are available in Supplementary Method.

PCAM-step 2, Bi-clustering: Bi-cluster identification in a binary matrix

PCAM uses our recently released bi-clustering R package – QUBIC-R to identify bi-clusters in discretized matrices, which was optimized based on the core algorithm of QUBIC for large-scale matrices (21,22). It is noteworthy that the number of rows ranges from 28,754 to 71,940 in this analysis. To the best of our knowledge, QUBIC is the most efficient bi-clustering method in the public domain that can handle input data of such large scale. The three parameters are set as follow: consistency level c=0.25, desired output number o=3000, and bicluster overlapping rate f is set at five different levels, 0.85, 0.875, 0.9, 0.95, and 1, depending on the input data size and number of 1s in each row. Detailed information for bi-clustering parameters determination and program running for each dataset are available in Supplementary Method.

By extending Xing Sun *et al.*'s work (27-29), we derived an analytical formula to evaluate the significance values for the BCs. For a random binary matrix M with m_0 rows and n_0 columns, the probability of being 1 for any element, namely, $p(M[i,j] = 1)$, is denoted as p_0. Then the upper bound of the probability that at least one submatrix M_1 exists in M could be assessed by the following formula, where M_1 has m_1 rows, n_1 columns, and z_0 total number of 0, and $n_1 \geq K$:

$$\mathrm{P}(\exists\, M_1\ with\ n_1 \geq \mathrm{K}) \leq \binom{\beta n_1^2}{z_0} {n_0}^{-(\beta+1)(K-s(n_1,n_0,\beta)} (\log_b n_0)^{\beta+1}, \text{when } n \to \infty,$$

where

$$\alpha = \frac{m_0}{n_0}, \beta = \frac{m_1}{n_1}, \mathrm{b} = \frac{1}{\mathrm{p}_0}$$

$$\mathrm{p}_0 = P(M[i,j] = 1) = 1 - P(M[i,j] = 0) \text{ for } \forall\, i,j$$

$$\mathrm{s}(\mathrm{n}_1, \mathrm{n}_0, \beta) = \frac{\beta+1}{\beta}\log_b \mathrm{n}_0 - \frac{\beta+1}{\beta}\log_b\left(\frac{\beta+1}{\beta}\log_b \mathrm{n}_0\right) + \log_b \alpha$$

$$+ \frac{(1+\beta)\log_b e - \beta\log_b \beta}{\beta}$$

More details of the derivation of this assessment formula is given in Supplementary Method. We have tested this significance assessment method on simulated data and compared its performance with the Chernoff's bound method (30), which is a popular measure for the effectiveness of biclustering methods. In detail, we conducted bi-clustering analysis on randomly generated gene expression matrices with same sizes. The analysis revealed that *p* values generated by our methods can more accurately recover the empirical *p* values comparing to the Chernoff's bound method. Particularly, our method offers a good control of false discover rate for the BCs that are highly enriched by 1s, hence it is more robust in picking out the significant ones from a large number of BCs identified in a large matrix. This is particularly key to large–scale matrix. Note that this significance test ensures that only BCs with sufficient width, height and number of 1's in it will be selected.

PCAM-step 3, Annotation: gene set enrichment and clinical association analysis

Enrichment analysis: Biological characteristics of each BC is assessed by whether genes in the BC significantly enrich a biology pathway or gene set. In total, 1,329 canonical gene sets including all pathways from KEGG, BIOCARTA, REACTOME databases and 1,472 GO (Gene Ontology) terms from MsigDB are used in the study (33). The enrichment analysis was computed by hypergeometric test, and for each BC in each dataset, genes in the BC were chosen as test set, while all genes in the dataset were chosen as the gene universe. Here p=0.005 is used as the cutoff for significance.

Single BC association analysis: Association analysis of each BC with clinical features was conducted using different tests based on the nature of the feature. For discrete clinical features including CMS classifications and pathological stages, we utilized Fisher's exact test; for continuous clinical features except for survival outcome, we compared the feature value for samples in and out of the BC by Mann Whitney test. p<0.005 was used as significance cutoff for all these tests. Notably, associations with CMS are conducted for only BCs containing more than five samples of the CMS class. For survival outcomes including DFS and OS, we compared the survival for samples in and out of the BC, using log-rank test with significance cutoff p<0.05.

Multiple BCs association analysis with prognosis: In order to identify the BCs that could best predict prognosis and drug resistance, we constructed multiple variable Cox-regression model between patients' survival and the BCs shown to be associated with survival with a variable selection procedure. Here, each BC is coded into one binary explanatory vector with 1's for samples in the BC and 0's for samples not in the BC. Specifically, we applied forward and backward stepwise variable selection approach to select the model with lowest AIC (Akaike information criterion) value by using SURVIVAL and MASS package R.

Multiple BCs associated with drug resistance: Among the BCs that are detected to show resistance to the chemo-drugs, we posit that each BC suggests one mechanism for the drug resistance. However, there may exist more than one BCs corresponding to the same mechanism. In order to identify the most unique set of resistance mechanisms, we use agglomerative clustering to cluster the BCs of similar resistance mechanisms into groups, and log-rank test is used to test each BC group with one drug resistance.

To do this, we first defined the distance between any two BCs as $\mathrm{D}\left(\mathrm{BC}_{\mathrm{I}}, \mathrm{BC}_{\mathrm{j}}\right) = 1 - \frac{|(Samples\ in\ BC_i)\cap(Samples\ in\ BC_j)|}{|(Samples\ in\ BC_i)\cup(Samples\ in\ BC_j)|}$, based on which an agglomerative clustering was performed. In each step of the clustering, two clusters X and Y are merged, if (1) samples in $X \cap Y$ is significantly associated with resistance to the drug, (2) neither samples in $X \backslash Y$ or $Y \backslash X$ is significantly associated with the drug resistance. A sample collection is defined as associated with resistance of a chemo-drug if the following two conditions are both met: (1) among drug treated samples, the overall survival of samples in the collection is significantly worse than those not in the collection (p<0.001); and (2) among samples in the collection, the overall survival of samples that are drug treated is significantly worse than those not treated (p<0.05). The agglomeration is stopped when no clusters could be merged.

Analysis of somatic mutations in TCGA data

TCGA COAD level 2 mutation profile of 429 samples predicted by *mutect* is retrieved from GDC database. A total of 932 genes with mutations in more than 5% (22/429) samples are selected. Considering high MSI (MicroSatellite Instability) causes the CRC genomes to be hyper-mutated, we exclude a majority of the 932 genes whose mutations are highly associated with MSI, and 73 gene mutations not associated with MSI are retained for further analysis. The association of a gene's mutation and MSI is calculated as the association between gene mutation and CMS class I—the class known to have high MSI, using Chi-square test ($p<0.1$).

Correction for multiple hypothesis testing

The p-value cutoff was set for significance test of identified BCs, pathway enrichment of each BC against 2,801 gene sets, and associations of BC with five types of phenotypic features. Among these, p-value was adjusted based on Benjamin and Hochberg method (31), when evaluating the significance of identified BCs, and the cutoff for the adjusted p-value is set at 0.05. However, we didn't apply the same criterion for the enrichment and association analysis. Rather, we set a fixed cut-off as 1e-6 for enrichment analysis, and 0.005 for associations analysis. The number of tests for enrichment and association analyzes are huge, which is the number of BCs multiple by the number of gene sets or phenotypes. Clearly, the current sample size is severely under powered, and we suspect a stringent Benjamin and Hochberg false discovery rate control would leave few tests to be significant. On the other hands, since these tests are highly dependent, while the level of dependency is impossible to track, we believe a lenient p-value cutoff could allow for more novel discoveries, that might be potentially interesting to experimentalists. Here, the more stringent p-value cutoff for enrichment analysis than association analysis is to control for higher false discovery rate due to the large number of gene sets analyzed.

Colon cancer consensus molecular subtype prediction

We applied the R package CMSclassifier to predict the CMS classification of each sample in the eight data sets (32), by which each sample will be predicted with four CMS scores representing its similarity to the four CMS classes. One sample is classified to one subtype if its CMS score of the subtype is larger than 0.5 and a sample is considered as with multiple-classification if both top two CMS scores are larger than 0.5 and the difference between the two scores is smaller than 0.1.

Results

We applied PCAM on eight colon cancer transcriptomics data sets with 1,440 samples. PCAM identified ~4,000 BCs on average in each data set (Table 2). We then evaluated each BC with its statistical significance, and annotated each BC by the pathways enriched by its genes, and clinical and prognostic outcomes associated with samples in the BC.

The overall pipeline of PCAM

Figure 1A shows a flowchart of PCAM, describing the analysis procedures we conducted on the selected datasets. Figure 1B details the bi-clustering analysis procedure. Each gene expression data set is was discretized such that the original $\mathrm{m} \times \mathrm{n}$ gene expression matrix with m genes and n samples is expanded to a $\mathrm{Km} \times \mathrm{n}$ binary matrix, as shown in Figure 1B and detailed in Methods section. Then, submatrices enriched by 1s in the discretized matrix are identified as BCs heuristically. Here, K is a hyperparameter that controls the granularity of the discretization. Clearly, the choice of K is very important: small K may blur the variability of gene expression across samples leading to insufficient capturing of the transcriptional regulatory states of the gene, and large K may severely undercut the power of bi-clustering and result in “narrow” bi-clusters that cover a very small percentage of samples. In all analyses, K=3 is selected because each gene could potentially be categorized into one of the three expression states: low/down-regulated, medium, and high/up-regulated. Each identified BC consists of a subset of samples and a group of genes, in which the genes are consistently expressed highly, moderately, or lowly by the subset of the samples.

The significant BCs will go through comprehensive annotation phases. PCAM examines whether genes in a BC enrich a certain pathway or gene set, and samples in a BC significantly over-represent a certain phenotype. Phenotypes of particular interests in this study include: 29 clinical features/outcomes in supplementary Table 1; 73 cancer-associated gene mutations (supplementary Table 1); and treatment responses to three chemo therapeutic drugs namely 5-Fluorouracil，Oxaliplatin, and the combination of 5-Fluorouracil, Oxaliplatin and Leucovorin. Functional annotation of the genes in each BC are conducted against 1,329 canonical pathways and 1,472 Gene Ontology sets in Msigdb (33).

PCAM was applied to transcriptomic data of 1,440 patient-derived CRC tissue samples including the TCGA COAD RNA-Seq data set, as well as seven microarray data sets (GSE14333, GSE17536, GSE29621, GSE33113, GSE37892, GSE383832 and GSE39582) measured by Affymetrix UA133 plus 2.0 array platform. (See detailed data information in Method). The computational pipeline of PCAM and key statistics for CRC are all provided in GitHub (https://github.com/changwn/BC-CRC). It is noteworthy that PCAM can be readily transplanted for similar analyzes in other disease scenarios. Below, we present the PCAM annotation results of the BCs identified in CRC datasets.

PCAM annotation of BCs with functional gene sets and phenotypic features

A total of 65,744 BCs were identified in the eight data sets. On average, ~4,000 BCs are found to be significant in each data set (Table 2) (adjusted $p<0.05$). Complete gene/sample information of all the significant BCs, are described in Supplementary Table 2. For each significant BC, we comprehensively investigated whether: (1) genes in the BC significantly enrich any of the 2,801 gene sets ($p<1e-6$), called PE BCs; (2) samples in the BC are significantly associated with any CMS class ($p<0.005$), called CMS I, II, III, IV and UC (unclassified) BCs; (3) samples in the BC are significantly associated with prognostic outcomes, namely patients’ overall and disease free survival ($p<0.005$), called Surv BCs or DFS BCs and OS BCs; (4) samples in the BC are significantly associated with clinical features such as age, gender, races and pathological stages ($p<0.005$), called Clin BCs; (5) samples in the BC are significantly associated with any of the 73 genomic mutation profiles ($p<0.005$), called Mut BCs; and (6) samples in the BC are significantly associated with the response to three selected chemo-drugs ($p<0.005$), called Drug BCs. The choice of p-value cutoffs is justified in Methods section. Figure 2A shows the proportion of BCs with significant findings in (1)-(4), in each of the eight data sets. On average, 71.79% (22,981/32,008) of the significant BCs can be significantly annotated by at least one of (1)-(4), with detailed numbers listed in Table 2. Note that (5) and (6) are specific to TCGA-COAD dataset, as mutation profiles and chemo-drug prognosis data are not available for the GEO datasets.

Figure 2B shows at different significance cutoff level (x-axis), the ratio (y-axis) of the BCs belonging to any one of the four kinds: PE BC, CMS BC, Surv BC, and Clin BC, among all significant BCs. The x-axis shows different significance levels of cutoff in ascending order, with leftmost the most stringent cutoff, and the y-axis shows the total number of annotatable BCs divided by the total number of significant BCs. It is obvious that not all significant BCs are annotatable, and interestingly, the most significant portion of the BCs are most likely to be annotatable, as indicated by the almost monotonically decreasing trend of all the eight curves. For example, we found if we only look at the top 20% of the significant BCs, then on average more than 80.7% of them are significantly annotatable; and the number drops to 66.4% if we look at all the significant BCs. This indicates BCs of higher significance tend to be more biologically/clinically relevant, demonstrating the rationality of our bi-clustering algorithm. Interestingly, by examining BCs of different significance levels, we found that the most significant BCs (p<1e-200) correspond to biological mechanisms that seem to be general to the whole population. Particularly, in these BCs, their genes tend to enrich pathways of low cell type specificity, including cell cycle, cell proliferation, cell death, biosynthesis and metabolism of nucleic acid, etc (Figure 3); and their samples don't seem to be associated with any phenotypic features. The biologically/clinically relevant BCs start to pop out in the next significance level (1e-200<p<1e-50). With higher sample specificity, these BCs have smaller sizes, and they tend to enrich pathways that are cell type specific, including immune response, extracellular matrix, O linked and N linked protein amino acid glycosylation, lipoprotein biosynthesis and lipid metabolism, etc. We have also seen that on average 44.7% of the DFS BCs and 33.9% of the OS BCs are also CMS BCs, particularly class I and IV, as shown in Figure 2C, and these BCs serve as possible CMS class specific prognosis markers. Other DFS BCs and OS BCs are found to be independent of the CMS class, suggesting the limitation of CMS in personalized prognosis prediction. In fact, the network complexity of the alternative pathways in cells and the uncertainty for cells to choose any of the alternative forms to maintain its viability in a perturbed microenvironment, has posed huge challenges for researchers to capture the heterogeneity of CRC withy any simple clinical stratifications. On the other hand, the large number of BCs presents us with comprehensive landscape of the alternative mechanisms, and potentially an increasing number of novel therapeutic targets.

The general trend of how BCs at different significance levels could be annotated by each category is shown in Figure 2D. Here, the ratio of PE BCs (left), CMS BCs (middle), and Surv BCs (right) among all significant BCs for all eight datasets, are shown as a function of the significance cutoff. While a stringent significance cutoff tend to produce BCs that significantly enrich biological pathways (PE BCs), this is not the case for CMS BCs or Surv BCs. Instead, a relatively lenient significance cutoff allows us to find more BCs associated with CMS and survival. Clearly, these novel patient subgroups contain far richer information than CMS. Below we will discuss in detail the BCs in relation to CMS. For all the eight data sets, on average 19.2% (12,641/65,744) of the BCs are CMS BCs. Among these, the proportion of BCs associated with each class is shown in Figure 2E. On average, the CMS BCs only cover 23.6%, 15.6%, 30.1% and 24.1% of the CMS I-IV samples, respectively (shown in Supplementary Figure 2). This suggests that there exists a large number of sample subgroups, that may not be aligned with CMS. The proportion of samples in the BCs that belong to different CMS class is shown in Figure 2F. There seems to be relatively more BCs aligning with CMS class I and IV, and unclassified, suggesting higher variations in patients of these classes. Of note, BCs associated with the four CMS classes, especially class III and IV, contain genes that highly overlap with the putative CMS marker genes; while the CMS marker genes rarely show up in BCs associated with the unclassified samples, as shown in Figure 2G. This indicates that the genes we identified in the BCs are indeed coherent with the marker genes of CMS class. Very few BCs are observed to have associations with the samples of multiple CMS classes, suggesting the exclusiveness of the CMS classes.

Among all the DFS BCs, 42.9% of them are also over-represented in certain CMS classes, while this rate is 49.5% for OS (See Figure 2H), on average. Particularly, 53.1% and 40.4% of these CMS-associated BCs belong to CMS IV class for DFS and OS respectively, on average. For DFS, the CMS IV associated BCs enrich the following pathways: glycosaminoglycan biosynthesis and metabolism, UDP glycosyltransferase, lipid, phospholipid and glycosphingolipid metabolism, mRNA splicing, and steroid hormone metabolism; while for OS, the pathways are: immune signaling, WNT and MYC signaling, VEGF signaling, tumor necrosis, notch signaling, cell proliferation and integrin pathways. This observation suggests that the extracellular matrix, glycosaminoglycan metabolism, lipid metabolism are prognostic markers for DFS if the patients are diagnosed with CMS class IV, while for OS, the markers are related to stromal infiltration. Similarly, we also observed a large proportion of CMS class I (19.1%) and CMS II associated (17.7%) BCs for DFS BCs, and CMS associated (25.1%) BCs for OS BCs. The CMS I associated DFS BCs enrich chemokine signaling, integrin signaling, chondroitin sulfate and sulfur metabolism, O linked glycosylation, and other immune and inflammation related pathways; CMS II associated DFS BCs enrich hypoxia response, O linked glycosylation, PI3K signaling, apoptosis, and immune response pathways; and CMS II specific OS associated BCs enrich cell cycle, nucleotide excision repair, and MYC signaling pathways.

We have also tested the association between BCs and 117 highly frequently mutated and non-MSI-associated genes in TCGA COAD data. Our analysis identified that 29.1% (550/1886) of the annotatable BCs and 22.5% (168/746) of the unannotated BCs are associated with at least one of the gene mutations. Interestingly, by looking at the mutation profiles of samples in the Mut BCs, a large proportion happen in genes including TMEM132D, BCL9L, NF-1, SCN10A, PCDHA10, DIP2C, GLI3, TET2, and ARFGEF2, while only a small number fall into key CRC associated gene including APC, TP53, KRAS, CTNNB1, and PIK3CA. The Mut BCs majorly enrich pathways of nucleotide and glucose metabolism and immune responses. Detailed pathway enrichment of the mutation BCs is provided through GitHub and described in Supplementary Table 2.

A consensus functional annotation of the BC landscape

The cellular system is sufficiently complex and robust that cells are able to deploy a variety of pathways to respond to perturbations in the microenvironment. Our analysis has revealed that BCs associated with different phenotypic features exhibit enrichment to distinct sets of pathways, as reflected by a consensus map that illustrates how different pathways are "favored" by the cellular systems under different phenotypic states in Figure 3. We call this a consensus functional annotation of the BC landscape in CRC. The BCs are examined with respect to biological pathway enrichment called the PE BCs, and 17 clinical phenotypes, including five CMS BCs, DFS BCs, DFS BCs that over-represent five CMS classes, OS BCs, and OS BCs that over-represent five CMS classes. In each setting, genes in the BCs are used for pathway enrichment, and in total, 43 most significant pathways consistent to all eight datasets are selected, shown as the left row-wise names of the consensus map in Figure 3. The right block row-wise names indicate one of the 18 categories the BCs are annotated. The 43 pathways are believed to represent the specific functions associated with the biological/phenotypic state. For each of the 43 pathways, its average activation level with regards to the 18 settings, shown as top column-wise names of the map, are calculated over all datasets. Clearly, the activation score matrix reflects the degree of similarity or dissimilarity among the 18 settings in relation to the 43 pathways.

This consensus map greatly helps us visualize the distinctions and similarities regarding different clinical phenotypes, using functional pathways derived from BCs. As shown in Figure 3, different CMS classes are characterized by different pathways/gene sets, but they also show certain continuity. CMS I BCs are also enriched by immune signaling pathways including IL-3, -5, -6, -12, -27, STAT, and interferon gamma signaling pathways, as well as nucleotide biosynthesis, WNT signaling, lipid metabolism, and glycolysis pathways, which are markers of CMS II and III (8). Considering that CMS I is a subtype with high MSI and strong immune cell activation (8), our observation clearly suggests that there are distinct subgroups inside CMS I with different immune activation status that display CMS II-like characteristics with high expression of epithelial and WNT signaling markers and CMS III-like characteristics of metabolism dysregulations. More intriguingly, the CMS IV BCs seem to fall into two categories: one enriched by integrin binding, epithelial cell cycle, cell death, cell-cell and cell-matrix adhesions pathways, while the other enriched by immune response, MYC and WNT signaling, and metabolism pathways. The first category show expression of cancer and stromal cell marker genes, suggesting different levels of stromal cell infiltration in CMS IV class. In contrast, the second category enriches marker genes of CMS class I-III, suggesting that there are subgroups within CMS IV class that resemble CMS I, II or III. CMS IV is a subtype with high stromal infiltration and angiogenesis (8). Our previous study has identified a dynamic population of mesenchymal-like cells with similar markers as CMS IV (34). With these observations, we suspect that CMS IV is a combination of CMS I-III but with higher proportion of stromal cells, hence higher expression of mesenchymal cell markers and lower rate of somatic mutations. However, it is noteworthy that the CMS IV cancers have generally poorer prognosis comparing to CMS I-III, indicating the level of stromal infiltration may serve as an important prognosis marker for all the CMS classes. We have also seen that a number of CMS II and III BCs show marker genes of other CMS classes. The CMS UC BCs enrich signaling pathways of MAPK, P38, GPCR, NOTCH, TGF-beta, ARF6 and other kinase receptors and pathways responsive to micro-environment stresses including ER stress, oxidative stress, dysregulated immune activation and extracellular matrix malfunction. We suspect that in response to the activation of specific signaling pathways and distinct micro-environment stresses, gene expression in these samples are highly volatile, and hence cannot be classified by CMS. Functional annotation of the genes in the CMS BCs are given in Supplementary Table 3.

Lastly, we employed a Cox regression model with variable selection using BCs to explain patients' prognosis (see Methods). Our analysis suggested that the DFS predictive BCs contain genes that enrich pathways including chemokine receptor, O-linked glycan biosynthesis, apoptosis, mitochondria, cell membrane, MAPK activity, tissue morphogenesis, VEGFR pathway, lipid homeostasis and cell surface receptor activity; while for OS, the BCs enrich cell death, cell proliferation, mitosis, glycosaminoglycan synthesis, integrin (possibly suggests stromal infiltration level), T cell activation, WNT beta-catenin signaling, leukocyte activation, extracellular region and glucose transport and VEGFR pathway.

PCAM annotations of BCs by alternative drug resistance mechanisms

Chemo-therapy is one of the standard cancer treatment methods that induces cell death of fast proliferating cancer cells (35). Usually, the administration of cytotoxic drugs may initially result in tumor shrinkage by destruction of non-resistant subclonal populations within a heterogeneic tumour, while leaving the resistant clones. With a selective advantage, these resistant clones can replicate to repopulate the tumour, and the repopulated tumor appears to be far more aggressive, called acquired drug resistance. The clinical information in TCGA provides patients' treatment response to three most prevalent CRC chemo-therapy plans, including 5-Fluorouracil (5-FU), Oxaliplatin (OXA), and the combination of OXA, 5-FU and Leucovorin (FOLFOX). In order to delineate the alternative drug resistance mechanism in CRC, we selected the drug associated BCs, called Drug BCs. A drug BC is defined if the following two conditions are both met: (1) among drug treated samples, the overall survival of samples in the BC is significantly worse than those not in the BC ($p<0.001$); and (2) among samples in the BC, the overall survival of samples that are drug treated is significantly worse than those not treated ($p<0.05$). Certainly, multiple drug BCs may correspond to the same resistance mechanism. We conducted a log-rank test coupled with agglomerative clustering to cluster the BCs into groups, each of which may be linked to one drug resistance mechanism (see details in Methods section). Complete information of Drug BC clusters are given in Supplementary Table 4.

5-FU is one of the most commonly used chemo-drugs in treating CRC (36). We identified 11 5FU BCs, and found that the 11 BCs form four groups, where each group consists of a number of genes tightly co-expressed, and a number of samples with 5FU resistance, as shown in Figure 4A. The first BC group contains genes enriching known chemo-resistance related mechanisms, including over expression of CFLAR involved in apoptosis and FAS signaling; CAPRIN2 related to cell proliferation and cancer multi-drug resistance; DNA excision repair gene XPA; cell cycle regulating proteins DMTF1 and SYCE2; killer cell activating receptor associated protein TYROBP; taurine metabolism gene CSAD; RNA processing proteins RBM6 and CLK1; DNA binding and transcriptional regulatory genes ZNF638, ZNF169, ZNF26, ZNF333, ZNF493, ZNF234 and ZNF33A; OGT, TAS2R5, LTB4R2 related to cellular response to chemical stimuli. It is noteworthy that a number of genes in this panel including CFLAR, CAPRIN2, XPA, TYROBP, CLK1, OGT, and LTB4R2 have been previously identified to be relate to chemo-resistance in other cancer types (37-42). The second group contains genes including SMAD2, SMAD4, TCF12, ELP2, ATG2B, PIGN, MBP, NCBP3 and PIK3C3, which enrich pathways of cell cycle, cell metabolism regulation, TGF-beta signaling, PI3K cascade, autophagy, immune responses and mRNA production regulation. The third BC group contains a large number of pseudo genes and also genes that enrich the translation regulation and viral infection pathways, among which genes TMA7, DEXI and EIF3CL have been previously reported as related to cisplatin and fluorouracil resistance in bladder and gastric cancer (43,44). Genes in the fourth group enrich two different groups of ribosome proteins, which are related to translational control and elongation of peptides.

OXA is a platinum-based antineoplastic chemo-drug used to treat colorectal cancer (36). We have identified 10 OXA BCs, which were further clustered into three groups as shown in Figure 4B. The first BC group shows an overlap with the first group in 5FU resistance, in that the genes are also involved in known chemo-resistance related mechanisms including CFLAR, CAPRIN2, TYROBP, CLK1, OGT and LTB4R2 as well as SYCE2, RBM6, ZNF638, ZNF169, ZNF26, ZNF333, ZNF493, ZNF234 and ZNF33A, related to cell cycle, mRNA processing and DNA binding. Meanwhile, this group also contains overly expressed DNA synthesis and cell cycle genes POLA1, CHFR, and TAF1; mRNA processing gene PCF11; EPHA7 and COL4A3 related to tissue development; and ITPR2 related to calcium dependent signaling transduction. The second group also contains CFLAR, CAPRIN2, SYCE2, and LTB4R2 identified in the first group. In addition, this group also contains cyclin-D binding transcription factor DMTF1; transcriptional regulation co-factor EP300; GTF2H4 related to RNA polymerase II transcription initiation; mRNA splicing gene DDX39B; and cell surface channel, transporter or exchanger genes PKD2, TRAPPC10, SMG1, and TRIO. The third group contains a number of nuclear ribonucleoproteins and HSPA5, where the latter has been previously identified as a chemo-resistance biomarker and molecular target in B-lineage acute lymphoblastic leukemia (45).

FOLFOX is a combinatorial therapy of 5Fu, OXA with Leu--a reduced folic acid based drug that is used in combination with other chemotherapies to enhance effectiveness or prevent side effects of the chemo-drugs (36,46). We have identified eight FOLFOX BCs forming four BC groups (Figure 4C). The first BC group shows strong overlaps with the first group of 5FU chemo-resistance, and the first and second group of OXA chemo-resistance, which includes CFLAR, CAPRIN2, SYCE2, CSAD, MSH5, XPA, OGT, LTB4R2, ZNF234, ZNF169, ZNF493, ZNF26, and ZNF333. The second group contains JAK2, which is involved in multiple cytokine receptor signaling pathways related to immune response; Rho GTPase Activating Protein DLC1 (tumor suppressor); cell death related genes NME1, BCL2L15 and RPSS3A; tissue development regulating gene FOXA2; TCA cycle and respiration electron transport genes ATP5C1 and COX7A2L; and mitochondrial inner membrane translocase TIMM23. In addition, this group also overly express ribosome proteins. The third group contains highly expressed CAPRIN2, cell proliferation regulating gene DMTF1 and mRNA processing proteins DDX39B and GTF2H4. The fourth group is composed of under expressed microRNA MIR3911 and antisense mRNA EIF1AX-AS1.

We collected drug screening data on colon cancer cell line to validate our identified possible resistance mechanism (see methods). To the best of our knowledge, 5-FU is the only one drug with a wide spectrum of sensitivity measure on cell lines among the three. 5-FU treatment was performed on 29 and 19 colon cancer cell lines for two independent datasets (47,48). In each dataset, we computed the correlations between the basal level expressions of all the genes and cell's response to 5-FU, measured by IC50 and GI50 (see Supplementary table 5). IC50 and GI50 are two metrics to evaluate drug treatment efficacy. Distribution of the correlations for genes in each BC group was compared with the distribution of the correlation for all genes, which serves as a random background. Density curves of the correlations of each BC group and the background are shown in Figure 4D and 4E. We have seen that, comparing to the background, genes in BC group 4 show much higher correlations to cells' resistance to 5-Fu, and BC groups 1-3 also contain a marked portion of genes that are more correlated with 5-Fu resistance than background. This serves as further validation of our observations of alternative drug resistance mechanisms. Detailed lists of the validation data are provided in Supplementary Table 5.

In summary, for each chemo-drug, we have identified a few potential drug alternative resistance mechanisms, presented in the form of BC groups, and some of which are novel to CRC. Further experimental validations are needed to confirm these findings. It is noteworthy that the genes CFLAR, CAPRIN2, SYCE2, OGT, and LTB4R2 are consistently observed as resistance associated for all the three drugs. Further investigation of the sample composition of the BC groups suggests that the first BC group of 5-Fu, OXA and the second BC group of FOLFOX highly overlap, which correspond to poor response of 5-Fu and OXA in CMS1 samples and FOLFOX in CMS2 samples (Figure 4F). The second BC cluster of OXA and the third BC cluster of FOLFOX overlap, which corresponds to poor response in CMS1 samples. In addition, the 5-Fu BC groups 2, 3 and 4 show that patients of CMS III, CMS III/IV and CMS II/III are particularly resistant to 5-Fu; OXA BC groups 2 and 3 show that OXA resistance is high in CMS II/III and CMS I/II/III; FOLFOX BC groups 1, 3, and 4 show that resistance of the drug prevalently happen to patients of CMS II/IV, CMS II and CMS IV. Interestingly, 5-Fu BC group 1 and FOLFOX BC groups 1 and 4 do not seem to show chemo-resistance mechanisms specific to any CMS classes. Among the identified BC groups, some of them point to known chemo-resistance mechanisms. Meanwhile, we have seen in 1-2 BC groups for each drug type there exists novel biomarkers, including overly expressed ribosome genes and under expressed ncRNAs. Further experimental validations are warranted.

Discussion and Conclusion

It has been widely recognized that cells have multiple alternative pathways to cope with microenvironmental perturbations, and the uncertainty surrounding the choice of a pathway under different circumstances contributes to cancer heterogeneity. In the case of drug resistance, multiple pathways are often altered to create a single off-target resistance mechanism (49-51). Molecular subtyping methods for CRC, such as CMS, have provided valuable information in understanding heterogeneity. However, due to the dynamic nature of the cancer microenvironment, novel alternative pathways can emerge under selective pressure, that may not have been captured by any disease stratifications. Limiting our computational analysis to a pre-defined molecular subtyping such as CMS would fail to capture a large number of alternative mechanisms (and their combinations) which are employed under different circumstances. Our bi-clustering based PCAM method is powerful in delineating a comprehensive collection of alternative mechanisms caused by the intrinsic heterogeneity within patients, and their associations with known phenotypic features. Each BC potentially contains a coherent gene module present in a subgroup of patients, and the gene subsets may enrich certain biological pathways that could lead to substantially deeper biological understanding for molecular stratification of cancer. More importantly, any existing sub-grouping methods, such as CMS, could be studied and integrated with the produced BCs.

We developed PCAM as an unsupervised exploratory approach with several advantages in identifying gene markers for certain phenotypes: (1) efficiently control false discoveries; (2) readily detect informative co-expressed prognostic markers; (3) conveniently handle the intricate relationships among different subtypes, and their interactions with various clinical outcomes. Of note, deriving prognostic or predictive markers from only BCs with high statistical significance could decrease the number of independent tests, and the resulted co-expressed gene modules are more relevant in the disease context. The sample compositions in each BC provides an easily comprehensible way to understand the underlying subtypes. Our analysis has clearly demonstrated that PCAM can effectively identify biomarkers for alternative prognosis related or drug resistance mechanisms from large scale transcriptomics data. We posit that bi-clustering is more sensitive to locate the biomarkers specific to small subset of samples and the inference on the multiple genes in the BC can provide more biologically coherent interpretations.

Nonetheless, we have seen a few more challenges that is beyond this study. Firstly, when several BCs are highly overlapping, only one will be retained, which may be problematic when consistency of BCs across different datasets are to be performed. This raises a demand for effective multi-tasking strategy to find bi-clusters with high consistency through multiple data sets. Secondly, currently PCAM lacks a predicative model using BCs, which largely limits its potential in practice of personalized treatment. Thirdly, the BC's statistical significance is estimated by an upper bound of *p* value, which works well for the BCs with small number of 0s in it, but not for BCs with large number of 0s. We fully anticipate future studies will address these challenges, and increase the feasibility of PCAM in characterizing the complexity of CRC heterogeneity, and aiding biomarker detection and personalized medicine.

STATEMENT OF INTERESTS

Statement of interests

We declare none of the authors have any competing interests.

Funding

This work was supported by the National Science Foundation (IIS-1850360, IIS-2145314), National Cancer Institute (5P30CA082709), and Indiana Clinical and Translational Sciences Institute Showalter Trust Young Investigator Award.

REFERENCES

1. Siegel, R.L., Miller, K.D., Wagle, N.S. and Jemal, A. (2023) Cancer statistics, 2023. *CA: A Cancer Journal for Clinicians*, **73**, 17-48.
2. Wolf, A.M.D., Fontham, E.T.H., Church, T.R., Flowers, C.R., Guerra, C.E., LaMonte, S.J., Etzioni, R., McKenna, M.T., Oeffinger, K.C., Shih, Y.T. *et al.* (2018) Colorectal cancer screening for

average-risk adults: 2018 guideline update from the American Cancer Society. *CA Cancer J Clin*, **68**, 250-281.

3. Inamura, K. (2018) Colorectal Cancers: An Update on Their Molecular Pathology. *Cancers (Basel)*, **10**.

4. Gonzalez-Garcia, I., Sole, R.V. and Costa, J. (2002) Metapopulation dynamics and spatial heterogeneity in cancer. *Proc Natl Acad Sci U S A*, **99**, 13085-13089.

5. Samowitz, W.S. and Slattery, M.L. (1999) Regional reproducibility of microsatellite instability in sporadic colorectal cancer. *Genes Chromosomes Cancer*, **26**, 106-114.

6. Giaretti, W., Monaco, R., Pujic, N., Rapallo, A., Nigro, S. and Geido, E. (1996) Intratumor heterogeneity of K-ras2 mutations in colorectal adenocarcinomas: association with degree of DNA aneuploidy. *Am J Pathol*, **149**, 237-245.

7. Marusyk, A. and Polyak, K. (2010) Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta*, **1805**, 105-117.

8. Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P. *et al.* (2015) The consensus molecular subtypes of colorectal cancer. *Nat Med*, **21**, 1350-1356.

9. Cancer Genome Atlas, N. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330-337.

10. Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M.H., Chresta, C.M., Rosenberg, R., Nitsche, U. *et al.* (2014) Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer*, **134**, 552-562.

11. Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K.O., Di Narzo, A.F., Yan, P., Hodgson, J.G., Weinrich, S. *et al.* (2013) Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol*, **231**, 63-76.

12. Schlicker, A., Beran, G., Chresta, C.M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Castro, D.A. *et al.* (2012) Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC Med Genomics*, **5**, 66.

13. Sadanandam, A., Lyssiotis, C.A., Homicsko, K., Collisson, E.A., Gibb, W.J., Wullschleger, S., Ostos, L.C., Lannon, W.A., Grotzinger, C., Del Rio, M. *et al.* (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med*, **19**, 619-625.

14. De Sousa, E.M.F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L.P., de Jong, J.H., de Boer, O.J., van Leersum, R., Bijlsma, M.F. *et al.* (2013) Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nat Med*, **19**, 614-618.

15. Marisa, L., de Reynies, A., Duval, A., Selves, J., Gaub, M.P., Vescovo, L., Etienne-Grimaldi, M.C., Schiappa, R., Guenot, D., Ayadi, M. *et al.* (2013) Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*, **10**, e1001453.

16. Perez-Villamil, B., Romera-Lopez, A., Hernandez-Prieto, S., Lopez-Campos, G., Calles, A., Lopez-Asenjo, J.A., Sanz-Ortega, J., Fernandez-Perez, C., Sastre, J., Alfonso, R. *et al.* (2012)

Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer*, **12**, 260.

17. Gong, Y. and Zhang, Z.J.F.l. (2005) Alternative signaling pathways: When, where and why? , **579**, 5265-5274.

18. Stelling, J., Sauer, U., Szallasi, Z., Doyle III, F.J. and Doyle, J.J.C. (2004) Robustness of cellular functions. **118**, 675-685.

19. Pontes, B., Giraldez, R. and Aguilar-Ruiz, J.S. (2015) Biclustering on expression data: A review. *J Biomed Inform*, **57**, 163-180.

20. Eren, K., Deveci, M., Kucuktunc, O. and Catalyurek, U.V. (2013) A comparative analysis of biclustering algorithms for gene expression data. *Brief Bioinform*, **14**, 279-292.

21. Zhang, Y., Xie, J., Yang, J., Fennell, A., Zhang, C. and Ma, Q. (2017) QUBIC: a bioconductor package for qualitative biclustering analysis of gene co-expression data. *Bioinformatics*, **33**, 450-452.

22. Li, G., Ma, Q., Tang, H., Paterson, A.H. and Xu, Y. (2009) QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Res*, **37**, e101.

23. Xie, J., Ma, A., Fennell, A., Ma, Q. and Zhao, J. (2018) It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Brief Bioinform*.

24. Wang, Y.K., Print, C.G. and Crampin, E.J. (2013) Biclustering reveals breast cancer tumour subgroups with common clinical features and improves prediction of disease recurrence. *BMC Genomics*, **14**, 102.

25. Fiannaca, A., La Rosa, M., La Paglia, L., Rizzo, R. and Urso, A. (2015) Analysis of miRNA expression profiles in breast cancer using biclustering. *BMC Bioinformatics*, **16 Suppl 4**, S7.

26. Liu, Y., Gu, Q., Hou, J.P., Han, J. and Ma, J. (2014) A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, **15**, 37.

27. Sun, X. (2007) *Significance and recovery of blocks structures in binary and real-valued matrices with noise*. The University of North Carolina at Chapel Hill.

28. Sun, X. and Nobel, A. (2006), *International Conference on Computational Learning Theory*. Springer, pp. 109-122.

29. Sun, X. and Nobel, A.B. (2008) On the size and recovery of submatrices of ones in a random binary matrix. *Journal of Machine Learning Research*, **9**, 2431-2453.

30. Hoeffding, W. (1963) Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association*, **58**, 13-30.

31. Hochberg, Y. and Benjamini, Y.J.S.i.m. (1990) More powerful procedures for multiple significance testing. **9**, 811-818.

32. Eide, P.W., Bruun, J., Lothe, R.A. and Sveen, A. (2017) CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Sci Rep*, **7**, 16618.

33. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545-15550.

34. Zhang, C., Cao, S. and Xu, Y. (2014) Population dynamics inside cancer biomass driven by repeated hypoxia-reoxygenation cycles. *Quantitative Biology*, **2**, 85-99.

35. DeVita, V.T., Jr. and Chu, E. (2008) A history of cancer chemotherapy. *Cancer Res*, **68**, 8643-8653.

36. Gustavsson, B., Carlsson, G., Machover, D., Petrelli, N., Roth, A., Schmoll, H.J., Tveit, K.M. and Gibson, F. (2015) A review of the evolution of systemic chemotherapy in the management of colorectal cancer. *Clin Colorectal Cancer*, **14**, 1-10.

37. Fraser, M., Leung, B., Jahani-Asl, A., Yan, X., Thompson, W.E. and Tsang, B.K. (2003) Chemoresistance in human ovarian cancer: the role of apoptotic regulators. *Reprod Biol Endocrinol*, **1**, 66.

38. Weaver, D.A., Crawford, E.L., Warner, K.A., Elkhairi, F., Khuder, S.A. and Willey, J.C. (2005) ABCC5, ERCC2, XPA and XRCC1 transcript abundance levels correlate with cisplatin chemoresistance in non-small cell lung cancer cell lines. *Mol Cancer*, **4**, 18.

39. Mochmann, L.H., Neumann, M., von der Heide, E.K., Nowak, V., Kuhl, A.A., Ortiz-Tanchez, J., Bock, J., Hofmann, W.K. and Baldus, C.D. (2014) ERG induces a mesenchymal-like state associated with chemoresistance in leukemia cells. *Oncotarget*, **5**, 351-362.

40. Zhang, L., Yang, H., Zhang, W., Liang, Z., Huang, Q., Xu, G., Zhen, X. and Zheng, L.T. (2017) Clk1-regulated aerobic glycolysis is involved in glioma chemoresistance. *J Neurochem*, **142**, 574-588.

41. Cheng, S., Mao, Q., Dong, Y., Ren, J., Su, L., Liu, J., Liu, Q., Zhou, J., Ye, X., Zheng, S. *et al.* (2017) GNB2L1 and its O-GlcNAcylation regulates metastasis via modulating epithelial-mesenchymal transition in the chemoresistance of gastric cancer. *PLoS One*, **12**, e0182696.

42. Park, J., Park, S.Y. and Kim, J.H. (2016) Leukotriene B4 receptor-2 contributes to chemoresistance of SK-OV-3 ovarian cancer cells through activation of signal transducer and activator of transcription-3-linked cascade. *Biochim Biophys Acta*, **1863**, 236-243.

43. Tanaka, N., Katayama, S., Reddy, A., Nishimura, K., Niwa, N., Hongo, H., Ogihara, K., Kosaka, T., Mizuno, R., Kikuchi, E. *et al.* (2018) Single-cell RNA-seq analysis reveals the platinum resistance gene COX7B and the surrogate marker CD63. *Cancer Med*, **7**, 6193-6204.

44. Kim, M., Jung, J.Y., Choi, S., Lee, H., Morales, L.D., Koh, J.T., Kim, S.H., Choi, Y.D., Choi, C., Slaga, T.J. *et al.* (2017) GFRA1 promotes cisplatin-induced chemoresistance in osteosarcoma by inducing autophagy. *Autophagy*, **13**, 149-168.

45. Uckun, F.M., Qazi, S., Ozer, Z., Garner, A.L., Pitt, J., Ma, H. and Janda, K.D. (2011) Inducing apoptosis in chemotherapy-resistant B-lineage acute lymphoblastic leukaemia cells by targeting HSPA5, a master regulator of the anti-apoptotic unfolded protein response signalling network. *Br J Haematol*, **153**, 741-752.

46. Tsai, Y.J., Lin, J.K., Chen, W.S., Jiang, J.K., Teng, H.W., Yen, C.C., Lin, T.C. and Yang, S.H. (2016) Adjuvant FOLFOX treatment for stage III colon cancer: how many cycles are enough? *Springerplus*, **5**, 1318.

47. Rees, M.G., Seashore-Ludlow, B., Cheah, J.H., Adams, D.J., Price, E.V., Gill, S., Javaid, S., Coletti, M.E., Jones, V.L., Bodycombe, N.E. *et al.* (2016) Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol*, **12**, 109-116.

48. Bracht, K., Nicholls, A.M., Liu, Y. and Bodmer, W.F. (2010) 5-Fluorouracil response in a large panel of colorectal cancer cell lines is associated with mismatch repair deficiency. *Br J Cancer*, **103**, 340-346.

49. Chang, R.L., Xie, L., Xie, L., Bourne, P.E. and Palsson, B.O. (2010) Drug off-target effects predicted using structural analysis in the context of a metabolic network model. *PLoS Comput Biol*, **6**, e1000938.

50. Schenone, M., Dancik, V., Wagner, B.K. and Clemons, P.A. (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat Chem Biol*, **9**, 232-240.

51. Mansoori, B., Mohammadi, A., Davudian, S., Shirjang, S. and Baradaran, B. (2017) The Different Mechanisms of Cancer Drug Resistance: A Brief Review. *Adv Pharm Bull*, **7**, 339-348.

Tables:

Table 1. Information of the eight CRC datasets.

Data ID	Sample#	Drug response	Follow-up	Platform	Normalization
GSE14333	290	No	DFS	Affy U133 Plus 2.0	RMA
GSE17536	177	No	OS/DFS	Affy U133 Plus 2.0	RMA
GSE29621	65	No	OS/DFS	Affy U133 Plus 2.0	RMA
GSE33113	90	No	DFS	Affy U133 Plus 2.0	RMA
GSE37892	130	No	DFS	Affy U133 Plus 2.0	RMA
GSE38832	122	No	OS/DFS	Affy U133 Plus 2.0	RMA
GSE39582	566	No	OS/DFS	Affy U133 Plus 2.0	RMA
TCGA-COAD	385	Yes	OS	RNA-Seq	RPKM

Table 2. PCAM identified Bi-clustering of the eight CRC data sets

Data ID	#BCs	#Sig BCs	#PE BCs	#CMS BCs	#Surv BCs	#Clin BCs
GSE14333	9631	6547	2597(39.7%)	2512(38.4%)	448(6.8%)	452(6.9%)
GSE17536	11255	4806	2187(45.5%)	1425(29.7%)	284(5.9%)	63(1.3%)
GSE29621	8167	1758	582(33.1%)	289(16.4%)	73(4.2%)	56(3.2%)
GSE33113	9238	2836	795(28%)	958(33.8%)	136(4.8%)	3(0.1%)
GSE37892	10644	4452	1600(35.9%)	1202(27%)	130(2.9%)	101(2.3%)
GSE38832	5845	4319	2603(60.3%)	1705(39.5%)	335(7.8%)	0(0%)
GSE39582	8267	4658	1200(25.8%)	2894(62.1%)	1068(22.9%)	1847(39.7%)
TCGA_COAD	2697	2632	1077(40.9%)	743(28.2%)	183(7%)	954(36.2%)

Figure legends:

Figure 1. (A) General analysis pipeline. The analysis was conducted on one TCGA RNA-seq and seven microarray datasets. BC identification from each high-dimensional data sets starts with a discretization step followed by a bi-cluster identification step (see details in B). The identified BCs are further annotated by their associations with biological pathways, CMS class, and patients clinical and prognostic features. Consensus analysis of the BCs throughout multiple data sets was further conducted. BCs were further associated with response to different chemo-drugs for identification of alternative chemo-resistance mechanisms. **(B) Data discretization and bi-clustering procedures.** The histogram on the left illustrates the distribution of a gene's expression. The gene expression is discretized into three levels, represented as three 0-1 vectors (D_high, D_moderate and D_low), corresponding to samples with top (blue), medium (green) and bottom (red) 1/3 expression level of the gene, respectively. The discretized data are then concatenated that expand an original $m \times n$ gene expression matrix to a $3m \times n$ binary matrix, as shown in the right panel. In the expanded matrix, rows represent different states of the gene, and columns represent cancer patient samples. BCs enriched by 1s are further identified by QUBIC-R.

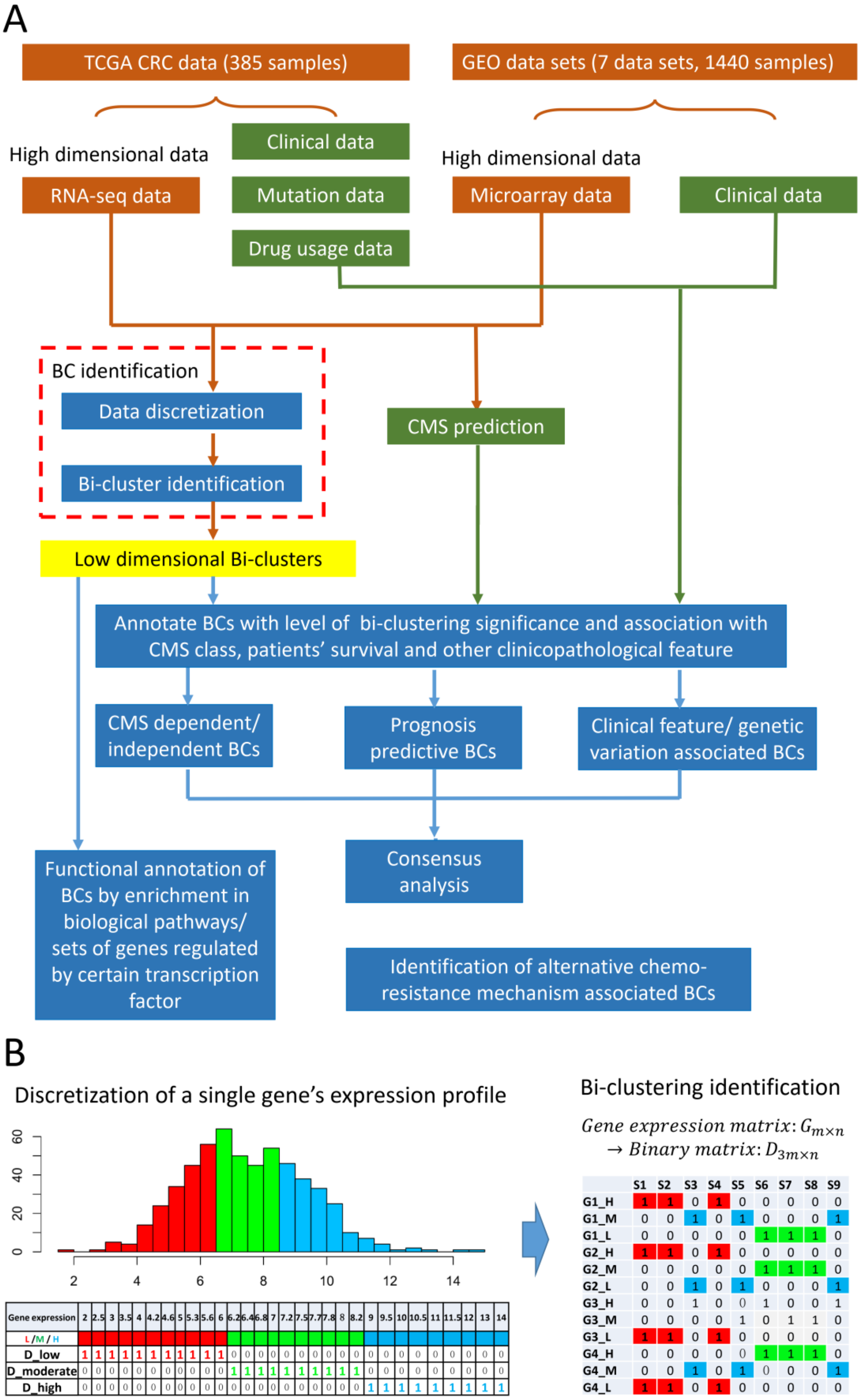
Figure 2. Statistics of the BC landscape in the eight data sets. (A) Proportions (y-axis) of PE BCs, CMS BCs, Surv BCs, Clin BCs, and their combinations (Multi) amongst all identified BCs in each data set (x-axis). (B) Rates of annotatable BCs (y-axis) as a function of significance cutoff of BCs at different levels (x-axis), most stringent on the left. (C) Among the DFS (left) and OS (right) BCs, the proportions (y-axis) of different CMS class BCs, in each dataset (x-axis). (D) Proportions (y-axis) of PE BCs, CMS BCs, Surv BCs amongst all significant BCs as a function of BC significance cutoff at different quantiles (x-axis), most stringent on the left. Here, a “0.2” quantile means the top 20% significant BCs. (E) Proportions of the BCs (y-axis) with significant associations to different CMS classes in each data set (x-axis). (F) Among the Surv BCs, the proportions of the BCs (y-axis) associated with CMS types in each data set (x-axis). (G) For BCs associated with different CMS class, the average overlapping rates (y-axis) between the genes in the BC and CMS marker genes in each dataset (x-axis). (H) Among all the DFS/OS BCs, the proportion of the BCs (y-axis) that significantly over-represent a (sub)sample class in each dataset (x-axis). In (C), (E) and (F): None: CMS unclassified samples; Multi-CMS: a class of samples falling into more than one CMS classes; Multi-class: a class of BCs significantly associated with more than one CMS classes. In (H): None: CMS unclassified samples; overall: the BCs associated with survival throughout all patients, but not with a particular CMS class; Multiple: the BCs associated with patients’ survival specific to the patients of more than CMS classes.

Figure 3. Functional annotation and concensus map of selected CMS classes and prognosis associated BCs. 43 pathways, shown on the left, that are most significantly and consistently over all datasets enriched by genes from PE BCs (or called Top BCs), CMS I, II, III, IV, UC BCs, DFS BCs, and OS BCs, shown on the right. The relative level of enrichment significance for these 43 pathways in the 18 settings, shown on the top, are shown in the color panels. For example, cell cycle is the pathway consistently enriched by BCs of top significance across all eight datasets, and the level of enrichment by genes in the BCs belonging to the 18 settings to cell cycle pathway is quite different, darker blue being the most significant.

Figure 4. Possible alternative chemo-resistance mechanism depicted by BC groups. (A-C) Discretized gene expression profile of the BC groups for 5FU (A), OXA (B), and FOLFOX (C). For (A-C), in the left-most panels, blue and white in the heatmap represent 1s and 0s in the discretized data matrix, while red marks the matrix element belonging to a certain BC group, framed in green dashed line. In the middle panels, the dendrograms show the results of agglomerative clustering of the resistance associated BCs. Each BC group is framed by a dashed rectangle. In the right-most panels, the survival curves represent the comparison of overall survival of the patients in a BC group (red) with those not (black), for the drug treated patients. (D-E) Distribution of the correlations calculated between expressions of genes in different groups with drug resistance measure IC50, in CTRP v2 dataset (D) and GI50 in K Bracht et al.’s dataset (E). The x-axis represents the correlations and the y-axis represents the density. (F) Relationships between chemo-resistance BCs and

different CMS classes. In columns 1-3, a “cross” sign indicates the drugs to that samples in the BCs show resistance; in columns 4-6, larger sizes of the sectors indicate higher significances that the BC’s resistance mechanisms is also exhibited in CMS I (blue), II (yellow), III (green), and IV (red); in columns 7-10, larger sizes of the squares indicate higher significances that the BC is positively (blue)/negatively (red) enriched by samples in each CMS class (only p<0.001 are shown); the last column shows for each BC, the type of drug and BC group it is linked to.

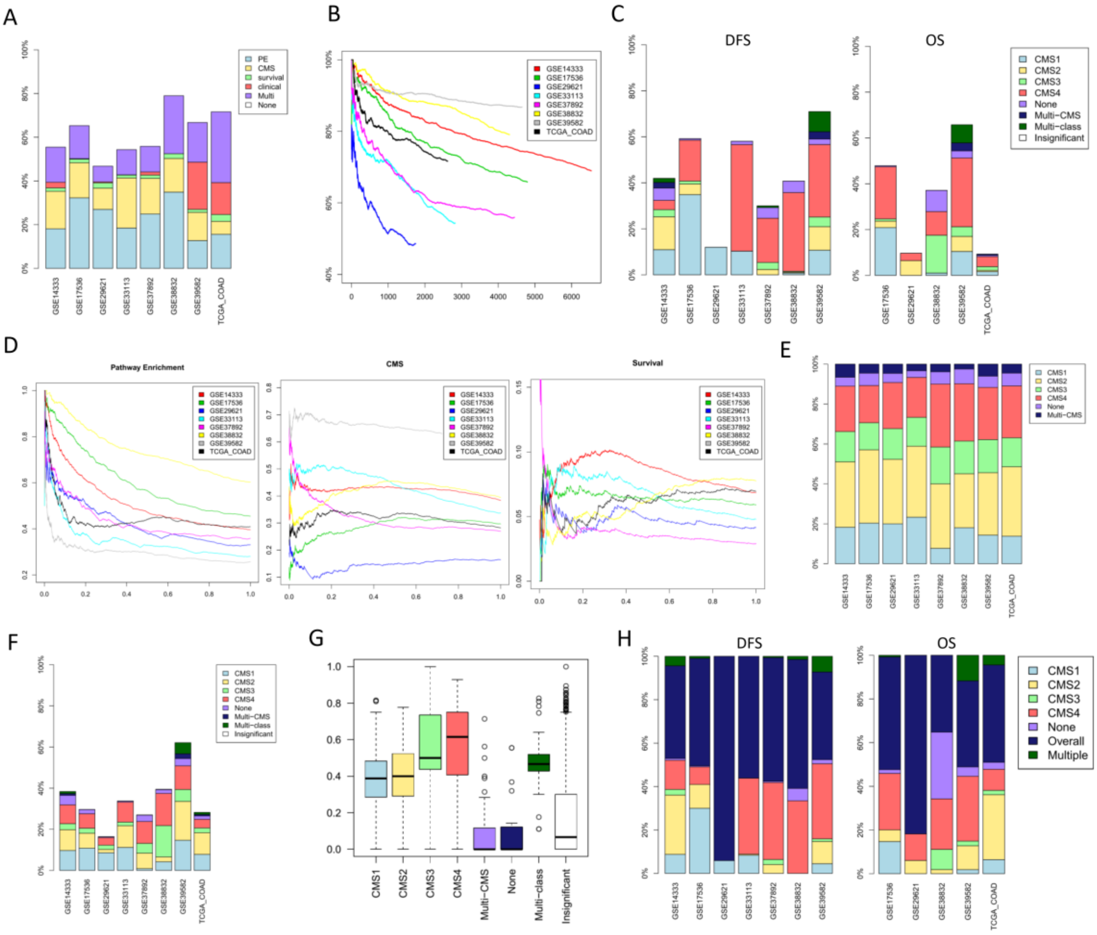
Figures:

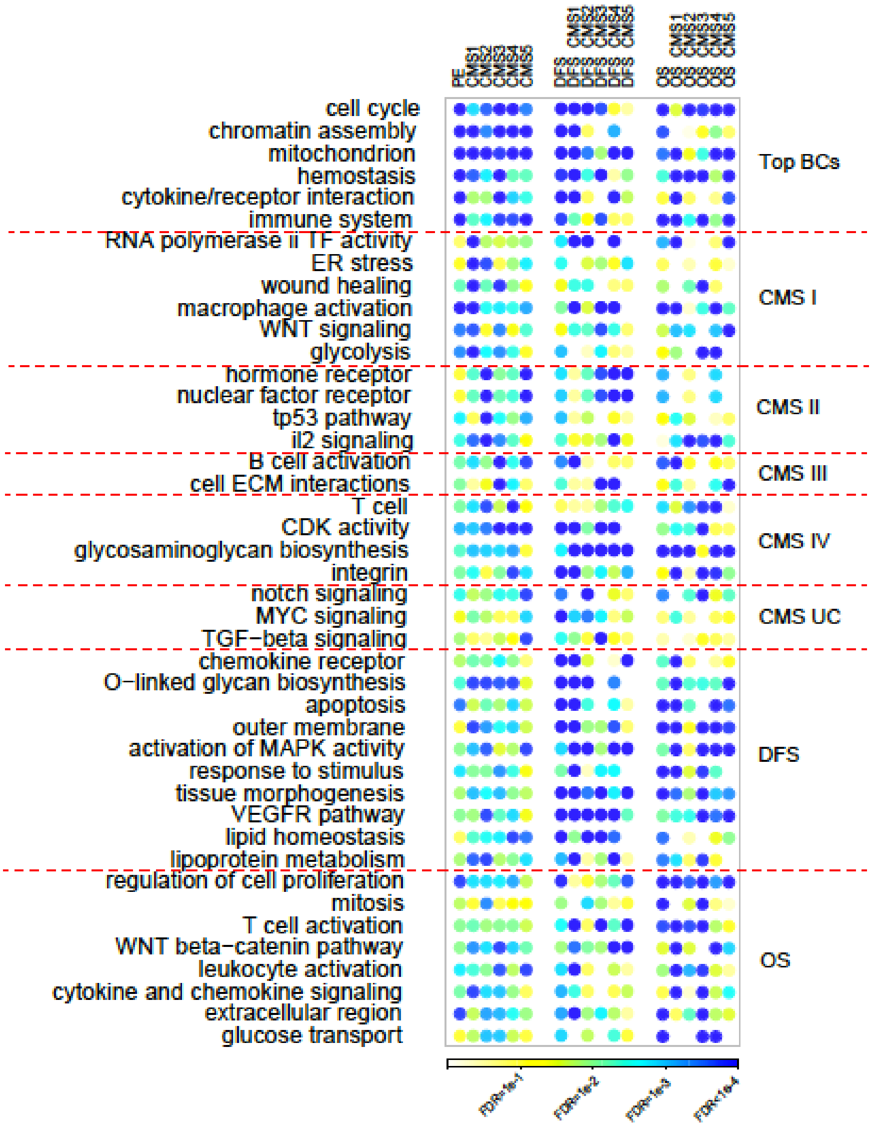

A
TCGA CRC data (385 samples)
GEO data sets (7 data sets, 1440 samples)
High dimensional data
Clinical data
High dimensional data
RNA-seq data
Mutation data
Microarray data
Clinical data
Drug usage data
BC identification
Data discretization
Bi-cluster identification
CMS prediction
Low dimensional Bi-clusters
Annotate BCs with level of bi-clustering significance and association with CMS class, patients' survival and other clinicopathological feature
CMS dependent/ independent BCs
Prognosis predictive BCs
Clinical feature/ genetic variation associated BCs
Functional annotation of BCs by enrichment in biological pathways/ sets of genes regulated by certain transcription factor
Consensus analysis
Identification of alternative chemo-resistance mechanism associated BCs
B
Discretization of a single gene's expression profile
0
20
40
60
2
4
6
8
10
12
14
Bi-clustering identification

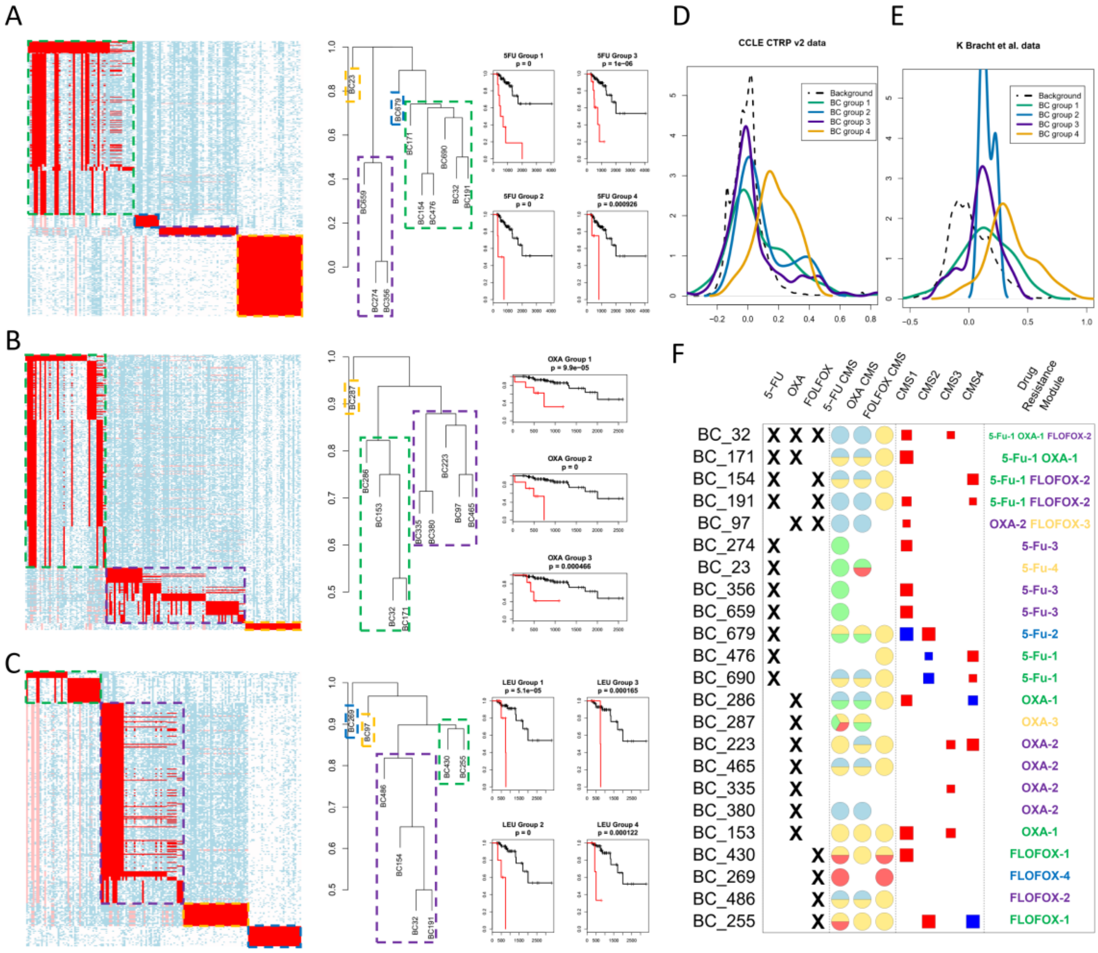
Gene expression	2	2.5	3	3.5	4	4.2	4.6	5	5.3	5.6	6	6.2	6.4	6.8	7	7.2	7.5	7.7	7.8	8	8.2	9	9.5	10	10.5	11	11.5	12	13	14
L / M / H	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
D_low	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D_moderate	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
D_high	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1

$Gene\ expression\ matrix: G_{m\times n}$
$\rightarrow Binary\ matrix: D_{3m\times n}$

	S1	S2	S3	S4	S5	S6	S7	S8	S9
G1_H	1	1	0	1	0	0	0	0	0
G1_M	0	0	1	0	1	0	0	0	1
G1_L	0	0	0	0	0	1	1	1	0
G2_H	1	1	0	1	0	0	0	0	0
G2_M	0	0	0	0	0	1	1	1	0
G2_L	0	0	1	0	1	0	0	0	1
G3_H	0	0	1	0	0	1	0	0	1
G3_M	0	0	0	0	1	0	1	1	0
G3_L	1	1	0	1	0	0	0	0	0
G4_H	0	0	0	0	0	1	1	1	0
G4_M	0	0	1	0	1	0	0	0	1
G4_L	1	1	0	1	0	0	0	0	0


A
B
C
D
E
F
G
H
PE
CMS
survival
clinical
Multi
None
GSE14333
GSE17536
GSE29621
GSE33113
GSE37892
GSE38832
GSE39582
TCGA_COAD
DFS
OS
CMS1
CMS2
CMS3
CMS4
None
Multi−CMS
Multi−class
Insignificant
Overall
Multiple
Pathway Enrichment
Survival
0%
20%
40%
60%
80%
100%
0
1000
2000
3000
4000
5000
6000
0.0
0.2
0.4
0.6
0.8
1.0


PE
CMS1
CMS2
CMS3
CMS4
CMS5
DFS
DFS CMS1
DFS CMS2
DFS CMS3
DFS CMS4
DFS CMS5
OS
OS CMS1
OS CMS2
OS CMS3
OS CMS4
OS CMS5
cell cycle
chromatin assembly
mitochondrion
hemostasis
cytokine/receptor interaction
immune system
Top BCs
RNA polymerase ii TF activity
ER stress
wound healing
macrophage activation
WNT signaling
glycolysis
CMS I
hormone receptor
nuclear factor receptor
tp53 pathway
il2 signaling
CMS II
B cell activation
cell ECM interactions
CMS III
T cell
CDK activity
glycosaminoglycan biosynthesis
integrin
CMS IV
notch signaling
MYC signaling
TGF−beta signaling
CMS UC
chemokine receptor
O−linked glycan biosynthesis
apoptosis
outer membrane
activation of MAPK activity
response to stimulus
tissue morphogenesis
VEGFR pathway
lipid homeostasis
lipoprotein metabolism
DFS
regulation of cell proliferation
mitosis
T cell activation
WNT beta−catenin pathway
leukocyte activation
cytokine and chemokine signaling
extracellular region
glucose transport
OS
FDR=1e-1
FDR=1e-2
FDR=1e-3
FDR<1e-4


A
BC23
BC679
BC171
BC690
BC659
BC154
BC476
BC32
BC191
BC274
BC356
5FU Group 1
p = 0
5FU Group 3
p = 1e−06
5FU Group 2
p = 0
5FU Group 4
p = 0.000926
B
BC287
BC286
BC153
BC32
BC171
BC223
BC335
BC380
BC97
BC465
OXA Group 1
p = 9.9e−05
OXA Group 2
p = 0
OXA Group 3
p = 0.000466
C
BC269
BC97
BC430
BC255
BC486
BC154
BC32
BC191
LEU Group 1
p = 5.1e−05
LEU Group 3
p = 0.000165
LEU Group 2
p = 0
LEU Group 4
p = 0.000122
D
CCLE CTRP v2 data
Background
BC group 1
BC group 2
BC group 3
BC group 4
E
K Bracht et al. data
Background
BC group 1
BC group 2
BC group 3
BC group 4
F
5−FU
OXA
FOLFOX
5−FU CMS
OXA CMS
FOLFOX CMS
CMS1
CMS2
CMS3
CMS4
Drug Resistance Module
BC_32 X X X 5-Fu-1 OXA-1 FLOFOX-2
BC_171 X X 5-Fu-1 OXA-1
BC_154 X X 5-Fu-1 FLOFOX-2
BC_191 X X 5-Fu-1 FLOFOX-2
BC_97 X X OXA-2 FLOFOX-3
BC_274 X 5-Fu-3
BC_23 X 5-Fu-4
BC_356 X 5-Fu-3
BC_659 X 5-Fu-3
BC_679 X 5-Fu-2
BC_476 X 5-Fu-1
BC_690 X 5-Fu-1
BC_286 X OXA-1
BC_287 X OXA-3
BC_223 X OXA-2
BC_465 X OXA-2
BC_335 X OXA-2
BC_380 X OXA-2
BC_153 X OXA-1
BC_430 X FLOFOX-1
BC_269 X FLOFOX-4
BC_486 X FLOFOX-2
BC_255 X FLOFOX-1