A functional module states framework reveals transcriptional states for drug and target prediction

Highlights
- FM states framework is developed to understand transcriptomic states
- FM states are discovered for cancer cell lines before or after intervention
- This tool facilitates drug and target prediction by integrating large data resources

In brief
Qin et al. develop a functional module states framework to understand the transcriptomic states of samples. This approach reveals transcriptional states of cancer cell lines before or after treatment. It provides valuable and biologically interpretable features for drug sensitivity and target prediction.
A functional module states framework reveals transcriptional states for drug and target prediction

Guangrong Qin,1,5,* Theo A. Knijnenburg,1,4 David L. Gibbs,1 Russell Moser,2 Raymond J. Monnat, Jr.,3 Christopher J. Kemp,2 and Ilya Shmulevich1,*

1Institute for Systems Biology, Seattle, WA 98109, USA
2Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA
3Department of Laboratory Medicine/Pathology & Genome Sciences, University of Washington, Seattle, WA 98195-7705, USA
4Present address: Allen Institute for Cell Science, Seattle, WA 98109, USA
5Lead contact
*Correspondence: guangrong.qin@isbscience.org (G.Q.), ilya.shmulevich@isbscience.org (I.S.)
https://doi.org/10.1016/j.celrep.2021.110269

SUMMARY

Cells are complex systems in which many functions are performed by different genetically defined and encoded functional modules. To systematically understand how these modules respond to drug or genetic perturbations, we develop a functional module states framework. Using this framework, we (1) define the drug-induced transcriptional state space for breast cancer cell lines using large public gene expression datasets and reveal that the transcriptional states are associated with drug concentration and drug targets, (2) identify potential targetable vulnerabilities through integrative analysis of transcriptional states after drug treatment and gene knockdown-associated cancer dependency, and (3) use functional module states to predict transcriptional state-dependent drug sensitivity and build prediction models for drug response. This approach demonstrates a similar prediction performance as approaches using high-dimensional gene expression values, with the added advantage of more clearly revealing biologically relevant transcriptional states and key regulators.

INTRODUCTION

Cells are complex systems that have been investigated at multiple levels, including genomic, epigenomic, transcriptomic, proteomic, and metabolomic levels. The varying concentrations and abundances of molecular species reflect and regulate the diverse processes that comprise cell function. On the genomic level, tumors are often stratified into different subtypes according to the mutation status of certain genes that are known to be predictive of clinical outcomes (Papaemmanuil et al., 2016; Schmitz et al., 2018). The transcriptome, however, can be affected by genetic and epigenomic alterations and is a more direct lens through which to investigate cell behavior and to provide clues to understand tumor heterogeneity or drug response.

Many methods have been developed to measure the transcriptome at different levels of resolution, such as gene expression microarrays, bulk RNA sequencing (RNA-seq), the L1000 platform (Subramanian et al., 2017), and single-cell RNA-seq (Zheng et al., 2017). These high-throughput techniques have been used to capture transcriptomes from thousands of primary tumor samples and cell lines. For example, The Cancer Genome Atlas (TCGA) (Hutter and Zenklusen, 2018; Knijnenburg et al., 2018; Thorsson et al., 2018) project measured gene expression on more than 10,000 tumor samples for 33 tumor types using RNA-seq. The Beat AML project measured gene expression for hundreds of acute myeloid leukemia (AML) samples from affected individuals (Tyner et al., 2018), whereas the Connectivity Map (CMap) (Subramanian et al., 2017) project provided more than one million transcriptomic profiles of different cell line samples after treatment with drugs or knockdown of genes using the L1000 platform. The Genomics of Drug Sensitivity in Cancer (GDSC) (Iorio et al., 2016) and Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012; Ghandi et al., 2019) projects measured transcriptional profiles of multiple cell lines before drug treatment using microarrays or RNA-seq. This growing archive of transcriptomics data is a rich source of information for defining transcriptional states and for understanding the functionality of cells that are perturbed by genetic alterations or drug treatment.

Large transcriptomics datasets are rich resources that can be used to define cellular states associated with cancer cell lines or more general transcriptional states for samples from affected individuals to understand the consequences of genetic alteration and the basis of drug response or predict drug sensitivity. Many module-based approaches have been developed for using gene expression data to characterize or stratify individuals or samples or to determine cellular states. These approaches reflect the use of context-specific or more general modules.
Among the first group, Segal et al. (2003) presented a probabilistic method to identify regulatory modules from gene expression data. These modules can be used to characterize clinical conditions according to a linear combination of modules that are activated and deactivated (Segal et al., 2004). Bild et al. (2006) focused on identifying oncogenic pathway signatures and understanding oncogenic states. Using experimentally derived pathway signatures, binary regression models, including principal components, were trained and used to predict pathway activation in tumor samples, which were then used in clustering of affected individuals (Bild et al., 2006). Onco-GPS first generates oncogenic activation signatures, followed by derivation of components through non-negative matrix factorization (NMF) decomposition of cell line data, and then uses a clustering procedure to arrive at cell states that are represented as a map (Kim et al., 2017). In the second group of methods, Pathifier is an algorithm that infers pathway deregulation scores for each tumor sample on the basis of expression data (Drier et al., 2013). The level of deregulation in a tumor sample is quantified by measuring the distance of the sample from normal samples. PAathway Representation and Analysis by Direct Reference on Graphical Models (PARADigm) infers specific genetic activities of affected individuals by incorporating interconnected variables encoding the expression and known activity of a gene and its products (Vaske et al., 2010).

The first group of context-specific methods, with a specific choice of signatures, provides a unique understanding of the biological system in a particular context but does not facilitate comparisons of different experiments. The second group of methods provides a more universal way to understand module activities using predefined modules. However, many of these approaches define a single score per module, which may introduce noise. In single-cell studies, several methods have been introduced to understand cellular states, such as Monocle (Trappnell et al., 2014), Monocle3 (Cao et al., 2019), single-cell energy path (scEpath) (Jin et al., 2018), Single Cell Lineage Inference Using Cell Expression Similarity and Entropy (SLICE) (Guo et al., 2017), and single-cell entropy (SCENT) (Teschendorf and Enver, 2017). The Monocle3 workflow performs dimensionality reduction using the top principal components (Cao et al., 2019). scEpath, SLICE, and SCENT, in contrast, quantify the energy landscape using the metric of single-cell energy (Guo et al., 2017; Teschendorf and Enver, 2017; Jin et al., 2018). The major limitations of these methods to define transcriptional states include (1) dimensionality reduction without using existing biological knowledge, which may limit the ability to interpret cell states; (2) failure to define cellular functionalities; and (3) use of only one factor or value to represent the transcriptomic states.

To address these points, we propose a functional module (FM) states framework (FM states) to define cell states by gene set- or pathway-derived factors (FM factors), numeric values that estimate the overall activity of the pathway or gene set. The vector of FM factors defined for each sample provides different ways to assess biologically interpretable quantifiers. Curated pathways, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2021; Kanehisa, 2019; Kanehisa and Goto, 2000; Ogata et al., 1999), cover a large number of genes and a diversity of biological processes, including DNA replication, transcription, energy metabolism, signaling, and others.

The expression of genes in these pathways may provide one way to estimate the activity of such FMs. We proposed that cell states or the average transcriptional states of samples can be defined by the overall activity profile of the FMs, which we define as FM factors. The vector of these so-called FM factors is then used to represent the FM states.

This approach, using FM factors to define FM states, has the ability to represent and interrogate diverse biological processes. In the remainder of this work, we demonstrate the utility of the newly established FM states framework by addressing the following questions: (1) are the transcriptomic states of a cancer cell line (MCF7) after drug treatment associated with drug concentration or drug targets? (2) Can we predict targetable vulnerabilities using the transcriptional cell states? (3) Can we use the functional states of cancer cell lines (multiple breast cancer cell lines) prior to drug treatment to predict the drug response? As an additional use demonstration, we also applied it to samples from individuals with AML to understand how the gene alterations or clinical features are associated with transcriptional states and how these transcription states are associated with drug response. Although we develop and answer the questions using breast cancer and AML as examples, our approach is general and can be readily applied to other tumor types as well as to non-neoplastic disease states.

RESULTS

Overview of the FM states framework

The goal of the FM states framework is to define biologically interpretable factors from high-dimensional gene expression data. The resulting FM factor matrix can be used for further clustering, annotation, detecting regulators for cell states, and machine learning. An overview of the FM states framework is shown in Figure 1. The first step is to select FMs. Different sets of modules can be used depending on the starting question. General criteria of module selection are the relevance of functionality for each module, wide coverage of functionalities across different modules, and wide coverage of genes across modules.

For the use cases presented here, we considered the 23 modules among metabolism, genetic information processing, environmental processing, and cellular process in the KEGG pathways (Kanehisa, 2019; Kanehisa et al., 2021; Kanehisa and Goto, 2000; Ogata et al., 1999) as FMs because they cover diverse cellular functional activities and many genes (over 5,000) and show mutual exclusivity across modules (Figure S1A and S1B).

In a second step, we defined four types of factors for each FM: a single-sample gene set enrichment analysis score (ssGSEA_score): up_strength, down_strength, and TF_strength. The ssGSEA_score represents the ranking of module-related genes given the expression levels in a sample (Hanzelmann et al., 2013). The up_strength and down_strength factors estimate the ratio of genes showing high expression (95th percentile or Z score > 1.6) or low expression (fifth percentile or Z score < −1.6) in one FM for one specific sample compared with all other samples. TF_strength estimates the weighted average expression level of transcription factors (TFs) that are predicted to
regulate genes in a given FM (STAR Methods). This approach differs from transcriptional regulation network inference approaches such as single-cell regulatory network inference and clustering (SCENIC) (Aibar et al., 2017), Network Reprogramming using EXPression (NetREX) (Wang et al., 2018), modular regulatory network learning with per gene information (MERLIN) (Roy et al., 2013), Inferelator (Bonneau et al., 2006), and network component analysis (NCA) (Liao et al., 2003) in that we are not aiming to estimate TF activities; rather, we are trying to indicate enrichment and correlation of TFs that map to a module. After defining the FM factor matrix, consensus clustering (Monti et al., 2003; Wilkerson and Hayes, 2010) can be used to determine the number of clusters (states) given the input samples. This framework provides a clear biological interpretation for each state by annotating states with different factors and enriched TFs (STAR Methods). The FM factors can also serve as features for machine learning methods to identify key features (i.e., FM factors) that distinguish different phenotypes or responses to perturbations, such as gene knockdown or drugs, as we demonstrate.

Functional transcriptional states of breast cancer cell lines reflect drug response

We applied the FM states method to the high-throughput transcriptional profiles of the MCF7 breast cancer cell line after drug treatment, available from CMap (STAR Methods; Subramanian et al., 2017), following the pipeline shown in Figure S2. MCF7 was derived from a luminal A subtype breast cancer (Subik et al., 2010) that was hormone receptor (estrogen and progesterone receptor) positive and HER2 negative and carried a PI3KCA p.E545K mutation (from CCLE and GDSC datasets) (Ghandi et al., 2019; Iorio et al., 2016). MCF7 cell drug sensitivity data (half-maximal effective concentration [EC50] values) from the Cancer Therapeutics Response Portal (CTRP v.2) were selected (Aksoy et al., 2017; Basu et al., 2013; Rees et al., 2016; Seashore-Ludlow et al., 2015), resulting in a dataset of 1,287 gene expression profiles that represent the transcriptome of the MCF7 cell line treated with 190 drugs or compounds at different concentrations (drug-treated samples). A reference gene expression dataset of 1,400 MCF7 samples treated with DMSO or H2O from CMap (Subramanian et al., 2017) was also used (drug-free samples).

Twenty-three FMs (Figure S1; Table S1) and the four categories of factors (ssGSEA_score, up_strength, down_strength, and TF_strength) were selected to define the functional states of MCF7 cells. FM factors associated with drug treatment were selected by comparing the drug-treated samples and drug-free samples (Mann-Whitney rank test followed by Benjamini-Hochberg adjustment, false discovery rate [FDR] < 1.310^{-6}, effect size > 0.2 [upregulated], effect size < −0.2 [downregulated]). The FM factors, including ssGSEA scores of cell cycle, replication and repair, nucleotide metabolism, transcription, and translation, were significantly downregulated after drug treatment, whereas membrane transport, signal transduction, and lipid metabolism were upregulated after drug treatment (Figure S3). FM factors for the drug-treated samples were normalized using reference samples without drug treatment. Drug treatment-associated FM factors were then selected for further analysis with consensus clustering (Monti et al., 2003) to define potential transcriptional states after drug treatment.

Five transcriptional states were detected for MCF7 cells after drug treatment (Figures 2A and S4). Compared with drug-free samples, state 1 (S1) shows upregulation of the modules involved in replication and repair, transcription, translation, and...
Figure 2. Five states were revealed for the MCF7 breast cancer cell line transcriptome after drug treatment using the FM factors and transcriptional factors

(A) Heatmap of relative FM factors for MCF7 breast cancer cell lines after treatment of different drugs at different concentrations normalized by the reference MCF cell line without drug treatment. FM factors associated with drug treatment were selected by comparing the drug-treated samples and drug-free samples (Mann-Whitney rank test followed by Benjamini-Hochberg adjustment, FDR < 1 \times 10^{-3} and |effect size| > 0.2). Five states are defined by the consensus clustering method. The x axis shows the 1,287 drug-treated MCF7 cells; the y axis shows the relative FM factors scaled by row.

(B) FMs with FM factors (ssGSEA, up-strength and down-strength) showing significant differences between one state to at least three other states using Wilcoxon rank-sum test in the scipy.stats library (p < 0.01 is considered the threshold for significance here) and with a size effect greater than 1 or smaller than \(-1\) (STAR Methods). FMs with an effect size between one state and all others greater than a threshold (effect size > 1) in the category of ssGESA up-strength or smaller than \(-1\) in the down-strength are considered upregulated; an effect size smaller than \(-1\) in ssGESA up-strength, or greater than 1 in down-strength is considered downregulated. The label “dysregulated” represents a module with alteration on both sides. To make a “dysregulated call,” the different factors are considered as a unit. Rules for making the call for each module are shown above. Primarily, for up- or downregulation calls, the factors must all agree (all up or all down for the significant differential factors). The dysregulated call is made when the factors do not agree and show up- and downregulation.

(C) Differential TF strength among different states using the same threshold as in (B).

(legend continued on next page)
cell cycle and downregulation of modules involved in transport and catabolism, signaling molecules and interaction, carbohydrate metabolism, and membrane transport (active cell cycle state; Figures 2A and 2B). S2 and S3 show slight differences compared with drug-free samples (basal state), although S3 shows a somewhat lower expression of cell cycle-related modules (Figures 2A and 2B). S4 and S5 featured upregulation of transport and catabolism, carbohydrate metabolism, signal transduction, amino acid metabolism, and apoptosis and downregulation of translation, transcription, replication and repair, nucleotide metabolism, cell cycle, and apoptosis. Specifically, S4 features significant downregulation of TFs that regulate replication and repair, cell cycle, and cellular senescence (high apoptosis/low cell cycle/cell cycle TF suppressed; Figure 2C). S5 features upregulation in TFs that regulate various FMs, including transport and catabolism, lipid metabolism, replication and repair, cell cycle, and cell motility (high apoptosis/low cell cycle/cell cycle TF activated). In this way, the FM states framework provides direct annotation for different states by measuring the FM factors among different states.

The FM states framework also provides annotations for transcriptional states by detecting the TFs that may regulate different states. We annotated the potential effect of one TF on a pathway by the enrichment of its target genes with positive or negative correlation in the module. We performed an enrichment analysis (Fisher’s exact test) to estimate whether there is significant enrichment of positively or negatively correlated target genes. If the positively correlated target genes for one TF are enriched in one module, then we define this TF as having a positive effect on this module. Otherwise, if the negatively correlated target genes for one TF are enriched in one module, then we define this TF as having a negative effect on this module. Figures 2D–2F show the TFs with (1) significant differential expression among different states (one-way ANOVA and Bonferroni multiple test correction) and (2) those with their target genes enriched in the differentially regulated FMs. For example, FOXO1 has a negative effect on the cell cycle, with lower expression in S4 and higher expression in S4 and S5 (Figures 2E, 2F, and S5). This is consistent with the observation that S4 and S5 are low-cell-cycle states. A previous study has also shown that FOXO1 is associated with cell cycle inhibition (Schmidt et al., 2002). The basal-like states (S2 and S3) did not show significant differences in TF expression. Expression of JUN is reduced in S1 and increased in S4 and S5 (Figures 2E, 2F, and S5). Previous time course studies of MCF7 cells after drug treatment showed that expression of the JUN gene was elevated 36 h after treatment with chemotherapy agents, such as doxorubicin and 5-fluorouracil (Troester et al., 2004). SOX2 also shows upregulation in S4 and S5 (Figures 2E, 2F, and S5). Upregulation of SOX2 has been reported to promote the cancer stem cell-like phenotype associated with resistance upon anti-cancer drug treatment (Huser et al., 2018).

MYC proto-oncogene target genes, which are positively correlated with MYC expression, are enriched in the cellular senescence module (Figure 2E, red lines). MYC shows reduced expression in S4 (Figures S5 and 2E, blue), which is consistent with previous studies suggesting that suppression of MYC induces cellular senescence in tumors (Wu et al., 2007). This result suggests that downregulation of MYC may drive S4 as a cellular senescence state. Similarly, FOXM1 and E2F1, which have a positive effect on the cell cycle and replication and repair FMs, show low expression in S4 and may drive formation of the S4 low cell cycle state. In contrast to S5, forkhead box O1 (FOXO1), transcription factor AP-2 alpha (TFAP2A), zinc finger and BTB domain containing 33 (ZBTB33), and nuclear factor I C (NFIC), which have a negative effect on the cell cycle, replication, and repair, are highly expressed. Previous studies have shown that overexpression of the TFAP2A-encoded TF AP-2α triggers apoptosis (Muller et al., 2004). The difference in regulators for the two apoptosis-associated states (S4 and S5) may suggest different biological mechanisms induced by drugs.

Functional transcriptional states reveal dose-dependent responses and mechanisms of action

We next set out to determine whether transcriptional states are associated with pharmacological effects by asking whether transcriptional states were associated with drug concentration. This analysis took advantage of data on MCF7 cell responses after treatment with different drugs over a range of concentrations. By mapping the drug concentration for a specific drug to the drug response curve measured in the CTRP v.2 project (STAR Methods; Rees et al., 2016), we categorized the MCF7 transcriptome profile following treatment with each drug to the high-dosage treatment group when the concentration was greater than the EC50 concentration or the low-dosage treatment group when less than the EC50 concentration. Five states induced by drugs were associated with drug concentrations (chi-square test, p < 0.001); e.g., S1 and S2 were enriched in cells treated with low drug concentrations, and S4 and S5 were enriched in cells treated with high drug concentrations (Figures 3A and 3B). Figure 3A shows the example of doxorubicin. These analyses collectively suggest that different drug concentrations induce different transcriptional states.

We next wanted to determine whether drug-induced transcriptional states are associated with drug targets or drug action in addition to drug concentration. MCF7 cells, as noted above, are classified as luminal A subtype, estrogen receptor (ER)/progesterone receptor (PR) positive, and HER2 negative (Subik et al., 2010). We grouped drugs according to their targets or activity as defined in the CTRP v.2 dataset (STAR Methods) and then analyzed the FM states of MCF7 cells treated with individual drugs (Figures...
The results show that S4 is enriched in cells treated with ER antagonists, topoisomerase II (TOP2) inhibitors, and histone deacetylases (HDAC) inhibitors (Fisher’s exact test, Benjamini-Hochberg multiple comparison [Benjamini-Hochberg adjustment (BH adjust)], FDR < 0.05; Figures 3C and 3E). The enrichment of ER antagonists and lower expression of estrogen receptor 1 (ESR1) in S4 suggests a drug-responsive state. HDAC inhibitors play important roles in epigenetic regulation, inducing death, apoptosis, and cell cycle arrest in cancer cells (Kim and Bae, 2011). The enrichment in approved anti-breast cancer drugs is also consistent with the observation that S4 features high apoptosis and a low cell cycle. S5 is enriched in cells treated with multiple kinase inhibitors, cyclin-dependent kinases inhibitors, inhibitors of RNA polymerase II, inhibitors of dihydrofolate reductase, and inhibitors of phosphatidylinositol 3-kinase (PI3K), and mechanistic target of rapamycin (mTOR) kinase activity (Fisher’s exact test, BH adjust, FDR < 0.05; Figures 3D and 3F). Because MCF7 cells contain the PI3KCA p.E545K mutation, the enrichment of PI3K inhibitors and alteration of signaling and transduction in S5 suggests that S5 is also a drug-responsive state. These results suggest that we are able to identify state changes associated with drugs or chemicals at different concentrations and with different actions. FM states annotation of the MCF7 transcriptomic states after drug treatment provides the advantage of greater mechanistic interpretability of the drug-induced cell states.

**FM states predict therapeutic vulnerabilities**

With classification of transcriptomic states of MCF7 cells after drug treatment, we then wanted to determine whether transcriptomic states could be used to predict therapeutic vulnerability. We made the assumption that knockdown or knockout of genes that drive drug-responsive transcriptomic states would identify these genes as potential therapeutic targets. Using the CMap dataset (Subramanian et al., 2017), which includes transcriptome data following shRNA knockdown, we generated FM factors and classified each sample into the pre-defined drug-induced states (S1–S5) using a K-nearest neighbor (KNN) classifier. We first assigned each transcriptome profile to one state using the KNN method (Figure S6). Because each sample represents the transcriptome after gene knockdown using one shRNA seed, we next performed an enrichment analysis (Fisher’s exact test, p < 0.05) of shRNA seeds for each gene in different states and assigned genes to different states (Figure 4A). To further validate the potential of target
dependency, we integrated cancer dependency data from the DepMap project (Meyers et al., 2017; Dempster et al., 2019), which measured the gene knockout effect using CRISPR. Our results show that the MCF7 cell line has greater sensitivity to depletion of genes that are associated with S3-, S4-, and S5-like states (Kolmogorov-Smirnov test, p < 0.05) (Figures 4B and 4C). We then used the perturbation effect and the proportion of seeds that drive the states as a threshold for selecting the potential gene targets for inducing each state (Table S2).

For example, for S4, our approach predicted that knockdown of the proteasome genes PSMA1 and PSMB2 would induce an S4-like state (Figure 4D). PSMB2 is a target of carfilzomib, a proteasome inhibitor and well-known antineoplastic agent. Knockdown of HDAC3 is also predicted to contribute to the S4-like state, which is consistent with our annotation that S4 is associated with HDAC inhibitors. The method also selected RUVBL1, MCM3, MCM7, RPS6, RPL7, and CCND1 as being linked to S4. Previous studies have suggested that a selective inhibitor of the RUVBL1/2 complex reduced growth in AML and multiple myeloma (Assimon et al., 2019). For S5, we predicted that knockdown of ribosome protein small subunit genes (such as RPS9) will drive the S5-like state (Figure 4E). Knockdown of ESR1 also drives the S5-like state, which is consistent with the genetic background of MCF7 as an ER+ breast cancer cell line. This approach also predicted SOAT1 as a potential target. SOAT1 has been investigated in hepatocellular carcinoma as a promising drug target (Jiang et al., 2019). Other genes predicted to be targetable vulnerabilities in MCF7-like breast cancer can

Figure 4. Prediction of potential targets for breast cancer (MCF7 cell line like)
(A) Number of genes that are predicted to be associated with each state that knock down of these genes will drive a similar FM state.
(B) Boxplot of the gene knockout (KO) effects of state-associated genes (data from DepMap 2019 Q3, Achilles_gene_effect). The median nonessential KO effect is scaled to 0, and the median essential KO effect is –1).
(C) Cumulative density curve for gene KO effects in different states. The curve for “ALL” represents the cumulative density curve for the gene knockout effects for all genes measured in the DepMap project. Kolmogorov-Smirnov test was used to test the difference of the gene knockout effect between each state-associated genes and “ALL” genes. With p value < 0.05 as significance threshold, S3-, S4-, and S5-like states show significance.
(D) Predicted targets associated with S4-like states. The x axis shows the KO effect as measured in DepMap data, and the y axis shows the proportion of shRNA seeds for this gene that are associated with this state.
(E) Predicted targets associated with S5-like states; the labels are similar to (D). See also Figure S6 and Table S2.
Figure 5. Predicting drug responses using FM factors
(A) Heatmap of Spearman correlation between the FM factors and drug-sensitivity values (log(IC50)), colored scale by Spearman correlation coefficient. An asterisk represents significant correlations with a threshold *p < 0.05.
(B–D) Volcano plot of Spearman correlation between FM factors and drug response. The x axis represents the Spearman correlation coefficient, and the y axis represents the p values (−log10 transformed).

(legend continued on next page)
Pre-existing transcriptional FM states are associated with drug response

We next wanted to determine whether the transcriptional states of cancer cell lines prior to drug treatment are associated with drug response. We applied the FM states method to define cell states using the basal transcriptomic profile from the GDSC project (Iorio et al., 2016) and analyzed the association between cell states and drug response. Using breast cancer cell line data, we generated the FM factors for all 49 breast cancer cell lines in the GDSC project. We selected potential effective drugs based on absolute IC50 values less than 1 μmol in at least 5 cell lines.

We estimated Spearman correlations between the FM factors and IC50 values for each selected drug, as shown in Figure 5A. The gene expression of cell cycle, replication and repair, metabolism of cofactors, and vitamin modules shows negative correlations with IC50 for most drugs, suggesting that breast cancer cell lines with higher expression of cell cycle or replication and repair pathways will show higher sensitivity to most drugs (Figure 5B). On the contrary, gene expression in the transport and catabolism pathway and in carbohydrate metabolism modules show positive correlations with the IC50 of most drugs (Figures 5A and 5C). In contrast, some FMs, such as cell motility, signaling molecules and interaction, and signaling transduction modules show different patterns of associations. Cells with overall high expression of these pathways are sensitive to drugs that target the ERK (extracellular signal-regulated kinase 1) mitogen-activated protein kinase (MAPK) signaling pathway. Pre-existing transcriptional FM states are associated with drug response.

Finally, we wanted to assess whether using FM factors alone, rather than all gene expression data, would provide comparable performance. For this purpose, we used the random forest (RF) algorithm to build predictive models for the sensitivity of cell lines for each drug. The median prediction accuracy for the bootstrapped RF models for all drugs is 0.8 (Figure 5E). The prediction accuracy using the FM factor-based RF models is comparable with the prediction models using the gene expression data (Figures 5E–5G).

FM states for individuals with AML

To address the generality of the FM states framework, we also analyzed FM states for individuals with AML. For this analysis, we selected the same 23 FMs and calculated the FM factors for samples from affected individuals using bulk RNA-seq from the Beat AML dataset (Tyner et al., 2018). Consensus clustering was applied to the FM factor matrix, and five FM clusters, S1–S5, are shown in Figure 6A. S1 and S2 show upregulation in transcription and translation; S3 features upregulation in the cell cycle, TP53 signaling pathway, and cellular senescence; S4 and S5 show upregulation in signaling molecules and interactions; and, more specifically, S4 shows lower expression of genes in the metabolism pathways (Figure 6B). AML has been reported to be driven by different genetic alterations, including gene mutations and cytogenetic alterations, and diagnosis of AML includes clinical factors such as blast percentage. We thus further analyzed whether each of the five transcriptional states showed enrichment with specific AML genetic and clinical features. The results show that S1 is enriched with mutation of CCAAT enhancer binding protein alpha (CEBPA) and fms related receptor tyrosine kinase 3 (FLT3); S2 is enriched with FLT3, nucleophosmin 1 (NPM1), and tet methylcytosine dioxygenase 2 (TET2); S3 is enriched with RUNX family transcription factor 1 (RUNX1), stromal antigen 2 (STAG2), and TP53; S4 with isocitrate dehydrogenase (NADP(+)) 2 (IDH2); and S5 with NRAS proto-oncogene, GTPase (NRAS), KRAS proto-oncogene, GTPase (KRAS), and protein tyrosine phosphatase non-receptor type 11 (PTPN11) (Figure 6C). Furthermore, ANOVAs revealed significant differences in blast percentage among the five states. S1 and S2, which show higher transcriptional levels of genes in transcription and translation, are associated with higher blast percentages compared with the other three groups (Figure 6D).

These results suggested that different mutations in AML may result in different transcriptional states that affect drug sensitivity. To explore this possibility further in AML, we applied the approach for drug sensitivity prediction described above for breast cancer to AML. When venetoclax, a BH3 mimic that blocks the anti-apoptotic BCL2 protein, was used as an example, the Spearman correlation between the FM factors and drug response showed that the FM factors transport and catabolism_ssGSEA, signaling transduction_ssGSEA, and others were significantly positively correlated with the AUC (area under the curve) values. On the other hand, the higher expression of modules such as transcription, translation, folding, sorting, and degradation was associated with higher venetoclax sensitivity. We further built random forest (RF) models for venetoclax drug sensitivity prediction using FM factors as features. By randomly selecting 80% of samples from the Beat AML dataset as training sets, we estimated the accuracy for the remaining 20% of test samples. The median accuracy in the test samples is about 78.9%, which is similar to the prediction accuracy in the cell line data from breast cancer cell lines. It is also comparable with but slightly lower than the prediction models using all gene expression values (82.9%). Again, a particular advantage of using FM states lies in the interpretability of the results, in contrast to classification using a large set of genes that may be found in Figures 4D and 4E and Table S2. These results suggest that the FM states framework, which integrates analyses of drug-induced cell states and functional screening data, can identify potential new drug targets.

Additional Notes:

(E) Histogram of prediction accuracy of drug response on the testing sets (the ratio of samples between the training set and testing set is 4:1) using the RF model and the FM factors as features. Each bin displays the bin’s raw count divided by the total number of drugs (n = 57) and the bin width (width = 0.1).

(F) Distribution of prediction accuracy of drug response using gene expression values as features.

(G) Comparison of prediction accuracy between the RF models using the FM factors and all gene expression values for each drug; each point represents the median prediction accuracy for the 100 bootstrapped RF models using the features of FM factors and the gene expression values for each drug.
Figure 6. FM states for acute myeloid leukemia (AML)
(A) FM states defined in samples from individuals with AML.
(B) Featured differentially expressed FMs across different states. FM factors show significant difference between one state to at least three other groups are shown. Ranksum test p value < 0.01, | Effect size | > 1.
(C) Enriched mutated genes in each state. Fisher’s exact test was used, p value < 0.05 was considered as significance.
(D) Distribution of blast percentage in groups of affected individuals with different FM states. ANOVA testing revealed significant differences in blast percentage among five states (p value < 0.01).
(E) FM factors that show significant Spearman correlation with the ex vivo drug sensitivity AUC values.
(F) Comparison of the RF model based on the features of FM factors or all gene expression.
be similar in accuracy but substantially harder to interpret. These results show that the FM states method can be generalized and can readily generate interpretable results that could drive additional experiments.

**DISCUSSION**

The FM states method considers preliminary knowledge (FMs or pathways) and defines the FM factors for any given sample to identify low-dimensional representative features and biologically relevant transcriptional states. Further, the FM states method determines key transcriptional regulators of different states. The FM states method thus differs from approaches that infer gene regulation networks, such as SCENIC (Aibar et al., 2017), NetREX (Wang et al., 2018), MERLIN (Roy et al., 2013), Inferelator (Bonneau et al., 2006), NCA (Liao et al., 2003), and others. Our approach, in contrast, is not trying to provide a precise gene regulatory network but, rather, clues regarding TFs associated with upregulation or downregulation of different FMs by using statistical enrichment analysis. Compared with other module-based scoring approaches, such as PARADIGM (Vaske et al., 2010) and Pathifier (Drier et al., 2013), our approach provides multiple factors for one single module with better direct biological interpretability.

Large datasets from consortium projects are rich resources for characterizing the transcriptional states of different cell lines before and after drug treatment or gene knockdown. We showed that the FM states method, a simple and biologically interpretable method, is able to predict drug targets and drug response through integrative analysis of multiple datasets. As an illustrative example, we used the FM states method to determine drug-induced transcriptional states for the breast cancer cell line MCF7 from public gene expression data from the CMap project. Different transcriptional states were associated with drug concentration and drug class. We further identified potential targets through integrative analysis of transcriptional states after drug treatment and gene knockdown. By combining the transcriptional states and gene knockdown efficacy, we predicted potential targets for MCF7-like breast cancer while recognizing the challenge of genetic and non-genetic intratumoral heterogeneity. Using the FM factors as features to represent transcriptional states, FM states revealed transcriptional state-dependency of drug sensitivity and resistance and exhibited similar performance in a drug response prediction test compared with models that use all gene expression data. A particular advantage of FM states is that consistent module selection makes it possible to directly compare different experiments.

This FM states approach can be used to analyze gene expression data from cell lines and samples from affected individuals. Care needs to be taken, however, when applying FM states to samples from affected individuals, where tumor heterogeneity and associated non-tumor cells will require more care when interpreting states; the functional states inferred from bulk gene expression data will represent the average states of all included cells or cell types. Despite these caveats, FM states and FM factors are a useful way to stratify individuals and to understand the average “activity” of different modules across a mixture of cell types.

This approach can also be applied to larger gene expression datasets. With an increase in sample size, we may expect a more fine-grained discovery of functional states. This work underscores the power and value of integrating large public data resources as well as preliminary knowledge to understand and predict cancer drug responses.

**STAR METHODS**

Detailed methods are provided in the online version of this paper and include the following:

- **KEY RESOURCES TABLE**
- **RESOURCE AVAILABILITY**
  - Lead contact
  - Materials availability
  - Data and code availability
- **METHOD DETAILS**
  - Selection of functional modules
  - Definition of the FM-factors
  - Use of FM states to predict targetable cancer vulnerabilities
  - Define basal FM-states of breast cancer cell lines to predict drug sensitivity
  - Define FM-states for acute myeloid leukemia patients
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.110269.

**ACKNOWLEDGMENTS**

G.Q., T.A.K., and I.S. are supported by National Cancer Institute (NCI), National Institutes of Health (NIH) grants U01 CA217883, P01 CA077852, and U24CA210992. R.J.M. was supported by NCI P01 CA077852. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank Dr. Jeffery W. Tyner from OHDSI for discussion of the AML analysis and Dr. Vésteinn Thorsson from the Institute for Systems Biology for the discussion during the revision stage. We would like to thank the Cancer Target Discovery and Development (CTD) consortium for discussion and providing resources for this project, the Broad Institute Connectivity Map project (CMap), the Broad Institute Cancer Dependency Map project (DepMap), and the Wellcome Sanger Institute Genomics of Drug Sensitivity in Cancer project (GDSC) for sharing datasets with the research community.

**AUTHOR CONTRIBUTIONS**

Conceptualization and methodology, G.Q., T.A.K., and I.S.; data analysis and implementation, G.Q.; writing – original draft, G.Q.; writing – review & editing, G.Q., I.S., D.L.G., C.J.K., R.M., R.J.M., and T.A.K.; supervision, I.S.

**DECLARATION OF INTERESTS**

The authors declare no competing interests.

Received: February 18, 2021
Revised: August 24, 2021
Accepted: December 23, 2021
Published: January 18, 2022
REFERENCES

Aibar, S., Gonzalez-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 14, 1083–1086.

Aksoy, B.A., Dancik, V., Smith, K., Mazerik, J.N., Ji, Z., Gross, B., Nikolova, O., Jaber, N., Califano, A., Schreiber, S.L., et al. (2017). CT2D2 Dashboard: a searchable web interface to connect validated results from the Cancer Target Discovery and Development Network. Database (Oxford) 2017, bax054.

Assimon, V.A., Tang, Y., Vargas, J.D., Lee, G.J., Wu, Z.Y., Lou, K., Yao, B., Menon, M.K., Pios, A., Perez, K.C., et al. (2019). CB-6644 is a selective inhibitor of the RUVBL1/2 complex with anticancer activity. ACS Chem. Biol. 14, 236–244.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.

Basu, A., Bodycombe, N.E., Cheah, J.H., Price, E.V., Liu, K.e., Schaefer, G.I., Ebright, R.Y., Stewart, M.L., Ito, D., Wang, Stephanie, et al. (2013). An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. Cell 154 (5), 1115–1161.

Bil, A.H., Yao, G., Chang, J.T., Wang, Q., Potti, A., Chasse, D., Joshi, M.B., Harpole, D., Lancaster, J.M., Berchuck, A., et al. (2006). Oncophenotypic signatures in human cancers as a guide to targeted therapies. Nature 439, 353–357.

Bonneau, R., Reiss, D.J., Shannon, P., Facciotti, M., Hood, L., Baliga, N.S., and Thorsson, V. (2006). The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol. 7, R36.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature 566, 496–502.

Dempster, J., Rossen, J., Kazachkova, M., Pan, J., Kugener, G., Root, D.E., and Tashemni, A. (2019). Extracting biological insights from the project achilles genome-scale CRISPR screens in cancer cell lines. BioRxiv. https://doi.org/10.1101/720243.

DepMap. B. (2019). DepMap 1902 Public (Fireshare.Datasset).

Drier, Y., Sheffer, M., and Domany, E. (2013). Pathway-based personalized analysis of cancer. Proc. Natl. Acad. Sci. U S A 110, 6386–6393.

Garcia-Alonso, L., Holland, C.H., Ibrahim, M.M., Turei, D., and Saez-Rodriguez, J. (2019). Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 29, 1363–1375.

Ghandi, M., Huang, F.W., Jane-Valbuena, J., Kryukov, G.V., Lo, C.C., McDougal, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). SCENIC: single-cell regulatory network inference and clustering. Nat. Methods 16, 79–88.

Guo, M., Bao, E.L., Wagner, M., Whitsett, J.A., and Xu, Y. (2017). SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res. 45, e54.

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7.

Hutter, C., Afshari, F., Knijnenburg, T.A., Vis, D.J., Bignell, G.R., Menden, M.P., Schubert, M., Aber, N., Goncalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A landscape of pharmacogenomic interactions in cancer. Cell 166, 740–754.

Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., Xing, B., Sun, W., Ren, L., Hu, B., et al. (2019). Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature 567, 257–261.

Jin, S., Maclean, A.L., Peng, T., and Nie, Q. (2018). scEPath: energy landscape-based inference of transition probabilities and cellular trajectories from single-cell transcriptomic data. Bioinformatics 34, 2077–2086.

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Sci. 28, 1947–1951.

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguo-Watanabe, M., and Tanabe, M. (2021). KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 49, D545–D551.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kim, H.J., and Bae, S.C. (2011). Histone deacetylase inhibitors: molecular mechanisms of action and clinical trials as anti-cancer drugs. Am. J. Transl. Res. 3, 166–179.

Kim, J.W., Abudayyeh, O.O., Yerema, H., Yeang, C.H., Stewart, M., Jenkins, R.W., Kitajima, S., Konieczkowski, D.J., Medetug-Emar, K., Cavazos, T., et al. (2017). Decomposing oncogenic transcriptional signatures to generate maps of divergent cellular states. Cell Syst. 5, 105–118.e9.

Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018). Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome Atlas. Cell Rep. 23, 239–254.e6.

Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., and Roychowdhury, V.P. (2003). Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. U S A 100, 15522–15527.

Meyers, R.M., Bryan, J.G., Mclfarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dhiria, N., Montgomery, P.G., Cowley, G.S., Pantel, S., et al. (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat. Genet. 49, 1779–1784.

Monti, S., Tamayo, P., Mesirov, J.P., and Golub, T.R. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learn. 52, 91–118.

Muller, F.U., Loser, K., Kiedeiter, U., Neumann, J., von Wallbrunn, C., Dobner, T., Scheld, H.H., Bantel, H., Engels, I.H., Schulze-Osthoff, K., and Schmitz, W. (2004). Transcription factor AP-1/zapala triggers apoptosis in cardiac myocytes. Cell Death Differ. 11, 485–493.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 27, 29–34.

Papaemmanuil, E., Gerstung, M., Bullinger, L., Gadzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Boll, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. N. Engl. J. Med. 374, 2209–2221.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Rees, M.G., Seashore-Ludlow, P., 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). Correlated sensing sensitivity and basal gene expression reveals mechanism of action. Nat. Chem. Biol. 12, 109–116.

Roy, S., Lagree, S., Hou, Z., Thomson, J.A., Stewart, R., and Gasch, A.P. (2013). Integrated module and gene-specific regulatory inference implicates upstream signaling networks. PLoS Comput. Biol. 9, e1003252.

Schmidt, M., Fernandez de Mattos, S., van der Horst, A., Klompmaker, R., Roy, S., Lagree, S., Hou, Z., Thomson, J.A., Stewart, R., and Gasch, A.P. (2013). Integrated module and gene-specific regulatory inference implicates upstream signaling networks. PLoS Comput. Biol. 9, e1003252.
Seashore-Ludlow, B., Rees, M.G., Cheah, J.H., Cokol, M., Price, E.V., Coletti, M.E., Jones, V., Bodycombe, N.E., Soule, C.K., Gould, J., et al. (2015). Harnessing connectivity in a large-scale small-molecule sensitivity dataset. Cancer Discov. 5, 1210–1223.

Segal, E., Friedman, N., Koller, D., and Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. Nat. Genet. 36, 1090–1098.

Segal, E., Shapira, M., Regev, A., Pe’er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 34, 166–176.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504.

Subik, K., Lee, J.F., Baxter, L., Strzepek, T., Costello, D., Crowley, P., Xing, L., Hung, M.C., Bonfiglio, T., Hicks, D.G., and Tang, P. (2010). The expression patterns of ER, PR, HER2, CK5/6, EGFR, Ki-67 and AR by immunohistochemical analysis in breast cancer cell lines. Breast Cancer (Auckl) 4, 35–41.

Subramanian, A., Narayan, R., Consello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell 177, 1437–1452.e17.

Teschendorff, A.E., and Enver, T. (2017). Single-cell entropy for accurate estimation of differentiation potency from a cell’s transcriptome. Nat. Commun. 8, 15599.

Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plasier, C.L., Eddy, J.A., et al. (2018). The immune landscape of cancer. Immunity 48, 812–830.e14.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Nonon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. 32, 381–386.

Troester, M.A., Hoadley, K.A., Parker, J.S., and Perou, C.M. (2004). Prediction of toxicant-specific gene expression signatures after chemotherapeutic treatment of breast cell lines. Environ. Health Perspect. 112, 1607–1613.

Tyner, J.W., Tognon, C.E., Bottomly, D., Wilmot, B., Kurtz, S.E., Savage, S.L., Long, N., Schultz, A.R., Traer, E., Abel, M., et al. (2018). Functional genomic landscape of acute myeloid leukaemia. Nature 562, 526–531.

Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., and Stuart, J.M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, 237–245.

Wang, Y., Cho, D.Y., Lee, H., Fear, J., Oliver, B., and Przytycka, T.M. (2018). Reprogramming of regulatory network using expression uncovers sex-specific gene regulation in Drosophila. Nat. Commun. 9, 4061.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 26, 1572–1573.

Wu, C.H., van Riggelen, J., Yetil, A., Fan, A.C., Bachireddy, P., and Felsher, D.W. (2007). Cellular senescence is an important mechanism of tumor regression upon c-Myc inactivation. Proc. Natl. Acad. Sci. USA 104, 13028–13033.

Zheng, G.X., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., Modermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. 8, 14049.
# STAR METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| Connectivity Map (gene expression) | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742 | GSE92742_Broad_LINCS_Level5_COMPZ.MODZ_n473647x12328.gctx.gz |
| Connectivity Map (perturbation information) | https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE92742 | GSE92742_Broad_LINCS_sig_info.txt.gz |
| CTRPv2 | https://ctd2-data.nci.nih.gov/Public/Broad/CTRPv2.0_2015_ctd2_ExpandedDataset/ | CTRPv2.0_2015_ctd2_ExpandedDataset.zip |
| DepMap (gene knockout effects) | https://depmap.org/portal/ | https://ndownloader.figshare.com/files/16757666 |
| GDSC (gene expression) | https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html | GDSC1000_WebResources//Data/preprocessed/Cell_line_RMA_proc_basalExp.txt.zip |
| GDSC (drug sensitivity) | https://www.cancerrxgene.org/gdsc1000/GDSC1000_WebResources/Home.html | GDSC1000_WebResources//Data/suppData/TableS4A.xlsx |
| Beat AML (gene expression) | https://www.nature.com/articles/s41586-018-0623-z#Sec38 | https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-018-0623-z/MediaObjects/41586_2018_623_MOESM3_ESM.xlsx |
| Beat AML (Variants) | https://www.nature.com/articles/s41586-018-0623-z#Sec38 | https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-018-0623-z/MediaObjects/41586_2018_623_MOESM3_ESM.xlsx |
| Beat AML (drug sensitivity) | https://www.nature.com/articles/s41586-018-0623-z#Sec38 | https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-018-0623-z/MediaObjects/41586_2018_623_MOESM3_ESM.xlsx |
| Transcription factor - target interactions | https://genome.cshlp.org/content/29/8/1363/suppl/DC2 | https://genome.cshlp.org/content/supp/2021/03/02/gr.240663.118.DC2/Revised_Supplemental_Table_S3.csv |
| KEGG | https://www.genome.jp/kegg/pathway.html | N/A |

## Software and algorithms

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Docker image        | https://hub.docker.com/r/jupyter/datascience-notebook | jupyter/datascience-notebookr-4.1.1 |
| R 4.1.1             | https://cran.r-project.org/ | N/A |
| Bioconductor 3.10   | https://www.bioconductor.org/install/ | N/A |
| Jupyter notebook    | https://jupyter.org/ | N/A |
| BiocParallel        | https://bioconductor.org/packages/release/bioc/html/BiocParallel.html | N/A |
| GSVA                | https://bioconductor.org/packages/release/bioc/html/GSVA.html | N/A |
| ConsensusCluster Plus | https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html | N/A |
| pheatmap           | https://cran.r-project.org/web/packages/pheatmap/index.html | N/A |

(Continued on next page)
RESOURCE AVAILABILITY

Lead contact
Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Guangrong Qin (guangrong.qin@isbscience.org).

Materials availability
This study did not generate new unique reagents.

Data and code availability
- Publicly available data used in the study are listed in the key resources table.
- All original code has been deposited and is publicly available as of the date of publication. DOIs are listed in the key resources table. Example datasets used for the case studies in this study can be found at https://osf.io/34xnm/?view_only=5b968aeb0ee14d4c97f9d7ce4c5b070.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Selection of functional modules
A general criteria of module selection is the relevance of functionality for each module, wide coverage of functionalities across different modules, and the wide coverage of genes across modules. For the use cases presented here, we consider KEGG processes or pathways as functional modules (Ogata et al., 1999; Kanehisa, 2019; Kanehisa et al., 2021; Kanehisa and Goto, 2000). The KEGG pathways were annotated in a hierarchical structure, with four main categories: ‘Metabolism’, ‘Genetic Information Processing’, ‘Environmental Information Processing’, and ‘Cellular Processes’. These categories include the functional gene sets that cover 20 cellular processes: ‘Replication and repair’, ‘Transcription’, ‘Translation’, ‘Folding, sorting and degradation’, ‘Cellular community’, ‘Cell growth and death’, ‘Transport and catabolism’, ‘Cell motility’, ‘Membrane transport’, ‘Signaling transduction’, ‘Signaling molecules and interaction’, ‘Amino acid metabolism’, ‘Metabolism of other amino acids’, ‘Lipid metabolism’, ‘Carbohydrate metabolism’, ‘Metabolism of cofactors and vitamins’, ‘Xenobiotics biodegradation and metabolism’, ‘Glycan biosynthesis and metabolism’, ‘Energy metabolism’, and ‘Nucleotide metabolism’. Each of the 20 cellular processes includes several different pathways, such as the process of ‘cell growth and death’, which includes the pathways ‘Cell cycle’, ‘Apoptosis’, ‘Cellular senescence’, and ‘p53 signaling’. The 23 functional modules in this manuscript include the 4 pathways in the process of cell growth and death and the other 19 cellular processes out of all the 20 processes shown above. The selected function modules cover diverse cellular activities, cover a large range of genes (over 5,000), and they show mutually exclusivity across modules (Figures S1A and S1B). The selected modules with diverse functionalities present distinct aspects of response, represented by orthogonal dimensions. Different sets of modules can be used depending on the researcher’s question.

Definition of the FM-factors
To define the FM-factors for each sample, we used four categories of factors: ssGSEA_score, TF_strength, up_strength, and down_strength. ssGSEA_score is measured using a modified script from the python package of gseapy 0.9.16 that embedded the GSVA R package (Hanzelmann et al., 2013). The ssGSEA_score measures the gene set enrichment score per sample as the normalized
difference between the empirical cumulative distribution function of gene expression ranks inside and outside the gene set. The \( TF_{\text{strength}} \) is defined as the average expression of transcription factors whose target genes are enriched in each functional module. It is determined by following these steps:

1. Transcription factor - target gene pairs (TF-Target) that show evidence in either literature-curated resources, ChIP-seq peaks or TF binding motifs on promoters were collected (Garcia-Alonso et al., 2019).
2. TF-target pairs with either curated/high confidence (confidence level A) or likely confidence were selected for further analysis (confidence level B) (Garcia-Alonso et al., 2019).
3. As the states of transcriptional regulatory networks exhibit tissue type or cell type specificity, we further measured the correlation between the transcription factors and their target genes using the input gene expression matrix (e.g., gene expression of MCF7 after drug treatment in the CMap data), and kept only those pairs that showed significant correlation in a given context, such as tumor type (absolute Pearson correlation coefficient >0.2, \( P \text{ value} <0.05 \)).
4. With the list of TF-target gene pairs, we used the one-tailed Fisher exact test to assess whether the target genes for one TF are enriched in a pathway, and only those TFs that showed significant enrichment in a module were selected as the signature TFs for that module.
5. We then measured the TF regulation weight using the following equations (Equations 1–3) for a specific TF, labeled TF_A:

\[
\text{RatioInPath} = \frac{N_{\text{TF_A target genes in one module}}}{N_{\text{TFs targeted genes in the module}}} \tag{Equation 1}
\]

\[
\text{RatioInTFs} = \frac{N_{\text{TF_A target genes in one module}}}{N_{\text{TF_A targeted genes}}} \tag{Equation 2}
\]

\[
\text{weight} = \text{RatioInPath} \times \text{RatioInTFs} \tag{Equation 3}
\]

where \( N_{\text{TF_A target genes in a module}} \) represents the number of target genes for TF_A in one module, \( N_{\text{TFs targeted genes in the module}} \) represent the number of genes that have been regulated by all the TFs, and \( N_{\text{TF_A targeted genes}} \) represents the number of target genes regulated by TF_A for all selected modules.

6. We normalized the weights in each pathway to make the sum of TF regulation weights equal to 1. We then calculated the average transcriptional strength by summing up the normalized weights, multiplied by the expression level for the master transcription factors (Equation 4), where \( M \) represents the number of TFs that are estimated to significantly regulate the pathway, and \( \text{Expr}(T_{Fi}) \) is the gene expression level for the Transcriptional factor \( i \) which regulates genes in the module.

\[
TF_{\text{strength}} = \frac{1}{M} \sum_{i=1}^{M} \text{weight}_i \times \text{Expr}(T_{Fi}) \tag{Equation 4}
\]

\( Up_{\text{strength}} \) (Up-regulation Strength) and \( Down_{\text{strength}} \) (Down-regulation Strength) are used for estimating the ratio of genes showing high or low expression (\( |z\text{-score}| > \text{threshold} \)) in one functional module for one specific sample compared to all other samples (Equations 5 and 6).

With the absolute gene expression matrix, z-score normalization was performed across all samples for each gene. After the z-score normalization, we define the \( Up_{\text{strength}} \) and \( Down_{\text{strength}} \) as follows (Equations 5 and 6):

\( Up_{\text{strength}} \) for each module is defined as the proportion of up-regulated genes in each module that have z-score above the upper threshold (threshold = 1.6 for our case studies):

\[
Up_{\text{strength}} = \frac{N_{\text{up-regulated genes in one module}}}{N_{\text{all genes in one module}}} \tag{Equation 5}
\]

where \( N_{\text{up-regulated genes in one module}} \) is the number of up-regulated genes in a module, and \( N_{\text{all genes in one module}} \) is the number of genes in that module.

\( Down_{\text{strength}} \) is defined as the negative fraction of down-regulated genes in a module, with a z-score below the lower threshold:

\[
Down_{\text{strength}} = -1 \times \frac{N_{\text{down-regulated genes in one module}}}{N_{\text{all genes in one module}}} \tag{Equation 6}
\]

Where \( N_{\text{down-regulated genes in one module}} \) is the number of down-regulated genes in a module.
**Annotation of transcriptional states**

**Consensus clustering.** We applied the R package of ConsensusClusterPlus (Wilkerson and Hayes, 2010) which uses the consensus clustering method (Liao et al., 2003) to classify the FM-matrix into different clusters, namely FM-states. We select the number of states (clusters) by inspection of the heatmap of the consensus matrix using different ‘K’s (K is the number of clusters), the empirical cumulative distribution function (CDF) corresponding to the entries of the consensus matrix, and the relative change in the area under the CDF with the increase of K. When $K_{true}$ is reached, further increase in the number of clusters does not lead to a corresponding marked increase in the CDF area (Liao et al., 2003).

**Annotation using the FM-factors.** The functional module factors can be easily translated into meaningful biological annotations for each state. To have a better understanding of the functional activity in each state, we detected the key features for each state. Using the Wilcoxon rank-sum test, we tested whether one FM-factor in one state is different from all of the other states. The difference of FM-factors between one state and all the other states are measured using Effect size (Equation 7).

$$\text{Effect size} = \frac{\text{Mean}(FMF_i) - \text{Mean}(FMF_j)}{\sqrt{\text{sd}(FMF_i) + \text{sd}(FMF_j)}} \left(\frac{N_i + N_j - 2}{N_i N_j}\right)$$

(Equation 7)

Where $FMF_i$ is the vector of functional module factors for state i, $FMF_j$ is the vector of functional module factors for all the other samples, $N_i$ is the number of samples in states i, and $N_j$ is the number of samples in all the other states.

We selected the FM-factor for each state showing significant difference ($P < 0.001$ for MCF7 data) and with an effect size greater than a threshold (|effect size| > 1 for MCF7 data) between this state and all the other states. The factors $ssGSEA_{score}$, $up\_strength$, and $down\_strength$ can be used for the annotation of high or low expression of each functional module for each state. $TF\_strength$ can be used for the annotation of the overall transcriptional regulation strength for each functional module for each state.

**Annotation using transcriptional factors.** To annotate which transcription factors contribute to the regulation of the states, we first identified which transcription factors show positive impact or negative impact for each functional module by 1) selecting the transcription factor – target gene pairs with positive correlation or negative correlation; 2) selecting transcription factors that show over-representation of their target genes in functional modules; 3) identifying transcription factors that show significant (adjusted $P$-value < 0.05) differences among different states using the one-way ANOVA test followed by Bonferroni multiple-test correction. Effect size was estimated as the difference of expression of the transcription factor between samples in one state and other states. The generated TF-functional module regulation network was then visualized using Cytoscape (Shannon et al., 2003).

**Defining functional module states for MCF7 cells following drug treatment**

FM-factors were generated that cover the four categories of factors for the twenty-three functional modules as described before (see section “Definition of the FM-factors”).

**Get the relative FM-factors for drug treated samples compared to reference samples.** MCF7 transcriptional profiles from the CMap (Subramanian et al., 2017) level 5 data following drug treatment which has reported drug response measurement in the CTRPv2 (Rees et al., 2016) were selected, resulting in 1287 drug treated MCF7 profiles. Transcriptional profiles for the DMSO or H2O treated MCF7 samples were selected as reference samples, resulting in 1400 reference samples. Functional module factors (FM-factors) were calculated for both drug treated MCF7 samples and reference MCF7 samples. The FM-factors for drug treated samples are then normalized using the distribution of reference sample FM-factors, which use the ranking method for the Up-strength, Down-strength and z-score method for $ssGSEA$ and $TF\_strength$. Specifically, for the Up-strength, each FM-factor for a sample from the experimental group was added to the vector of the same FM-factor from all samples in the reference dataset. Then the new FM-factor vector is ordered in an ascending order, and the ranking percentile for the FM-factor in the experiment group is considered as the normalized score for the Up-strength. For the Down-strength, the normalization procedure is similar to the up-strength except for the descending order of the FM-factor vector. For the $ssGSEA$-score and TF-strength, we normalized it using a z-score approach, using the drug treated FM-Factor, the reference set mean was subtracted from the FM-factor and divided by the reference set standard deviation.

**Clustering of the functional module-based factors.** Consensus clustering method (Monti et al., 2003; Wilkerson and Hayes, 2010) was used to cluster the FM-factor matrix into stable clusters by inspection of the CDFs’ shape and progression as the number of clusters, K, increases. The most stable classification is considered as the K-solution with the smallest proportion of ambiguous clustering.

**Annotation.** Annotate cell states using the functional module factors. We tested whether a given FM-factor in one state is different from other states using the Wilcoxon rank-sum test. Effect size of the difference between one state and the other states was measured. We selected the FM-factor for each state which showed a significant difference ($P$-value < 0.01) and with effect size greater than a threshold (|Effect size| >1) between this state and the other states for the interpretation of each state.

**Select the most discriminative transcriptional factors.** Context specific pairs of transcription factors and target genes were defined by correlation analysis (positive correlation ($p$ value < 0.05, $cor >0.2$), negative correlation ($P$-value < 0.05, $cor < -0.2$)). Transcription factors that show an over-representation of their positively correlated or negatively correlated target genes in each functional module were then selected for the annotation of each module. One way ANOVA test followed by Bonferroni multiple-test correction was then performed to identify whether the transcription factors show a significant (adjusted $P$-value<0.05) difference among different states. We used Cytoscape (Shannon et al., 2003) to visualize the network of transcription factors and functional modules that show differences among different states. The effect size measures the extent to which the TF is high- or low-expressed in each state.
Annotate cell states with external factors. Drug responses (EC50) of MCF7 for different drugs were extracted from the Cancer Therapeutics Response Portal v2 (CTRPv2) (Seashore-Ludlow et al., 2015; Rees et al., 2016; Aksoy et al., 2017). We considered drug concentrations greater than its EC50 values as high concentrations, and drug concentrations smaller than its EC50 values as low concentrations. We measured the ratio of high drug concentration samples to low drug concentration samples and statistically tested whether the distribution of high drug concentration samples and low concentration samples differs among different states using the Chi-squared test. The target or action of compounds (drugs) were also from the CTRPv2 (Seashore-Ludlow et al., 2015; Rees et al., 2016; Aksoy et al., 2017). Pie plots were used to visualize the distribution of samples with different targets or activity for each state. Fisher’s exact test was used to assess whether one class of drug target is overrepresented in a specific state.

Use of FM states to predict targetable cancer vulnerabilities

Association of gene knockdown to different states
Gene expression profiles of MCF7 after shRNA knockdown are taken from CMap (Subramanian et al., 2017), with each sample treated with one shRNA seed mapping to a specific gene. FM-factors (ssGSEA_score, TF_strength, up_strength and down_strength) for all the 23 modules were generated for each sample using the FM-States method. K-nearest neighbors (KNN) classifier (KNeighborsClassifier in the sklearn library) (Pedregosa et al., 2011) was used to assign each sample to one of the defined states after drug treatment based on the FM-factors for each sample. FM-factors (features) which show significant differences (Wilcoxon rank-sum test, \( P \)-value < 0.01 and \(|\text{effect size}| > 1\)) across five states were selected for the KNN models. To select a proper \( K \) for the prediction model, we compared the prediction accuracy using five-fold cross validation with \( K \) from 1 to 30, increased by 2. \( K = 5 \) was selected as the model shows relatively higher prediction accuracy (mean = 0.73) and smaller derivation (standard deviation = 0.05). We then used the KNN model to assign the shRNA treated sample to one of the defined states after drug treatment. Samples predicted as each state with a probability equal to or greater than 0.6 were selected for further analysis and the rest samples with ambiguous prediction results were excluded. As one gene is targeted by multiple shRNAs, we further assign one gene to one state by estimating the enrichment of shRNAs for this gene in one specific state using Fisher’s exact test. Genes with shRNAs enriched (Fisher’s exact test, \( p < 0.05 \)) in one state are assigned to this state.

Cancer dependent analysis
Gene knockout effect was downloaded from the DepMap portal (version 2019 Q3) (Meyers et al., 2017; Dempster et al., 2019, DepMap, 2019), which contains the results of genome-scale CRISPR knockout screens for 18,333 genes in 625 cell lines. Gene knockout effect data for MCF7 was selected from this dataset. Kolmogorov-Smirnov statistical test was performed to compare the distribution of target genes that drive each state and all genes in the genome-scale CRISPR knockout screen.

Selection of potential target genes
For each gene, we calculated the proportion of seeds which are predicted in one state, and we also used the gene knockout effect for the MCF7 cell line. Genes which are predicted to drive a specific state with a gene knockout effect smaller than 0 (median gene effect values for the non-essential genes) are selected as target genes that could drive to this state.

Define basal FM-states of breast cancer cell lines to predict drug sensitivity
We collected gene expression data for all the 49 breast cancer derived cell lines from the GDSC project (Iorio et al., 2016), and applied the FM-States method to this dataset to generate the functional module based factors (FM-factors) for all the 49 breast cancer cell lines. The functional module factors include ssGSEA_score, TF_strength, up_strength, and down_strength for all the 23 pathways were calculated.

Correlation analysis between drug response and FM-factors
Drug response data (log(EC50)) was derived from the GDSC datasets (Iorio et al., 2016). As not all of the drugs show efficiency to the breast cancer cell lines, we select potential effective drugs with the filtering criteria: 1) number of cell lines with log(EC50) smaller than zero is greater or equal to 5; 2) number of cell lines tested is greater or equal to 25 (more than half of the samples). Using the selected drugs, we then performed correlation analysis to compare the association between the FM-factors and drug response (log(EC50)) using Spearman correlation.

Drug sensitivity prediction models using the FM-factors and gene expression for breast cancer cell lines
For each drug, we binarized the drug response for each sample to either sensitive (log(EC50) < -1) or resistant (log(EC50) > -1) response and built a Random Forest (RF) classifier using the python library sklearn (Pedregosa et al., 2011) to predict the drug response using either the feature of FM-factors or gene expression for all the genes. 100 bootstrapped RF models were built, with 80% of the data used for training, 20% for testing. Feature importance scores greater than zero are selected for the final models for each drug.

Define FM-states for acute myeloid leukemia patients
Gene expression profiles (log-transformed RPKM) from the patient derived samples in Beat AML study (Tyner et al., 2018) were used for generating FM-factors. The same functional module factors include ssGSEA_score, TF_strength, up_strength, down_strength for all 23 pathways were calculated. Consensus clustering was performed to classify the AML samples based on the FM-factors matrix. Similar to the previous analysis, consensus clustering method (Monti et al., 2003; Wilkerson and Hayes, 2010) was used to cluster the FM-factor
matrix into clusters by inspection of the CDFs’ shape and progression as the number of clusters, K, increases. The most stable classification is considered as the K-solution with the smallest proportion of ambiguous clustering.

**Association between the FM-states and mutations in AML**

Genetic variants for the AML patients from the Beat AML study were used to analyze the association between the FM-states and the genetic mutations. Fisher exact test was used to estimate the enrichment of one mutated gene in one state. Only samples with RNA-Seq data and genetic data were considered in this analysis. Genes which are highly frequently mutated are considered in this analysis, that including FLT3, NPM1, DNMT3A, IDH2, RUNX1, IDH1, ASXL1, STAG2, TET2, TP53, CEBPA, WT1, PTPN11, NRAS and KRAS.

**Association between the FM-states and blast count percentage in AML**

Blast count percentages for the AML patients were also downloaded from the Beat AML study (Tyner et al., 2018). A preprocessing procedure is used to set blast percentage greater than 95% to 95%, greater than 50% to 50%, and less than 5% to 5%. One-way ANOVA was performed to test whether the blast count percentage in different states have the same population mean. Violin plot was then used to visualize the distribution of blast count percentage for each FM-states.

**Correlation analysis between FM-factors and drug sensitivity for AML patients**

To address the association between FM-factors and drug sensitivity, we used a FDA-approved drug venetoclax as an example, and performed Spearman correlation between the FM-factors and drug response data AUC values which were derived from the Beat AML datasets (Tyner et al., 2018).

**Drug sensitivity prediction models for AML using the FM-factors**

Drug sensitivity prediction models for venetoclax were built using the FM-factors and gene expression. We assigned the AML sample to two groups, 1) Venetoclax sensitive group with IC50 smaller than 0.1 uM, and 2) Venetoclax resistant group with IC50 > 0.1 uM. We then built a Random Forest (RF) classifier using the python library sklearn (Pedregosa et al., 2011) to predict the venetoclax response using either the feature of FM-factors or gene expression for all the genes. 100 bootstrapped RF models were built, with 80% of the data used for training, 20% for testing.

**QUANTIFICATION AND STATISTICAL ANALYSIS**

Pearson correlation analysis between the transcription factors and target genes was performed using the pearsonr function in the scipy.stats python library. A threshold of absolute Pearson correlation coefficient greater than 0.2 and P-value less than 0.05 was used to select significantly correlated transcription factors and target gene pairs. Fisher exact test (scipy.stats.fisher_exact, alternative hypothesis was set to ‘greater’) was used to assess whether the target genes for one TF are enriched in a module. To compare the FM-factors between samples with (sample size = 1287) or without drug treatment (sample size = 1400), we performed a Mann-Whitney rank test using the scipy.stats python package, followed by Benjamini-Hochberg multiple test correction using the multitests.multipletests function in the statsmodels python package. Effect sizes were measured as described in the above method details. The threshold of FDR smaller than 1e-6 and effect size greater than 0.2 or smaller than −0.2 were used to select the FM-factors associated with drug treatment. The Wilcoxon rank-sum test (scipy.stats.ranksums) was used to compare the FM-factors between different clusters (states). Transcription factors with adjusted P-value smaller than 0.05 were used for further analysis and visualization. The Chi-squared test was used to test whether the distribution of high drug concentration samples and low concentration samples differed among different states. Fisher’s exact test was used to assess whether one category of drug target is overrepresented in a specific state. To assign a potential target gene to different states, we used Fisher’s exact test to estimate the enrichment of shRNAs for this gene in one specific state (group) by comparing the rest of states (groups). The Kolmogorov-Smirnov statistical test (KS test) was performed to compare the distribution of target genes that drive each state and all genes in the genome-scale CRISPR knockout screen. Fisher’s exact test was used to test the significance of enriched mutated genes among different FM-states clusters for the AML analysis. One-way ANOVA was performed to test whether the blast count percentage in different states have the same population mean. Except for the thresholds explicitly described, we used a P-value smaller than 0.05 as significance level.