

© Copyright 2018

Safiye Celik

High-dimensional machine learning for drug target discovery and precision medicine

Safiye Celik

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Su-In Lee, Chair

Paul K. Crane

R. David Hawkins

Ali Shojaie

Program Authorized to Offer Degree:

Paul G. Allen School of Computer Science and Engineering

University of Washington

Abstract

High-dimensional machine learning for drug target discovery and precision medicine

Safiye Celik

Chair of the Supervisory Committee:
Assoc. Prof. Su-In Lee
Paul G. Allen School of Computer Science and Engineering

Identifying mechanisms of complex diseases such as Alzheimer's disease or cancer is essential to discover molecular targets for therapeutic intervention, ideally in a personalized manner. Gene expression measured in a tissue sample is an informative proxy to the protein levels in that tissue and known to drive disease outcomes. Machine learning offers promising computational and analytical solutions for extracting robust and relevant information from gene expression data. However, there are three main challenges with the application of a standard machine learning approach to a gene expression dataset. First, for most expression datasets, the number of variables is much greater than the number of samples, which makes models highly likely to overfit the training data (*high-dimensionality*). Second, there are technical and/or experimental confounders in any one study that make the features learned from an individual

expression dataset not necessarily generalizable to other datasets (*study-specific confounders*).

Finally, any complex disease mechanism originates from an interplay among different molecular elements and failing to incorporate prior information regarding a number of these elements is likely to result in an incomplete understanding of the underlying complex biology (*different molecular elements*). In this dissertation, we address these challenges by introducing four novel machine learning techniques, namely MGL, INSPIRE, MERGE, and EMBARKER. In fact, each of these techniques leverages the prior information readily available on *different molecular elements* to overcome the overfitting problem caused by the *high-dimensionality* and *study-specific confounders*. Our extensive statistical and biological evaluations demonstrate the superiority of each of the proposed methods over the alternative methods and reveal important disease mechanisms as well as potential therapeutic targets for disease intervention.

TABLE OF CONTENTS

Chapter 1. Introduction.....	16
Chapter 2. Unsupervised Learning of Module Networks From Gene Expression Data	20
2.1 Related Work	21
2.2 Computational Model of MGL	23
2.2.1 Preliminaries on Gaussian Graphical Models.....	24
2.2.2 MGL Formulation.....	25
2.2.3 Properties of MGL	26
2.2.4 MGL Learning Algorithm.....	27
2.3 Results from Application of MGL to Ovarian Cancer Data	29
2.3.1 Cross-validation test log-likelihoods	30
2.3.2 Functional enrichment of modules.....	31
2.3.3 Survival prediction using latent variables as features.....	32
2.3.4 Interesting findings	33
2.4 Discussion	34
Chapter 3. Unsupervised Learning of A Shared Module Network from Multiple Expression Datasets	36
3.1 Related Work	41
3.2 INSPIRE Learning Algorithm	45
3.3 Experimental Results	50
3.3.1 Overview of the INSPIRE framework.....	50

3.3.2	INSPIRE learns underlying modules and their dependencies from simulated data more accurately than 13 other methods	51
3.3.3	Evaluation on two genome-wide ovarian cancer expression datasets	59
3.3.4	INSPIRE learns a statistically more robust low-dimensional representation model than alternative approaches	61
3.3.5	INSPIRE modules are more significantly enriched for functional gene sets than alternative methods	62
3.3.6	INSPIRE modules are more significantly enriched for putative targets of the same TF than the alternative approaches	65
3.3.7	Application to nine genome-wide ovarian cancer expression datasets.....	67
3.3.8	Negatively correlated modules show distinct pathways and potential regulatory TFs enrichment.....	69
3.3.9	INSPIRE latent variables are significantly associated with clinical and histologic phenotypes in cancer	70
3.3.10	INSPIRE latent variables are more predictive of clinical and histologic phenotypes in cancer than other kinds of LDRs and all genes.....	74
3.3.11	Subtypes inferred based on INSPIRE latent variables are highly relevant to genomic abnormalities in ovarian cancer	76
3.3.12	Subtypes revealed by INSPIRE and their relationships with the TCGA subtypes	78
3.3.13	INSPIRE provides novel insights into molecular basis for ovarian tumor resectability	81
3.3.14	INSPIRE modules and the conditional dependence network among them	83
3.3.15	INSPIRE reveals a previously unknown stroma-associated marker HOPX.....	85

3.3.16	HOPX is a putative driver for the tumor-associated stroma/mesenchymal module (module 5)	88
3.3.17	HOPX is a molecular marker of aggressive tumor stroma	90
3.4	Discussion	93
Chapter 4. Supervised Learning of cancer Expression Markers Using Multi-Omics Prior Information		97
4.1	Related Work	100
4.2	MERGE Model	102
4.3	Application of MERGE to Identify AML Drug Response Markers.....	105
4.3.1	Expression hubness significantly determines the MERGE scores	107
4.3.2	Consistency of gene-drug associations in left-out data.....	109
4.3.3	Drugs in the same classes exhibited similar response pattern	114
4.3.4	Specificity of gene-drug associations to drug classes.....	116
4.3.5	Drug sensitivity prediction performance	117
4.3.6	MERGE identifies the roles of several drug response markers	120
4.3.7	Experimental validation of the marker role of SMARCA4.....	122
4.4	Discussion	126
Chapter 5. Supervised Learning of Expression Markers by a Pathway-guided Hierarchical Bayesian approach		129
5.1	Related work	132
5.2	EMBARKER Model	136
5.3	Experimental Results	140

5.3.1	Overview of the EMBARKER framework.....	140
5.3.2	EMBARKER’s way to incorporate pathway information leads to improved robustness and relevance of the identified gene and pathway markers.	141
5.3.3	EMBARKER incorporates the prior information based on the amount of relevant information available in the data.....	147
5.3.4	EMBARKER’s natural extension to a multi-dataset setting is applied to nine brain regions to learn expression markers for AD neuropathology.	148
5.3.5	EMBARKER-M learns more robust markers than the alternative methods to combine multiple regions.	150
5.3.6	Markers from EMBARKER-M are highly enriched for AD-related pathways.....	152
5.3.7	In vivo validation of EMBARKER-M identifies mitochondrial Complex I as a potential AD therapeutic target.	154
5.4	Discussion	156
Chapter 6. Conclusions.....		157
Bibliography		159

LIST OF FIGURES

Figure 1. GGM vs. MGL representation.....	21
Figure 2. Performance of GGM vs. MGL in capturing the underlying modular data structure on synthetic data.....	24
Figure 3. Performance of Toh method (k-means followed by graphical lasso) vs. MGL.	32
Figure 4. Comparison among MGL latent variables, k- means cluster centroids, and individual genes based on survival prediction performance.	33
Figure 5. A small portion of the pathway structure identified by MGL on Tothill dataset.	34
Figure 6. Overview of the INSPIRE framework.	38
Figure 7. Overview of the evaluation and application of INSPIRE procedure.	40
Figure 8. Test log-likelihood of INSPIRE vs. alternative models on the synthetic data.	56
Figure 9. Module recovery performance of INSPIRE vs. alternative methods on the synthetic data.	57
Figure 10. Feature dependency recovery of INSPIRE vs. alternative models on the synthetic data.	59
Figure 11. Illustration of the two OV datasets used for evaluating INSPIRE.	60
Figure 12. Negative test log-likelihood of INSPIRE vs. alternative models on real data.	62
Figure 13. Functional enrichment of the modules identified by INSPIRE vs. alternative methods.	64
Figure 14. ChEA enrichment of the modules identified by INSPIRE vs. alternative methods....	66
Figure 15. Phenotype-based characterization of the INSPIRE modules.	72

Figure 16. Prediction performance of INSPIRE vs. four alternative methods for six prediction tasks in a cross-validation setting.....	75
Figure 17. Association of INSPIRE subtypes with the CNV levels.	78
Figure 18. The modules that differentiate the subtypes based on INSPIRE features and the interactions among these modules as learned by INSPIRE.	79
Figure 19. Performance of different models for resectability prediction.....	82
Figure 20. Pan-cancer survival association of modules 5 and 6 genes.	87
Figure 21. Potential targets of HOPX.	90
Figure 22. Fluorescent staining of ovarian tumors from sub-optimally bulked and optimally debulked patients.....	91
Figure 23. Histopathology of a patient with low vs. high HOPX expression.....	92
Figure 24. Conventional statistical methods vs. MERGE framework.....	99
Figure 25. Contributions of different drivers feature to the MERGE score.	108
Figure 26. Comparison of MERGE with four alternative methods in terms of the percentage of the significant associations replicated in the left-out test data.....	111
Figure 27. Drug class specificity of patients' drug response and genes' MERGE scores.....	115
Figure 28. Prediction performance of MERGE compared to alternative methods.....	118
Figure 29. Drug class specificity and drug response association heat maps of the top 44 genes selected by our MERGE framework.	121
Figure 30. SMARCA4 plasmid transfection experiments on two cell lines for comparison of response to etoposide and mitoxantrone between original vs. transfected cells.	124
Figure 31. The EMBARKER methodology.....	131

Figure 32. Statistical consistency of the gene markers learned by EMBARKER vs. alternative methods.	144
Figure 33. Statistical consistency of the pathway markers learned by EMBARKER vs. alternative methods.	146
Figure 34. Statistical consistency of the gene markers learned by EMBARKER-M vs. alternative meta-analysis methods.	152
Figure 35. In-vivo experimental validation of NDUFA9 and the other Complex-I genes.	155

LIST OF TABLES

Table 1. Methods we compared with our INSPIRE framework.....	53
Table 2. FDR-corrected Fisher’s exact test (FET) p-values for the top pathways for which the identified markers from EMBARKER-M were enriched.	153

ACKNOWLEDGEMENTS

First and foremost, I am truly thankful to my advisor Su-In Lee for all her help throughout my PhD journey. She has been an exceptional advisor putting a lot of effort to be available whenever I needed advice, feedback, or just brainstorming. I remember several late nights of working together on our paper in her office. She patiently taught me how to conduct research, always motivated me to be productive, guided me to the right direction at those times when I had little or no idea of how to move forward.

I wish to express my sincere gratitude to my quals and defense committee members, Paul Crane, David Hawkins, Ali Shojaie, Tara Madhyastha, Meliha Yetisgen, and Larry Ruzzo for helping me improve my dissertation with their constructive feedbacks. Paul, I am especially thankful to you for reading both my general exam report and dissertation line by line, noting even simple grammar mistakes, and helping me become not only a better researcher but also a better writer in my second language.

I am extremely grateful to the American Association of University Women (AAUW) for funding me in my second year of PhD and all the amazing work they do to empower women including international students like me who have fewer funding opportunities.

I am thankful to the staff of Paul G. Allen School of Computer Science and Engineering for being accommodating and helping me focus on my research. I especially thank our former graduate advisor Lindsay Michimoto and our current graduate advisor Elise DeGoede Dorough, for writing me so many letters for my visa applications to attend international conferences, sometimes at a very short notice.

I am grateful to my terrific collaborators, especially Pamela Becker, Matt Kaeberlein, Josh Russell, and Dirk Keene, from whom I learned tremendous amount of biology and medicine.

I will truly miss Lee Lab members who been wonderful peers. My Friday afternoons will be incomplete without the lab meetings full of stimulating discussions and free lunch.

I feel so blessed to have a second family in Seattle: Zubeyde Sisman, Havva Gurbuz, Ezgi Mercan, Betul Gokce, Merve Merakli, and Busra Celikkaya. Seattle was a city full of sunshine thanks to you. Zubeyde, your warm hugs, the delicious meals you offered any time I arrived at your place, and your adorable kids made grad school much more enjoyable. Havva, 8 pm was never late to make a dinner plan with you after a crazy workday. Ezgi, my dear colleague and friend, I so much appreciate all those coffee breaks full of inspiring conversations. I also would like to thank my first roommates Semra Oner, Sabina Tair, Halida Lomanova, and Anu Battsengel, and past colleagues Bilge Esra Soran and Kivanc Muslu for helping me adjust to a new culture when I was a newbie to this country. I am thankful to my old but gold friends Sinem Aykanat, Hilal Guneren, and Canan Ersan who never gave up asking me when I was finishing PhD from 6,000 miles away.

I am so grateful to my dear friend Betul Durmaz. Although it was just for a year, I feel so blessed to have you as my roommate. Thank you for always being there when I needed you, for giving me rides and bringing me bites during my long nights working in the office.

I cannot thank enough to my (in-two-days-to-be) husband Ahmet Erdem Yavuz, for entering my life when I needed him the most, for making a long-distance relation as easy as possible, for never complaining about my lack of time for a video chat, for his therapeutic voice over the phone at times when life and grad school were so overwhelming.

I am so grateful for having two elder sisters. Sumeyye Akin and Seyma Tas have been great examples of educated women as well as my mentors and tutors throughout my middle and high school years. I am thankful to my mentally challenged younger sister, Zeynep Celik, although unconsciously, for always being pure and innocent, and reminding me to be grateful for intelligence.

The last but not the least, I would like to express my deepest gratitude to my parents Fatma Celik and Mehmet Tahir Celik, for being so open-minded in a Middle Eastern country, for believing in the value of higher education for anyone and everyone, and for always respecting my freedom. I would especially thank my mother for always encouraging me to do a PhD while she did not have the opportunity to study after elementary school due to the huge gender gap in education when she was a young girl. I feel so fortunate to be a part of such a wonderful family.

Chapter 1. INTRODUCTION

Cell is the building block of all human tissues, organs, and systems. It contains the genetic material named *deoxyribonucleic acid* (DNA), a double helix structure composed of around 3 billion pairs of four types of chemical bases: adenine (A), thymine (T), guanine (G) and cytosine (C). *Genes*, specific portions of the DNA sequence, encode a *protein* through two main stages of information flow: (1) *transcription*, which is the production of messenger RNA (mRNA) from DNA, and (2) *translation*, which is the production of a peptide for a protein based on the mRNA. There are around 20,000 human genes that encode proteins. Genes are known to regulate each other; that is, they increase or decrease each other's expression levels in different stages of the information flow from DNA to protein. Most biological processes occur because of the regulatory interactions among a specific group of genes (i.e., a pathway) that modify the protein levels.

While the DNA sequence is the same in different cells in a person's body, differences in gene activity levels lead to difference in cell functionalities. Proteins dictate nearly all cell functions; therefore, examining the abundance and type of proteins a tissue manufactures is crucial for identifying mechanisms occurring in that tissue. This examination is specifically important for understanding the disease mechanisms based on cells retrieved from a diseased tissue. Due to the complex structure of the proteins, mRNA levels (i.e., gene expression), which is relatively easier to measure than protein levels, are generally used as proxy for protein levels. Gene expression is a highly informative tool to learn about the how much functional product is produced (i.e., how much protein is encoded) by the gene. Gene expression in a diseased tissue usually differ from the one in its counterpart normal tissue, and therefore, gene expression is commonly used to identify

disease mechanisms. In this thesis, we are following this approach and we focus on gene expression data to better understand the processes in the cell and make new discoveries.

Machine learning methods are commonly used to extract information (i.e., informative *features*) from gene expression data. We examine these methods in two broad classes: (1) *supervised* methods that extract a set of features that is discriminative of different class labels (e.g., cancer vs. control) or varying outcome levels (e.g., survival time), and (2) *unsupervised* methods that extract a set of features purely based on the underlying structure of the gene expression data with no class labels or outcome variables. The most common application of supervised methods is to identify a gene or a set of genes that can differentiate the samples in different classes or can predict a particular phenotype. Unsupervised methods are often used to identify groups of genes that work together to perform a particular function (i.e., pathways) or regulatory interactions among the genes or pathways. While unsupervised learning methods extract a set of features *without* the knowledge of class labels or an outcome variable, learned features may subsequently be used for classification purposes.

The success of machine learning algorithms for biology and genomics heavily relies on the representation of input data. Different representations may capture different explanatory factors of variation underlying the data¹. A representation is composed of extracted *features* that can be effectively used for subsequent tasks such as classifying cancer patients into subtypes. Extracting biologically intuitive features is an important step to identify robust and relevant information from data².

There are three main challenges with applying existing feature learning approaches to gene expression data. The first challenge is *high-dimensionality*: most molecular datasets are within the

$p \gg n$ regime (high-dimensional), where p is the number of variables and n is the number of samples. In a high-dimensional setting (e.g., a gene expression dataset where we have tens of thousands of transcripts but only a few hundreds or less samples), the observed data are statistically not sufficient to retrieve reliable information about the similarity between the variables or the role of the variables in production of an outcome. This phenomenon of low statistical power in a high-dimensional variable space is known as the *curse of dimensionality*. If the goal is to learn a network structure among the variables (e.g., a gene regulatory network), the problem is further aggravated because the number of parameters (i.e., the number of all potential edges among the variables) is much larger than the number of variables (p) which is much larger than the number of samples (n). A second challenge with standard feature learning approaches is *study-specific confounders*. Any one study may not be generalizable in that there may be either technical (e.g. sample ascertainment) or experimental (e.g. batch effects) confounders that make the data from different datasets look very different than each other. Therefore, features extracted from an individual dataset in a naïve way does not necessarily capture the relevant aspects in other datasets. And finally, there are *different molecular elements* which we need to incorporate to the learning problem, and just relying on the provided samples from a single study/dataset would not provide a comprehensive understanding of the underlying biological and disease mechanisms.

In this dissertation, we introduce algorithms to learn reliable and robust features from molecular data while addressing the aforementioned challenges: *high-dimensionality*, *study-specific confounders*, and *different molecular elements*. Specifically, we incorporate information from *different molecular elements* to the problems in hand to address the *high-dimensionality* and *study-specific confounders*. The algorithms we introduce, namely MGL, INSPIRE, MERGE, and EMBARKER, assist feature extraction and marker identification tasks using biologically

reasonable assumptions or prior knowledge extracted from literature. By this, they enable inference of features that can reliably explain the mechanisms underlying the data. Two of these (MGL and INSPIRE) are unsupervised machine learning methods, and the other two are supervised. We note that all these methods are general machine learning frameworks that can be used in several disease settings beyond the limited number of the applications we included in our experiments. Where relevant, we refer to the paper that was published or in review for each work. Additional experiments as well as the additional details on the experiments contained in this dissertation can be found in the referred papers.

The rest of the dissertation is organized as follows. Chapter 2 introduces MGL, a novel module network-based feature learning method that extracts biologically relevant features from a single expression dataset. Chapter 3 provides the INSPIRE method extending this idea to multiple expression datasets with partially overlapping sets of variables. Chapter 4 introduces the MERGE algorithm, which also learns expression markers for complex disease phenotypes in a supervised manner. Chapter 5 introduces the EMBARKER method we developed to identify expression markers in a supervised manner like MERGE but using known pathways rather than multi-omic gene driver features as prior information. EMBARKER specifically focuses on Alzheimer's disease gene expression marker identification. We conclude with a summary and discussion in Chapter 6.

Chapter 2. UNSUPERVISED LEARNING OF MODULE NETWORKS FROM GENE EXPRESSION DATA

[The work presented in this chapter was published in the Proceedings of the 31st International Conference on Machine Learning.³]

In this chapter, we propose the Module Graphical Lasso (MGL) algorithm that aims to extract a set of statistically robust and informative features from a single high-dimensional dataset using the *biological pathway* assumption for the genes (i.e., genes usually function in groups of closely related genes). MGL is an unsupervised method dividing genes into modules and jointly learning module-level dependencies as well as a set of features that are the inferred latent variables (each associated with a module).

As mentioned in the previous section, genes are involved in biological processes in groups called *pathways*, and the gene expression levels of the genes in a pathway are regulated by a number of genes inside and/or outside of the pathway. In this chapter, we accommodate those two pieces of prior information to learn robust and reliable features from a high-dimensional gene expression dataset. This approach also helps us understand how genes are organized into functional modules and how these modules interact with each other, where the interaction among the modules is represented by a probabilistic graphical model.

Probabilistic graphical models provide a powerful framework to represent rich statistical dependencies among nodes. An edge in a graphical model represents a conditional dependence between the two nodes the edge connects. In a Gaussian graphical model (GGM), edges are parameterized by elements of the inverse covariance matrix (precision matrix) of the variables

(e.g., gene expression levels in an expression dataset). MGL is an aggressive dimensionality reduction and network estimation technique for a high-dimensional GGM. MGL achieves scalability, interpretability and robustness by exploiting the modularity property of many real-world networks. Variables are organized into tightly coupled modules and a graph structure is estimated to determine the conditional independencies among modules. **Figure 1** compares a standard gene-level GGM to module-level GGM learned by MGL. MGL iteratively learns the module assignment of variables, the latent variables, each corresponding to a module, and the parameters of a GGM of the latent variables. When applied to gene expression data from ovarian cancer, MGL outperforms standard clustering algorithms in identifying functionally coherent gene sets and predicting survival time of patients. The learned modules and their dependencies provide novel insights into cancer biology as well as identifying possible novel drug targets.

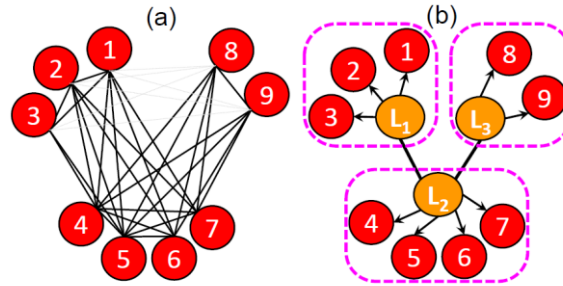


Figure 1. GGM vs. MGL representation.

(a) GGM representation of $X = \{X_1, \dots, X_9\}$. **(b)** MGL representation of X .

2.1 RELATED WORK

MGL learns a low-dimensional representation (latent variables) from high-dimensional data by jointly clustering variables into modules and learning a network among the modules through an iterative procedure. This key aspect differentiates MGL from previous approaches that can be

organized into four categories: The first category includes latent factor models, such as latent factor analysis or probabilistic PCA⁴, which do not learn the network among latent factors. Second, Toh & Horimoto (2002)⁵ cluster variables first and then learn the dependency structure among the cluster centroids, instead of jointly clustering and learning the network. This method can achieve improved scalability and interpretability; however, we showed through our extensive experiments that MGL outperforms this approach based on all of the evaluation criteria we incorporated. Third, He et al.⁶ models each latent variable as a linear combination of variables and estimates the network among k latent variables. Although this approach also learns a network of k latent variables instead of p variables, it does not explicitly cluster variables, which results in a vastly different learning algorithm from MGL. Clustering of variables is a key feature of MGL reducing the number of parameters and increasing the model's interpretability, which enables interesting analyses shown in Sec. 4.2.2. Finally, many authors attempted to incorporate latent variables into GGMs. However, they do not explicitly cluster variables into modules, and require the learning of Σ^{-1} of p variables instead of k latent variables ($k \ll p$), which drastically increase the number of parameters. Chandrasekaran et al.⁷ assume that Σ^{-1} of observed variables decomposes into a sparse matrix and a low-rank matrix, and the low-rank matrix represents the effect of unobserved latent variables. They proposed a convex optimization algorithm that utilizes both L1 and nuclear norm as penalty terms. The SIMoNe⁸ method uses an Expectation-Maximization (EM) approach⁹ for variational estimation of the latent structure while inferring the network among the entire set of variables. In contrast, MGL performs a more aggressive dimensionality reduction by learning a network of k latent variables instead of p observed variables ($k \ll p$). Guo & Wang (2010)¹⁰ proposed an algorithm consisting of three steps: 1) apply the graphical lasso to compute an adjacency matrix of the variables; 2) partition variables into disjoint clusters; and 3) estimate a

sparse Σ^{-1} with a modified penalty term such that within-cluster edges are less strongly penalized. Given the module assignment of variables, Duchi et al.¹¹ proposed to penalize the infinity-norm and Schmidt et al.¹² proposed to penalize the two-norm of the inverse covariance matrix block corresponding to each module in the network. Marlin et al.¹³ and Marlin & Murphy¹⁴ make use of these methods^{11,12}, after first identifying the groups of the variables when the modular structure is unknown.

2.2 COMPUTATIONAL MODEL OF MGL

Our approach MGL is characterized by the incorporation of latent variables (low-dimensional features) into the GGM. **Figure 1b** illustrates a toy example where three latent variables L_1 , L_2 and L_3 have mutual dependencies in addition to connections to observed variables by directed edges. Each of L_1 , L_2 and L_3 represents the aggregate activity levels of specific functional modules as defined by a core of tightly coupled genes. The undirected edges among latent variables represent the dependencies among these functional modules. As can be seen in **Figure 1**, MGL provides a more compact representation of conditional independence relationships compared to the equivalent gene-level GGM. By modeling the conditional independence relationships among k latent variables instead of p ($k \ll p$), we show that MGL scales better than standard graphical lasso (i.e. the common gene-level GGM learning method) when $p \gg n$, enabling us to efficiently learn a GGM with thousands of variables. We considered a toy example with five latent variables L and 15 observed variables X with the inverse covariance matrix of the latent variables (Σ_L^{-1}) illustrated in **Figure 2**. Given the same data consisting of 100 observations on X , MGL almost perfectly estimates (Σ_X^{-1}), whereas graphical lasso fails to reveal the latent structure among the five groups of variables (**Figure 2**).

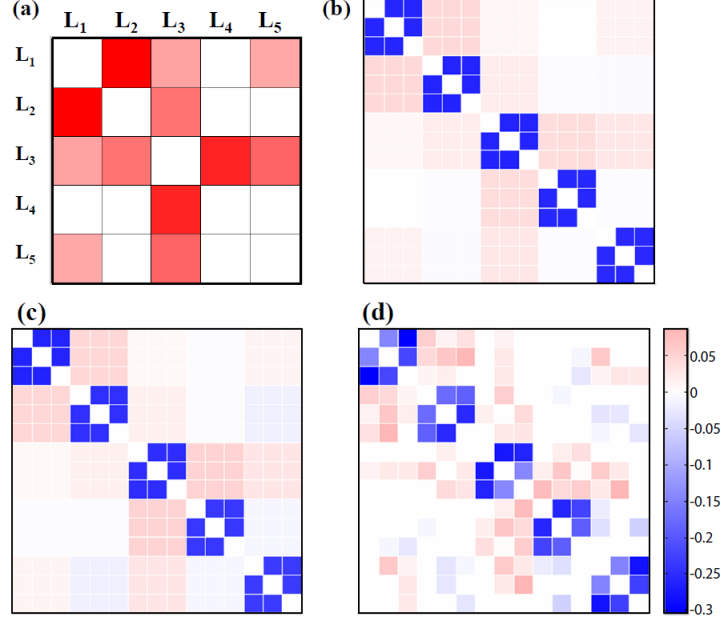


Figure 2. Performance of GGM vs. MGL in capturing the underlying modular data structure on synthetic data.

(a) Heat map of Σ_L^{-1} . White elements are zero and colored ones are nonzero; thus, colored elements correspond to the edges in the graph. **(b)** Heat map of Σ_X^{-1} . **(c)** MGL estimate of Σ_X^{-1} . **(d)** GGM (graphical lasso) estimate of Σ_X^{-1} .

2.2.1 Preliminaries on Gaussian Graphical Models

Assume that we wish to learn a GGM of p variables based on n observations $x[1], \dots, x[n] \sim N(0, \Sigma)$ where Σ is a $p \times p$ covariance matrix. It is well-known that the sparsity pattern of Σ^{-1} represents the conditional independence relationships among the variables^{15,16}. Specifically, $(\Sigma^{-1})_{jj'} = 0$ for some $j \neq j'$ if and only if X_j and $X_{j'}$ are conditionally independent given X_k where $k = \{1, \dots, p\} \setminus \{j, j'\}$. Hence, the nonzero pattern of Σ^{-1} corresponds to the graph structure of the GGM. To obtain a sparse estimate for the GGM, a number of authors^{17–19} have considered maximizing the penalized log-likelihood, a method called graphical lasso:

$$\max_{\theta > 0} \left\{ \log \det \theta - \text{tr}(S\theta) - \lambda \sum_{j \neq j'} |(\theta)_{jj'}| \right\} \quad (1)$$

where S is empirical covariance matrix, λ is a positive tuning parameter, the constraint $\theta > 0$ restricts the solution to the space of positive definite matrices of size $p \times p$, and the last term is the element-wise L_1 penalty. We denote by θ the estimate of inverse covariance matrix throughout this section. When λ is large, the resulting estimate will be sparse.

The probabilistic interpretation of the L_1 penalty term defines the optimization parameters θ as random variables rather than fixed parameters. This interpretation requires that we optimize the joint probability density $\log P(S, \theta) = \log P(S|\theta) + \log P(\theta)$. The use of the Laplacian prior $P(\theta_{jj'}) = \frac{\lambda}{2} \times \exp -\lambda |\theta_{jj'}|$ leads to the optimization problem described in Equation (1). The hyperparameter λ adjusts the sparsity of the optimization variable θ .

2.2.2 MGL Formulation

Let $L = \{L_1, \dots, L_k\}$ be a set of *latent variables*: $L \sim N(0, \Sigma_L)$ where Σ_L is a $k \times k$ covariance matrix. Let $X = \{X_1, \dots, X_p\}$ be a set of *observed variables*, each having the distribution $X_i | L_{Z_i}, \sigma^2 \sim N(L_{Z_i}, \sigma^2)$, where Z_i refers to the index of the latent variable which X_i is associated with. Here, we refer to a set of observed variables that correspond to the same latent variable as a module. As an example, the j th module M_j can be defined as $M_j = \{X_i | Z_i = j\}$ for $1 \leq j \leq k$. Thus, $Z = \{Z_1, \dots, Z_p\}$ defines the module assignment of p variables into k modules. Then, the joint probability distribution function $P(X, L, Z, \Sigma_L)$ of the MGL has the following form:

$$\begin{aligned}
P(X, L, Z, \Sigma_L) &= \prod_{i=1}^p P(X_i | L_{Z_i}) P(L | \Sigma_L) P(\Sigma_L^{-1}) P(Z) P(Z) \\
&= \prod_{i=1}^p \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(X_i - L_{Z_i})^2}{2\sigma^2} \\
&\quad \cdot \frac{1}{\sqrt{(2\pi)^k |\Sigma_L|}} \exp \left\{ -\frac{1}{2} L^T \Sigma_L^{-1} L \right\} \\
&\quad \cdot \prod_{j \neq j'} \frac{\lambda}{2} \exp \{ -\lambda |(\Sigma_L^{-1})_{jj'}| \} P(Z). \tag{2}
\end{aligned}$$

MGL can be seen as generalized k -means clustering that takes into account the Mahalanobis distances between latent variables (2nd term in Equation (2)), in addition to the distances between each variable and the corresponding latent variable (1st term in Equation (2)).

Given n observations $x[1], \dots, x[n] \in \mathbb{R}^p$ in X , MGL aims to estimate values on the latent variables L , module assignment variables Z , and the inverse covariance matrix of the latent variables Σ_L^{-1} . To estimate the inverse covariance matrix over the observed variables, Σ_X^{-1} , we can use the relationship between Σ_L^{-1} and Σ_X^{-1} , as described in *Lemma 3* below.

2.2.3 Properties of MGL

Lemma 1. The joint distribution of $X = \{X_1, \dots, X_p\}$ and $L = \{L_1, \dots, L_k\}$ is Gaussian: $(X, L) \sim N(0, \Sigma_{XL})$ where Σ_{XL} is a $(p + k) \times (p + k)$ covariance matrix.

Lemma 2. The marginal probability distribution of the observed variables $X = \{X_1, \dots, X_p\}$ is Gaussian: $X \sim N(0, \Sigma_X)$ where Σ_X is a $p \times p$ covariance matrix.

Lemma 3. Let Σ_L be a $k \times k$ covariance matrix of L . The relationship between Σ_X and Σ_L is as $\Sigma_X = \{A - C^T B^{-1} C\}^{-1}$ where A is a $p \times p$ diagonal matrix with diagonal elements $A_{ii} = 1/\sigma^2$; C is a $k \times p$ matrix whose element $C_{ij} = -(1/\sigma^2)$ if $X_j \in M_i$ and 0 otherwise; and

$$B = \Sigma_L^{-1} + \begin{bmatrix} |M_1|/\sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & |M_k|/\sigma^2 \end{bmatrix},$$

$|M_k|$ meaning the number of observed variables in the module k .

2.2.4 MGL Learning Algorithm

Here, we present our learning algorithm that optimizes the likelihood function based on the joint distribution described in Equation (2). Given $X(\in \mathbb{R}^{p \times n})$ that contains n observations $x[1], \dots, x[n] \in \mathbb{R}^p$, MGL aims to learn the following:

- $L \in \mathbb{R}^{k \times n}$ containing the values $l[1], \dots, l[n]$ on k latent variables in n samples in X
- $Z \in \{1, \dots, k\}^p$ specifying the module assignment of $X_1, \dots, X_p$ into k modules; and
- $\theta_L (\in \mathbb{R}^{k \times k})$ denoting the estimate of the inverse covariance matrix of the latent variables, i.e. Σ_L^{-1} . Using *Lemma 3*, we can obtain $\theta_X (\in \mathbb{R}^{p \times p})$, the precision matrix estimate of X once we learn θ_L .

We choose to address our learning problem by finding the joint maximum a posteriori (MAP) assignment to all of the optimization variables – L , Z , and θ_L . This means that we optimize the following objective function with respect to L , Z , and $\theta_L (> 0)$:

$$\log P(X, L, Z, \theta_L; \lambda, \sigma) = \log P(\theta_L) + \log P(L|\theta_L) + \log P(X|L, Z) + \log P(Z) \quad (3)$$

$$= \frac{n}{2} \{\log \det \theta_L - \text{trace}(S_L \theta_L)\} - \lambda \sum_{j \neq j'} |(\theta_L)_{jj'}| - \frac{1}{2} \sum_{i=1}^p \frac{\|X_i - L_{Z_i}\|_2^2}{\sigma^2} + \text{const},$$

where $S_L = \frac{1}{n} LL^T$ is the empirical estimate of the covariance matrix Σ_L , X_i denotes the i th row of the matrix X , L_i denotes the i th row of the matrix L , and λ is a positive tuning parameter that adjusts the sparsity of θ_L . We assume a uniform prior distribution over Z , which makes $\log P(Z)$ constant.

We use a coordinate ascent procedure over the three sets of optimization variables – L , Z , and θ_L . We iteratively estimate each of the optimization variables until convergence.

Learning L To estimate L from Equation (1) given Z and θ_L , we solve the following problem:

$$\max_{L_1, \dots, L_k} \left\{ -\text{trace}(LL^T \theta_L) - \sum_{i=1}^p \frac{\|X_i - L_{Z_i}\|_2^2}{\sigma^2} \right\}. \quad (4)$$

Setting the derivative of the objective function in Equation (4) to zero with respect to L_m for $m \in \{1, \dots, k\}$ leads to:

$$L_m = \frac{\sum_{X_i \in M_m} X_i - \sigma^2 \sum_{i \neq m} (\theta_L)_{im} L_i}{|M_m| + \sigma^2 (\theta_L)_{mm}} \quad (5)$$

where M_m means a set of X_i that belongs to the m th module: $M_m = \{X_i | Z_i = m\}$ and $|M_m|$ means the number of variables belonging to M_m . We update L_m for each m ($1 \leq m \leq k$), based on the current values of the other latent variables.

If all elements in θ_L are equal to zero, L_m would be set to the centroid of the m th module. This leads to a nice interpretation of the MGL learning algorithm with respect to the k-means clustering. The k-means clustering algorithm is the special case of the MGL when no network structure is assumed to exist among the latent variables (cluster centroids). More specifically, the MGL is a

generalization of the k-means with the distance metric determined by the sparse estimate of the latent structure (Θ_L).

Learning Z To estimate Z given L , we solve the following optimization problem:

$$\min_{Z_1 \dots Z_p} \sum_{i=1}^p \|X_i - L_{Z_i}\|_2^2. \quad (6)$$

In the hard assignment paradigm that we follow, Equation (6) finds the module for X_i that minimizes the Euclidean distance between X_i and the latent variable. We used $\sigma = 1$ such that we evaluate MGL in a simple and efficient setting.

Learning Θ_L To estimate Θ_L given L , we solve the following optimization problem:

$$\max_{\Theta_L > 0} \left\{ \log \det \Theta_L - \text{tr}(S_L \Theta_L) - \lambda \sum_{j \neq j'} |(\Theta_L)_{jj'}| \right\}, \quad (7)$$

where $S_L = \frac{1}{n} L L^T$ is the empirical estimate of the covariance matrix Σ_L . Based on the estimated value of L , Equation (7) can be solved by the standard graphical lasso algorithm¹⁹ applied to L . We implemented MGL in C, and we used the C source code of CRAN R package *huge*²⁰ for the step of learning Θ_L , that is, to estimate the inverse covariance matrix of the latent variables. We iteratively estimate each of the optimization variables until convergence.

2.3 RESULTS FROM APPLICATION OF MGL TO OVARIAN CANCER DATA

Ovarian cancer is the 5th leading cause of cancer death among US women and has a 5-year survival rate of 30%.²¹ Learning the features from tumor gene expression data is an effective strategy to identify novel disease mechanisms^{22,23}. Thus, we experimented MGL on three gene

expression datasets containing 10,404 ovarian tumor gene expression levels in a total of 909 patients with ovarian serous carcinoma – Tothill (269 samples)²⁴, TCGA (560 samples)²³, and Denkert (80 samples)²⁵. We mainly used Tothill for training, and TCGA and Denkert for testing. Given the data, MGL estimates Z , L and θ_L (see Equation (3)), which describe a gene module network characterized by the assignments of genes to modules and the latent structure among the modules (**Figure 1b**). We evaluated MGL based on: 1) how well the learned model fits unseen data; 2) how significantly the inferred modules are coherent in terms of gene functions; and 3) how well the inferred latent variables are predictive of survival time. Those evaluations are explained in the next three subsections respectively. We finally present some of the biologically interesting findings that we obtain from the MGL results. Since this application requires learning a network with >10K variables, the methods that attempt to learn the network of all individual variables do not scale. Therefore, we compared MGL with only Toh that first clusters the variables and then learns the network of cluster centroids.

2.3.1 *Cross-validation test log-likelihoods*

We applied k-means clustering and used the resulting clusters as a starting point for MGL and Toh. We compared between MGL and Toh in terms of the cross-validation (CV) test log-likelihood of the estimated p-dimensional precision matrices. Cluster count (k) was determined as 150 by Bayesian Information Criterion (BIC). Since the data are high-dimensional, we performed 2-fold CV. We used a wide range of values such that solutions for the module precision matrices range from a full matrix to an empty matrix. The results were averaged over 10 iterations due to the non-deterministic nature of the k -means. **Figure 3a** shows the test log-likelihoods of each method. MGL clearly outperforms Toh, meaning that the learned model by MGL fits unseen data better

than the one by Toh. Moreover, the standard deviation of the test log-likelihoods of the folds is smaller for MGL than Toh, indicating the robustness of MGL. In the subsequent sets of experiments, we use $k = 150$ (as determined by BIC) and $\lambda = 0.004$ (as chosen by CV).

2.3.2 *Functional enrichment of modules*

Genes assigned to the same module are likely to share similar functions, and those in the connected modules are likely to be involved in similar cellular processes as well. We define a super-module (or a super-cluster) as the set of genes in two connected modules (or clusters). We compared super-clusters from the learned network by Toh to super-modules from the learned network by MGL in terms of functional coherence. For each of the 4722 Curated gene sets from the Molecular Signatures Database²⁶, we computed the significance of the overlap between the gene set and super-modules (or superclusters). We applied Bonferroni correction to the p -values and only considered gene sets with $p < 0.05$ in either MGL or Toh. We repeated this process 50 times with different random initial points for k-means. As can be seen in **Figure 3b**, for each of 50 runs, there are a larger number of gene sets that are more significantly overlapped with MGL super-modules than with Toh super-clusters. Thus, MGL improves the initial network of Toh, resulting in far more shared processes between modules that are connected in the estimated network.

Additionally, we observed that in each independent run, MGL improves the enrichment p -values. In **Figure 3c**, for each of 4722 functional categories, the smallest p -value achieved by MGL supermodules is plotted (y-axis) against that achieved by Toh super-clusters (x-axis). The results for all 50 runs were aggregated in **Figure 3c**. Most of the dots in **Figure 3c** lie above the diagonal, meaning that for most of the functional categories, MGL super-modules achieve better enrichment than Toh super-clusters. Moreover, 6 gene sets were observed to be enriched by MGL

super-modules with p -values not only smaller than 10^{-90} , but also smaller than the best p -values for Toh super-clusters (10^{-20}). These gene sets were related to cell differentiation and increased cell growth, which are core processes relevant to cancer progression. This shows MGL's power to detect core cancer modules.

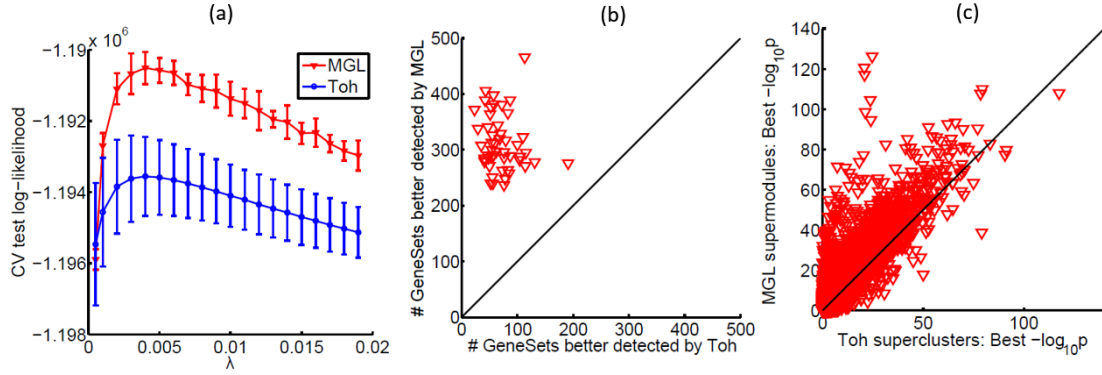


Figure 3. Performance of Toh method (k -means followed by graphical lasso) vs. MGL.

(a) Comparison between MGL and Toh in terms of cross validation test log-likelihood for varying λ values and for $k = 150$ (which was determined by BIC). Standard deviation between the test log-likelihoods of the folds are shown by the error bars. **(b)** Each dot represents a run with a random k -means starting point. For each of 50 runs, the number of gene sets more significantly overlapped with MGL super-modules (y-axis) is compared to that with Toh super-clusters (x-axis). **(c)** Each dot represents a gene set. For each gene set, the smallest enrichment p -value achieved by Toh (x-axis) vs. MGL (y-axis) is compared. We note that the plot for each run was consistent with the aggregated plot.

2.3.3 Survival prediction using latent variables as features

The latent variables (features) learned by MGL could represent activity levels of pathways relevant to the disease processes and clinical outcomes. We evaluated how well the features learned by MGL from Tothill are predictive of survival time of ovarian cancer patients in TCGA and Denkert datasets. After learning Toh and MGL on the Tothill dataset, we trained the Cox

regression model using the inferred latent variables as features in the same dataset, and then tested the model on a separate test dataset. In the test dataset, we computed the concordance index (c-Index) which is a standard evaluation metric estimating the accuracy of survival prediction based on the ‘censored’ survival data. **Figure 4** shows the c-Index achieved for varying sparsity levels for the Cox regression model (x-axis) on the two training-test dataset pairs. It compares latent variables (modules) from MGL to clusters from Toh and individual genes. In both of the settings, the c-Index values for modules are larger than those for Toh clusters or individual genes, for a wide range of sparsity levels. The maximum c-Index for modules is also higher than those for clusters and individual genes. The c-Index values were averaged over 50 runs for MGL and Toh.

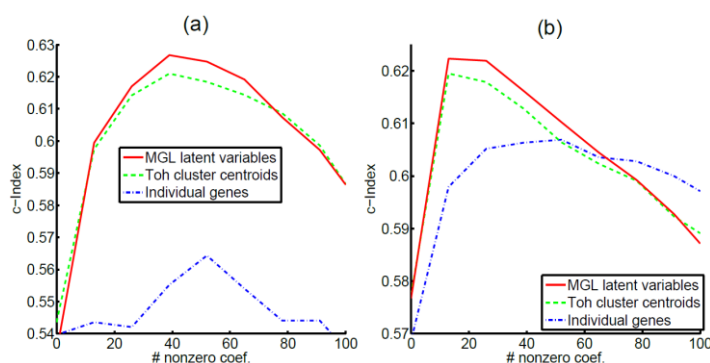


Figure 4. Comparison among MGL latent variables, k - means cluster centroids, and individual genes based on survival prediction performance.

x-axis gives the number of nonzero coefficients selected by penalized Cox regression model and *y*-axis gives *c*- Index values. Two pairs of training-test data are considered: **(a)** Tothill-Denkert, **(b)** Tothill-TCGA.

2.3.4 Interesting findings

A handful of modules identified by MGL are enriched for processes relevant to tumor biology, drug metabolism, and response to drug therapy. **Figure 5** shows a small portion of the module

network learned by MGL on Tothill dataset. It is a network among immune system, cell cycle and drug metabolism processes. Edges between modules 1 through 4 indicate conditional dependencies among cytokines, inflammation, and immune signaling, which play important roles in tumor biology.²⁷ There are suggestive edges between module 4 and modules 7 and 8 (cell cycle modules), since innate immune response can stimulate cell division in neoplastic cells.²⁷ Finally, module 5 is significantly enriched for PDGF for signaling. PDGF receptor agonists, such as the popular drug Gleevec, have succeeded in treating chronic myelogenous leukemia patients.²⁸

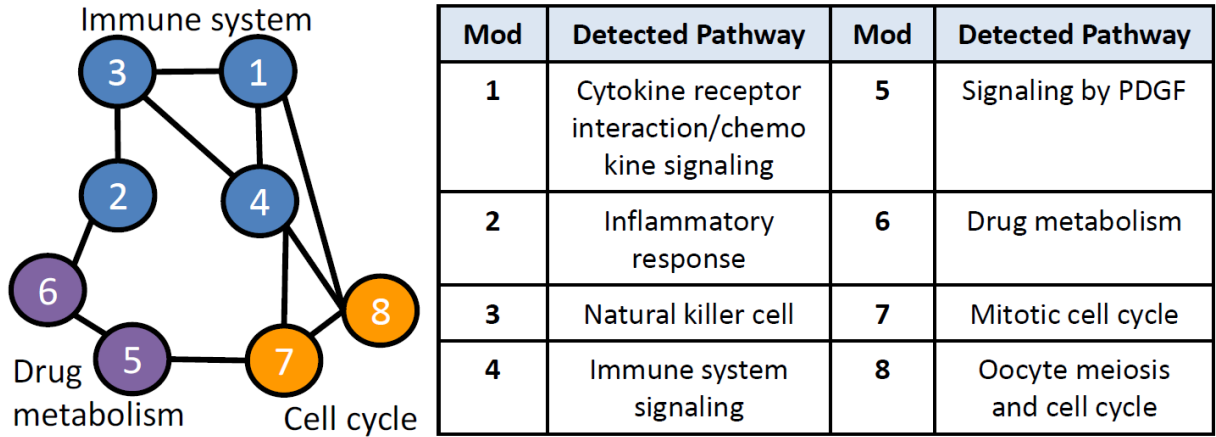


Figure 5. A small portion of the pathway structure identified by MGL on Tothill dataset.

2.4 DISCUSSION

In this chapter, we proposed the module graphical lasso, a novel high-dimensional GGM representation of conditional independencies among tightly coupled sets of variables (modules). The MGL algorithm is a novel high-dimensional clustering algorithm that is a generalization of k-means clustering, with Mahalanobis distances between variables. The full joint probability distribution function defines a non-Euclidean distance metric between the latent variables

L based on θ_L .

There are several possible extensions. First, MGL could be extended to other graphical models, such as Markov random fields, with novel distance metrics and clustering properties. Second, the assumptions about relationships between latent and observed variables could be relaxed. For instance, we could apply soft assignments of variables to modules, and learn sub-networks within modules. Third, we could add learning of module variances (σ) as an inference step to the MGL algorithm. MGL could be applied to gene expression data across multiple healthy and cancerous tissues to identify conserved and differential latent molecular networks driving tumor biology.

Chapter 3. UNSUPERVISED LEARNING OF A SHARED MODULE NETWORK FROM MULTIPLE EXPRESSION DATASETS

[The work presented in this chapter was published in Genome Medicine²⁹ and featured in Greene²]

As datasets increase in size, scope, and generality, the possibility to infer potentially *relevant* and *robust* features from data increases. Extracting a biologically intuitive low-dimensional representation (LDR) of data in an *unsupervised* fashion (i.e., based on the underlying structure in the data, not with respect to a particular prediction task) has become an important step to identify robust and relevant information from data. Development of unsupervised LDR learning methods is a very active area of modern research in machine learning and high dimensional data analysis^{30–32}. Specific machine learning approaches that have recently had noted success include deep learning algorithms³², where authors demonstrate enormous increases in performance on difficult tasks such as image and text classification^{33,34}. Analogously, in cancer transcriptomics unsupervised LDR learning has seen success on very difficult problems, such as predicting patient outcome in breast cancer in the DREAM7 breast cancer prognosis challenge³⁵. The winning team leveraged an unsupervised LDR extraction method on independent transcriptomic data from multiple cancer types, and significantly outperformed the other contestants in the challenge by a large margin³⁶ along with all other known prognostic signatures in breast cancer.

There are three main challenges with applying existing unsupervised LDR learning approaches to cancer transcriptomic data. First, any one study may not be generalizable in that there will be either technical (e.g. sample ascertainment) or experimental (e.g. batch effects) confounders that make an LDR of data extracted from an individual dataset in a naïve way not

necessarily generalizable to other datasets. Second, identifying simple modules (co-expressed sets of genes) using methods such as WGCNA³⁷ or simple clustering approaches^{38,39} will not necessarily capture complex dependence structures among the modules. Appropriately accounting for rich dependencies among these modules will improve their biological coherence. It has been shown that modeling the dependencies among modules improves the quality of the inferred modules from gene expression data³. Finally, and most importantly, most cancer transcriptomic data are within the $p \gg n$ regime (high-dimensional), i.e. we usually have tens of thousands of genes, but only hundreds of samples at most. This means that a successful method must include a very aggressive dimensionality reduction mechanism that allows generalization across datasets, since the potential for overfitting is high. This implies that models that allow for arbitrarily rich dependencies among variables (such as those used in deep learning methods) cannot necessarily be applied without overfitting the data.

We present a novel unsupervised LDR learning method, called INSPIRE (**IN**fering **S**hared modules from multi**P**le gene exp**R**ession datasets), to infer highly coherent and robust modules of genes and their dependencies on the basis of gene expression datasets from multiple independent studies (**Figure 6**). INSPIRE is an unconventional and aggressive data dimensionality reduction approach that extracts highly biologically relevant and coherent modules from gene expression data, where the number of samples is much less than the number of observed genes – the norm for cancer expression data. INSPIRE addresses the three aforementioned challenges. First, INSPIRE naturally integrates many datasets by modeling the latent (hidden, unobserved) variables in a probabilistic graphical model⁴⁰, where the latent variables are modeled as a Gaussian graphical model, the most commonly used probabilistic graphical model for continuous-valued variables

(**Figure 6**). Each observed gene is treated as a noisy and independent observation of these underlying latent variables.

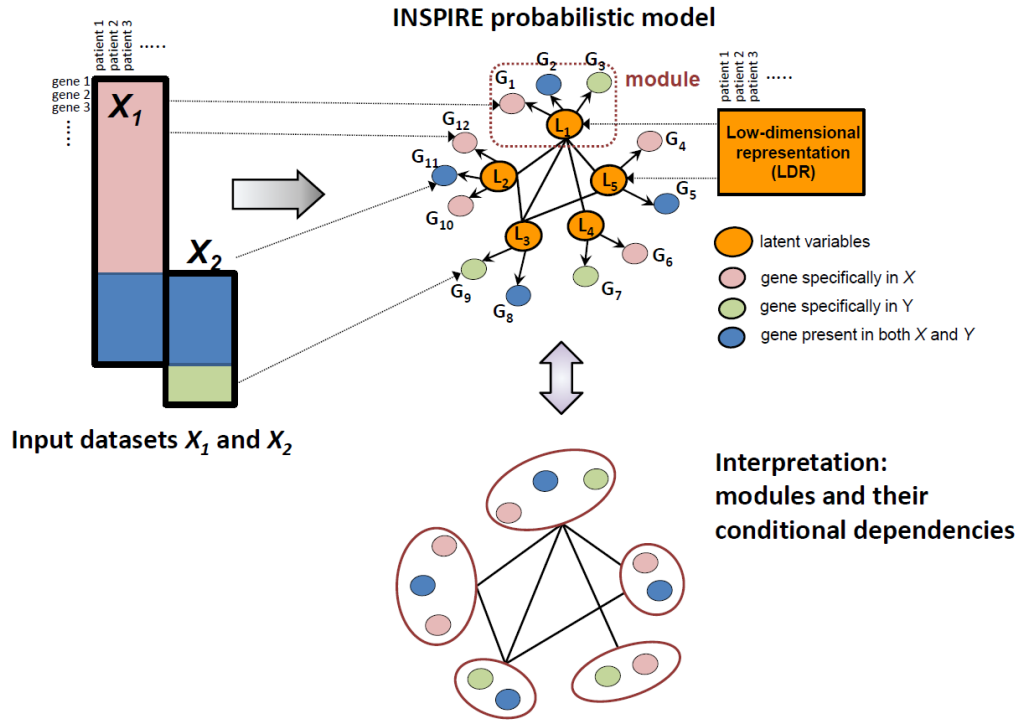


Figure 6. Overview of the INSPIRE framework.

INSPIRE takes as input multiple expression datasets that potentially contain different sets of genes and learns a network of expression modules (i.e., co-expressed sets of genes) conserved across these datasets. *INSPIRE* is a general framework that can take any number of datasets as input; two datasets (X_1 and X_2) are shown in representation for simplicity. **Top left:** Two input datasets are represented by rectangles with black solid lines. Rows represent genes and columns represent samples. The blue region contains the data for the genes that are contained in both datasets. The pink and green regions contain the data for the genes which are contained by only one of the datasets. **Top right:** The features (latent variables), each corresponding to a module, are shown by the orange matrix as learned by *INSPIRE*. These are used as a low-dimensional representation of the expression datasets. **Top middle:** As an example, five *INSPIRE* features $L_1, \dots, L_5$ (orange-shaded circles), 12 genes $G_1, \dots, G_{12}$ associated with those features, and the conditional dependency network among the *INSPIRE* features are represented. The

*dependencies among features are conserved across the datasets. **Bottom middle:** Five modules, each corresponding to an INSPIRE feature, and the dependency network among them are represented as the interpretation of the INSPIRE features and their conditional dependencies.*

By jointly inferring the assignment of observed genes to latent variables and the structure of the Gaussian graphical model among these latent variables, we can naturally capture both modules and their dependencies that generalize across multiple datasets (**Figure 6**). This addresses the issue with generalizability of modules across datasets. Second, our method naturally models the dependencies among the modules, which allows us to capture more complicated dependencies among pathways, cell populations, or other biologically driven modules than naïve approaches such as hierarchical clustering. In a previous study³, we have shown that modeling the dependencies among modules directly improves the biological coherence of the modules we learn, and their generalizability across datasets. Finally, by modeling the data as noisy observations from a much lower dimensional subset of modules, we are able to overcome the curse of dimensionality, and have better power to both learn the modules and their dependencies, even when the number of genes is much greater than the samples size. Through extensive simulated and real data analysis (**Figure 7**) we demonstrate our approach is a great practical tradeoff between model complexity and model parsimony when understanding biological pathways characterizing the cancer transcriptome across ovarian cancer patients.

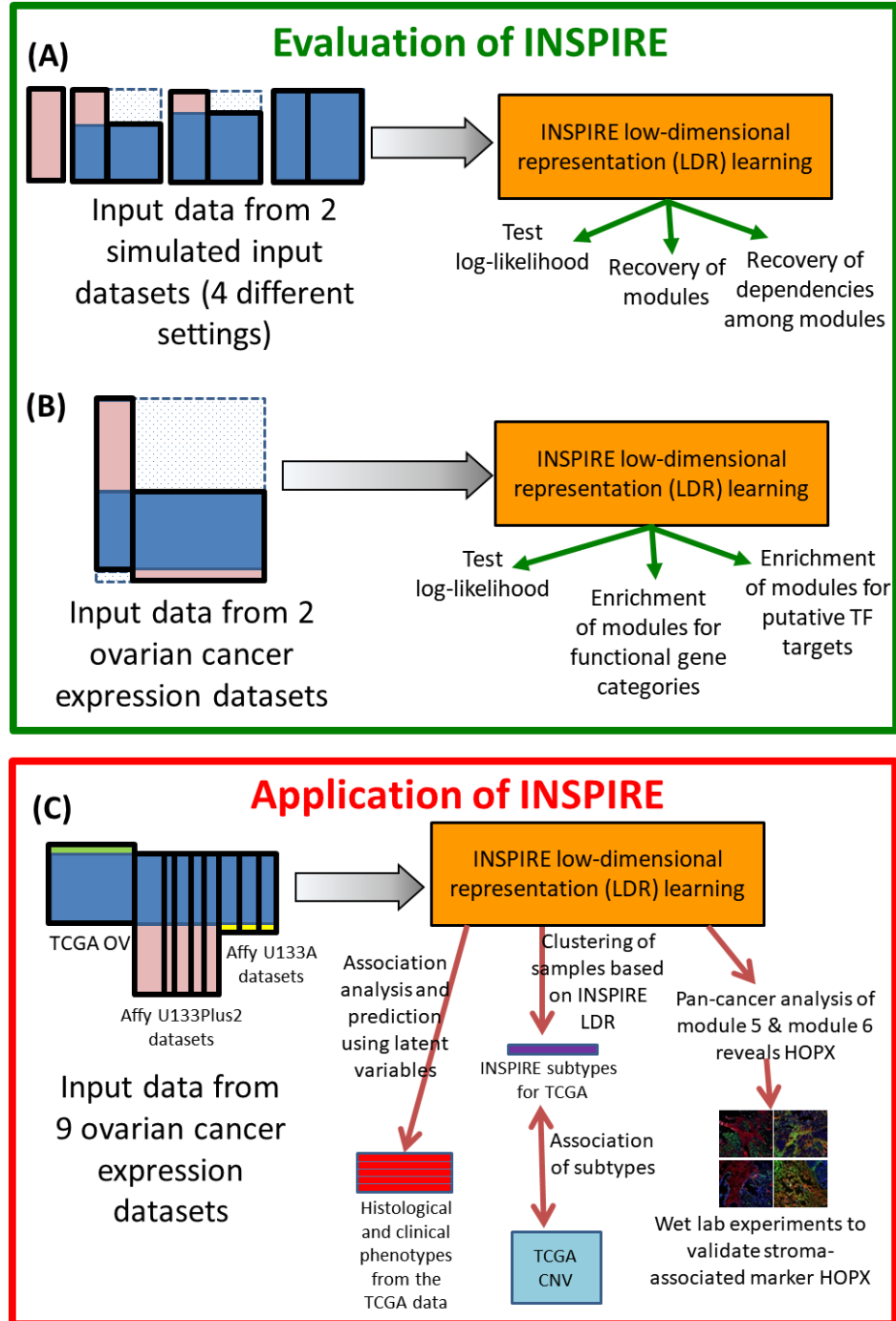


Figure 7. Overview of the evaluation and application of INSPIRE procedure.

The procedure takes as input $K \geq 2$ datasets, and the method is an iterative procedure that determines both the assignment of the genes to modules, the features each corresponding to a module, and the dependencies among the features which are conserved across the datasets. (A)

Evaluation of INSPIRE using simulated data. Two simulated datasets in four settings corresponding to different amount of gene overlap are provided as input to the INSPIRE learning algorithm, and the learned modules and network are evaluated in terms of three different metrics. (B) Evaluation of INSPIRE using two ovarian cancer expression datasets. Two expression datasets from different platforms are provided as input to the learning algorithm and the learned modules and network are evaluated in terms of three different metrics. (C) Application of INSPIRE on nine real-world ovarian cancer expression datasets. As an application of INSPIRE, we first check the association of the learned INSPIRE features with six histological and clinical phenotypes, which is followed by subtyping the patients into groups based on the learned INSPIRE features. Observing that INSPIRE features have high association with the histological and clinical phenotypes in cancer, and the subtypes learned based on the features can predict CNV abnormalities well leads us to do a deeper analysis of two modules (modules 5 and 6), which are good predictors of many phenotypes and good differentiators of learned ovarian cancer subtypes.

3.1 RELATED WORK

Previous approaches to extract LDR from expression data can be divided into two categories; 1) *supervised* methods that extract an LDR that is discriminative of different class labels in the training samples, and 2) *unsupervised* methods (including INSPIRE) that extract an LDR purely based on the underlying structure of the data.

Several authors developed methods that use known pathways or biological networks along with gene expression data to extract an LDR (“pathway markers”) whose activity is predictive of a given phenotype^{41–44}. Chuang et al.⁴¹ proposed a greedy search algorithm to detect subnetworks in a given protein-protein interaction (PPI) network, such that each subnetwork contains genes whose average expression level is highly correlated with class labels

(metastatic/non-metastatic) measured by the mutual information. The authors claim that subnetwork markers outperform individual genes for predicting breast cancer metastasis. Lee et al.⁴² developed a similar algorithm to select subsets of genes from MSigDB (Molecular Signatures Database) C2 (curated) pathways that give an optimal discriminative power for the classification of leukemia/breast cancer phenotypes. Both Chuang et al.⁴¹ and Lee et al.⁴² determined LDR as the average expression levels of genes in each subnetwork and pathway, respectively. Taylor et al.⁴³ proposed a similar approach that uses a PPI network, but instead of computing the LDR by averaging gene expression levels within a subnetwork (or a pathway), they compute the expression difference between a hub protein and all of its neighbors in the PPI network. Ravasi et al.⁴⁴ used a similar approach to extract subnetwork features as hub transcription factors (TFs) from TF PPI networks in human and mouse. Besides the methods that infer an LDR by averaging (or aggregating) expression levels of subsets of genes, there have been methods to select a subset of genes. For example, Herschkowitz et al.⁴⁵ used 106 genes selected by an *intrinsic analysis* for a classification problem (122 mouse breast tumors/232 human breast tumors). This analysis aims to select genes that are relevant to tumor classification by identifying genes whose expression levels show relatively low within-group variation and high between-group variation for known groups of tumors in each of human and mouse datasets⁴⁵. Although supervised methods would be useful to infer an LDR relevant to a particular prediction problem, there are several disadvantages over unsupervised methods. First, we need to have a particular prediction problem with class labels, which may not be available. Second, supervised methods usually rely on the assumption that the same genes are differentially expressed in all samples within a class, which is unlikely to be true in heterogeneous diseases such as cancer.

On the other hand, unsupervised methods extract an LDR without knowing about the class labels, while the learned LDR can be used for classification purposes later. Development of unsupervised feature learning methods is a very active area of modern research in machine learning and high dimensional data analysis³⁰⁻³². One of the most commonly used methods is principal component analysis (PCA)⁴⁶ which sequentially extracts most of the variance (variability) of the data. Another is independent component analysis (ICA)^{39,47}, a statistical technique for identifying latent factors that underlie sets of random variables, measurements, or signals. However, each PC (principal component - or eigengene) or IC is a linear combination of all genes not a small subset of genes, which makes it difficult to biologically characterize it. Clustering algorithms⁴⁸, on the other hand, generate explicit gene clusters, and they define an LDR as a set of mean or median expression levels of the genes in each cluster. In the seminal work by Langfelder and Horvath (a technique called WGCNA)³⁷, the adjacencies retrieved from Pearson's correlation of expression levels of gene pairs is transformed into a topological overlap measure (TOM), which is a network interconnectivity measure that takes into account the shared neighbors of each gene pair. TOM is then used in a hierarchical clustering to define modules. While WGCNA³⁷ defines its similarity measure (i.e., TOM) based on the *marginal correlations* between genes, other authors have used *partial correlations (conditional dependencies)* to model gene relationships^{3,6,7}. Chandrasekaran et al.⁷ incorporated latent variables into a Gaussian graphical model among individual genes, while Celik et al.³ divided variables into modules and learned module-level dependencies (module graphical lasso - MGL). He et al.⁶ defined an LDR as a set of latent factors, and modeled each latent factor as a linear combination of genes (structured latent factor analysis - SLFA). While similar to Celik et al.³ in modeling a higher-level dependency structure, He et al.⁶ does not form explicit clusters. Finally, Cheng et al.³⁶ identified 12 metagenes, each of which is a weighted

average of the genes that are co-expressed across multiple cancer types. They showed that the prediction model they derived based on these metagenes is highly predictive of survival in breast cancer within the context of the DREAM7 Challenge, leading to the top scoring model³⁵.

There are three major differences between INSPIRE and previous approaches. First, none of the previous methods to learn LDR can accommodate multiple datasets containing different sets of genes (e.g., different microarray platforms), while INSPIRE directly addresses this challenge. One naïve way to run previous methods on datasets that contain different sets of genes with a partial overlap is to treat the values on the genes that are not observed in each dataset as missing data. We could use missing value imputation techniques to fill in missing data and learn a single statistical model from the imputed data. However, most imputation methods perform poorly when a large number of values are missing (**Figure 6**). We demonstrate that INSPIRE outperforms the imputation-based approaches. Second, INSPIRE uses a novel probabilistic model that can describe more complex relationships (i.e., conditional dependencies) than pairwise marginal correlations among genes. We show that INSPIRE outperforms a correlation-based method, WGCNA. Finally, INSPIRE uses a novel learning algorithm to make use of all samples in multiple datasets, which increases the statistical power to detect a statistical robust model (**Figure 6**). Our extensive experiments show that these key properties of INSPIRE lead to biologically more relevant and statistically more robust features than alternative methods.

When we apply INSPIRE to nine gene expression datasets from ovarian cancer studies (**Figure 7C**), we identify a novel tumor-associated stromal marker *HOPX*, which additional analyses suggest may be a molecular driver for a conserved module in the network that contains known epithelial-mesenchymal transition (EMT) inducers and is significantly associated with percent stroma in ovarian tumors from The Cancer Genome Atlas (TCGA). This module is one of the two

modules that best represent one of the predominant subtypes of ovarian cancer, ‘mesenchymal’ subtype identified in the TCGA ovarian cancer study²³. These multiple lines of evidence suggest that *HOPX* may be a great target for further functional validation to understand the maintenance of tumor-associated stroma along with understanding the clinically relevant ‘mesenchymal’ subtype in ovarian cancer.

The implementation of INSPIRE, the data used in the study, and the resulting INSPIRE models are freely available on our website⁴⁹.

3.2 INSPIRE LEARNING ALGORITHM

We present the INSPIRE (Inferring Shared modules from multiPle gene expRession datasets) method to extract a compact description of high-dimensional gene expression data by learning a set of *k modules and their dependencies* from *Q* gene expression datasets. The technical novelty of the INSPIRE is that it provides a flexible model that does not require the *Q* datasets to have exactly the same set of genes (e.g., different microarray platforms). INSPIRE takes *Q* expression matrices as input and learns how genes are assigned to modules, the latent (unobserved) variables each representing a module, and the dependencies among the latent variables, through an iterative procedure described in detail below. Each latent variable represents the activity level of a certain biological process or a regulatory module. In the sections that describe the probabilistic model and the learning algorithm, we will refer them to as ‘latent variables’ because that is a commonly used term to refer to hidden, unobserved variables in the statistical domain. Inferring the latent variables by using the INSPIRE method is an effective way to obtain a *low-dimensional representation* for prediction tasks (e.g., predicting histopathological phenotypes) or clustering (e.g., patient stratification) (**Figure 6**).

INSPIRE uses a formal probabilistic graphical model, specifically the Gaussian graphical model (GGM), to model relationships between genes and latent variables, and the conditional dependence relationships among the latent variables. A GGM is a popular probabilistic graphical model for representing the *conditional dependency network* among a set of continuous-valued random variables. In a GGM, the variables connected by an edge are *conditionally dependent* to each other given all the other variables in the model^{15,16}. For example, in a simple latent network shown in **Figure 6**, five *latent variables* ($L_1, \dots, L_5$) have mutual dependencies. So, let $L = \{L_1, \dots, L_5\} \sim N(0, \Sigma_L)$, then non-zero pattern of Σ_L^{-1} corresponds to the conditional dependencies among the latent variables, namely the topology of the network. That means, since L_1 and L_2 are connected to each other for example, knowing L_1 's expression level gives information about L_2 's expression level, even when we know the expression levels of all the other latent variables, which indicates a *direct* dependency between L_1 and L_2 . We refer to the observed variables that stem from the same latent variable as a *module*. As an example, genes G_1, G_2, G_3 in **Figure 6** form a module since they are associated with the same latent variable L_1 . Below, we provide a mathematical formulation of the INSPIRE probabilistic model and the learning algorithm.

Let $X^1, \dots, X^Q$ be a set of Q expression datasets where the q th dataset $X^q = \{X_1^q, \dots, X_{p_q}^q\}$ contains the expression levels of p_q genes across n_q samples and each of X_i^q is a row vector of size n_q . Let $L^1, \dots, L^Q$ be a set of matrices where each L^q is associated with a dataset and consists of k latent variables. $L^q = \{L_1^q, \dots, L_k^q\} \sim N(0, \Sigma_L)$, where Σ_L is a $k \times k$ covariance matrix. These latent variables can be viewed as a lower-dimensional representation (LDR) of expression data and Σ_L represents the dependencies among the features. We assume that Σ_L is conserved across the Q datasets. Each gene is associated with exactly one of the k latent variables as represented by the

directed edge between a gene and a latent variable in **Figure 6**. The total number of unique genes across all Q datasets is p_T ; and each data matrix X^q contains samples from a different subset of p_q genes ($p_q \leq p_T$). Let Z be a $p_T \times k$ matrix indicating which of the k modules each of p_T genes belongs to, such that $\forall i, j \ Z_{ij} \in \{0, 1\}$ and $\forall i, \sum_{c=1}^k Z_{ic} = 1$. Each observed dataset X^q is generated by the multivariate Gaussian distribution $X^q | Z^q L^q, \sigma^2 \sim N(Z^q L^q, \sigma^2)$, where Z^q is a $p^q \times k$ matrix composed of the rows of Z corresponding to the p_q genes contained by the dataset X^q . Here, we refer to a set of genes that correspond to the same latent variable as a *module* where σ determines the module tightness. As an example, the j th module M_j can be defined as $M_j = \cup_{\{q=1\}}^Q \{X_i^q | Z_{ij}^q = 1\}$. Thus, Z defines the module assignment of all unique genes in all Q datasets into k modules. Each gene belongs to exactly one module. We choose hard assignment of genes to modules ($\forall i, \exists! c : Z_{ic} = 1$) to reduce the number of parameters. Soft assignment is a straightforward extension where we relax the constraint $\forall i, j \ Z_{ij} \in \{0, 1\}$ to $\forall i, j \ 0 \leq Z_{ij} \leq 1$. INSPIRE jointly learns the latent variables $L = [L^1, \dots, L^Q]$ each corresponding to a module; the module assignment indicator Z ; and the feature dependence network Σ_L^{-1} . Given Q datasets $X^1, \dots, X^Q$, where $X^q (\in \mathbb{R}^{p_q \times n_q})$ contains n_q observations on p_q genes and $n_T = \sum_{q=1}^Q n_q$, INSPIRE aims to learn the following:

- $L^q \in \mathbb{R}^{k \times n_q}$ for each $q (\in \{1, \dots, Q\})$ containing the values on k features in n_q samples in X^q
- $Z | \sum Z_i = 1$, a binary vector for each $i (\in \{1, \dots, p_T\})$ specifying the module membership of the i th gene in one of the k modules; and

- $\theta_L (\in \mathbb{R}^{k \times k})$ denoting the estimate of the inverse covariance matrix of the features, i.e. Σ_L^{-1}

We address our learning problem by finding the joint maximum a posteriori (MAP) assignment to all of the optimization variables – L , Z , and θ_L . This means that we optimize the joint log-likelihood function of the Q data matrices, with respect to L , Z , and $\theta_L (> 0)$. Given the statistical independence assumption that genes in a dataset X^q are statistically independent to one another given the latent variables L^q , the joint log-likelihood can be decomposed as follows:

$$\begin{aligned} & \log P(X^1, \dots, X^Q, L^1, \dots, L^Q, Z, \theta_L; \lambda, \sigma) \\ &= \sum_{q=1}^Q \log P(X^q | L^q, Z^q) + \sum_{q=1}^Q \log P(L^q | \theta_L) + \log P(\theta_L) + \log P(Z) \\ &= \frac{n_T}{2} \{ \log \det \theta_L - \text{tr}(S_L \theta_L) \} - \lambda \sum_{j \neq j'} |(\theta_L)_{jj'}| - \frac{1}{2} \sum_{q=1}^Q \frac{\|X^q - Z^q L^q\|_2^2}{\sigma^2} + \text{const}, \quad (8) \end{aligned}$$

where $S_L = \frac{1}{n_T} \sum_{q=1}^Q L^q L^{qT}$ is the empirical estimate of the covariance matrix Σ_L and λ is a positive tuning parameter that adjusts the sparsity of θ_L . We assume a uniform prior distribution over Z , which makes $\log P(Z)$ constant.

We use a *coordinate ascent procedure* over three sets of optimization variables – L , Z , and θ_L .

We iteratively estimate each of the optimization variables until convergence.

Learning L To estimate $L^1, \dots, L^Q$ from Equation (8) given Z and θ_L , we solve the following problem:

$$\max_{L^1, \dots, L^Q} \left\{ -\text{tr}(L^q L^{qT} \theta_L) - \frac{\|X^q - Z^q L^q\|_2^2}{\sigma^2} \right\}. \quad (9)$$

Setting the derivative of the objective function in Equation (9) to zero with respect to L^q_c for $q \in \{1, \dots, Q\}$ and $c \in \{1, \dots, k\}$ leads to:

$$L^q_c = \frac{Z^{qT}_c X^q - \sigma^2 \sum_{i \neq c} (\Theta_L)_{ic} L^q_i}{\|Z^{qT}_c\|_2^2 + \sigma^2 (\Theta_L)_{cc}}. \quad (10)$$

Learning Z To estimate Z given $L^1, \dots, L^Q$, we solve the following optimization problem:

$$\min_{Z_1 \dots Z_{p_T}} \sum_{q=1}^Q \|X^q - Z^q L^q\|_2^2. \quad (11)$$

In the hard assignment paradigm that we follow throughout this work, Equation (11) assigns gene p_i to module $c \in \{1, \dots, k\}$ that minimizes the Euclidean distance computed using all samples from the datasets containing the gene p_i .

Learning Θ_L To estimate Θ_L given $L^1, \dots, L^Q$, we solve the following optimization problem:

$$\max_{\Theta_L > 0} \left\{ \log \det \Theta_L - \text{tr}(S_L \Theta_L) - \lambda \sum_{j \neq j'} |(\Theta_L)_{jj'}| \right\}, \quad (12)$$

where the constraint $\Theta_L > 0$ restricts the solution to the space of positive definite matrices of size $k \times k$, and $S_L = \frac{1}{n_T} \sum_{q=1}^{q=Q} L^q L^{qT}$ is the empirical covariance matrix of L . Based on the estimated value of L , Equation (12) can be solved by the graphical lasso¹⁹, a well-known algorithm for learning the structure of a Gaussian graphical model (GGM).

We iteratively estimate each of the optimization variables until convergence. Since our objective is continuous on a compact level set, based on Theorem 4.1 in Tseng (2001)⁵⁰, the solution sequence is defined and bounded. Every coordinate group reached by the iterations is a stationary

point of INSPIRE objective function. We also observed that the value of the objective likelihood function monotonically increases.

3.3 EXPERIMENTAL RESULTS

3.3.1 Overview of the INSPIRE framework

INSPIRE extracts a low-dimensional representation (LDR) from multiple gene expression datasets by inferring k latent (unobserved) variables and the dependencies among the latent variables captured by a probabilistic graphical model (**Figure 6**). INSPIRE uses a standard iterative learning algorithm to optimize the joint log-likelihood objective function, Equation (8), by iteratively updating its model parameters until convergence. INSPIRE iterates the following three steps until convergence: i) inferring the values of latent variables with all the other parameters held fixed, as described in Equation (10), ii) assigning genes into latent variables as described in Equation (11), and iii) learning a network of latent variables as described in Equation (12). In each iteration, latent variables are computed based on the current assignment of genes into modules and the estimated dependency network among the latent variables, as described in Equation (8). If there are no dependencies among latent variables, each latent variable would be an average expression level of the genes in the module. Thus, latent variables can be viewed as module centers adjusted for the estimated dependency network among latent variables.

A set of genes assigned to the same latent variable is referred to as a *module* (**Figure 6**). To focus on identifying a parsimonious, independent set of modules from high-dimensional gene expression data, we design our model such that each gene is assigned to only one module, although it would be a simple extension to assign each gene to multiple modules. However, when we

implemented an extension of INSPIRE which allows each gene to be assigned to more than one module, the functional coherence of modules significantly decreased (**Appendix Figure 4-Appendix Figure 5**). This could be because the model with genes assigned to multiple modules has a significantly increased number of parameters.

The number of modules k is determined based on the standard Bayesian Information Criterion (BIC), although users can determine k in a different way depending on the problem. The INSPIRE framework simultaneously infers the assignment of genes into k latent variables and the dependency network among k latent variables by fitting the probabilistic model across multiple gene expression datasets that can potentially have different sets of genes (e.g., different platforms). The INSPIRE model provides a biologically intuitive LDR model for gene expression data where many biological networks are *modular*, and genes involved in similar functions are likely to be connected more densely with each other. How genes are organized into modules and how these modules are connected with each other may provide improved insights into the underlying disease processes, as discussed below.

After evaluating INSPIRE by comparing with alternative methods on simulated data and a small set of genome-wide expression datasets (**Figure 7A-B**), we applied INSPIRE to many ovarian cancer expression datasets, which led to a novel marker and potential driver of tumor-associated stroma (**Figure 7C**).

3.3.2 *INSPIRE learns underlying modules and their dependencies from simulated data more accurately than 13 other methods*

We first evaluate INSPIRE on data simulated using a probabilistic model of (unobserved) latent variables, gene expression levels, and the dependencies among the latent variables captured by a

probabilistic graphical model (**Figure 6**). To simulate the situation in which we are given expression datasets that contain different sets of genes (e.g., different microarray platforms), we generated two datasets (Dataset1 and Dataset2) with the same genes and included all genes in Dataset1 and varying percentages of the genes in Dataset2 such that varying numbers of genes are present in the overlapping portion of the datasets. This leads to three settings (**Appendix Figure 1a**, **Figure 8(ii)-(iv)** left): (ii) 60% of the genes are present in Dataset2, (iii) 80% of the genes are present in Dataset2; and (iv) all genes are present in Dataset2. The total number of genes in each of these settings is 250, and the number of modules is 10, with an average of 25 genes in a module (see APPENDIX A for details of synthetic data generation).

We compare INSPIRE with the following five state-of-the-art methods: **i) GLasso** - standard graphical lasso¹⁹ that learns a gene-level conditional dependence network with no LDR or module assumption; **ii) UGL** - unknown group L_1 regularization¹⁴ that learn sparse block-structured inverse covariance matrices with unknown block structure; **iii) SLFA** - structured latent factor analysis⁶ that learn an LDR of the data as well as the relationship between the latent factors; **iv) WGCNA** - weighted gene co-expression network analysis³⁷ that allows to define modules based on a special metric derived from the correlations of the gene pairs; **v) MGL** - module graphical lasso³ which simultaneously learns an LDR and the conditional dependencies among the latent variables (**Table 1**). Since all those methods work on a single dataset, to enable the application of these methods to multiple datasets with variable discrepancy, we adapt the input data to those five methods in three ways (**Appendix Figure 1b**): (1) using only Dataset1 that contains all genes, (2) using data on the genes that are present in both datasets (blue-shaded region in **Figure 6**), and assigning the rest of the genes to the learned modules based on the Euclidean distance between the gene's expression and the expression of each of the modules, and (3) imputing missing values in

Dataset2 and using both datasets as if they were a single dataset. This leads to 13 methods (**Table 1**). InterMGL, ImpMGL and INSPIRE represent different ways of handling missing data: INSPIRE uses a novel learning algorithm that does not require the missing portion when learning; ImpMGL imputes missing variables in the datasets before learning; and InterMGL ignores missing variables in the datasets. We ran each method on 20 different instantiations of the synthetic data and reported the average over the results. For each method, we also computed p -values of significance of difference from INSPIRE (APPENDIX A; **Figure 8-Figure 10**). We evaluated INSPIRE and 13 competitors in terms of how well they explain unseen data measured by the test-set log-likelihood, gene-module assignment accuracy, and the module dependency network accuracy. To make a comparison with WGCNA variant methods possible, we applied a standard graphical lasso algorithm to the modules learned by a WGCNA variant method. INSPIRE, SLFA and MGL are iterative algorithms with non-convex objective functions, so their results may depend on the initialization of the parameters. To reduce the possibility to make a conclusion based on a particular set of initial parameters, we performed the variants of those algorithms multiple times with different starting points (see APPENDIX A for details on initialization).

Table 1. Methods we compared with our INSPIRE framework.

*To our knowledge, there are no published methods for learning modules and their dependencies that can handle variable discrepancy. We adapted the following five state-of-the-art methods that can run on a single dataset: **GLasso** - standard graphical lasso¹⁹, **UGL** - unknown group L_1 regularization¹⁴, **SLFA** - the structured latent factor analysis⁶, **WGCNA** - weighted gene co-expression network analysis³⁷, and **MGL** - module graphical lasso³ (see APPENDIX A for details). We adapted the input datasets such that we can apply these methods to datasets with variable discrepancy (**Appendix Figure 1b**): ‘**---I**’ - learning a model from only Dataset1 that contains all genes; ‘**Inter---**’ - learning a model from the data on the overlapping and assigning the rest of the genes to learned modules by using the k -nearest neighbor approach (i.e., based on the Euclidean distance between the gene’s expression and each*

module's expression), and '**Imp---**' - imputing missing values in Dataset2 and learning a model from the imputed data (see APPENDIX A for details on imputation) (**Appendix Figure 1b**). These adaptations lead to 13 competitors: 1) GLasso1, 2) ImpGLasso, 3) UGL1, 4) ImpUGL, 5) WGCNA1, 6) InterWGCNA, 7) ImpWGCNA, 8) SLFA1, 9) InterSLFA, 10) ImpSLFA, 11) MGL1, 12) InterMGL, and 13) ImpMGL. In the experiments on synthetic data, we compared to all 13 methods, while in the experiments with two genome-wide ovarian cancer gene expression datasets which we will discuss in the subsequent sections, we only used the methods that are scalable (see **Appendix Figure 2**), as indicated by the purple-shaded region in the table. The '**Inter---**' method is not applicable to GLasso and UGL, because GLasso and UGL learn a network of genes, not modules, and it is not obvious how to connect the genes that are present only in Dataset1 to the learned network. We do not consider an adaptation that applies the methods to Dataset2 only ('---2'). This is because, other than the genes in the overlap, Dataset2 has no genes (in the synthetic data experiments) or a very small number of genes (in the experiments with genome-wide expression data), which makes '---2' that uses only the samples from Dataset2 unlikely to outperform '**Inter---**' that uses all samples.

Method	Description	Different ways to deal with missing data			Scalability (see Appendix Figure 2)
		---1	Inter---	Imp---	
GLasso	standard graphical lasso ¹⁹	GLasso1	X	ImpGLasso	No
UGL	unknown group L_1 regularization ¹⁴	UGL1	X	ImpUGL	No
SLFA	structured latent factor analysis	SLFA1	InterSLFA	ImpSLFA	No
WGCNA	weighted gene co-expression network analysis ³⁷	WGCNA1	InterWGCNA	ImpWGCNA	Yes
MGL	module graphical lasso ³	MGL1	InterMGL	ImpMGL	Yes

Test log-likelihood. The *test log-likelihood* that measures how well the learned models fit unseen data is a widely used evaluation metric for probabilistic models^{3,11,14}. We generated test data Y containing 100 samples, which was created in the same way as the training data X (see APPENDIX A). The 13 learned models are tested based on the same unseen data Y . Each method selects its own regularization parameter using the standard cross-validation (CV) test⁵¹ selecting λ with the best average CV test log-likelihood measured on Dataset1 in X (see APPENDIX A). We used the

test set of Dataset1 to compute the test log-likelihoods for all methods since Dataset1 contains all genes. **Figure 8** shows the average negative test log-likelihood per sample (lower the better) in (i) - (iv): (i) shows the methods that use only Dataset1, and (ii)-(iv) show Imp---, Inter--- and INSPIRE methods that use Dataset2 as well with varying numbers of genes in Dataset2 (**Appendix Figure 1**). Each bar (except INSPIRE) displays a p -value from Wilcoxon signed rank test that measures how significantly INSPIRE is better than the corresponding method across 20 instantiations of the data (see APPENDIX A). The bars for the methods that use only Dataset1 display three p -values, each for comparison to INSPIRE in (ii) - (iv). INSPIRE has significantly better test log-likelihoods than the methods that utilize one dataset ($p \leq 2.4 \times 10^{-5}$) and all the other 8 methods that can utilize multiple datasets ($p \leq 4.3 \times 10^{-4}$). This indicates that making use of multiple datasets by using INSPIRE has a great potential to increase the chance to infer the underlying model. In (iv), ImpMGL, InterMGL and INSPIRE perform similarly as expected, and they are better than the other methods that utilize multiple datasets. The methods that utilize only Dataset1 (i) achieve worse average test log-likelihood than their multiple-dataset counterparts (ii)-(iv); and the test log-likelihoods of most methods increase with the increasing number of overlapping variables, from (i) to (iv).

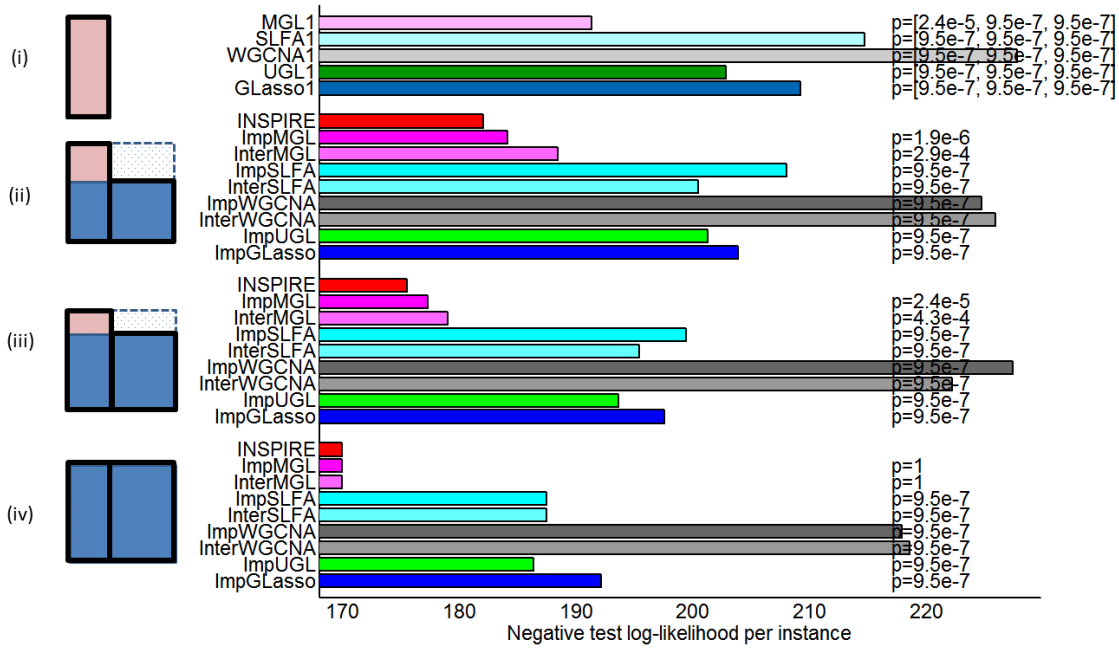


Figure 8. Test log-likelihood of INSPIRE vs. alternative models on the synthetic data.

Negative test log-likelihood per instance is averaged over 20 different instantiations of the synthetic data (lower is better). In the four synthetic data illustrations on the left, rows represent genes and columns represent samples. The Wilcoxon signed rank test p-value represented on each bar (except the bars for INSPIRE) measures the statistical significance of the difference between the method and INSPIRE.

Module recovery. We then evaluated based on how well important aspects of the underlying model are recovered by each method. We first checked whether pairs of genes that are assigned to the same module in the true model are in the same modules in the learned model. We used the rand index⁵² that measures how well pairs of genes agree on being in the same or different modules between two models - the true model and a learned model. A rand index of 0 means that none of the genes agree on being in the same/different groups, while 1 means a perfect recovery of the modules. The evaluation based on module recovery is not applicable for GLasso1 and ImpGLasso, since they do not learn modules. As shown in **Figure 9**, the module recovery performance of

INSPIRE is significantly better than its 13 competitors. INSPIRE has significantly higher rand index than (i) the methods that utilize a single dataset ($p \leq 4.9 \times 10^{-2}$), and (ii)-(iv) the methods that use multiple datasets ($p \leq 6.6 \times 10^{-2}$).

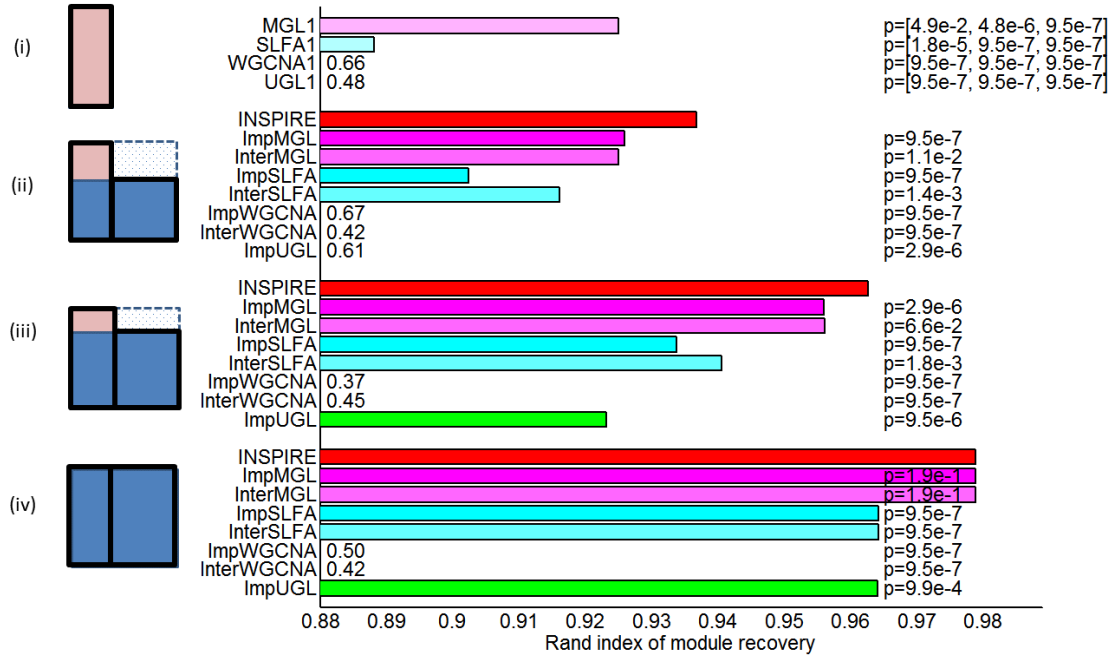


Figure 9. Module recovery performance of INSPIRE vs. alternative methods on the synthetic data.

Rand index is averaged over 20 different instantiations of the synthetic data (higher is better). In the four synthetic data illustrations on the left, rows represent genes and columns represent samples. The Wilcoxon signed rank test p-value represented on each bar (except the bars for INSPIRE) measures the statistical significance of the difference between the method and INSPIRE.

Module dependencies. Then, we evaluated based on how well the inferred module dependencies by each method are consistent with those in the true model. Since it is not clear how to map a module in the true model to the corresponding module in the learned model, we converted each module-based network model into the equivalent gene-based probabilistic model, using a well-established method³. It is not enough to get only high precision or recall, so we used the *F-measure* $= 2 \frac{(prec*rec)}{(prec+rec)}$ as an evaluation metric. As shown in **Figure 10**, INSPIRE has the highest average F-measure that measures the accuracy of the dependencies learned by each method in all (i)–(iv). INSPIRE is significantly better than methods that utilize a single dataset ($p \leq 2.4 \times 10^{-4}$) and other methods that use multiple datasets ($p \leq 2.7 \times 10^{-2}$).

The methods that use only one dataset tend to have lower average rand-index (modules) and F-measure (module dependencies) than their multiple-dataset counterparts; and as the number of genes shared across datasets increases, the overall performance of the methods that utilize multiple datasets increases. This indicates that combining multiple datasets reveals the underlying modules and their dependencies better, and INSPIRE is better than 13 alternative approaches in revealing the underlying model.

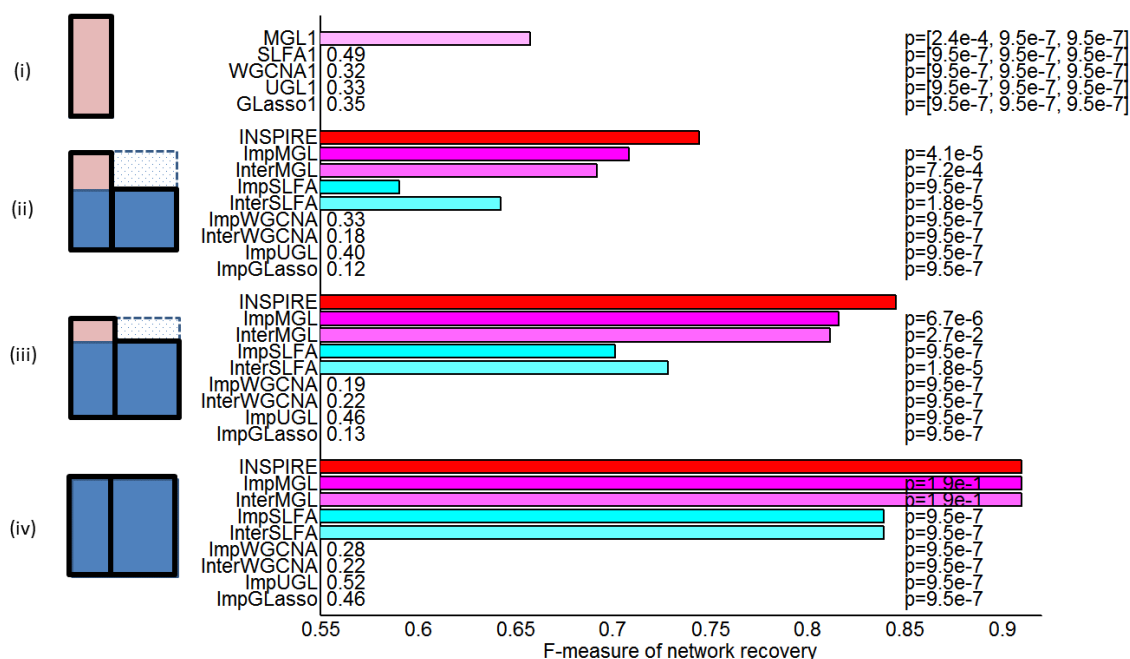


Figure 10. Feature dependency recovery of INSPIRE vs. alternative models on the synthetic data.

F-measure is averaged over 20 different instantiations of the synthetic data. In the four synthetic data illustrations on the left, rows represent genes and columns represent samples. The Wilcoxon signed rank test p-value represented on each bar (except the bars for INSPIRE) measures the statistical significance of the difference between the method and INSPIRE.

3.3.3 Evaluation on two genome-wide ovarian cancer expression datasets

Next, we evaluate INSPIRE based on the statistical robustness and biological relevance of the learned modules on two publicly available ovarian cancer gene expression datasets⁵³ (**Figure 11**): (1) OV1 that contains 18,113 genes and 28 patients (Affy U133 Plus 2.0 platform), and (2) OV2 that contains 8,331 genes in a total of 42 patients (Affy U95Av2 platform) (see APPENDIX A; **Appendix Table 2**).

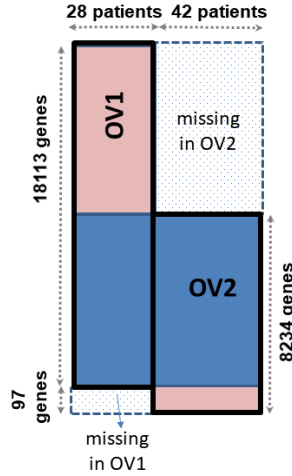


Figure 11. Illustration of the two OV datasets used for evaluating INSPIRE.

Rows represent genes and columns represent samples.

We compared INSPIRE with six alternative methods that are scalable to genome-wide data (**Table 1; Appendix Figure 2**). The runtime of all the other methods when $p = 3,000$ is >10 hours, which means that running these methods on genome-wide data would be very inefficient. 8,234 genes are presented in both datasets (rows in the blue-shaded region in **Figure 11**). As a preprocessing step, we standardized each dataset so that each gene has zero mean and unit variance across the samples within each dataset. We used $k = 91$, where k is the number of modules, as selected by BIC on the k -means clustering applied to the imputed data matrix. We also present the results when $k = 182$ based on the biological plausibility of having on average 100 genes per module, in order to show that the outperformance of INSPIRE does not depend on one specific k value (**Appendix Figure 6-Appendix Figure 7**).

In the next three subsections, we show the results of the following evaluations (**Figure 7B**): (1) how well the INSPIRE model fits unseen data measured by test log-likelihood, (2) the statistical significance of the overlap between the learned modules (i.e., gene-module assignment) and

known functional gene sets, and 3) how well the learned modules reflect putative regulatory relationships between TFs and targets based on the ChEA database⁵⁴.

3.3.4 *INSPIRE learns a statistically more robust low-dimensional representation model than alternative approaches*

We first evaluated the learned low-dimensional representation (LDR) model based on the test-set log-likelihoods that measure how well the learned model can explain left-out test data in OV1 through the standard 5-fold cross-validation (CV) tests (see APPENDIX A). We used the test set of OV1 for computing the test log-likelihoods for all compared methods since OV1 contains almost all of the genes contained by either of the datasets. In **Figure 12**, the best average test log-likelihood per sample across the tested λ values is plotted for each method. As can be seen in **Figure 12**, INSPIRE achieves better test log-likelihood than six alternative methods, WGCNA1, InterWGCNA, ImpWGCNA, MGL1, InterMGL and ImpMGL (**Table 1**) for both $k = 91$ chosen by the BIC score (left panel), and $k = 182$, an alternative k value that results in modules with average size of 100 (right panel). Since MGL and INSPIRE may depend on the initialization of the model, the standard deviation across 10 runs of those methods with different initializations are represented by the error bars on the bottom panel in **Figure 12**.

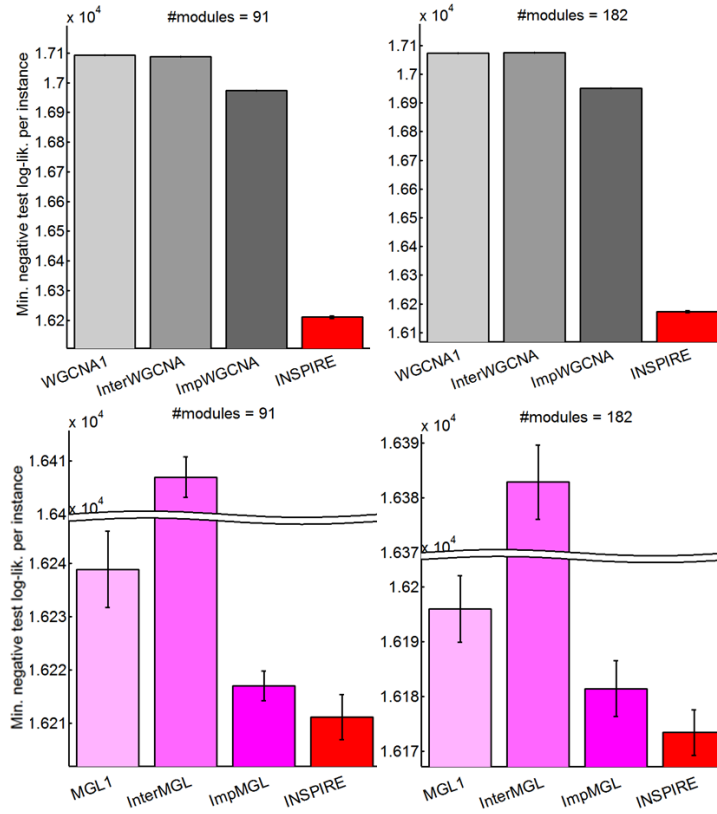


Figure 12. Negative test log-likelihood of INSPIRE vs. alternative models on real data.
For $k = 91$ (left) and $k = 182$ (right), INSPIRE is compared to WGCNA variants (top) and MGL variants (bottom) in terms of the best CV negative test log-likelihood (lower is better) across all tested sparsity tuning parameters (λ).

3.3.5 INSPIRE modules are more significantly enriched for functional gene sets than alternative methods

INSPIRE uses a biologically intuitive low-dimensional representation (LDR) model for expression data, in which genes are assigned to k modules, and each module can be interpreted as biological processes performed by the genes in that module. Thus, whether each module is enriched for genes that are known to be in the same functional categories can be a way to evaluate

the biological relevance of the LDR inferred by INSPIRE. Here, we evaluated INSPIRE based on whether the learned modules are significantly enriched for known pathways from MSigDB²⁶. We compared INSPIRE with six alternative methods, WGCNA1, InterWGCNA, ImpWGCNA, MGL1, InterMGL and ImpMGL (**Table 1**), using $k = 91$ chosen by the BIC score and $k = 182$, an alternative k value that results in modules with average size of 100. For each method, we chose λ that achieves the best CV test log-likelihood, a standard technique⁵¹.

We considered 1,077 GeneSets (pathways) from the C2 collection (curated gene sets from online pathway databases) of the current version of the MSigDB²⁶ based on Reactome⁵⁵, BioCarta and KEGG⁵⁶. We excluded the pathways based on computational predictions from this collection. We computed the significance of the overlap between each GeneSet and each module measured by Fisher's exact test p -value, followed by Bonferroni multiple hypothesis correction. **Figure 13** and **Appendix Figure 6** show results of the functional enrichment analysis for $k = 91$ (chosen based on BIC) and $k = 182$, respectively. In each scatter plot, a larger portion of the dots lie above the diagonal, which implies that the INSPIRE modules are more significantly enriched for known pathways than those inferred by the alternative approaches. This indicates that INSPIRE is better at identifying biologically coherent modules as defined by prior knowledge more accurately than the alternative methods.

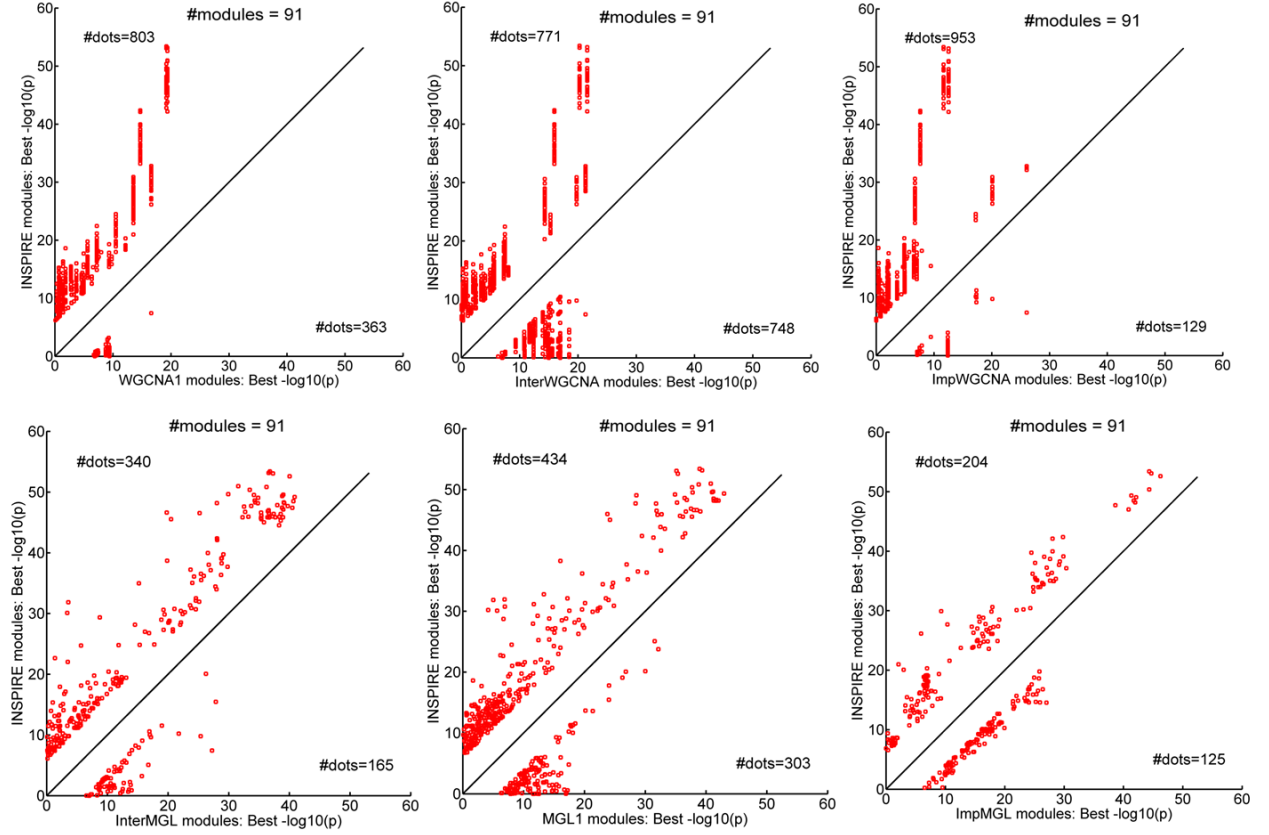


Figure 13. Functional enrichment of the modules identified by INSPIRE vs. alternative methods.

For $k = 91$, INSPIRE (y-axis) is compared to each of the six competing methods (x-axes) in terms of the best $-\log_{10} p$ from the functional enrichment of the learned modules. Each dot is a KEGG, Reactome or BioCarta GeneSet, and only the GeneSets with a Bonferroni corrected $p < .05$ in at least one of the compared two methods are shown on each plot. For MGL variants and INSPIRE, results from multiple runs are shown. We only considered the GeneSets with sufficiently different significance between the two methods, i.e., $|\log_{10} p(\text{INSPIRE}) - \log_{10} p(\text{ALTERNATIVE_METHOD})| \geq \delta$. $\delta = 6$ here and the results were consistent for varying δ .

3.3.6 *INSPIRE modules are more significantly enriched for putative targets of the same TF than the alternative approaches*

As an alternative way to evaluate the biological coherence of the learned modules, we checked how significantly the modules are enriched for the genes that have been shown to be bound by the same TFs. The ChEA database⁵⁴ provides a large collection of TF-target interactions captured in previously published ChIP-chip, ChIP-seq, ChIP-PET and DamID (referred herein as ChIP-X) data. For each of 107 TFs in the ChEA database⁵⁴, we computed the significance of the overlap between each module and each TF's putative targets from ChEA database measured by Fisher's exact test p -value followed by Bonferroni correction. **Figure 14** and **Appendix Figure 7** show the results of our ChEA enrichment analysis for $k = 91$ (chosen based on BIC) and $k = 182$, respectively. In each scatter plot, a much larger portion of the dots lie above the diagonal, which indicates that INSIRE modules are biologically more coherent, i.e., more significantly enriched for putative targets of the same TF. In **Figure 14** and **Appendix Figure 7**, we indicate with a blue dot a TF that resides in the same module as the module that is enriched for the TF's putative targets. We do not expect all dots to be blue-colored (i.e., all TFs being in the same modules as their putative targets), because the protein level of TF may not be correlated with its mRNA expression level. It is still interesting to see that INSPIRE modules are more significantly enriched for the genes that have been shown to be bound by the same TFs in ChIP-X data.

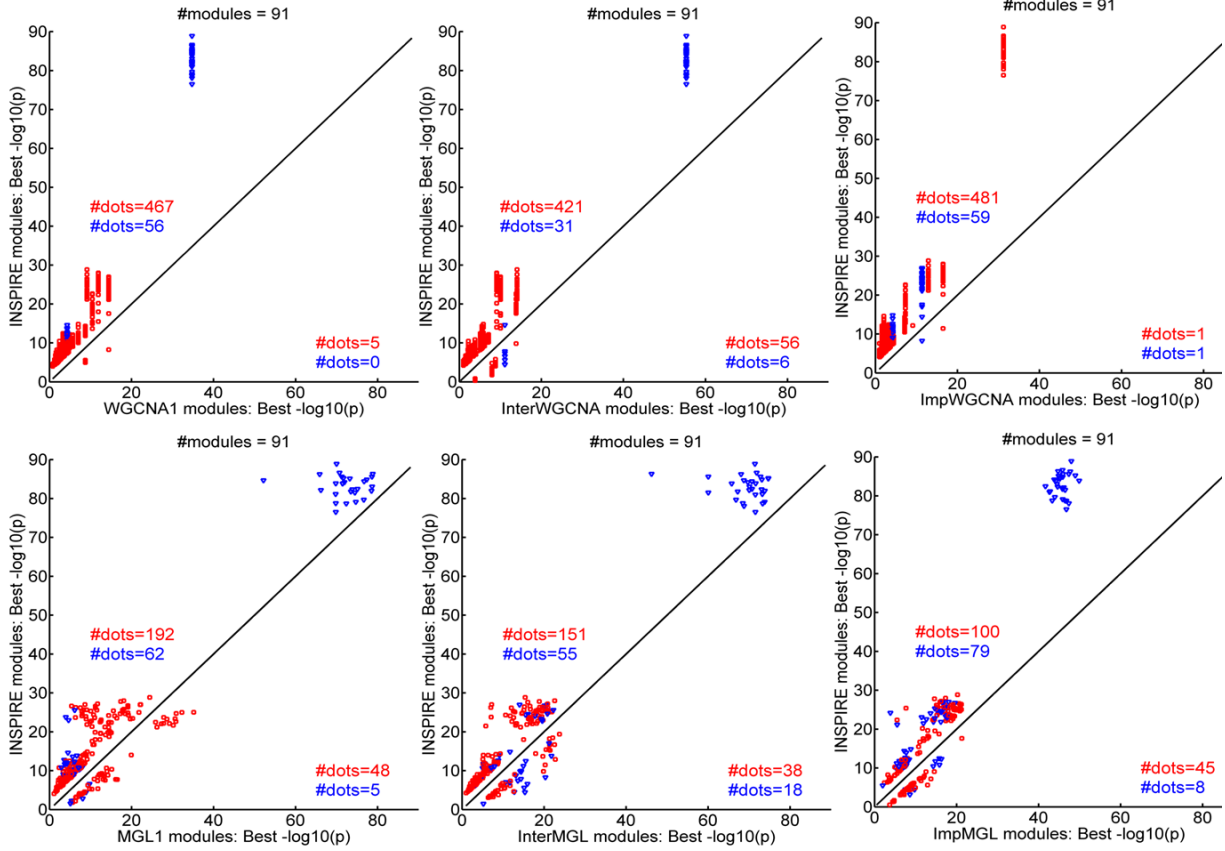


Figure 14. ChEA enrichment of the modules identified by INSPIRE vs. alternative methods.

For $k = 91$, INSPIRE (y-axis) is compared to each of the six competing methods (x-axes) in terms of the best $-\log_{10} p$ from the ChEA enrichment of the learned modules. Each dot is for a gene set composed of a TF and its targets, and only the sets with a Bonferroni corrected $p < .05$ in at least one of the compared two methods are shown on each plot. For MGL variants and INSPIRE, results from multiple runs are shown. We only considered the TFs with sufficiently different significance between the two methods, i.e., $|\log_{10} p(\text{INSPIRE}) - \log_{10} p(\text{ALTERNATIVE_METHOD})| \geq \delta$. $\delta = 3$ here and the results were consistent for varying δ . Each blue dot corresponds to a TF which sits in the INSPIRE module that is significantly enriched for its targets, and each red dot corresponds to a TF which sits in an INSPIRE module different than the one that is significantly enriched for its targets.

3.3.7 Application to nine genome-wide ovarian cancer expression datasets

Encouraged by the in-depth evaluation described above, we applied INSPIRE to nine expression datasets that comprise 1,498 ovarian cancer patient samples downloaded from the TCGA project website and the Gene Expression Omnibus (GEO)⁵⁷ (**Figure 7C**). This corpus of data consists of publicly available transcriptomic characterizations of ovarian cancer across nine distinct studies where gene expression data collected in different studies come from distinct platforms. These data therefore make a perfect corpus to apply the INSPIRE method for a variety of reasons. First, there is a sufficient sample size across studies to resolve distinct modules that are robust across datasets. Second, our method will outperform more naïve approaches by imputing missing genes through leveraging shared structure across the data and will therefore increase the resolution to detect robust modules. Finally, there are known subtypes in ovarian cancer as identified by the TCGA ovarian cancer study²³, and we anticipate that our approach will not only re-identify these subtypes based on the expression of our inferred modules, will also further resolve potential molecular drivers of these subtypes through ancillary analyses of INSPIRE-inferred modules. These ancillary analyses are described below. We repeated our analyses for this application using varying module counts $k = \{90, 129, 181\}$ that correspond to the average number of 200, 140 and 100 genes respectively in each module and for varying sparsity tuning parameters $\lambda = \{0.01, 0.03, 0.1\}$; and we observed that all results were highly robust for the varying values of k and λ . We reported results from our biological analysis for $k = 90$, as selected by BIC for the k -means clustering applied to the imputed data matrix, and $\lambda = 0.1$ which leads to the sparsest network of modules, given that sparsity is of key importance in learning and interpretation of a high-dimensional conditional dependence network.

We evaluated the learned low-dimensional representation (LDR) consisting of 90 modules and the corresponding latent variables using three evaluation metrics:

1) We performed gene set enrichment analysis to characterize each module based on its associated genes (see **Appendix Table 3** for the gene set enrichment analysis results together with the significance).

2) We analyzed the associations between the learned latent variables, each representing a module, and six important phenotypes in cancer, including resectability which was defined by the residual tumor size after surgery, survival, and four histopathological phenotypes manually curated based on the histopathology in the TCGA ovarian cancer data (see **Appendix Table 3**), and we used inferred INSPIRE latent variables as features for predicting those phenotypes. **Figure 15** shows the association between the learned latent variables with the six important phenotypes, and **Figure 16** compares INSPIRE to the following based on the prediction of those phenotypes: (i) principal component analysis (PCA)⁴⁶ – an unsupervised LDR method; (ii) subnetwork analysis⁴¹ – a supervised LDR method; and (iii) all genes when no LDR is learned. The histopathological phenotypes are provided in **Appendix Table 4**, and residual tumor size and survival are available on the TCGA web site.

3) We used the inferred latent variables to identify new subtype definitions in ovarian cancer. We compared INSPIRE subtypes to i) subtypes recently described by the TCGA ovarian cancer study²³, and ii) subtypes learned by a method that uses mutation profiles for the network-based stratification of cancer patients (NBS)⁵⁸, based on how relevant they are to genomic abnormalities in ovarian cancer. Detailed information concerning expression datasets used in the INSPIRE

analysis is presented in **Appendix Table 1**, and the processing of the expression data is described in APPENDIX A.

4) We perform both statistical and biological experiments to show that *HOPX* is a potential molecular driver from tumor-associated stroma in a module that differentiates the patients with increased percent stroma, infiltrative stroma, and desmoplastic stroma.

3.3.8 *Negatively correlated modules show distinct pathways and potential regulatory TFs enrichment*

We emphasize that the key goal of INSPIRE is to reduce the dimensionality of expression data in a biologically intuitive way and in such a way as to capture important dependencies. Given that the gene regulatory network is known to be highly modular⁵⁹ and dimensionality reduction is the key goal, we chose to focus on module-level dependencies rather than gene-level dependencies. The ability to capture the high-level abstraction of the dependencies among gene expression levels is a key goal and advantage of INSPIRE. As a result of the INSPIRE model assumptions, expression of genes in the same INSPIRE module would tend to be positively correlated, and positive correlation in expression levels across patients is an important property - expression activated or deactivated within similar sets of patients. Genes with strong negative correlations are likely to be highly related functionally, however they would have completely different regulatory mechanisms (e.g., different TF binding) and biological interpretation. In **Appendix Figure 8**, we show scatter plots in which each dot corresponds to a GeneSet (from pathway databases or TF binding information) and we plot the maximum $-\log_{10}(p)$ obtained by each model (axis).

Appendix Figure 8 (top) demonstrates that the modules that are strongly negatively correlated with each other show very distinct pathway enrichment (left) as well as TF binding enrichment (right). **Appendix Table 5** shows the significance of enrichment for five negatively correlated module pairs and the biggest absolute correlation for five pathways or TFs for which the highest enrichment difference between the modules is observed.

We also compared between following two models in terms of functional enrichment of the modules: I) two negatively correlated modules are defined as two separate modules as in the original work; and II) instead of the two negatively correlated modules, there is one hypothetical module that contains all genes in the two negatively correlated modules. **Appendix Figure 8** (bottom) compares between model I (y-axis) and model II (x-axis) in terms of functional coherence based on the pathway database (left) and putative TF binding targets (right). Model I reveals more functionally coherent modules than model II, which justifies our modeling assumption that negatively correlated genes need to be in separate modules.

3.3.9 *INSPIRE latent variables are significantly associated with clinical and histologic phenotypes in cancer*

To gain relevant biological insight from ovarian cancer (OV) transcriptome data, we used the 90 inferred latent variables from the INSPIRE model as a lower dimensional representation (LDR) of transcriptomic profiles across patients (**Figure 7C**), that captures robust cross-dataset patterns of gene expression. We evaluated the clinical relevance of these latent variables by measuring the statistical association between these latent variables and histopathological phenotypes of tumor. The morphological interpretation of histologic sections of tumor forms the basis of diagnosis,

aggressiveness assessment, and prognosis prediction. Pathologists examine the tumor diagnostic images based on semi-quantitative histologic phenotypes of the tumor such as invasion pattern and percent stroma to predict the aggressiveness of cancer. Identifying the molecular basis for these histologic phenotypes will advance the understanding of the molecular biology of ovarian cancer. We manually examined five histologic phenotypes for 98 randomly selected patient images from TCGA: percent stroma, percent tumor, vessel formation, stroma type, and invasion pattern (details in APPENDIX A; **Appendix Table 4**). For each pair of a histologic phenotype and a latent variable from the INSPIRE model, we performed Pearson's correlation test that produces a correlation coefficient and a *p*-value. **Appendix Table 3** lists the *p*-values from these association tests of INSPIRE latent variables, with each of the five histologic phenotypes. **Figure 15** shows the correlation of each latent variable with each of the histologic phenotypes. Since percent stroma and percent tumor phenotypes are almost perfectly (anti-) correlated, we only included percent stroma in **Figure 15**. We used *p*-values from a likelihood ratio test for a Cox proportional hazards model to determine the significance of association of a gene with patient survival, and we used *p*-values from Pearson's correlation test for tumor resectability.

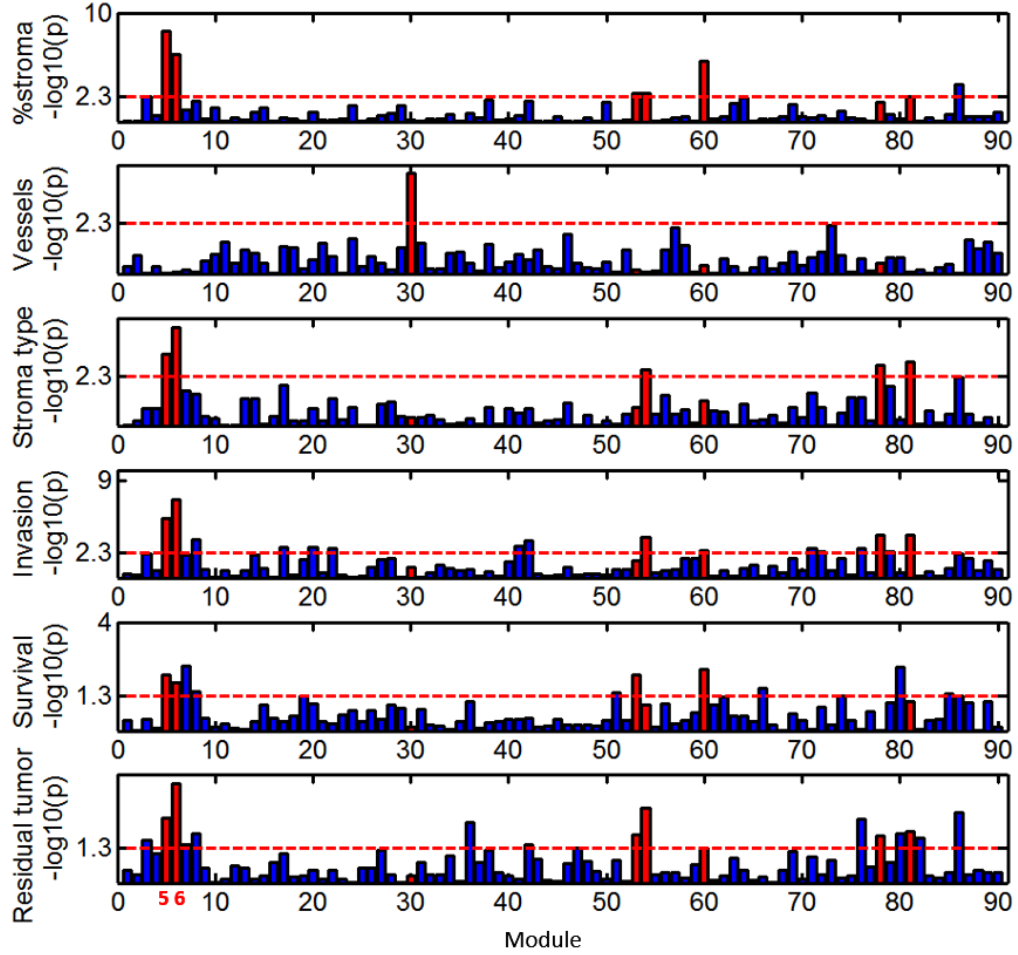


Figure 15. Phenotype-based characterization of the INSPIRE modules.

For each of 90 INSPIRE latent variables (x -axis), the $-\log_{10} p$ of the statistical association is shown (y -axis) for six phenotypes. The p -value threshold (shown by red dotted horizontal lines) is 5×10^{-3} for histological phenotypes, and 5×10^{-2} for clinical phenotypes, which are harder to predict. We highlight modules 5, 6, 53, 54, 60, 78 and 81 that are significantly correlated with at least three of the six phenotypes by red. We also highlight module 30 by red since it is the only module that has a significant correlation with the vessel formation phenotype. Modules 5 and 6 achieve the first or second rank in terms of the significance of correlation with five of the six phenotypes.

Modules 5 and 6 show high correlations with the histopathological phenotypes, such as percent stroma, stroma type and invasion pattern. As shown in **Figure 15**, those modules are also

associated with patient survival and tumor resectability. We observed that the quantity of residual tumor after surgery is positively correlated with the amount of tumor-associated stroma, where increased residual tumor, i.e. low resectability, is an important and a previously known indicator of poor patient prognosis. Although the latent variables of module 5 and module 6 show high expression correlation (the correlation coefficient between the module 5 latent variable and the module 6 latent variable is 0.84), these two modules are functionally fairly different. **Appendix Figure 9** compares module 5 and module 6 in terms of the pathways and putative TF targets that are enriched in these modules. There are handful of dots that are distant from the diagonal line implying that module 5 and module 6 exhibit several distinct biological properties. In **Appendix Table 6**, the significance of enrichment from modules 5 and 6 are listed for five pathways or TFs for which the highest enrichment difference between the modules is observed.

To examine the differences between module 5 and module 6 in terms of the phenotypes associated with them, we compared between the following two models in an experiment where the latent variables are used as *features* in predicting six different phenotypes (percent stroma, stroma type, vessel formation, invasion pattern, resectability and survival): I) module 5 and module 6 exist as two separate modules as in the original work; and II) instead of module 5 and module 6, there is one hypothetical module that contains all genes in modules 5 and 6. As shown in **Appendix Table 7**, we observed that modules 5 and 6 are significantly predictive of distinct sets of phenotypes, and interestingly, either module 5 or module 6 is always better in terms of predictability of phenotypes than the hypothetical module containing all genes in modules 5 and 6, which means model I is a better predictor of all six phenotypes than model II. Thus, even if modules 5 and 6 are highly correlated with each other, the genes in these modules need to be separated into the two modules.

3.3.10 *INSPIRE latent variables are more predictive of clinical and histologic phenotypes in cancer than other kinds of LDRs and all genes*

Many biological processes are performed by a group of genes rather than individual genes, and as a result, many *complex* phenotypes and clinical outcomes can be explained based on module activity levels rather than individual genes. Moreover, expression level of an individual gene is often noisy and even if it was not, it still may not be perfectly correlated with a protein level of a true regulator for a phenotype.

To test this hypothesis and further demonstrate the effectiveness of INSPIRE as an LDR of gene expression data, we used the INSPIRE latent variables as *features* in prediction tasks, and we compared INSPIRE with the following methods: i) principal component analysis (PCA)⁴⁶ – the most widely used unsupervised LDR method; ii) subnetwork analysis⁴¹ – a powerful supervised LDR method that extracts network markers; and iii) all genes when no LDR is learned. The subnetwork analysis method⁴¹ learns small sub-networks of genes in a given large PPI network, based on expression data and a particular prediction task. For example, for a stroma type prediction (fibroblast/desmoplastic), it learns sub-networks of genes in a given PPI network such that the average expression level of each sub-network significantly differentiates the two patient groups based on the classes of stroma type. This method is a supervised method in that the subnetworks are learned such that they can explain a particular phenotype well. On the other hand, INSPIRE is an unsupervised method in that the result does not depend on a particular prediction task. Each of INSPIRE latent variables, subnetworks, principal components, and all genes is considered as a set of features in predicting six different phenotypes; percent stroma, stroma type, vessel formation, invasion pattern, resectability and survival (see APPENDIX A for details). The result of the

comparison shows that the features learned by INSPIRE show the best prediction performance measured among all methods considered (**Figure 16**). This result strengthens our claim that the INSPIRE latent variables provide informative lower-dimensional features for prediction tasks.

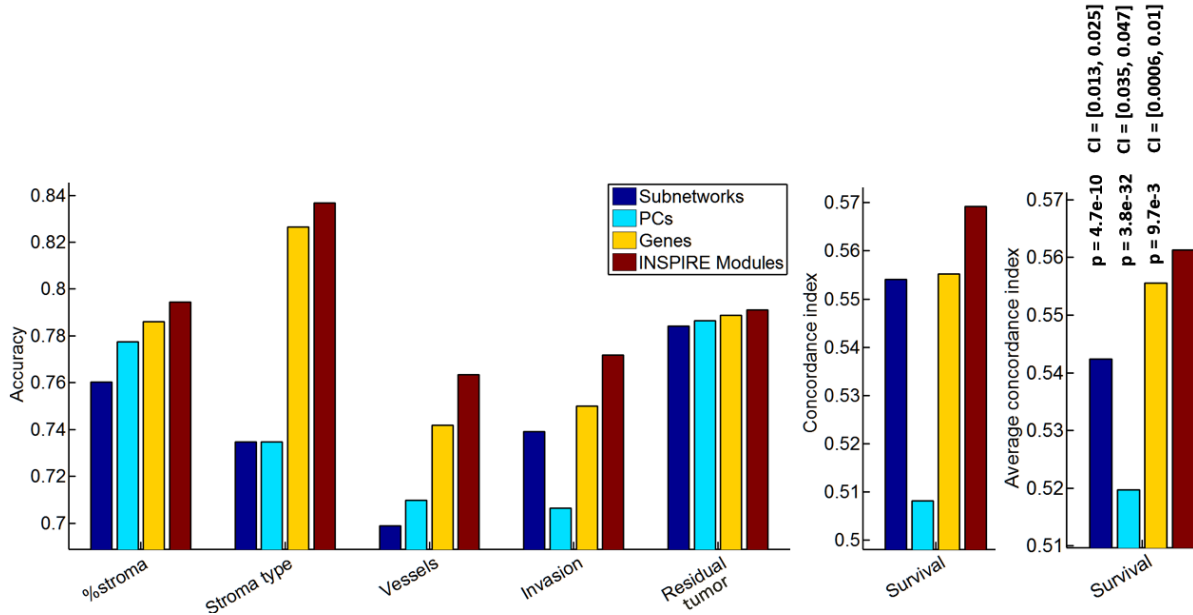


Figure 16. Prediction performance of INSPIRE vs. four alternative methods for six prediction tasks in a cross-validation setting.

In all bar charts except for the last one, a single accuracy (or concordance index for survival) is reported based on the predicted phenotype vector formed by pulling together the predictions for all folds. For survival, an additional analysis is presented where the mean concordance index is reported across all 500 folds in 50 rounds of 10-fold CV tests. For this bar chart (right), the p -value from the Wilcoxon signed rank test and the 95% confidence interval for the mean of a standard normal distribution fitted to the difference are reported to show the significance of the difference between INSPIRE and its alternative method.

Because INSPIRE groups genes in multiple datasets into a set of modules, most modules may include a significant number of genes whose expression is not correlated with the predicted phenotype. To examine the effect of those genes in phenotype prediction tasks, we generated four

hypothetical module sets by excluding 20%, 40%, 60%, and 80% of the genes whose expression levels in training samples are least significantly associated with the respective phenotype from each of 90 modules, and repeated the phenotype prediction experiments for those four hypothetical module sets. **Appendix Table 8** shows that the original INSPIRE latent variables which correspond to the module set including non-discriminative genes perform the best and in most cases, the performance even decreases when top 20% of the most discriminative genes are left. This result indicates that latent variables resulting from the contribution of all genes make robust features informative of the phenotypes.

3.3.11 *Subtypes inferred based on INSPIRE latent variables are highly relevant to genomic abnormalities in ovarian cancer*

Cancer is a heterogeneous disease with multiple distinct genetic drivers, where identifying subtypes of cancer relevant to potential genetic drivers is a primary goal of the field of cancer biology. Here, we cluster ovarian cancer patients from the TCGA study²³ (560 samples) into four subtypes by using the latent variables learned by the INSPIRE method as *features for clustering* patients (details in APPENDIX A). **Appendix Table 9** lists the assignment of the patients in the TCGA ovarian cancer data to four INSPIRE subtypes.

To examine the relevance of the INSPIRE-based subtypes to the potential drivers of ovarian tumor, we checked the significance of the association between the subtypes and the copy-number variation (CNV) of genes, an important genomic abnormality that can drive cancer (**Figure 17** and **Appendix Figure 10**). We focused on CNV for this test instead of mutation since ovarian cancer has been characterized as a c-class cancer (as opposed to m-class, where ‘m’ represents mutation) in which CNV is more prevalent than mutations⁶⁰. For each CNV (as

quantified by the CNV level), we performed a multivariate linear regression using the INSPIRE subtypes, where we computed a p -value (from the regression f -statistic) to ascertain how well the INSPIRE subtype regression model fits a given CNV. We then compared the number of CNVs with significant INSPIRE p -values (determined by varying thresholds; see **Figure 17**) to the number of CNVs with significant p -values from the following two approaches: (1) the subtypes learned by using a method that uses mutation profiles for the network-based stratification (NBS) of cancer patients⁵⁸, and (2) the subtypes inferred from a recent TCGA ovarian cancer study²³. **Figure 17** shows that INSPIRE results in subtypes that are more associated with CNV based genomic abnormalities than alternative approaches. In **Appendix Figure 10**, we show the comparison for varying numbers of modules (k), for varying sparsity tuning parameters (λ), and for varying p -value thresholds, which shows that the results are robust to varying hyperparameters. **Figure 17** and **Appendix Figure 10** indicate that INSPIRE further resolves subtypes as defined by the potential genomic drivers of ovarian cancer when compared to alternative approaches. In APPENDIX A, we list the CNV levels that are significantly correlated with each of the four subtypes. The enrichments of those CNV levels with the MSigDB²⁶ C2 (curated gene sets) categories and the corresponding $-\log_{10} p$ are also listed for each subtype.

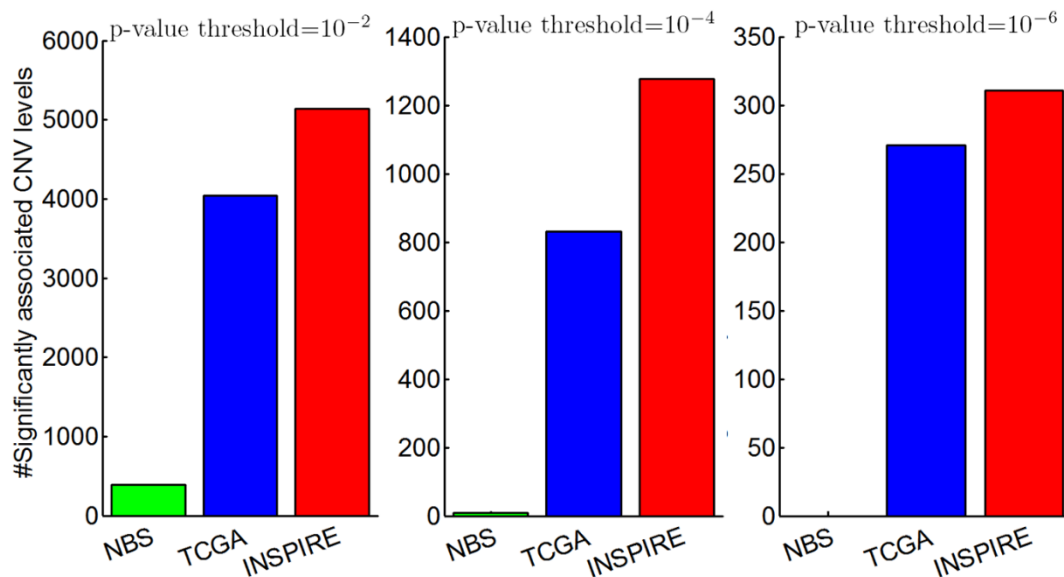


Figure 17. Association of INSPIRE subtypes with the CNV levels.

For three different Pearson's correlation p-value thresholds (10^{-2} , 10^{-4} , 10^{-6} respectively from left to right), the number of CNV levels that are significantly associated with the learned subtypes are shown for two published methods and INSPIRE.

3.3.12 Subtypes revealed by INSPIRE and their relationships with the TCGA subtypes

Figure 18 reveals a sub-network learned by modules from an INSPIRE model using parameters $\lambda = .1$ and $k = 90$ (chosen based on BIC). This subnetwork contains modules that are differentially expressed in one of the four subtypes, as represented by the heatmaps in **Figure 18**. The differentially expressed modules, termed *marker modules*, are determined for each subtype by comparing the subtype versus the other three subtypes, using the Significance Analysis of Microarrays (SAM) algorithm⁶¹ implemented in the R package *siggenes*. **Appendix Table 10** lists the enrichment of the marker modules with the MSigDB²⁶ C5 (GO gene sets) and the corresponding $-\log_{10} p$. We observed that the set of marker modules (**Appendix Table 10**) have

a significant overlap ($p = 2.4 \times 10^{-3}$) with the set of modules that have significant associations with at least three of the six phenotypes (the modules colored in red in **Figure 15** and **Appendix Table 3**, except module 30). Not surprisingly, the INSPIRE subtypes show diverse histologic features across subtypes, and we accordingly termed the INSPIRE subtypes ‘vascular’, ‘stromal’, ‘immunoresponse’, and ‘fibrous’. See **Figure 18** and **Appendix Table 10**, where the marker modules for the vascular, stromal, immunoresponse, and fibrous subtypes are colored in green, blue, red and orange, respectively.

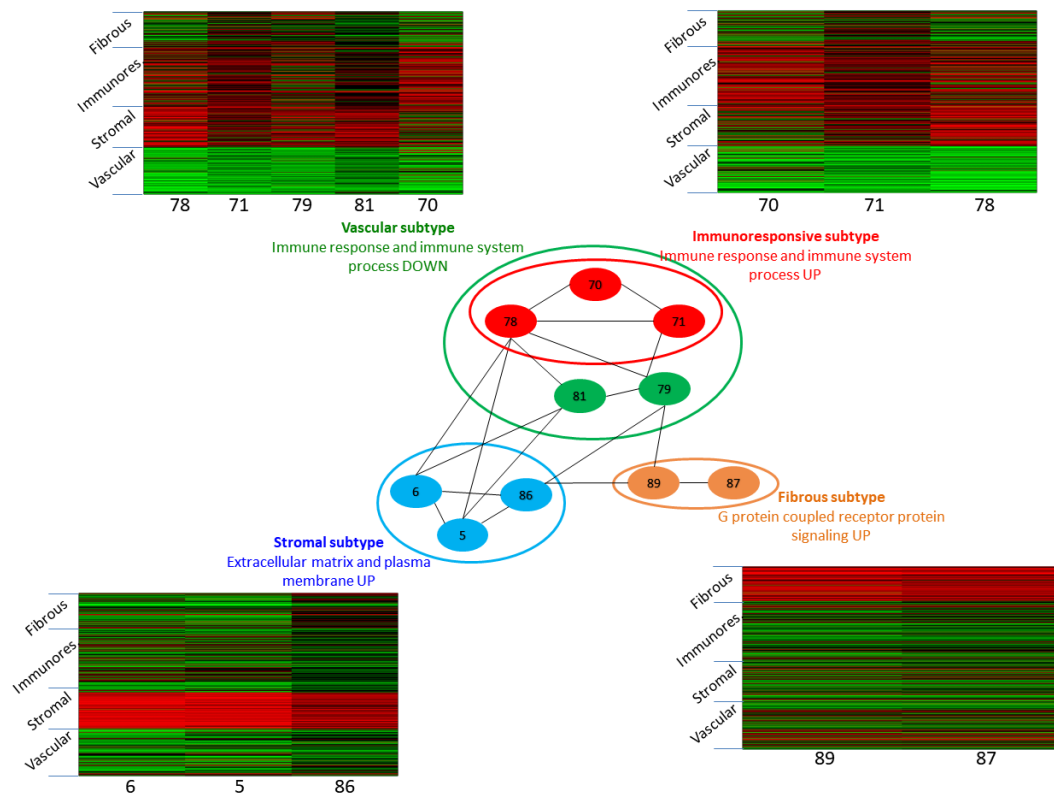


Figure 18. The modules that differentiate the subtypes based on INSPIRE features and the interactions among these modules as learned by INSPIRE.

The modules are grouped and colored according to the subtypes they differentiate. Next to each one of the four module groups, there is the heat map of the features corresponding to the modules in this module group.

Appendix Table 11 shows a confusion matrix that describes the overlap between the INSPIRE subtype assignments and the TCGA subtype assignments²³ together with the *p*-values for the significance of the overlap for the highly-overlapping subtypes. There is a more significant overlap for the vascular-proliferative pairs and stromal-mesenchymal pairs, which implies that the proliferative-like and mesenchymal-like subtypes are highly conserved across different OV datasets, which is consistent with the finding of Way et al.⁶². Although the INSPIRE subtypes have a statistically significant overlap with the TCGA subtypes, the INSPIRE subtypes show much stronger association with genomic abnormalities, as mentioned above (see **Figure 17**). We further include the description of the stromal subtype here since it is characterized by high expression of modules 5 and 6, which are strongly associated with the six important phenotypes in cancer (**Figure 15**). See APPENDIX A for the characterization of the other three (‘vascular’, ‘immunoresponse’ and ‘fibrous’) subtypes.

The stromal subtype is characterized by high expression of modules 5, 6 and 86 (**Figure 18**) and associated increased percent stroma, infiltrative growth pattern, and desmoplastic stroma (**Appendix Figure 11**). Modules 5 and 6 are significantly enriched for proteinaceous extracellular matrix gene sets (**Appendix Table 10**), which is likely due to increased percent stroma. In **Figure 18**, there are quite a few edges between modules associated with the vascular subtype and those associated with stromal subtype, which suggests a strong association between the increased stromal components and neovascularization of the tumor. This likely reflects the known tumor neovascular niche in cancer that involves proangiogenic factor release from tumor stroma along with the vasculature itself⁶³. This is supported by multipotent mesenchymal stromal cells having unique immunoregulatory and regenerative properties⁶⁴. A substantial amount of the tumor stroma

is composed of immune cells, and the net effect of the interactions between these various immune cell types and the stroma participates in determining anti-tumor immunity and neovascularization potential⁶⁵. We note that the immune system modules 78 and 81 that are connected to extracellular matrix modules 5 and 6 are also up-regulated in the stromal subtype (**Figure 18**). Stromal subtype is a significant predictor of poor patient survival (Cox proportional hazards model log-rank $p = 8.8 \times 10^{-2}$) with a median survival of 914 days. Cancers associated with a reactive stroma is typically diagnostic of poor prognosis⁶⁶, and we observed that median survival of the stromal subtype is the smallest among all subtypes. Stromal subtype has a significant overlap ($p = 1.03 \times 10^{-35}$) with the mesenchymal subtype discovered by TCGA²³ (**Appendix Table 11**).

3.3.13 *INSPIRE provides novel insights into molecular basis for ovarian tumor resectability*

Riester et al. identified POSTN as a candidate marker for tumor resectability in ovarian cancer⁶⁷, where the resectability phenotype was defined by the residual tumor size after surgery. The authors showed that high POSTN expression is strongly associated with poor tumor resectability, even more so than a multi-gene model chosen by leave one out cross-validation across 1,061 samples in 8 datasets including the TCGA²³ and Tothill²⁴ datasets. POSTN is a member of module 6 that shows the most significant association with resectability among all 90 modules (**Figure 15**). We, therefore, compared our supervised prediction model using the INSPIRE latent variables corresponding to modules 5 and/or 6 to a model that contains just POSTN to determine whether the genes in module 5 and the genes in module 6 other than POSTN provide any information to the prediction of resectability in addition to the information provided by POSTN expression. We observed that when training on TCGA data²³ including the clinical covariates, the models trained

using (1) modules 5 and 6 together, (2) module 6 only, (3) module 5 and POSTN together, and (4) module 5 only, outperformed the model with the known marker for resectability, POSTN, when tested in the Tothill²⁴ dataset (see AUC values in **Figure 19**). TCGA data²³ was used for training because of its large sample size. Tothill²⁴ was used for testing, because it has the largest sample size except TCGA data (**Appendix Table 1**) and contains the most fine-grained information on the residual tumor size. Additionally, the proportion of optimally and sub-optimally debulked patients was similar between TCGA and Tothill data. We used a stringent definition of resectability (0 cm vs. >0 cm) (see APPENDIX A).

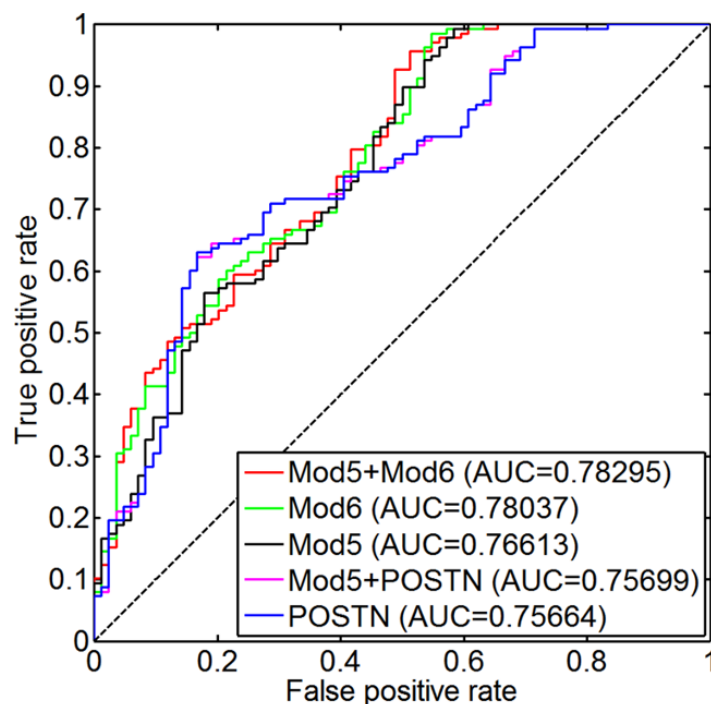


Figure 19. Performance of different models for resectability prediction.

ROC curves of the supervised models are shown for resectability prediction trained in TCGA and tested in Tothill data. Different combinations of POSTN and the INSPIRE features corresponding to modules 5 and 6 are used for training each model. The clinical covariates age and stage are also included in all models. AUC of each model is shown in the legend.

Since module 6 contains POSTN, outperformance of (1)-(3) means that modules 5 and 6 representing the expression of genes in module 5 and/or module 6, which are significantly predictive of stromal histology features and resectability, add information to the prediction of resectability by POSTN in a cross-dataset analysis. Outperformance of (4) means that the module 5 representing the gene expression levels in module 5, which does not contain POSTN, is a better predictor of resectability than POSTN. When we repeated this experiment with no clinical covariates (age and stage) in the training, the models including module 6 outperformed the model that includes only POSTN, which means the genes in module 6 other than POSTN add information to the prediction of resectability by POSTN (see AUC values in **Appendix Figure 3**). Modules 5 and 6 with strong stromal and mesenchymal properties (see below) provide potential novel molecular basis for tumor resectability.

3.3.14 *INSPIRE modules and the conditional dependence network among them*

Here, we discuss the modules that show significant correlations with many of the histological and clinical phenotypes in the TCGA ovarian cancer data or that achieve the only significant correlation with a phenotype among all modules (see **Figure 15** and **Appendix Table 3**).

Module 5 contains known EMT inducers *ZEB1*, *SNAI2* and *TCF4* (*E2.2*)⁶⁸, as well as multiple other genes known to be important in focal adhesion⁶⁹, extracellular matrix interaction⁷⁰, extracellular matrix organization⁷¹, and markers of cancer-associated fibroblasts (*PDGFRB*, *PDGFRA*)⁷² (see **Appendix Table 3**). Similarly, module 6 contains EMT inducer *TWIST1*⁶⁸, many extracellular matrix genes, as well as genes associated with senescence and autophagy, collagen genes, and the well validated predictor of tumor resectability, *POSTN*⁶⁷ (see **Appendix Table**

3). These two modules are prime candidates for genes driving EMT associated tumor aggression. Although modules 5 and 6 have many shared GO categories and pathways, data are consistent with the hypothesis that they may represent fairly different biological processes (**Appendix Figure 9**). When we combined these two modules and used one latent variable that represents the two modules, the overall prediction results became worse (**Appendix Table 7**).

While modules 5 and 6 contain known drivers of EMT and extracellular matrix genes and these modules are also associated with tumor-associated stroma/mesenchymal phenotypes, we found other modules with significant correlations with most of the histological and clinical phenotypes. Additionally, an active area of research in cancer biology is to identify pathways and genes driving tumor aggression. This includes genes associated with cancer stem cells (i.e., tumor-initiating cells)^{73–76}. Module 78 contains genes indicative of hematopoietic cell lineages likely because it includes many innate immune response genes, as well as multiple innate immune response signaling pathways including cytokine-cytokine receptors, toll like receptors, and TCR signaling. Module 78 also contains a known EMT inducer ZEB2⁶⁸. This indicates that module 78 may capture aspects of tumor associated inflammation, a known contributing factor to EMT⁷⁷. Module 81 includes genes that regulate the MAPK and ERK cascades, signal transduction pathways that are known to be upstream of multiple oncogenic processes⁷⁸. Module 54 represents genes involved in pro-apoptotic and cell cycle regulation. GADD45 genes, known to be upstream of JNK signaling⁷⁹, are present along with JUN and FOS. In addition, this module contains KLF4 and KLF6, which like GADD45, are known to repress cell cycle arrest and associated cyclin-dependent kinase inhibitors⁸⁰. Modules 30 and 54 are may reflect the likely metabolic shift that cancer cells undergo as these modules are enriched in metabolic and biosynthesis pathways. When considering these modules jointly, we get a picture of multiple processes (**Appendix Figure 12**),

and potential tumor cell subpopulations, that populate the tumor microenvironment and perpetuate aggressive tumor states in sub-populations of patients.

One of the advantages of the INSPIRE framework over naïve clustering algorithms is that it suggests potentially biologically relevant interactions or couplings between the modules. These interactions can be used to motivate higher-level hypotheses about the coupling of disease specific processes.

3.3.15 *INSPIRE reveals a previously unknown stroma-associated marker HOPX*

Given the association of the genes in modules 5 and 6 with aggressive stroma and patient prognosis and the significance of modules 5 and 6 in differentiating the stromal subtype, we were interested in understanding if modules 5 and 6 capture prognostic signatures that generalize across other cancers. Prognostic genes are more likely to be shared by distinct tumor types than would be expected by random chance likely because of prognostic mechanisms that generalize across cancers (e.g. metastatic potential or immune system evasion), and conversely, cancer-specific prognostic genes are less frequent than would be expected by random chance⁸¹. Therefore, to further annotate modules 5 and 6, we performed a pan-cancer analysis to check whether the genes contained in those modules are significantly associated with survival in six publicly available datasets^{24,35,76,82–84} from five cancer types: ovarian cancer, breast cancer, acute myeloid leukemia, glioblastoma and lung cancer (see **Appendix Table 12** for the details of these datasets). We used p -values from the likelihood ratio test for a Cox proportional hazards model to determine the significance of association of a gene with patient survival, and we considered a p -value ≤ 0.05 to be significant. We observed that the genes in module 5 and the genes in module 6 are significantly

associated with survival in at least three of the six datasets (Fisher's test statistic p -value = 1.68×10^{-8} for module 5 and = 4.44×10^{-8} for module 6). For breast cancer, we used the Osloval (the test data) but not Metabric (the training data with 1981 samples from the same study³⁵ with Osloval) for breast cancer because we need the sample sizes to be similar across datasets such that the meta-analysis is not dominated by a single cancer type.

To further investigate the specific genes that are associated with patient survival across cancer types in these modules, we computed a combined p -value statistic using Fisher's combined probability test for the association of each gene with patient survival in a meta-analysis of the six datasets from these five cancer types. *HOPX*, which is in module 5 achieved the lowest combined p -value among all genes in module 5 or module 6, and the third lowest combined p -value genome-wide ($p = 1.32 \times 10^{-10}$). The top two genes that yield smaller p -values than *HOPX* genome-wide are *CD109* ($p = 2.49 \times 10^{-11}$) and *SKAP2* ($p = 3.55 \times 10^{-11}$) neither of which is in the module 5 or module 6 (**Figure 20**). As shown in the previous sections, module 5 (containing 183 genes) is highly associated with percent stroma (**Figure 15**), and is significantly enriched ($p = 8 \times 10^{-5}$) for the known drivers of EMT that has been shown to contribute to poor patient survival. Not all 183 genes in module 5 would play a key role in the formation of tumor-associated stroma or EMT, and in fact, many of the genes in module 5 would simply have correlated expression pattern with key genes in these processes. We hypothesize that such genes have robust association with survival enough to be conserved across different cancer types, given the previously known association between tumor-associated stroma and patient survival. We note that known EMT drivers *ZEB1*, *SNAI2* and *TCF4* in module 5 have significant associations with survival in our pan-cancer analysis (p -values 8.5×10^{-6} , 5.3×10^{-4} , 1.3×10^{-3} and rankings 153, 749 and 1098

respectively out of 11119 total genes). Thus, our pan-cancer analysis that highlights *HOPX* in module 5 led to us to consider *HOPX* as a potential molecular marker strongly associated with percent stroma and tumor aggression. Additionally, *HOPX* is one of the 15 genes in module 5 (out of 183 genes) that have been classified as ‘candidate regulators’⁸⁵. Gentles et al. have defined a list of about 3,000 genes as candidate regulators, those that have a potential regulatory role in the broad sense (not specific to cancer): TFs, signaling proteins and translational initiation factors that may have transcriptional impact⁸⁵. This implies that *HOPX* could be a regulator in the stroma-associated processes.

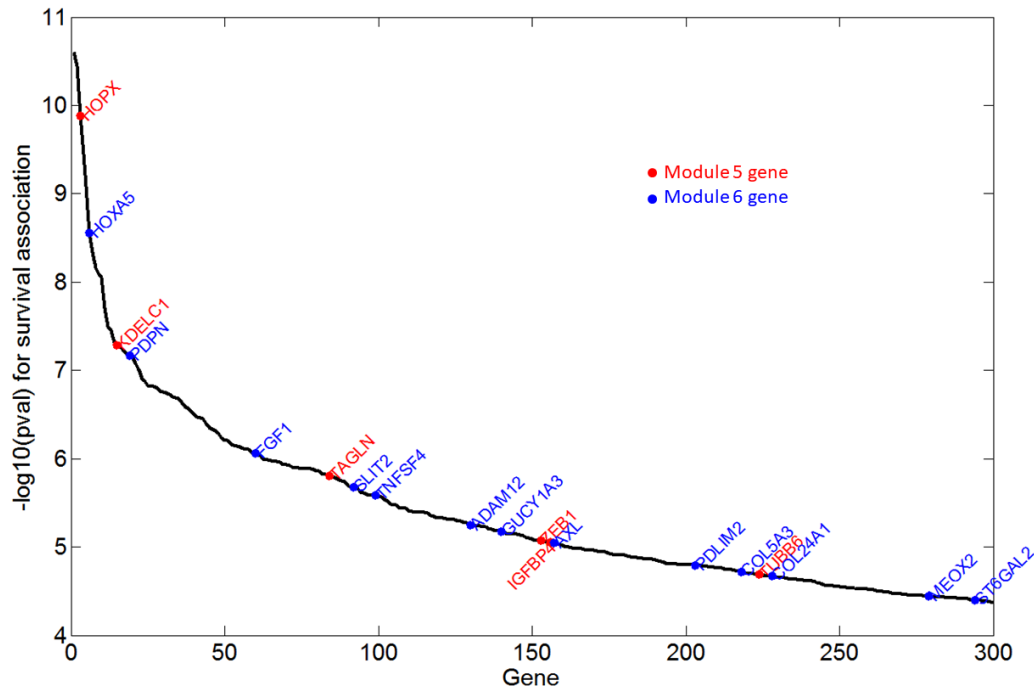


Figure 20. Pan-cancer survival association of modules 5 and 6 genes.

Fisher’s combined p-values for survival (y-axis) are shown for the top 300 genes (x-axis) which achieve the most significant survival association in the pan-cancer survival analysis.

Module 5 genes are shown by red and module 6 genes are shown by blue.

3.3.16 *HOPX is a putative driver for the tumor-associated stroma/mesenchymal module (module 5)*

HOPX is an unusual HOX protein that does not contain a DNA binding domain, and has been implicated in multiple aspects of cardiac and skeletal muscle development through recruitment of histone deacetylases^{86–88}. It has been suggested to have tumor suppressive function in other cancer types^{89–91}, which confounds how its expression in OV is associated with several poor outcomes. This may also reflect different roles for *HOPX* in ovarian tumor-associated stromal tissue.

Previous studies characterize *HOPX* as a mediator of canonical Wnt and Bmp signaling, and may play key roles in maintaining a stem cell-like state⁹². In our further analysis of *HOPX*, we observed that *HOPX* is one of the top candidate expression regulators for ovarian cancer^{85,93}. To understand how *HOPX* is associated with the genes in the tumor-associated stroma/mesenchymal module (module 5), we compared these genes with those down-regulated in *HOPX*^{−/−} mice compared to *HOPX*^{+/-} control mice⁹² and found a significant enrichment based on Fisher's exact test ($p = 1.5 \times 10^{-3}$). Those two results together suggest *HOPX* is a good candidate driver for the tumor-associated stroma/mesenchymal module, as many of the other genes in module are putative downstream targets of *HOPX*, either directly or indirectly. **Appendix Figure 14** shows the enrichment p -value and the fold enrichment of the genes in the tumor-associated stroma/mesenchymal module with the down-regulated genes in *HOPX*^{−/−} mice for varying fold change of expression of the down-regulated genes (x-axis).

Furthermore, genes down-regulated in *HOPX*^{−/−} mice after addition of XAV939, a potent inhibitor of Wnt signaling to *HOPX*^{−/−} mice⁹², are even more significantly enriched ($p = 1 \times 10^{-8}$) for genes in the tumor-associated stroma/mesenchymal module. The *HOPX* protein is a potent Wnt

inhibitor⁹², therefore in the HOPX^{-/-} mice Wnt is activated, and genes inhibited by Wnt are also turned off. When the Wnt inhibitor is applied to the HOPX^{-/-} mice the genes inhibited by Wnt are no longer turned off, and the down-regulated genes are more specific to genes specifically activated by HOPX, instead of being a mixture of genes activated by HOPX and inhibited by Wnt. In addition, it is not surprising to see a higher enrichment upon Wnt inhibition, because canonical Wnt signaling has been implicated in the regulation of the stromal activity of MSCs^{94,95}. **Appendix Figure 15** shows the enrichment p-value and the fold enrichment of the stroma/mesenchymal module genes that are down-regulated in Wnt-inhibited HOPX^{-/-} mice for varying fold change of expression of the down-regulated genes (x-axis).

These results suggest that the genes in the tumor-associated stroma/mesenchymal module which are down-regulated in both HOPX^{-/-} mice and Wnt-inhibited HOPX^{-/-} mice are good candidates as downstream targets of HOPX. **Figure 21** shows those 32 potential targets of HOPX. The purple-colored genes are the targets that are down-regulated in HOPX^{-/-}, and their expression does not change significantly ($|FC \text{ change}| \leq 0.55$) upon Wnt inhibition. On the other hand, the red-colored genes are the targets of HOPX which are down-regulated in HOPX^{-/-}, and they are down-regulated further upon Wnt inhibition ($|FC \text{ change}| \geq 0.93$). It is highly likely that the expression of the red-colored genes in **Figure 21** are driven by both HOPX and Wnt signaling pathway. We note that *HOPX* is, therefore, a potential driver for *SNAI2*, which is involved in EMT⁹⁶ and *AEBP1*, which is a *stromal* adipocyte enhancer-binding protein.

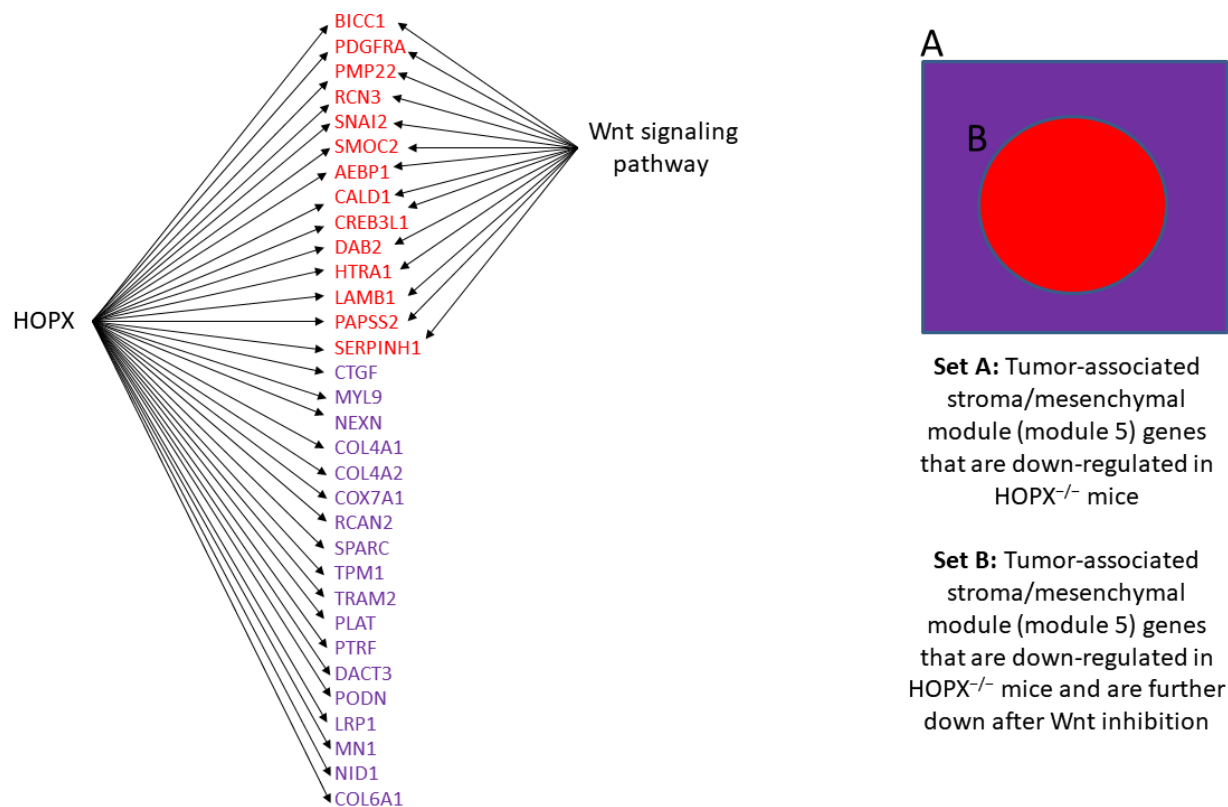


Figure 21. Potential targets of HOPX.

A total of 32 genes that are potential targets of HOPX are shown. The purple-colored genes are the targets whose expression does not depend on Wnt signaling, and the red-colored genes are the targets of HOPX which are down-regulated in HOPX^{-/-}, and further down upon Wnt inhibition in HOPX^{-/-}. It is highly likely that the expression of the red-colored genes is driven by both HOPX and Wnt signaling pathway.

3.3.17 HOPX is a molecular marker of aggressive tumor stroma

To further disentangle the molecular underpinnings of the tumor-associated stroma/mesenchymal module, we stained tumor sections with antibodies against HOPX. We co-stained with E cadherin, a tumor epithelial cell marker. Patient samples were selected based on patient survival and optimal

debulking (see APPENDIX A for details). As shown in **Figure 22**, there is no overlap between HOPX and E cadherin. Given localization outside of epithelial regions, we tested if there was overlap with stromal tissue. To do so, we co-stained with CD73, a known mesenchymal stem cell (MSC) marker, as MSCs play an important role in the generation of cancer-associated fibroblasts and stroma⁹⁷. Combining these results with corresponding tumor sections with H-E staining indicate that HOPX and CD73 are uniquely localized to the tumor stroma. Representative images depicting HOPX, CD73 and HOPX, E cadherin staining for additional samples are shown in **Appendix Figure 13**.

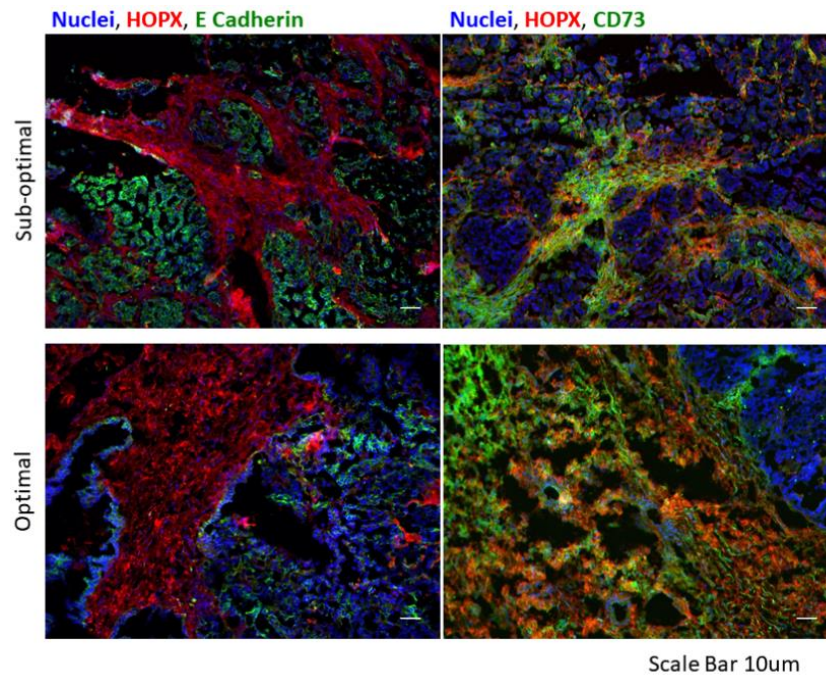


Figure 22. Fluorescent staining of ovarian tumors from sub-optimally bulked and optimally debulked patients.

Each row is a single patient. HOPX is localized to the stroma and does not overlap with E Cadherin positive cancer cells. HOPX does however overlap with CD73, a mesenchymal stem cell marker.

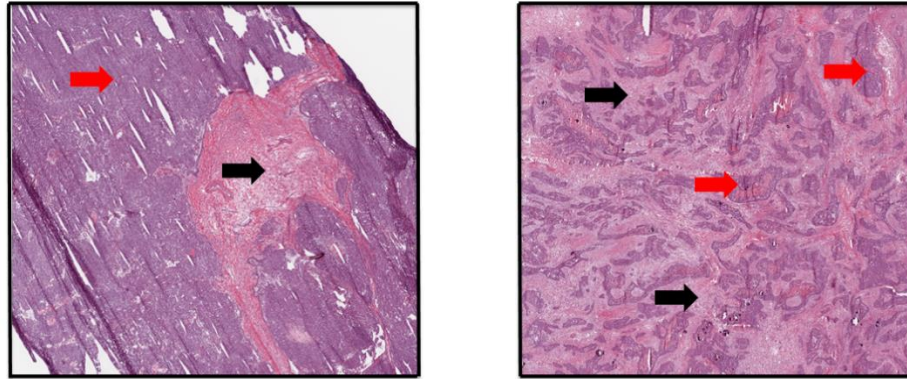


Figure 23. Histopathology of a patient with low vs. high HOPX expression.

Left: Expansile growth pattern of high-grade serous carcinoma associated with optimal resectability and low HOPX expression from the TCGA ovarian cancer study. Note high percentage of carcinoma (red arrow) and low percentage of stroma (black arrow). Hematoxylin and Eosin, 100X. **Right:** Infiltrative growth pattern of high grade serous carcinoma associated with low resectability and high HOPX expression from the TCGA ovarian cancer study. Note high percentage of stroma (black arrows) compared with carcinoma (red arrows). Hematoxylin and Eosin, 100X.

It is not surprising that HOPX potentially marks MSCs. Several recent studies have shown HOPX to be associated with other stem cell populations and to play a role in their hierarchy and more importantly maintenance of a stem-cell like state through integration of canonical Wnt and Bmp signaling^{92,98,99}. Nonetheless, these results indicate HOPX as a putative novel marker for tumor-associated MSCs. In patients with poor tumor resectability and prognosis, CD73 and HOPX expression is riddled throughout tumor tissue (**Figure 22**). A typical patient with optimal resectability and low HOPX expression is shown on the left in **Figure 23**, whereas a patient with low resectability and high HOPX expression is shown on the right. As can be seen, the tumors with strong evidence of HOPX have very distinct histopathology from those without. This aggressive stromal tumor phenotype provides evidence that patients with poorly resectable tumors

have higher levels of stroma that cannot be disentangled from the tumor tissue itself. This suggests one histopathological mechanism for why some tumors are harder to remove from the surrounding stromal tissue. Additionally, the HOPX-CD73 staining indicates that the presence of tumor-associated MSC populations are highly informative of the development of an aggressive stromal phenotype.

3.4 DISCUSSION

We propose the INSPIRE (INferring Shared modules from multiPle gene expREssion datasets) framework for learning a low-dimensional representation (LDR) of multiple gene expression datasets. INSPIRE infers a conserved set of modules and their dependencies across multiple molecular datasets (e.g., gene expression datasets) that contain different sets of genes with a small overlap. We show that INSPIRE outperforms alternative approaches in both synthetically generated datasets and gene expression datasets from ovarian cancer patients. When we applied INSPIRE to nine expression datasets from ovarian cancer studies, which comprises 1,498 patient samples, we identified the stroma/mesenchymal module highly associated with percent stroma and patient survival in the TCGA samples. Our follow-up analysis on this module identifies the *HOPX* gene, which we experimentally validated to be expressed in mesenchymal stem cells (MSCs). *HOPX* is an unusual HOX protein that does not contain a DNA binding domain, and has been implicated in multiple aspects of cardiac and skeletal muscle development through recruitment of histone deacetylases^{86–88}. *HOPX* has recently emerged as a marker of numerous stem cell types^{92,98,99}. Our results indicate that MSCs are yet another stem cell population marked by *HOPX*. It has been shown that in response to inflammatory cytokines, MSCs release a myriad of growth factors including *FGF*, *EGF*, *PDGF*, and *VEGF*, which promote fibroblasts and endothelial cell

differentiation and growth¹⁰⁰. The tumor MSCs are known contributors to tumor-associated stroma via differentiation to cancer-associated fibroblasts (CAFs)⁹⁷, and may also promote metastasis¹⁰¹. *HOPX* could play an important role in this process by acting as a driver given that expression data from *HOPX* knockout mice reveals that many genes in the tumor-associated stroma/mesenchymal module are downstream of *HOPX*. Given the importance of *HOPX* in maintaining a stem cell like state⁹², it is suggestive that *HOPX* expression in the cancer-associated stroma may be maintaining the cancer-associated stroma niche, and could be an attractive target for further functional validation and therapeutic intervention – e.g. if loss of *HOPX* expression in the tumor stroma leads to differentiation of the cancer-associated mesenchymal stem cells.

We demonstrate thorough multiple analyses that modules identified by INSPIRE are more biologically coherent across a wide battery of tests of biological significance, including MSigDB pathway enrichment, ChEA TF regulatory networks, and enrichment for known OV CNV tumor drivers. Importantly, the INSPIRE latent variables can be used to predict disease phenotypes or clinical outcome, identify patient subtypes, and when integrated with multiple data modalities, resolve the importance of a specific gene expression module for understanding the mesenchymal subtype in ovarian cancer. Furthermore, when integrated with functional studies of *HOPX* in mice along with immunohistochemistry on multiple patient samples, our analysis suggests an important role for the *HOPX* associated module in maintaining a population of tumor associated mesenchymal stem cells in patients with aggressive stromal components to their tumors.

The effective joint learning strategy of the INSPIRE algorithm makes it possible to integrate datasets containing different sets of genes into a single network framework, which was impossible in the existing network inference approaches. This component of INSPIRE should greatly increase

the applicability of LDR learning algorithms to genomics problems where sample size provided by a single dataset is not large enough to learn a robust set of modules and module dependencies. In addition, inferring a network structure among pathways from high-dimensional molecular data is an important and open problem in biology, but is hampered by the need for very large sample sizes. INSPIRE would increase the applicability of network analysis by leveraging existing data and eliminate the cost of regenerating data from the same samples using different platforms.

INSPIRE is a general computational framework and can be applied to various diseases and different types of molecular data. For example, as we used INSPIRE to integrate mRNA expression datasets from different studies, we can apply it to integrate proteomic data from multiple studies. A future work is to extend INSPIRE such that it can integrate different types of molecular data such as transcriptomic, proteomic, epigenomic and metabolomics data in the same model. In this manuscript, we are applying INSPIRE to integrate microarray data. Since RNA-Seq has been emerging as an important platform for gene expression data profiling, one may want to combine microarray data and RNA-Seq data using INSPIRE. We recommend applying the *voom* normalization method¹⁰² to read counts when RNA-Seq data are used as input. The *voom* method estimates the mean-variance relationship of the log-counts, generates a precision weight for each observation and enters these into the *limma* (*Linear Models for Microarray and RNA-Seq Data*) empirical Bayes analysis pipeline. This makes the distributions of the read count data more similar to a normal distribution, and will make it possible to combine array data with RNA-Seq data using INSPIRE. The authors have shown that the *voom* normalization method has improved statistical properties when applying correlation or linear modeling, which are assumptions in most of the methods being applied to the processed microarray data¹⁰².

INSPIRE provides a great, effective starting point to learn complex dependencies between genes. After learning the lower-dimensional module network, we may learn a gene-level conditional dependence network by using the *graphical lasso*¹⁹ algorithm within each module. There are several other potential next steps to make technical improvement on the proposed INSPIRE framework. One of those is to extend INSPIRE to the case where the latent network is not perfectly conserved across datasets. We could allow for structured differences characterized by a small subset of modules while we encourage the latent network estimates to be quite similar to each other across datasets. This could be appropriate in many problems where different datasets involve biologically meaningful differences. Another technical improvement is to extend INSPIRE to the setting in which there are no overlapping genes across datasets. For example, one dataset measures mRNA expression levels of genes and the other dataset measures protein levels. In this case, we will need to develop a novel method for discovering correspondences between variables/modules across datasets. Finally, we could exploit the INSPIRE module network information inferred by INSPIRE for imputing the missing variable values in the datasets.

Chapter 4. SUPERVISED LEARNING OF CANCER EXPRESSION MARKERS USING MULTI-OMICS PRIOR INFORMATION

[The work presented in this chapter was published in Nature Communications.¹⁰³]

The research and development process for new drugs remains challenging and expensive. Nonetheless, the repertoire of potential cancer drugs continues to expand, with more than 1,200 cancer medicines in clinical development in the U.S.¹⁰⁴. However, cancers that appear pathologically similar often respond differently to the same drug regimens. Thus, methods to better match patients to the existing chemotherapy drugs are in high demand.

Conventionally, one tests for associations between gene expression levels and drug sensitivity measures by: (1) measuring pairwise association between each gene and each drug, or (2) performing a penalized regression for each drug using all genes as potential molecular markers, as was done in the CCLE and CGP drug sensitivity studies (**a**). However, drug response could be associated with gene expression levels that do not reflect the underlying drug's biological mechanism (i.e., false positive associations), and therefore, many results do not replicate in another data set¹⁰⁵. This discrepancy can happen due to biological confounders (disease sub-types or heterogeneity), experimental confounders (sample ascertainment), or technical confounders (e.g., batch effects). Previous studies also raised concerns with respect to drug sensitivity assay robustness¹⁰⁶. The *high-dimensionality* of data (i.e., when the number of gene-drug pairs greatly exceeds the number of samples) increases the multiple hypothesis testing burden and the chance of false positive gene-drug associations.

We present MERGE (**M**utation, **E**xpression hubs, known **R**egulators, **G**enomic CNV, and **m**Ethylation), a novel computational method that identifies reliable gene expression markers using a principled way of integrating multi-omic prior information relevant to disease processes (**Figure 24b**). MERGE learns from data how much each of the following *driver features* contributes to genes' potentials to drive cancer progression: (1) Mutations based on the AML study from The Cancer Genome Atlas (TCGA)¹⁰⁷, (2) *hubness* in a gene Expression network inferred from publicly available expression datasets, (3) the gene's known Regulatory role based on gene annotation databases⁸⁵, (4) Genomic CNV status based on the AML study from TCGA, and (5) mEthylation status based on the AML study from TCGA. We model each gene's *marker potential* (i.e., the prior probability that the gene is a reliable molecular marker for drugs) as a weighted combination of the gene's driver features. Through an iterative procedure, the MERGE algorithm jointly learns these weights and the degree of impact that genes' marker potentials have on observed gene-drug associations (**Figure 24b**). We show that MERGE outperforms several state-of-the-art approaches in identifying molecular markers replicated in validation data and predicting drug sensitivity accurately. We experimentally validate *SMARCA4* as a molecular marker and driver of sensitivity to topoisomerase II inhibitors, mitoxantrone and etoposide, in AML by showing that cell lines transduced to have high *SMARCA4* expression show dramatically increased sensitivity to these agents.

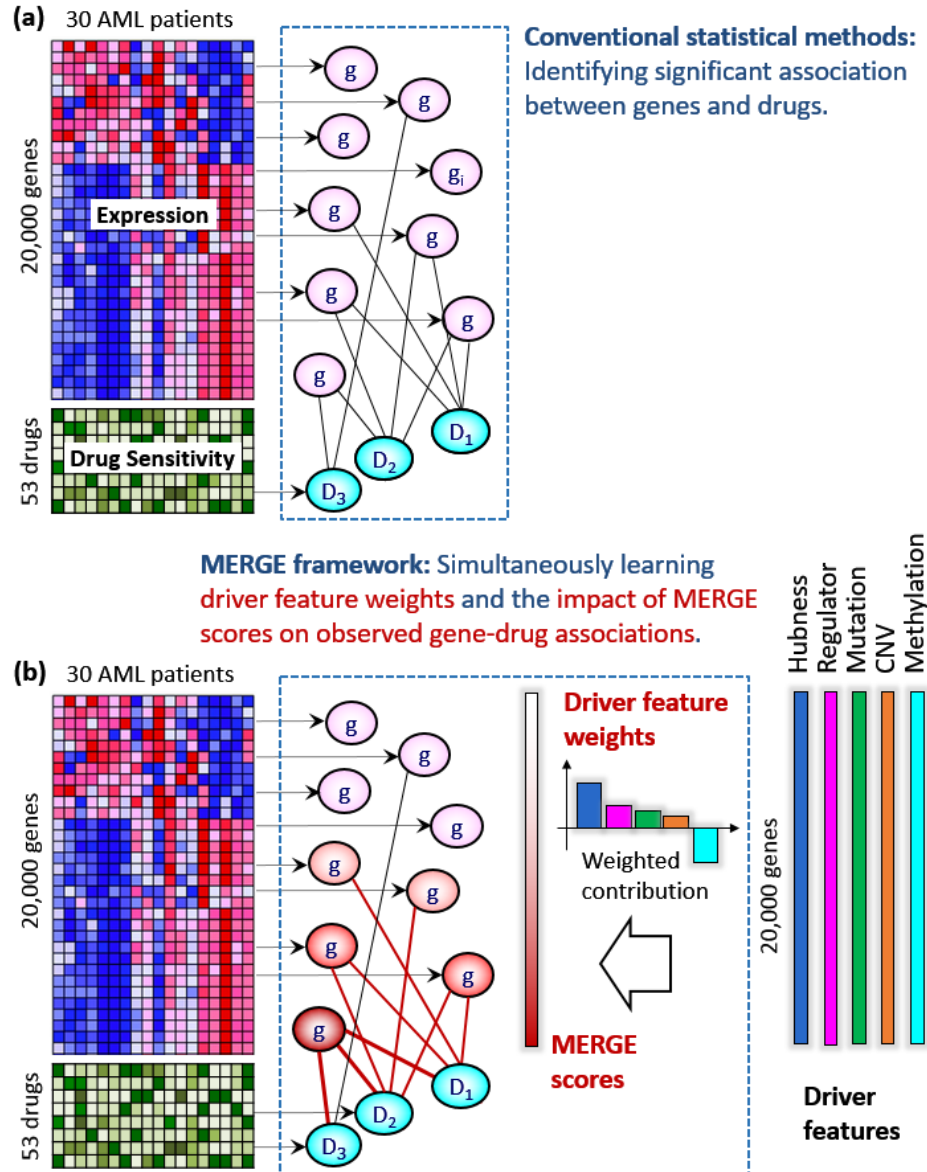


Figure 24. Conventional statistical methods vs. MERGE framework.

(a) Conventional statistical methods identify gene expression markers for drugs on the basis of expression data and drug sensitivity data. The goal is to measure the statistical significance of the association between each gene expression level and the sensitivity measure of each drug. **(b)**

Overview of the MERGE framework.

Our MERGE algorithm provides a new way to prioritize gene-drug associations by incorporating prior information on genes' relevance to AML and potential to drive it. MERGE learns a priority score for each gene, called a MERGE score, based on the gene's *driver features*:

(1) Mutation, (2) Expression hubness, (3) whether the gene has a known Regulatory role, (4) Genomic CNV, and (5) mEthylation. These driver features were extracted from publicly available sources such as TCGA AML study¹⁰⁷, AML expression studies⁷⁶, and gene annotation databases. The details on the sources and the preprocessing of the driver features are included in APPENDIX B.

The MERGE score represents a prior probability that the gene is a reliable (i.e., likely driven by biological mechanisms, not confounders) molecular marker for response to drugs, modeled as a weighted combination of the gene’s MERGE features (**Figure 24b**). The MERGE algorithm jointly learns these driver feature weights and how the MERGE score of genes explains the observed gene-drug associations. Genes with high MERGE scores (i.e., high marker potentials) tended to have many observed associations with drugs. The learned driver feature weights provide new insights into what kind of molecular data is most informative of a gene’s potential to be a reliable marker for therapeutic response. After the MERGE score of each gene is estimated, we considered the top N genes based on MERGE scores as a prioritized subset and then selected the gene-drug pairs with significant association p -values (genome-wide FDR corrected p -value < 0.1).

4.1 RELATED WORK

The growing availability of genome-wide expression data and in-vitro drug sensitivity data from cancer cell lines has enabled a *data-driven approach* to identifying molecular markers by finding robust statistical associations between genes and drugs. The Cancer Genome Project (CGP) tested 130 drugs in 639 cell lines, with a mean of 368 cell lines tested for each drug¹⁰⁸. The Cancer Cell Line Encyclopedia (CCLE) tested 479 cell lines for sensitivity against a panel of 24 drugs¹⁰⁹. These studies employed a penalized (elastic net) regression method¹¹⁰ to identify novel associations

between gene expression levels and drug sensitivity measures. While both CGP and CCLE evaluated large numbers of cell lines, some of the most interesting associations were detected by focusing analyses within, rather than across, tumor types. Consistent with this, a study by Heiser et al. (2012) was able to identify novel associations using a much smaller panel of 49 breast cancer cell lines with sensitivity to a panel of 77 compounds.

Successful attempts to reduce false positives by incorporating prior information have occurred in genome-wide association studies. Li et al. (2008) proposed a *prioritized subset analysis*: they pre-selected a prioritized subset of single nucleotide polymorphisms (SNPs) from candidate genes or regions and applied false discovery rate (FDR) correction within this subset to make it more likely that these SNPs would be selected. Roeder et al. (2006) and Genovese et al. (2006) up- or down- weighted the association p -value for each SNP, and the resulting weighted p -values were subsequently used in FDR corrections; the p -values of SNPs with significant associations in prior linkage analyses were lowered. Although these methods increase the power to detect causal SNPs in genome-wide association studies, they do not apply when it is not obvious how to define candidate genes or prioritize a subset of genes, especially when multiple sources of prior information are available.

One simple way to use prioritized subset analysis¹¹² considers the genes frequently mutated in AML as candidate genes. However, relying on only mutation information is unlikely to prove successful. Many cancer mutations are *passengers* (i.e., not drivers), and some genes' driver roles may not be reflected in mutation data but in other data types, such as epigenomic, copy number variation (CNV) and even gene expression data. For example, Logsdon et al. suggested that *hub* genes in a gene network inferred purely based on gene expression data might indicate tumor driver

events. To reliably identify gene expression markers for drug response, we must precisely determine how much each piece of relevant prior information contributes to the gene’s *marker potential*.

4.2 MERGE MODEL

To learn the MERGE scores, we developed a probabilistic graphical model approach that provides a statistical model to represent relationships among variables in input data and a principled way of learning the parameters of the model from data⁴⁰. Let w_{ij} represent the magnitude and sign of the impact of gene i on drug j . This relationship can be modeled as: $y_j \sim \mathcal{N}(w_{ij}x_i, \mathcal{E}^2)$, where x_i represents the expression level of gene i and y_j represents the drug response measure (i.e., AUC) of drug j . A conventional statistical model widely used in a genome-wide association analysis does not model marker potential (**Figure 24a**). In the conventional model, we estimate w_{ij} for each combination of a gene and a drug by finding w_{ij} , which minimizes the squared error of prediction across all samples (here, patients):

$$\underset{w_{ij}}{\text{minimize}} \sum_{n \in \{\text{all patients}\}} (y_j^{(n)} - w_{ij}x_i^{(n)})^2, \quad (13)$$

where (n) means the n th sample. Then, the learned w_{ij} value would equal Pearson’s correlation coefficient between the two variables, x_i and y_j , when both x_i and y_j were scaled to have the same standard deviance (specifically through the common standardization procedure).

MERGE’s key innovation is the modeling of the *prior probability distribution* over each w_{ij} such that it incorporates prior knowledge on gene i in terms of its potential to be a molecular marker. We model the prior distribution over each w_{ij} as a normal distribution $P(w_{ij}) \sim \mathcal{N}(0, \sigma_i^2)$,

where σ_i^2 represents a prior variance specific to gene i . If gene i has a low value on σ_i^2 , the weight value w_{ij} would be inclined to be close to zero (i.e., low marker potential); if gene i has a high value on σ_i^2 , then, since there is more probability mass away from zero, the weight value w_{ij} would be more inclined to deviate from zero (i.e., high marker potential). We model σ_i^2 as a function of a weighted combination of its driver features $\sigma_i^2 = \frac{1}{\lambda - \sum_{k=1}^5 v_k d_{ik}}$, where d_{ik} is the value of the k th driver feature on gene i . Genes with a high value of $(\sum_{k=1}^5 v_k d_{ik})$ would have a high variance σ_i^2 (i.e., high marker potential); thus, we define the MERGE score of gene i as $(\sum_{k=1}^5 v_k d_{ik})$. λ is a regularization parameter: when it is high, each σ_i^2 would take a lower value, which means a lower variance on the weight (w_{ij}) values associated with each gene.

We learn the parameters – weight values w_{ij} for each gene i and drug j , and the driver feature weights v_k – that optimize the joint log-likelihood function $\log P(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{V})$, where $\mathbf{X}$ is the expression data matrix (p genes $\times$ n patients), $\mathbf{Y}$ is the drug response data matrix (q drugs $\times$ n patients), $\mathbf{W}$ is the gene-drug weight matrix (p genes $\times$ q drugs) that contains w_{ij} for each gene i and drug j , and $\mathbf{V}$ is a vector that contains 5 driver feature weights. The MERGE algorithm can be seen as a process of projecting high-dimensional gene-drug associations (p genes $\times$ q drugs) onto a lower-dimensional space by constraining weight values based on prior information on genes' potential to drive the disease.

The objective function $\log P(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{V})$ can be decomposed as

$$\begin{aligned} & \log[P(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \mathbf{V})P(\mathbf{X}|\mathbf{W}, \mathbf{V})P(\mathbf{W}|\mathbf{V})P(\mathbf{V})] \\ &= \log P(\mathbf{Y}|\mathbf{X}, \mathbf{W}) + \log P(\mathbf{X}) + \log P(\mathbf{W}|\mathbf{V}) + \log P(\mathbf{V}) \end{aligned} \quad (14)$$

$$= \log P(\mathbf{Y}|\mathbf{X}, \mathbf{W}) + \log P(\mathbf{W}|\mathbf{V})$$

assuming a uniform prior distribution over $\mathbf{X}$ and $\mathbf{V}$ (i.e., $P(\mathbf{X})$ and $P(\mathbf{V})$ are constant). The conditional log-likelihoods are:

$$\begin{aligned} \log P(\mathbf{Y}|\mathbf{X}, \mathbf{W}) &= \log \prod_{i,j,n} \mathcal{N}(w_{ij}x_i^{(n)}, \mathcal{E}^2) \\ &= - \sum_{i,j,n} \left\{ \log \sqrt{2\mathcal{E}^2\pi} + \frac{(y_j^{(n)} - w_{ij}x_i^{(n)})^2}{2\mathcal{E}^2} \right\} \end{aligned} \quad (15)$$

and

$$\begin{aligned} \log P(\mathbf{W}|\mathbf{V}) &= \log \prod_{i,j} \mathcal{N}\left(0, \frac{1}{\lambda - \sum_{k=1}^5 v_k d_{ik}}\right) \\ &= - \sum_{i,j} \left\{ \log \sqrt{\frac{2\pi}{\lambda - \sum_{k=1}^5 v_k d_{ik}}} + \frac{(\lambda - \sum_{k=1}^5 v_k d_{ik})w_{ij}^2}{2} \right\} \end{aligned} \quad (16)$$

for each gene i , drug j , and a sample n . Maximizing $\log P(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{V})$ with respect to $\mathbf{W}$ and $\mathbf{V}$ leads to the following optimization problem when constants are dropped and $\mathcal{E}^2$ is set to 1 (since we standardize $\mathbf{X}$ and $\mathbf{Y}$ before we apply the algorithm, $\mathcal{E}^2 = 1$ is a reasonable choice that simplifies the optimization).

$$\begin{aligned} &\underset{w_{ij} \in \mathbf{W}, v_k \in \mathbf{V}}{\text{minimize}} \sum_{i,j} \left\{ \sum_n (y_j^{(n)} - w_{ij}x_i^{(n)})^2 \right\} + \sum_{i,j} \left(\lambda - \sum_{k=1}^5 v_k d_{ik} \right) w_{ij}^2 \\ &\quad - \sum_{i,j} \log \left(\lambda - \sum_{k=1}^5 v_k d_{ik} \right) \\ &\text{subject to } \lambda - \sum_{k=1}^5 v_k d_{ik} > 0 \text{ for each gene } i. \end{aligned} \quad (17)$$

The first term is the loss function for learning the value of w_{ij} (for each gene i and drug j) that captures the degree of association between gene i and drug j . The second term can be viewed as a weighted $L2$ regularization term that favors small values on w_{ij} (which improves generalizability of the learned model) with a different strength for each gene depending on the value of $(\lambda - \sum_{k=1}^5 v_k d_{ik})$. A gene i with a high value of $(\sum_{k=1}^5 v_k d_{ik})$ (i.e., MERGE score) would be regularized weakly and more inclined to have a weight value w_{ij} with a large magnitude. The last term requires the regularization parameter $(\lambda - \sum_{k=1}^5 v_k d_{ik})$ to be positive and encourages it to take on relatively larger values.

We iteratively estimate the optimization variables $\mathbf{W}$ and $\mathbf{V}$ using a block coordinate descent procedure^{50,115,116} until convergence. The objective function of MERGE is not jointly convex with respect to $\mathbf{W}$ and $\mathbf{V}$, though it is convex with respect to each set of parameters with the other set held fixed. When $\mathbf{V}$ is held fixed, the objective is convex with respect to $\mathbf{W}$; when $\mathbf{W}$ is held fixed, the objective is convex with respect to $\mathbf{V}$. This means that each learning step in the block coordinate descent algorithm (learning an element in $\mathbf{V}$ or in $\mathbf{W}$) is a convex optimization problem with a single local minimum that is also the global minimum. In each of these experiments, we chose the value of the hyperparameter λ by cross-validation (see APPENDIX B for details).

4.3 APPLICATION OF MERGE TO IDENTIFY AML DRUG RESPONSE MARKERS

Cancer is very well-known to be influenced by the MERGE driver features we use in our algorithm. Thus, we use MERGE to identify markers that have a role in determining response of cancer patients to chemotherapy drugs. To test the MERGE algorithm, we used acute myeloid

leukemia (AML) patient and cell line gene expression and *in vitro* drug sensitivity data which we generated and published in our MERGE paper¹⁰³ and provided in **Appendix Table 17**.

The patient data contained measurements of genome-wide gene expression and *in vitro* drug sensitivity (see APPENDIX B for details) to a panel of 160 chemotherapy drugs and targeted inhibitors across 30 AML patient samples. The customized drug panel contained 62 drugs approved by the U.S. Food and Drug Administration (FDA) as well as investigational agents that have been studied in cancer patients, and encompassed a broad range of drug action mechanisms. We chose 53 drugs that exhibited activity (cell viability $\leq 50\%$) against at least half of the patient samples (see APPENDIX B for details). As was done previously in the CCLE study¹⁰⁹, we processed the drug sensitivity data by fitting the response curve followed by extracting the summary statistics. We used the area under the curve (AUC) statistic throughout the paper because it represents an average of drug sensitivity across a range of drug concentrations; indeed, AUC showed by far the strongest association with gene expression levels (see APPENDIX B for details). APPENDIX B summarizes the clinical information and describes our analysis on the consistency between clinical data and our *in vitro* drug sensitivity data.

We utilized cell lines to intensify our focus on the hypotheses for which we could provide additional experimental evidence, since it is easier to perform over-expression or knock-out experiments on cell lines than on primary patient tissues. However, we noted a very small overlap between our 160 drugs and the drugs tested on the 14 AML cell lines in the CCLE data (2 drugs overlapping, each tested on 3 AML cell lines). Thus, for effective computational and experimental validation of the significant gene-drug associations discovered in the patient data, we measured *in vitro* drug sensitivity of 14 AML cell lines to the same set of 160 drugs in our high throughput

assay while we used publicly available expression data of the 14 AML cell lines from CCLE. We observed a statistically significant overlap (Fisher’s exact test $p\text{-value} = 3 \times 10^{-6}$) of gene-drug pairs with significant association $p\text{-values}$ between our discovery data from 30 AML patient samples and the CCLE validation data. Our unique validation setting let us measure the *testability* of discovered associations.

We first statistically evaluated our MERGE algorithm in several different ways and observed that the results were very promising in all categories of evaluation we performed. We then examined top-ranked genes for several drug classes based on our MERGE prioritization method and interpreted their associations with the corresponding drugs. Finally, we described results of the experimental validation on one of the top-ranked genes, *SMARCA4*, whose expression is significantly associated with increased sensitivity to mitoxantrone and etoposide.

4.3.1 *Expression hubness significantly determines the MERGE scores*

The driver feature weights learned by the MERGE algorithm indicate the relative importance of each driver feature on the MERGE scores (**Figure 24b**). The MERGE score of gene i is defined as $(\sum_{k=1}^5 v_k d_{ik})$, where v_k is the k th driver feature weight and d_{ik} is the k th driver feature of gene i . *Expression hubness* (i.e., number of neighbors in a gene network estimated based on publicly available AML expression data) has the highest weight (**Figure 25a**) and makes the largest contribution to the MERGE scores (**Figure 25b; Appendix Table 20**).

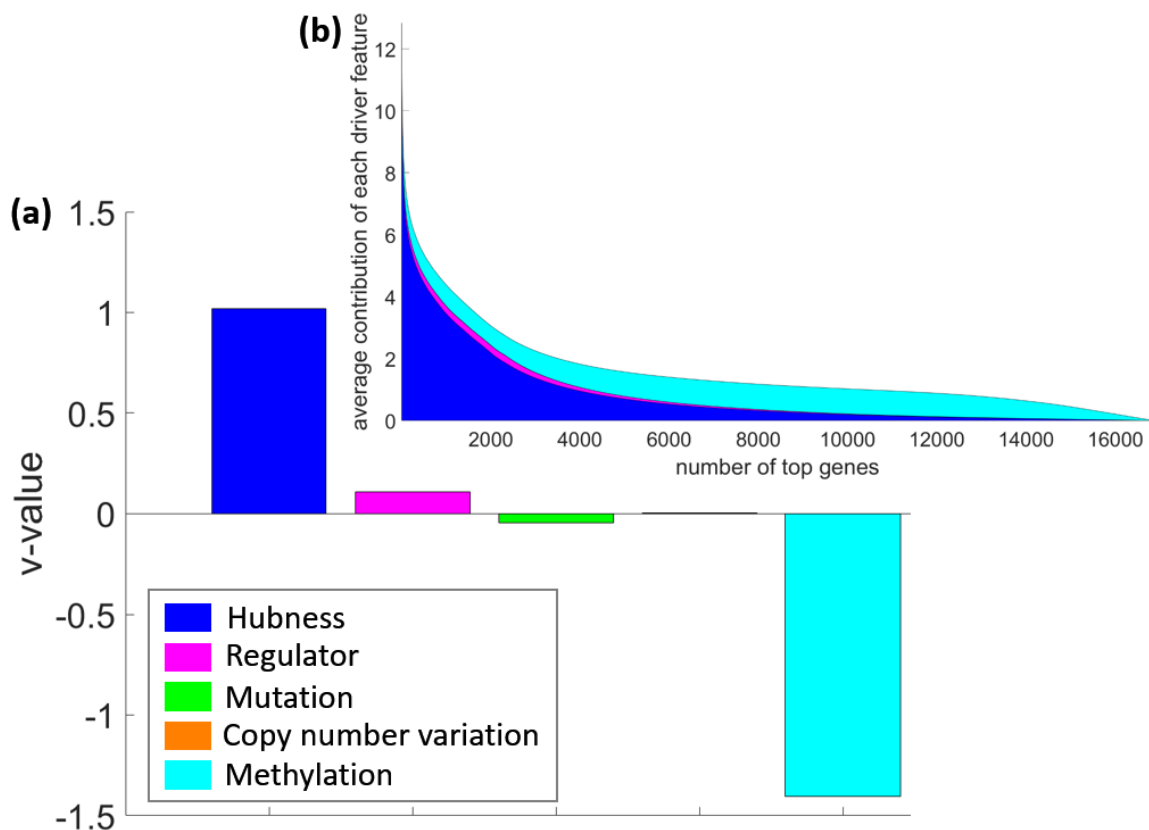


Figure 25. Contributions of different drivers feature to the MERGE score.

(a) Learned driver feature importance (i.e., weight values for each feature). The methylation feature has a negative weight, consistent with our prior knowledge that when DNA is methylated in the promoter region, the corresponding genes are inactivated and silenced. **(b)** We sort genes based on the MERGE score (x-axis) and plot the sum of the contribution of all driver features to the MERGE score (i.e., weighted combination of driver features) (y-axis). We decomposed the weighted combination into the five driver features and indicated the magnitude of the contribution of each feature (driver feature weight $\times$ driver feature value) with different colors. Expression hubness contributes the most to the score, followed by regulatory function and (lack of) methylation.

This implies that gene expression data provide more information about a gene's potential to predict drug response than other types of data, perhaps for the following reasons. (1) Gene expression data can reflect downstream effects of genetic or epigenetic changes that may not have

been detected by existing mutation, CNV and methylation profiles. (2) Expression hubness was estimated from a larger number of patients because expression data are the most common type of molecular data from disease studies. (3) Expression hubness has been considered likely to indicate selective pressure in tumor genome evolution and to drive events⁹³.

Identifying expression hubs has therefore been considered a powerful complementary way to identify candidate tumor drivers that are hard to detect from sequence data due to a large number of passenger mutations. The importance of the expression hubness feature (**Figure 25**) suggests that these candidate tumor drivers might prove promising markers for drug response. The methylation feature is negatively weighted, consistent with prior knowledge that methylation in a promoter region silences the corresponding gene¹¹⁷.

4.3.2 *Consistency of gene-drug associations in left-out data*

We compared MERGE to four conventional methods in terms of consistency of significant gene-drug associations, where we used all 30 patient samples for discovery and validated the discovered associations using (i) 14 cell lines, (ii) independent patient data from 12 new AML specimens. The four other methods we compared MERGE to are: (1) Pearson's P-value of correlation. For each gene-drug pair, we computed the t statistics and the associated *p*-value, measuring the correlation in a univariate linear regression model. We then selected the gene-drug pairs with the smallest association *p*-values. (2) Spearman P-value of correlation. For each gene-drug pair, we measured the rank association using the Spearman correlation and selected the gene-drug pairs with the smallest association *p*-values. (3) ElasticNet. For each drug, we solved the ElasticNet optimization problem¹¹⁰ using gene expression as input, as done by Barretina et al. (2012) and Garnett et al. (2012), and selected the gene-drug associations with the strongest weight.

(4) Multi-task learning. We used the multi-task learning method¹¹⁸ implemented by Pong et al. (2010), which considers each drug as a different task. We then selected the gene-drug associations with the strongest weight.

We discovered gene-drug associations within the data from all 30 samples and tested them on two independent datasets: (a) 14 CCLE cell lines (**Figure 26a**), (b) 12 additional AML patients who had relapsed or were refractory to at least 3 (up to 6) prior regimens (**Figure 26b**). The clinical information on the additional 12 relapsed patients can be found in APPENDIX B. **Appendix Table 21** presents the gene expression (processed RNA-seq) and the drug sensitivity (AUC) data from the same patients.

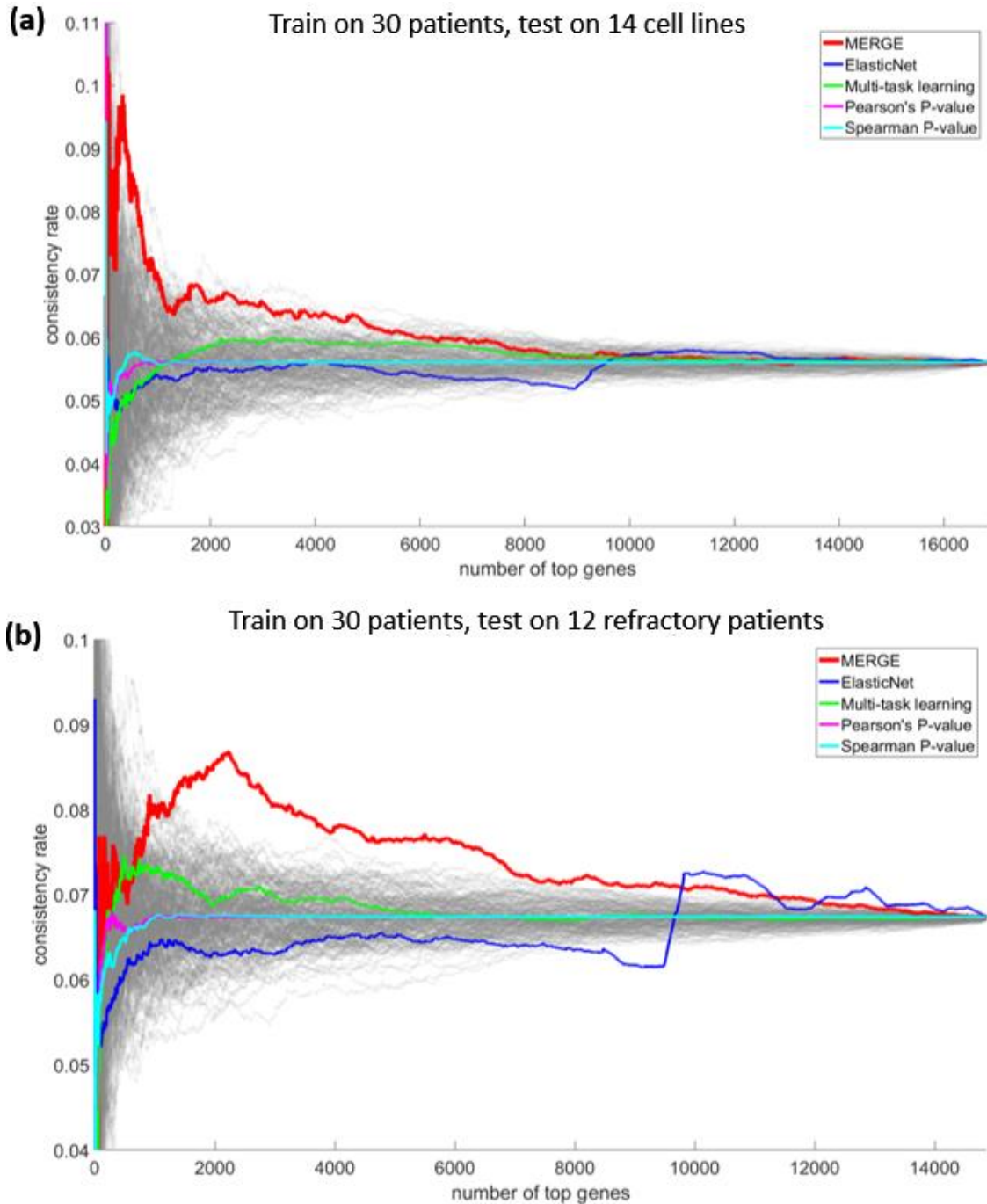


Figure 26. Comparison of MERGE with four alternative methods in terms of the percentage of the significant associations replicated in the left-out test data.

We discovered gene-drug associations within the data from all 30 samples, and we tested on (a) the 14 cell line samples, and (b) the data from the additional 12 refractory patient samples. Each gray line corresponds to a random ordering of all genes. We note that the x-axis in (b) contains

a lower total number of genes; some of the genes existing in the microarray gene expression data from the 30 patient and cell line samples did not exist in the RNA-seq data from the 12 refractory patient samples.

Each method prioritized gene-drug pairs differently (details in APPENDIX B). For evaluation, we computed the *true discovery rate* (that is, how many significant associations were replicated in the left-out test data) when considering the top N genes per drug on average (i.e., $53 \times N$ gene-drug pairs in total). Specifically, we computed the *consistency rate* (y-axis) – defined as the number of significant gene-drug associations replicated in the left-out test data divided by the total number of significant gene-drug associations within the selected ($53 \times N$) gene-drug pairs – for varying values of N (x-axis) from 1 to all genes. In both settings (**Figure 26a** and **b**), MERGE showed a much higher consistency rate for high-scoring genes (small N values) than the other two methods and the random ordering of genes (gray lines). As N increased, methods had more similar consistency rates because their top N genes became more similar.

Unlike the initial 30 samples, the additional 12 samples were highly refractory, and many of them exhibited extremely poor risk features. This could potentially account for the lower consistency rate in **Figure 26b** even though both the discovery and validation samples were obtained from patient specimens. We did not expect that highly refractory patients (refractory after 3-6 prior regimens) would exhibit the same drug sensitivity patterns as the 30 patients with newly diagnosed disease or 1st or 2nd relapses. In addition, we surmised that the refractory patients may have activated other survival pathways and mechanisms of drug resistance. Still, even in this challenging validation setting, MERGE outperformed four alternative methods.

Three reasons accounted for the poor performance of the ElasticNet and Multi-task learning methods in **Figure 26a** and **b**. First, these multiple regression methods, intended to solve a prediction problem, were not specifically designed to capture robust gene-drug associations. Conversely, MERGE was designed to aggressively decrease the number of false positive gene-drug associations by incorporating prior knowledge about the genes' potential to drive the disease. Second, since this problem was ultra-high-dimensional (12 samples and ~17K variables), multiple regression methods were likely to learn models too complex to identify robust gene-drug associations, even with regularization (e.g., elastic net penalty). Finally, since multiple regression methods model each response as an aggregated effect of multiple features, highly correlated features would share a fixed amount of weight. This would assign a small magnitude of weight to each of many correlated features. Many robust gene-drug pairs whose associations would have been replicated in validation data were likely to have been eliminated in this way.

MERGE uses a strong prior and this might have been an important factor for MERGE's strong feature consistency result. To show that the training data are also critical for the high consistency rate of MERGE, we performed MERGE on 100 different permutations of the data where the training samples in the drug response data are shuffled. We then tested the learned models on both 14 cell lines and the 12 refractory samples, similarly to **Figure 26**. As shown in **Appendix Figure 16**, the MERGE run using the original sample ordering achieved a higher consistency rate than most of the permutation runs, and many of the permutation runs perform worse than the competing methods, which shows that MERGE makes use of the information in the training data and the prior information alone is not very helpful when the training data are random.

4.3.3 *Drugs in the same classes exhibited similar response pattern*

Different drugs with similar mechanisms of action are expected to show similar response across patients and show similar patterns of gene expression association. Thus, if a gene X were associated with a drug Y, X would also be likely associated with another drug Y' with the same mechanism of action (e.g., sunitinib and tandutinib are in the 'Flt3 inhibitor' class) (**Appendix Table 18**). The 53 drugs we considered were classified into 24 broad classes based on their mechanism of action; 15 classes contained >1 drug (**Appendix Table 18**). Sixteen of the 53 drugs were shared across two classes, while the other 37 drugs were in a single class. We compared MERGE to four other methods as in the previous section in terms of within drug class consistency – the extent to which the gene-drug association was conserved across drugs in the same functional category. To test this hypothesis in an unbiased way, we first applied an agglomerative hierarchical clustering approach to compare the AUC values of different drugs across 30 patient samples. **Figure 27a** shows that drugs in at least 10 of 15 classes (with >1 drug) were clustered, which indicates that drugs with the same mechanism of action tended to have similar response patterns (details in APPENDIX B).

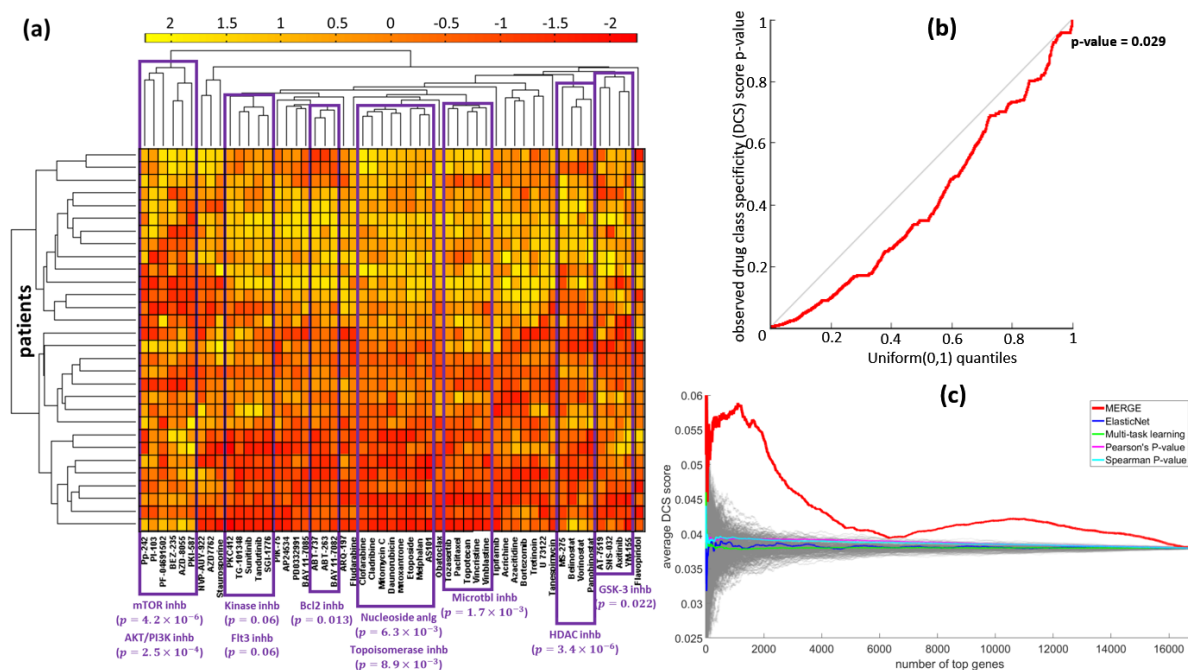


Figure 27. Drug class specificity of patients' drug response and genes' MERGE scores.

(a) The result of the agglomerative hierarchical clustering that is applied to the AUC values of all drugs across 30 patient samples. AUC values are standardized before applying the hierarchical clustering, and the color bar at the top represents the standardized AUC values.

For several branches in the resulting dendrogram, we report the drug classes that have significant enrichment with the drugs in that branch; we also report the corresponding Fisher's exact test p-values. The result shows that drugs in at least 10 of 15 classes (that contain >1 drug) are expectedly grouped together in the dendrogram; for each group, Fisher's exact test p-value associated with the overlap between the drug class and the dendrogram group is ≤ 0.06 .

(b) A QQ (quantile-quantile) plot of the observed DCS score p-values from 1,000 random permutation tests for all genes with at least one significant gene-drug association. In each permutation test, drug labels were shuffled. Dots falling below the diagonal imply that the overlap with known drug classes is more significant than would be expected by random chance (permutation p-value = 0.029). **(c)** For varying N (x-axis), we plotted the average DCS score over the N genes (y-axis) associated with the top ($53 \times N$) gene-drug pairs based on the five methods in comparison.

For each gene, we also examined whether the drugs with which the gene was significantly associated tended to be clustered into the drug mechanism classes. We computed each gene's specificity measure, which we referred to as its drug class specificity (DCS) score (details in APPENDIX B). For each gene significantly associated with at least one drug, we then estimated the significance of its DCS score by comparing it with 1,000 random DCS scores for the same gene, each from a permutation test where we shuffled the class labels of all drugs. **Figure 27b** shows a QQ plot of the empirical DCS score p -values from 1,000 random permutation tests (y-axis) against the Uniform(0,1) distribution quantiles (x-axis) for each gene. The empirical DCS p -values were significantly lower than random (permutation test p -value: 0.029), indicating the statistical significance of the DCS in gene expression dependency of the drugs overall.

4.3.4 *Specificity of gene-drug associations to drug classes*

If a gene's expression level were *specifically* associated with drugs in the same mechanism class, we would have higher confidence in the observed associations resulting from the underlying biological mechanisms. Genes with expression levels associated with too many drugs in a diverse set of classes would be less likely to be true markers, and the observed associations would more likely be due to confounders. Here, we used the specificity of gene-drug associations to a drug mechanism class as an evaluation metric, and we used the DCS score as the evaluation metric for each gene (as described above). We took the average DCS score (y-axis) over the genes associated with the top gene-drug pairs (i.e., $53 \times N$ gene-drug pairs in total) for varying N values (x-axis) (**Figure 27c**). MERGE showed a much higher degree of drug mechanism class specificity than the alternative methods.

4.3.5 *Drug sensitivity prediction performance*

We compared MERGE to three other methods for evaluating consistency of patient rankings based on actual vs predicted drug sensitivity. We tested in two ways: (i) We used one batch containing 12 patient samples for training and a different batch containing 12 patient samples for validation (and vice versa). (ii) We used a leave-one-out cross validation (LOOCV) test to obtain the predicted drug sensitivity across 30 patient samples. We compared MERGE to the following methods that predict patient drug response: (1) ElasticNet¹¹⁰, (2) multi-task learning^{118,119}, and (3) Bayesian multi-task multiple kernel learning (MKL), the winner of the NCI-DREAM Drug Sensitivity Prediction Challenge¹²⁰. Like MKL, we compared these methods in terms of the consistency among the ranking of patients based on their actual versus predicted drug sensitivity. We considered the following two settings. (a) Training a prediction model based on 12 samples in one batch and testing on the 12 samples in another batch, and vice versa. We averaged prediction accuracy results (**Figure 28a**). (b) Measuring the prediction performance via LOOCV using all 30 patient samples (**Figure 28b**).

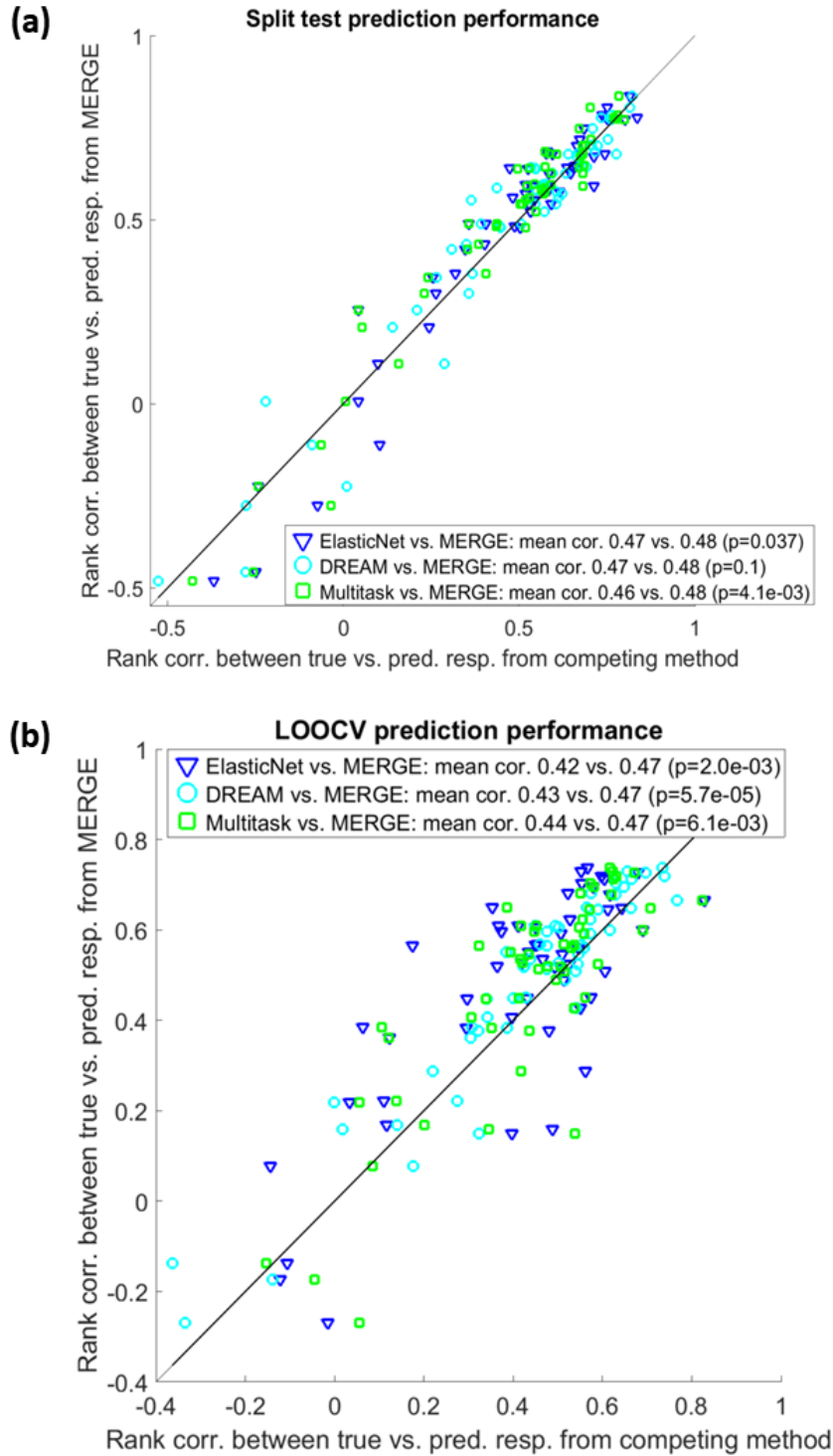


Figure 28. Prediction performance of MERGE compared to alternative methods.

Comparison of prediction performance of MERGE to the prediction performances of three other methods: ElasticNet, Bayesian multi-task MKL and Multi-task learning. **(a)** One batch of

the samples is used for training, and a different batch is used for validation. **(b)** LOOCV setting. Performance is measured in terms of the Spearman correlation of the predicted response with the actual response. This evaluation metric was used in the NCI-DREAM Drug Sensitivity Prediction Challenge. Each dot corresponds to a different drug, and each color to a different method's prediction on the x-axis compared to the MERGE prediction on the y-axis. The mean correlation from each of the methods in comparison and the associated p -values from a one-sided Wilcoxon signed rank test are reported in the legend.

In **Figure 28a** and **b**, we compared MERGE (y-axis) to the three other methods (x-axis) in terms of prediction performance measured by rank correlation between predicted and actual drug response in the test set across patients. Each dot corresponds to one of the 53 drugs, and each color to one of the methods compared to MERGE. In both experimental settings (a and b), MERGE performed competitively with the alternative methods in terms of prediction performance averaged over all drugs. We performed a one-sided Wilcoxon signed rank test to show the significance of the outperformance of MERGE relative to the methods in comparison. In fact, the p -values are significant at a $p \leq 0.007$ level for four out of six comparisons and at a $p \leq 0.1$ level for the other two. We show the p -values in the legend of **Figure 28a** and **b** for each comparison, and **Appendix Figure 17a** and **b** show the detailed performance for each of the drugs. As shown in the legend of **Appendix Figure 17a** and **b**, MERGE achieves the best prediction performance for a higher number of drugs than each of the three alternative methods. Indeed, in the LOOCV test (**Appendix Figure 17b**), MERGE achieved the best prediction performance for 62% of the drugs.

4.3.6 MERGE identifies the roles of several drug response markers

We now interpret the gene-drug associations identified by MERGE, which offer the potential to make novel discoveries about molecular markers (**Figure 7**). We seek to derive hypotheses likely to lead to discoveries or experimental validation targets. Therefore, we further narrowed significant gene-drug associations to: (1) those that were consistently significant in cell lines, so we could perform experimental validations, and (2) those that showed a high degree of specificity for a drug mechanism class. For each of the 24 drug mechanism classes, we considered the high MERGE-scoring genes whose associations with drugs were specific to that class (Fisher's exact test p -value < 0.05) and whose associations to that class were conserved in cell lines. **Appendix Table 14** lists these genes for the 20 drug classes that had at least one class-specific gene. **Appendix Table 22** shows the entire list of the genes for each class and the corresponding results in detail. **Figure 7a** depicts a heat map that shows the level of specificity of each gene (row) to each drug class (column), measured by $-\log_{10}[\text{Fisher's exact test } p\text{-value}]$ for the top three MERGE-scoring genes in each drug class. **Figure 7b** highlights the drugs associated with each gene. In **Figure 7b**, the red color indicates a negative association between gene expression and drug AUC measure (i.e., high expression indicates low AUC and hence sensitivity), while green indicates a positive association (i.e., high expression indicates resistance). We show in **Appendix Figure 18** the heat maps for the four alternative methods with which we compared MERGE (in **Figure 26** and **Figure 27c**). **Appendix Figure 19** shows the amount of contribution of each driver feature on the MERGE score for the genes shown in **Figure 7a** and **b**.

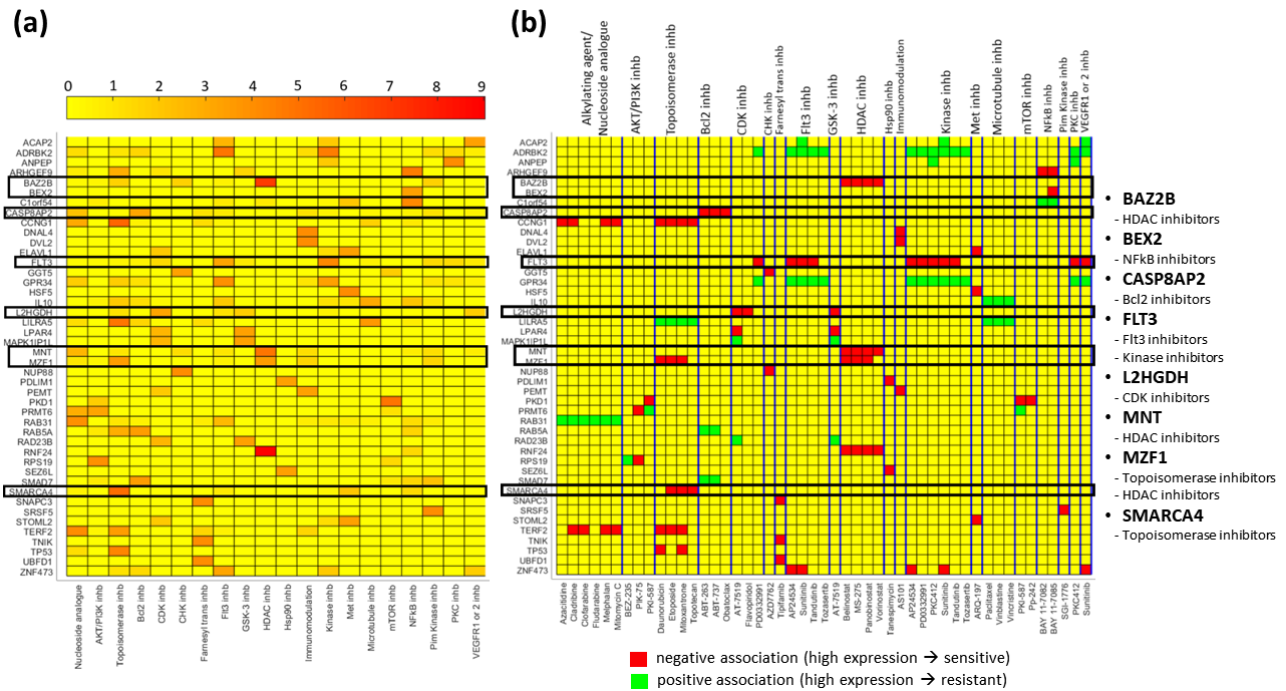


Figure 29. Drug class specificity and drug response association heat maps of the top 44 genes selected by our MERGE framework.

(a) A heat map that shows the level of specificity of each of the 44 genes (row) to each drug class (column) measured by $-\log_{10}[\text{Fisher's exact test } p\text{-value}]$. For clarity, we considered only Fisher's exact test $p\text{-value} < 0.05$ to be significant; other values are indicated in yellow. The drug classes that are not assigned by MERGE any genes with associations specific to the class and consistent in the cell line data are not shown. We highlighted the genes whose biological significance with black-colored boxes (see APPENDIX B for their biological interpretation). **(b)**

A heat map that shows the gene-drug association for genes and drug classes shown in (a). Yellow indicates that the corresponding gene-drug pair does not have a statistically significant association (genome-wide FDR corrected $p\text{-values} < 0.1$), while green indicates a positive and red a negative association. The drugs are grouped by blue lines based on their classes, and the class names for each group are written on top of the heat map. Drugs that are members of more than one drug class (e.g., sunitinib) are shown multiple times for each class to which the drug belongs. The list on the right shows the genes whose biological significance (APPENDIX B), and the drug classes they are specific to.

APPENDIX B summarizes the 8 top-ranked genes in some of the major drug classes (*FLT3*, *CASP8AP2*, *L2HGDH*, *MNT*, *BAZ2B*, *MZF1*, *BEX2*, and *SMARCA4*) that are highly likely to have notable biological significance in leukemia. We observed that for 5 of these 8 genes (*SMARCA4*, *FLT3*, *MNT*, *BAZ2B* and *MZF1*), MERGE provided a unique path towards identifying their roles that the other 4 methods simply could not identify (**Appendix Table 15**). Except for ElasticNet, the other 3 methods identified only 1 of these 8 genes, *L2HGDH*.

4.3.7 Experimental validation of the marker role of *SMARCA4*

The association MERGE identified between *SMARCA4* and the topoisomerase inhibitor class exhibited a high level of specificity and prior knowledge supported a potential relationship between *SMARCA4* and topoisomerase II (details in APPENDIX B). We experimentally validated the association between *SMARCA4* and the topoisomerase II inhibitors (etoposide and mitoxantrone) by overexpressing *SMARCA4*. The U937 leukemia cell line expressed high levels of *SMARCA4* protein, while the KG1 leukemia cell line expressed very low levels of the protein, as analyzed by Western blot (**Figure 30e**, **Appendix Figure 21**) and flow cytometry (**Figure 30g**). *In vitro* high throughput drug sensitivity testing showed that U937 was much more sensitive to mitoxantrone and etoposide than KG1 (**Appendix Table 17**). After selection of stably transfected KG1 cells overexpressing *SMARCA4*, confirmed by Western blot of whole cell lysates (**Figure 30e**, **Appendix Figure 22**) and flow cytometry (**Figure 30g**), *in vitro* chemo sensitivity testing demonstrated enhanced cytotoxicity (reduced viability) compared to non-transfected cells, more pronounced at the 72h times (**Figure 30a and b**) than 48 hours or 24 hours (**Appendix Figure 22**) after addition of

mitoxantrone or etoposide. Transfected U937 cells did not show significantly more sensitivity to these drugs compared to non-transfected cells (**Figure 30c** and **d**), likely because U937 already showed a high expression of *SMARCA4*. **Appendix Table 16** summarizes the AUC and IC₅₀ values of transfected/non-transfected cells for each cell line and each drug. This result supports the hypothesis that *SMARCA4* expression drives sensitivity to etoposide and mitoxantrone.

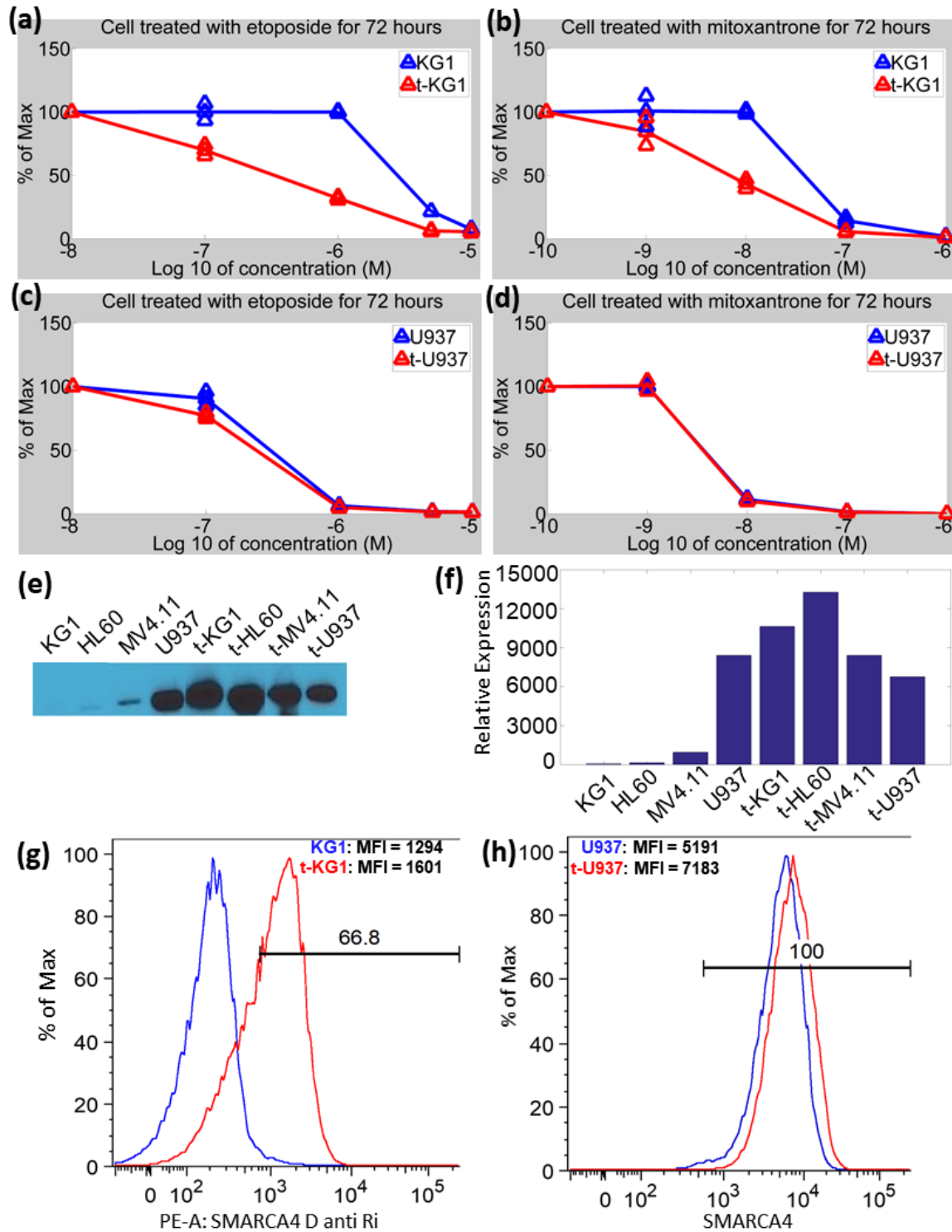


Figure 30. *SMARCA4 plasmid transfection experiments on two cell lines for comparison of response to etoposide and mitoxantrone between original vs. transfected cells.*

(a), (b): Comparison of the 72-hour dose-response curves between KG1 cells (blue) and transfected KG1 cells (red) when cells are treated with **(a)** etoposide, and **(b)** mitoxantrone. **(c), (d):** Comparison of the dose-response curves between U937 cells (blue) and transfected U937 cells (red) when cells are treated with **(c)** etoposide and **(d)** mitoxantrone. Three triangular marks at each point on the line indicate individual data points in duplicates and the average among them. The line connects averages of duplicates in each concentration measured. **(e)** Representative cropped Western blot of control and transfected AML cell lines: KG1, U937, HL60, and MV4.11. Uncropped version is shown in **Appendix Figure 21**. **(f)** Quantifications of the SMARCA4 protein expression pattern of each of AML cell lines in **(e)**. **(g)** Flow cytometry of SMARCA4 surface expression data confirms the overexpression for KG1 (blue) vs. transfected KG1 (red). **(h)** Flow cytometry of SMARCA4 surface expression data for U937 (blue) vs. transfected U937 (red). We note that U937 already strongly expresses SMARCA4, while KG1 exhibits minimal expression until after transfection. Abbreviations in **(d)** and **(f)** are as follows: PE-A: P-phycoerythrin area, MFI: Mean Fluorescence Intensity, D anti Ri: Donkey anti-Rabbit.

We repeated these experiments using two more cell lines -- HL60 and MV4.11 -- that show higher SMARCA4 expression levels than KG1 but lower levels than U937 (**Figure 30f**). HL60 shows lower SMARCA4 expression than MV4.11; thus, we would expect to see more increased sensitivity in HL60 than in MV4.11 when overexpressing SMARCA4. As expected, we saw increased sensitivity to etoposide when HL60 was transfected to overexpress SMARCA4 (**Appendix Figure 23i**) and no sensitivity change for MV4.11 (**Appendix Figure 23k** and **l**). HL60 had already been very sensitive to mitoxantrone before transfection to overexpress SMARCA4; thus, it was not surprising to see no change in sensitivity after transfection (**Appendix Figure 23j**). HL60 had been as sensitive to mitoxantrone as U937, which is why overexpressing SMARCA4 did not lead to a change in mitoxantrone response of

HL60.

4.4 DISCUSSION

Due to the small sample size and the potential confounding factors in the gene expression and the drug sensitivity data, standard methods to discover gene-drug associations usually fail to identify replicable signals. We present a new way to identify robust gene-drug associations by prioritizing genes based on the multi-dimensional information on each gene's potential to drive cancer. We demonstrate that our method increases the chance that the identified gene-drug associations are replicated in validation data. This leads us to a short list of genes which are all attractive biomarkers for different classes of drugs. Our results – including the expression, drug sensitivity data, and association statistics from patient samples – have been made freely available to academic communities.

Our results suggest that high *SMARCA4* expression could be a molecular marker for sensitivity to topoisomerase II inhibitors in AML cells. These results offer a potentially enormous impact to improve patient response. Mitoxantrone is an anthracycline, like daunorubicin or idarubicin, and one of the two component classes of drugs included in nearly all upfront AML treatment regimens. It is also included (the "M") in the CLAG-M regimen¹²¹, a triple drug component upfront regimen now being studied as GCLAM¹²². Mitoxantrone and etoposide (also a topoisomerase II inhibitor) are 2 of the 3 drugs in the MEC regimen¹²³, used together with cytarabine, as a common regimen for relapsed/refractory AML. Many modern regimens are in clinical trials that add an investigational drug to the MEC backbone, for example, an antibody to CXCR4 (NCT01120457) or an E selectin inhibitor (NCT02306291) in combination, or decitabine priming preceding the MEC regimen¹²⁴. Identifying a predictor

of response to mitoxantrone based on clinically available biospecimens, such as leukemic blast gene expression measured prior to treatment, could potentially increase median survival rates for patients with high expression of *SMARCA4* and indicate alternative therapies for patients with low *SMARCA4* expression.

The AML patients used in our study were consecutively enrolled on a protocol to obtain laboratory samples for research. They were selected solely based on sufficient leukemia cell numbers. As the patient samples were consecutively obtained and not selected for any specific attribute, we postulated that they were representative of patients seen at a tertiary referral center and that the results would be relevant to a larger, more general clinical population. Moreover, since each of the datasets from which we collected prior information (driver features) contained many more than 30 samples (e.g., TCGA AML data), it would be highly likely that MERGE results would be more generalizable to larger clinical populations than the methods that retrieve results specifically based on the 30 AML samples. In fact, **Figure 26a** and **b** imply higher generalizability of MERGE compared to alternative methods.

While we have genotype information on *FLT3* and *NPM1* and the cytogenetic risk category for most of the 30 patients, the current version of the MERGE framework did not take these features into account: our main focus sought to build a general framework that could address the high-dimensionality challenge (i.e., the number of samples being much smaller than the number of genes) and make efficient use of expression data to identify robust associations. However, to consolidate our findings, we performed a covariate analysis to confirm that the top-ranked gene-drug associations discovered by MERGE remained significant when the risk group/cytogenetic features were considered in the association analysis. We checked whether the gene-drug

associations shown in the heat map in **Figure 29b** (highlighted as red or green) were conserved when we added each of the following as an additional covariate to the linear model: (1) cytogenetic risk, (2) *FLT3* mutation status, and (3) *NPM1* mutation status. In **Appendix Figure 20**, each dot corresponds to a gene-drug pair, and each color to a different covariate. Most of the dots being closer to the diagonal indicates that the associations did not decrease significantly after adding the covariates. Moreover, of 357 dots, only 8 were below the horizontal red line; this indicates that 98% of the gene-drug associations MERGE uncovered were still significant ($p \leq 0.05$) after modeling the covariate.

Chapter 5. SUPERVISED LEARNING OF EXPRESSION MARKERS BY A PATHWAY-GUIDED HIERARCHICAL BAYESIAN APPROACH

[The work presented in this chapter is under peer review.]

Gene expression measured in a tissue sample provides a detailed molecular profile of genes and may drive complex phenotypes such as cancer survival or lesions like plaques or tangles in Alzheimer's disease (AD). Accurate inference of meaningful associations between gene expression and phenotypes is of paramount importance to make biological discoveries and identify therapeutic targets. There are two challenges with expression-phenotype association analysis: First, standard approaches, such as Pearson's correlation or regression models,^{110,125} produce gene-level relevance scores (e.g., association p -values or regression coefficients), which do not *explain* the association between a gene and a phenotype. Given that genes function as part of pathways, pathway-level relevance scores in addition to the gene-level scores may help (i) gain insights into the etiology of complex diseases which involve multiple disturbed pathways and (ii) pinpoint specific genes likely relevant to the phenotype, instead of false positives. Second, due to the high-dimensional nature of gene expression data (i.e., the number of genes is much greater than the number of samples), false positive associations are common. Incorporating appropriate prior knowledge has been a successful approach in various kinds of inference tasks in computational biology which involves high-dimensional data with experimental noise and confounders.

To resolve these challenges, we propose EMBARKER (EMpowering Big data with prior Knowledge for Expression markeR discovery), a probabilistic model-based framework to identify reliable gene expression markers for a phenotype by leveraging growing availability of gene-

pathway membership information, where a pathway is as a set of genes involved in the same biological process. EMBARKER takes as input gene expression and phenotype data collected from a set of samples and a gene-pathway membership matrix, and it outputs both gene-level and pathway-level relevance scores to the phenotype (**Figure 31**). Genes known to be involved in the same pathway (e.g., according to KEGG) are more likely to show similar association levels with a phenotype than those not. This assumption has been used by many authors (e.g., GSEA²⁶ or pathway-based GWAS studies¹²⁶; see the ‘Related Work’). However, simply aggregating the association statistics of all genes assigned to the same pathway is unlikely to be successful because: (1) not all genes within the same pathway would exert on the phenotype, and (2) knowledge on gene memberships of pathways is not complete. We need a “soft” way to integrate pathway information (i.e., gene-pathway membership) with observed gene-phenotype association signals. EMBARKER uses principled probabilistic modeling and learning techniques to explicitly model the pathway-level association signals as ‘prior’ and observed associations as ‘data-likelihood.’ It automatically determines how much prior information to take and optimizes the overall probabilistic objective function by estimating gene-level associations and pathway-level associations — α and β , respectively, in **Figure 31**. Our work introduces a novel hierarchical Bayesian model that utilizes known pathways and enables pathway-phenotype associations improving interpretability, accuracy, robustness, and clinical translatability. Our computational experiments demonstrate that EMBARKER dramatically improves statistical robustness of the identified markers in 42 different genome-wide expression dataset-complex phenotype pairs.

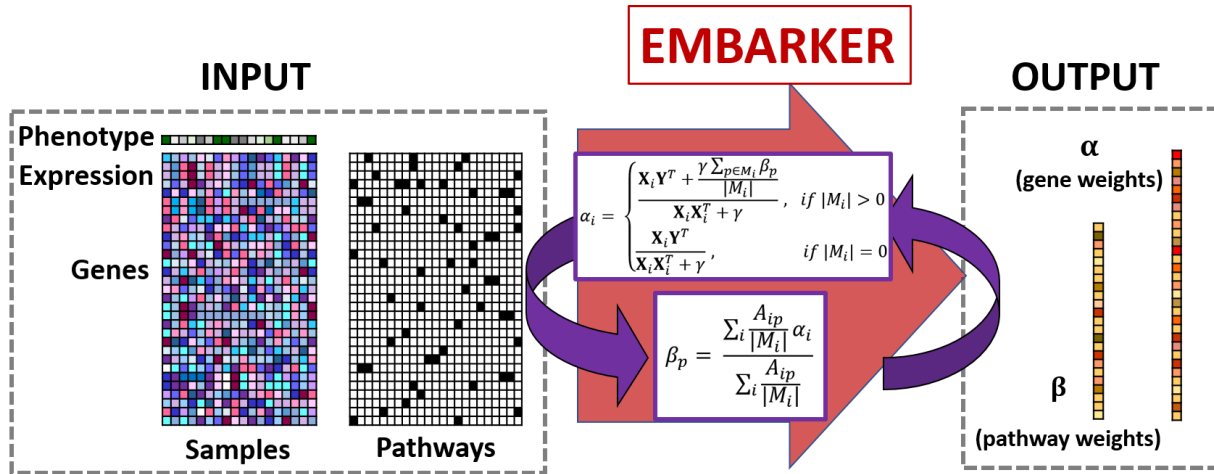


Figure 31. The EMBARKER methodology.

Based on a hierarchical Bayesian probabilistic model of the observed gene expression and phenotype data, we identify scores for the genes as represented by α and the pathways as represented by β .

EMBARKER is a general framework which can be applied to many diseases or phenotypes that involve identifying marker genes and pathways for a phenotype. We apply the EMBARKER approach to the problem of identifying the molecular basis for neuropathological phenotypes in AD. AD is the 6th most common cause of death in the United States with no effective therapy to delay or prevent its onset or progression. The currently known genetic modifiers account for only a small fraction of AD risk^{127,128} and provide no clear picture of AD neuropathology mechanisms. The major bottleneck is the lack of molecular understanding of neuropathological phenotypes, such as the levels of A β , the building block of plaques. Recently, multiple AD studies^{129,130,131,132,133} have provided gene expression and neuropathology measures which can be used to identify processes underlying AD neuropathology. However, this involves significant challenges because (1) the datasets include data from different brain regions with different cell type compositions, different functions, and different relevance to AD, (2) each study uses slightly

different methods to measure the neuropathology phenotype, and (3) there are confounding factors due to differences in subject compositions or experimental procedures used to generate data (i.e., batch effects). These various sources of noise in addition to the high-dimensionality problem create a challenging setting in which EMBARKER's power to leverage prior knowledge can address. Thus, we extended EMBARKER to handle multiple datasets, namely EMBARKER-M, and applied it to 1,742 autopsied human brain tissue samples from 9 brain regions in 3 studies, which is, to our knowledge, the largest expression meta-analysis for AD. We demonstrate that EMBARKER-M outperforms 19 alternative single dataset or meta-analysis approaches in identifying robust expression patterns for AD.

In addition to these computational innovations, here we also present new biological findings from our subsequent *in vivo* experimental validation of the top gene-phenotype associations identified by EMBARKER-M for AD neuropathology. The results from these experiments in a transgenic *C. elegans* model expressing AD-associated pathology is aligned with our hypothesis that the computationally identified markers are potential modifiers of tolerance to toxicity resulting from AD neuropathology. This successful *in vivo* validation pinpoints mitochondrial Complex I as a critical mediator of proteostasis and a potential target for treating AD.

5.1 RELATED WORK

Related work on gene marker identification methods **using a single dataset:**

A simple yet commonly used approach to identify genes that are associated with a specific phenotype is to focus on the gene expression levels resulting in a high absolute coefficient from a

Pearson's correlation with the phenotype. However, due to the small sample size and confounding factors, the marker genes identified based on PCC are poorly reproduced and the associated p -values usually become insignificant after multiple hypothesis testing correction.²⁶ MERGE was developed to learn a global ranking of the genes for acute myeloid leukemia (AML) drug response by using a number of gene features (e.g., mutation frequency, copy number variation, regulatory potential) to reduce false positive associations.¹⁰³ While this approach was shown to be effective for cancer, it is not clear which gene features to use for poorly understood diseases such as AD. Moreover, since MERGE learns a global ranking of the genes based on multiple phenotypes (e.g. response to >50 chemotherapy drugs), it is not obvious how to adapt it to a setting with a single phenotype. L_1 - or L_2 - penalized multiple regression methods (lasso, and ridge regression, respectively) or the elastic net¹¹⁰ which linearly combines the L_1 and L_2 penalties could be used to learn an ordering of the genes for the phenotype. Group lasso and overlapping group lasso (OGL)¹³⁴, multiple regression methods that encourage the genes in the same pathway to be assigned similar association levels (i.e., weights), were developed for better interpretability of the learned gene weights. However, multiple regression methods were shown to produce poorly reproducible top genes in high-dimensional settings.¹⁰³

Unreproducible and uninterpretable genetic association results drove researches to devise ways to incorporate prior pathway membership information of the genes to genome-wide association studies (GWAS)¹²⁶. Because gene expression constitutes intermediate processes between genotype and phenotype, there have also been efforts that incorporate pathway information in association analyses for gene expression. Gene Set Enrichment Analysis (GSEA)²⁶ is a well-known technique that determines whether the member genes of a pathway are randomly distributed throughout *the association vector* or primarily found at the two tails of it, the latter

being the expectation if the pathway is related to the phenotypic distinction. EMBARKER is distinct from GSEA in that it is a probabilistic model-based approach that enables simultaneous learning of gene-level and pathway-level associations with a phenotype. A recent model-based approach that incorporates pathway information to gene expression analysis is PLIER¹³⁵ which is a matrix factorization method that uses pathway information to learn an informative pathway-guided low-dimensional representation (LDR) from gene expression data. PLAGE¹³⁶, another unsupervised method that aims to identify pathways associated with a phenotype based on gene expression data, determines the activity level of a pathway based on the first eigenvector from a singular value decomposition (SVD) performed on the expression levels for the genes in the pathway. While the learned LDR could be useful for subsequent association analysis for complex disease phenotypes, PLIER or PLAGE, being unsupervised by nature, do not assign phenotype association statistics to the genes or pathways. Voyle et al.¹³⁷ reported that the pathway-level blood gene expression learned by PLAGE did not better classify AD vs. control samples than the actual gene expression levels. EMBARKER, on the other hand, is an interpretable supervised method that enables identification of phenotype association statistics for gene expression in both gene level and pathway level. Some approaches like PARADIGM¹³⁸, Pathifier¹³⁹, iPAS¹⁴⁰, and individPath¹⁴¹ identify deregulated pathways in individual tumor samples using a number of control samples and prior pathway information. EMBARKER is different than these approaches in that rather than identifying disease-associated pathways in specific patient samples, it aims to provide an improved general understanding of gene-level and pathway-level processes, particularly for multifactorial and poorly understood diseases like AD.

Related work on gene marker identification methods using a meta-analysis approach:

In addition to modeling gene expression-phenotype relation in a probabilistic framework using pathway prior information to identify robust and interpretable markers, another innovation of this work is providing an extension of this model to multiple datasets from different studies that include heterogeneous samples. Combining data or estimations from multiple studies to increase the confidence of association results (i.e., meta-analysis) is a well-studied research problem. Fisher's method¹⁴² is a widely used statistical procedure for meta-analysis that computes a summary p -value based on the average logarithm of the study-specific p -values. Stouffer's method¹⁴³ is based on Z-scores rather than p -values where the Z-score for study i is computed by $Z_i = \Phi^{-1}(1 - p_i)$ where Φ is the standard normal cumulative distribution function. Stouffer's method allows weighting studies equally or differently. For the marker identification problem, Pearson's correlation p -values for the gene-phenotype associations are supplied to Fisher's or Stouffer's method and the genes with resulting lowest meta-analysis p -value are considered reliable markers. Olkin-Pratt method¹⁴⁴ is a fixed effect model based on where the standard errors of the study-specific estimates are adjusted based on the sampling variance. Dersimonian-Laird method¹⁴⁵ is the most commonly used random effects model that accounts for both the heterogeneity among the studies and the sampling variance of each study. Both Olkin-Pratt and Dersimonian-Laird methods can work with Pearson's correlation coefficients as effect sizes. EMBARKER-M, the extension of our EMBARKER framework to the meta-analysis setting, is distinct from all the aforementioned meta-analysis approaches in that it embodies the two proposed innovations, i.e., (1) incorporation of pathway information to improve robustness and

interpretability of identified markers, and (2) combination of multiple datasets in an end-to-end interpretable Bayesian framework followed by a successful experimental validation *in vivo*.

In our experiments, we extensively compare EMBARKER to 15 previous approaches including Pearson’s correlation, elastic net, OGL, PLIER, PLAGE, GSEA, Fisher’s method, Stouffer’s method, Olkin-Pratt method, and Dersimonian-Laird method.

5.2 EMBARKER MODEL

Bayesian probabilistic models provide a principled way to incorporate prior knowledge into the learning process. They enable modeling the uncertainty over the model parameters by using a posterior distribution, which consists of prior beliefs about the parameters (‘prior’) and the empirical evidence (‘data-likelihood’), based on the Bayes’ theorem. A hierarchical Bayesian model represents the posterior distribution in multiple levels (i.e., a hierarchical form) and it enables modeling the complex structure in many real-world problems. We utilize hierarchical Bayesian probabilistic modeling to incorporate pathway prior knowledge to the evidence from observed data through the EMBARKER framework. As discussed earlier, genes known to be involved in the same pathway (e.g., KEGG) are more likely to show similar association levels with a phenotype than those not. This assumption has been used by several authors (e.g., GSEA²⁶ or pathway-based GWAS studies¹²⁶), as mentioned in the ‘Related Work’ section. The idea is that we model the prior distribution of gene-level associations based on the pathway-level associations, and jointly learn these associations for all genes and for all pathways.

Let $\mathbf{X} \in \mathbb{R}^{G \times S}$ be a matrix representing observed gene expression measurements on G genes for S samples, and $\mathbf{Y} \in \mathbb{R}^{1 \times S}$ be a row vector representing an observed disease phenotype measured for the same samples. $\mathbf{X}$ is standardized such that each row has a zero-mean and unit variance, which is a standard data pre-processing procedure for any method which uses prior information or regularization on the variables (i.e., genes). The phenotype vector $\mathbf{Y}$ is also standardized. Let $\boldsymbol{\alpha} \in \mathbb{R}^{G \times 1}$ be a parameter vector that represents the relevance of each gene to the phenotype (gene-level scores), and $\boldsymbol{\beta} \in \mathbb{R}^{P \times 1}$ a parameter vector that represents the relevance of each pathway to the phenotype (pathway-level scores). Let $\mathbf{A} \in \mathbb{R}^{G \times P}$ be a binary matrix given to the algorithm, representing the assignment of genes to pathways based on pathway databases.

Let the disease outcome $\mathbf{Y}$ depend on each gene's expression level ($\mathbf{X}_i$) and its relevance to the disease outcome (α_i), which is represented by the 'data-likelihood' term $P(\mathbf{Y}|\mathbf{X}_i, \boldsymbol{\alpha})$. We put a prior on α_i that is dependent on the relevance of the pathways the gene is included in to the disease outcome ($\boldsymbol{\beta}$). To learn the parameters of this model ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$), we solve an optimization problem minimizing the error term $\sum_{i=1}^G \|\mathbf{Y} - \mathbf{X}_i \alpha_i\|_2^2$ subject to a regularization:

$$\min \sum_{i=1}^G \{\|\mathbf{Y} - \mathbf{X}_i \alpha_i\|_2^2 + \gamma F(\alpha_i, \boldsymbol{\beta})\}$$

where

$$F(\alpha_i, \boldsymbol{\beta}) = \begin{cases} \frac{1}{|M_i|} \sum_{p \in M_i} (\alpha_i - \beta_p)^2, & |M_i| > 0 \\ \alpha_i^2, & |M_i| = 0 \end{cases}$$

is the regularization term. Here, $M_i = \{p \mid A_{ip} = 1\}$ is the set of pathways gene i is present in, and $|M_i|$ represents the number of elements in M_i (i.e., $|M_i| = \sum_{p=1}^P A_{ip}$). The regularization term

encourages each gene's relevance to be close to the relevance of the pathways the gene belongs to (if there is any). The factor $1/|M_i|$ in $F(\alpha_i, \boldsymbol{\beta})$ performs normalization on the regularization term such that the genes with many functions (i.e., that are involved in many pathways) are not regularized too heavily. γ is a regularization parameter where a higher γ value means more weight on the prior pathway information. When chosen by a principled statistical test, we expect the resulting γ value to be higher when the statistical signal available from the data is low; for example, when the data are noisy, the sample size is too small, or the brain region is less affected by AD disease processes. Such an approach allows EMBARKER to adapt its learning process to observed data characteristics, and it is likely to lead to a better performance than the standard data-driven approaches in identifying true expression markers.

Equation (18) gives the solution for α_i as:

$$\alpha_i = \begin{cases} \frac{\mathbf{X}_i \mathbf{Y}^T + \frac{\gamma \sum_{p \in M_i} \beta_p}{|M_i|}}{\mathbf{X}_i \mathbf{X}_i^T + \gamma}, & |M_i| > 0 \\ \frac{\mathbf{X}_i \mathbf{Y}^T}{\mathbf{X}_i \mathbf{X}_i^T + \gamma}, & |M_i| = 0 \end{cases}, \quad (19)$$

For the genes assigned to no pathways by the A matrix (i.e., $|M_i| = 0$), $F(\alpha_i, \boldsymbol{\beta})$ is equivalent to an L_2 (ridge) penalty and does not depend on $\boldsymbol{\beta}$. For the genes with $|M_i| > 0$, the value of $\boldsymbol{\beta}$ is used while computing α_i . Maximizing the joint log-likelihood in Equation (18) with respect to β_p for each pathway p leads to the following optimization problem:

$$\beta_p = \frac{\sum_i \frac{A_{ip}}{|M_i|} \alpha_i}{\sum_i \frac{A_{ip}}{|M_i|}}. \quad (20)$$

which is equal to the weighted average α_i value over all genes assigned to the pathway p .

We initialize α with $\mathbf{X}_i \mathbf{Y}^T$ and iteratively update the values of α and β until convergence.

Adapting EMBARKER model to meta-analysis of multiple datasets

With the growing availability of expression and phenotype data, we often can find multiple datasets from multiple studies focused on the same disease. However, it is usually not clear how to combine these data effectively to identify gene expression markers for the disease. Taking AD neuropathology as an example, available data are from different studies and/or brain regions involving study-specific confounders and/or region-specific biological processes. EMBARKER framework naturally extends to multiple datasets to do meta-analysis in such a challenging multi-study or multi-region setting by incorporating the pathway information to the meta-analysis setting.

Equation (21) provides the EMBARKER update rule for each gene's phenotype association (α_i) in this meta-analysis setting as:

$$\alpha_i = \begin{cases} \frac{\frac{\sum_{r=1}^R \mathbf{X}_i^{(r)} \mathbf{Y}^{(r)T}}{R} + \frac{\gamma \sum_{p \in M_i} \beta_p}{|M_i|}}{\frac{\sum_{r=1}^R \mathbf{X}_i^{(r)} \mathbf{X}_i^{(r)T}}{R} + \gamma}, & |M_i| > 1 \\ \frac{\frac{\sum_{r=1}^R \mathbf{X}_i^{(r)} \mathbf{Y}^{(r)T}}{R}}{\frac{\sum_{r=1}^R \mathbf{X}_i^{(r)} \mathbf{X}_i^{(r)T}}{R} + \gamma}, & |M_i| = 0 \end{cases}, \quad (21)$$

where R is the number of available datasets (i.e., number of brain regions in our AD application), superscript (r) means the region from which the data are collected.

5.3 EXPERIMENTAL RESULTS

5.3.1 Overview of the EMBARKER framework

Here we present EMBARKER which attempts to identify robust expression markers for complex phenotypes, which can be used to explain the molecular basis for disease processes based on the known pathways and identify potential therapeutic targets for complex diseases. In addition to using genome-wide gene expression and a disease phenotype, EMBARKER leverages prior knowledge on gene groups responsible for functional pathways to filter out noise and extract information likely to be enriched for true signals from *big* high-throughput data (**Figure 31**).

Given G genes, S samples, and P pathways, EMBARKER takes as input: (1) $\mathbf{X}$ – expression data matrix ($\in \mathbb{R}^{G \times S}$) and $\mathbf{Y}$ – phenotype data vector ($\in \mathbb{R}^{1 \times S}$), and (2) $\mathbf{A}$ – a binary matrix indicating gene-pathway membership ($\in \mathbb{1}^{G \times P}$), and produces two association results: (1) $\boldsymbol{\alpha}$ – gene-level associations ($\in \mathbb{R}^{G \times 1}$), and (2) $\boldsymbol{\beta}$ – pathway-level associations ($\in \mathbb{R}^{P \times 1}$) (**Figure 31**). EMBARKER naturally extends to a setting where multiple $\mathbf{X}$ and $\mathbf{Y}$ pairs are provided as input. This extension, named EMBARKER-M, enables an effective utilization of samples from multiple tissues and/or studies and leads to a higher statistical power and a further improvement in robustness of identified gene-phenotype association signals.

EMBARKER can use any pathway database that provides a grouping of genes. Throughout this paper, we used the 674 REACTOME pathways from the current version of the Molecular Signatures Database (MSigDB) C2²⁶ (curated gene sets from online pathway databases) to create the pathway membership indicator matrix $\mathbf{A}$. We chose REACTOME because it is the largest pathway database in MSigDB C2 providing a wide range of pathways representing molecular

functions and cellular processes including signal transduction, transport, DNA replication, and metabolism.¹⁴⁶

5.3.2 *EMBARKER's way to incorporate pathway information leads to improved robustness and relevance of the identified gene and pathway markers.*

A notable strength of EMBARKER is that it learns *both* gene-level and pathway-level phenotype relevance scores, while most methods focus on only one of these two aspects. Therefore, we compared EMBARKER to relevant alternative methods based on the generalizability and robustness of (i) identified gene-level markers (**Figure 32**) and (ii) identified pathway-level markers (**Figure 33**). We performed experiments in two different case studies involving a total of 42 genome-wide gene expression datasets. APPENDIX C provide the details on data collection and preprocessing.

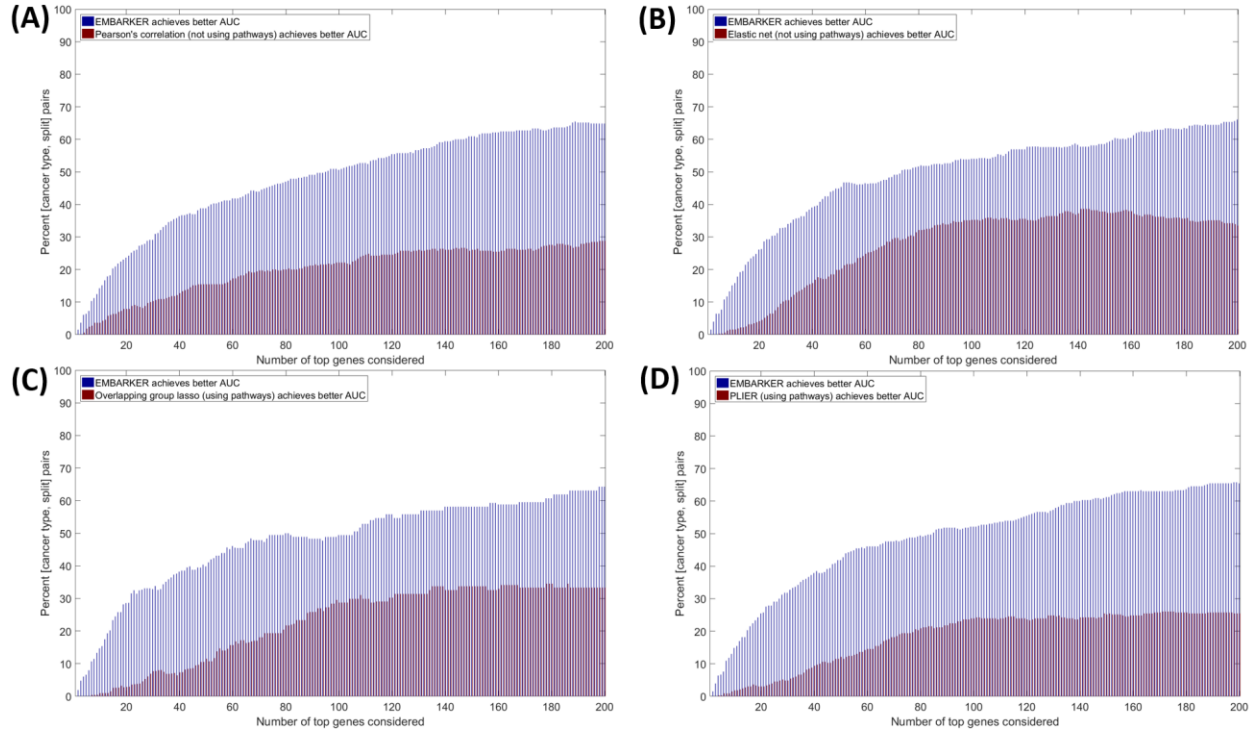
- **Case Study 1:** Identifying markers for survival of cancer patients in each of 33 different cancer types from TCGA study (**Appendix Table 23**; data available on suinlee.cs.washington.edu/projects/embarker); and
- **Case Study 2:** Identifying markers for AD-related neuropathology of human individuals in 9 brain regions from 3 studies (**Appendix Table 24**; data available on suinlee.cs.washington.edu/projects/embarker).

We first sought to test whether our model could identify robust gene-level expression markers (learned α vector). We compared our method to four alternative methods to identify the top relevant genes for the phenotype. Two of these methods, (3) and (4), utilize pathway information as EMBARKER does.

- (1) **Pearson’s correlation** between each gene expression level and the phenotype;
- (2) **Elastic net**.¹¹⁰ A sparse multiple regression method which linearly combines the L_1 and L_2 penalties to learn gene weights for predicting the phenotype;
- (3) **Overlapping group lasso (OGL)**.¹³⁴ An extension of the L_1 -regularized multiple regression where genes in predefined overlapping groups are allowed to have zero or nonzero weights together; and
- (4) **Pathway-Level Information ExtractoR (PLIER)**.¹³⁵ An unsupervised method which reconstructs gene expression data by learning a low-dimensional representation (LDR) guided by the pathway information. We adapted this method as a marker identification method by computing Pearson’s correlation between each reconstructed gene expression level (i.e., product of the learned Z and B matrices) and the phenotype.

We used the area under the ‘statistical consistency curve’ (AUC) (details in APPENDIX C) as the robustness metric. We first tested the statistical consistency of the top genes identified by EMBARKER for Case Study 1 (patient survival in 33 different cancer types). **Figure 32A-D** show the result of this experiment for $M = 1$ to 200 top markers out of 15,097 genes. Each panel corresponds to a comparison of EMBARKER to an alternative method. Blue bars show the percent runs of EMBARKER with a better consistency than the alternative method, and red bars show the percent runs of the alternative method with a better consistency than EMBARKER. Each *split* for a cancer type corresponds to one of the 10 different random half-half training-test splits of the data for that cancer type. The blue bars are always higher than the red bars, which means that EMBARKER identifies more robust gene-level markers than the methods in comparison. We show the detailed scatter plot that compares EMBARKER with the four methods in terms of the actual AUC value for each cancer type and split pair for $M = 200$ in **Appendix Figure 29A**. We then repeated the same experiment for our Case Study 2 (9 problems: AD neuropathology in 9 brain regions from 3 studies). The outperformance of EMBARKER was consistent for this task as shown

for $M = 1$ to 200 top markers out of 14,912 genes in **Figure 32E-H** and as a detailed scatter plot for $M = 200$ in **Appendix Figure 29B**.



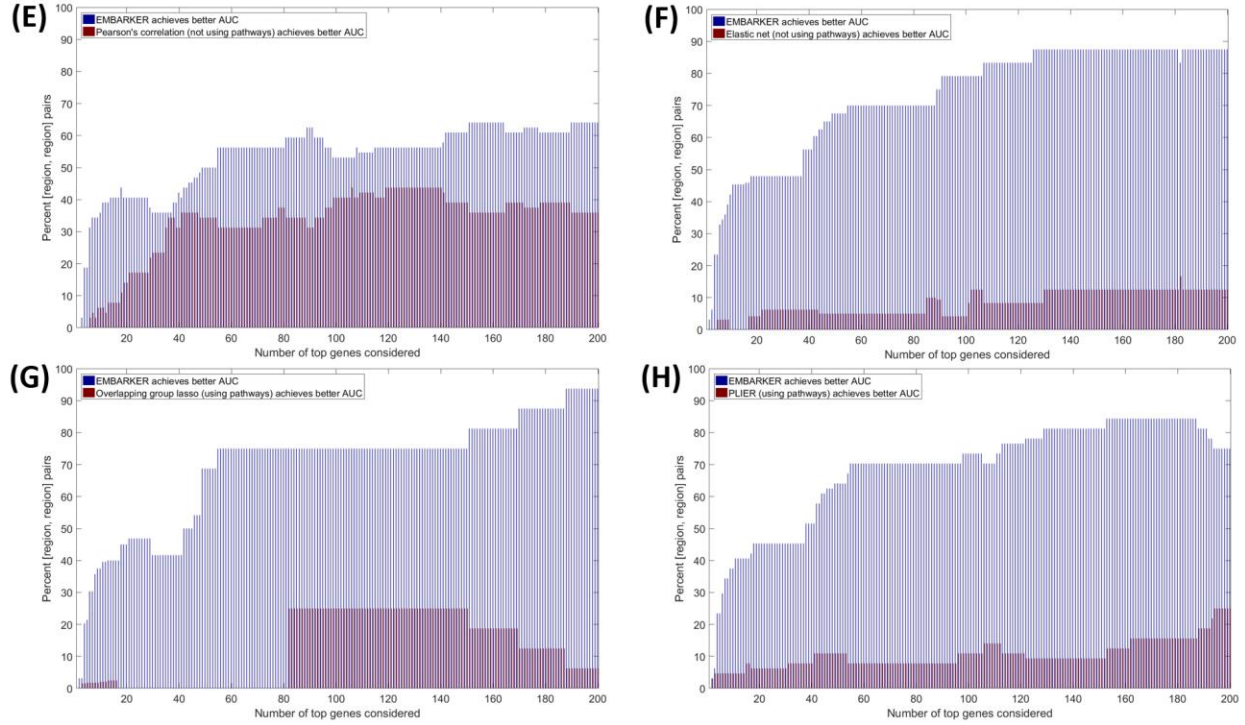


Figure 32. Statistical consistency of the gene markers learned by EMBARKER vs. alternative methods.

Consistency plots comparing gene-level marker statistical consistency achieved by EMBARKER (blue bars) vs. the four alternative methods (red bars) in **(A-D)** TCGA survival case study, and **(E-H)** AD neuropathology case study. For varying number of top gene markers considered (x-axis), the percent runs that achieved a higher consistency AUC (**Appendix Figure 27**) is reported for each method in comparison.

In addition to the gene-level markers, EMBARKER also learns pathway-level expression markers (learned β vector). Identifying pathway-level markers is an important strength of EMBARKER since they might provide better insights in underlying disease mechanisms than the individual gene-level markers. We compared EMBARKER to four alternative methods that learn pathway-level marker scores:

- (1) **Mean expression:** Pathway expression levels are computed by averaging the expression levels of the genes in the pathway, and the pathway markers are identified based on the coefficients from Pearson's correlation of each pathway's expression with the phenotype.
- (2) **Pathway Level Analysis of Gene Expression (PLAGE):**¹³⁶ Pathway expression levels are computed by using SVD performed on the expression levels of the genes in the pathway, and the pathway markers are identified by computing Pearson's correlation of each pathway's expression with the phenotype.
- (3) **Gene Set Enrichment Analysis (GSEA):**²⁶ Each pathway is assigned a p -value based on a permutation test that tests whether the member genes of the pathway are randomly distributed throughout the Pearson's correlation coefficients (PCC) vector or primarily found at the two tails of it.
- (4) **Pathway-Level Information ExtractoR (PLIER):**¹³⁵ We adapted PLIER method as a pathway marker identification method by computing Pearson's correlation between each reconstructed pathway expression level (i.e., product of the learned U and B matrices) and the phenotype.

We again used AUC as the robustness metric, where a higher AUC means a higher robustness (*Appendix Figure 27*). *Figure 33* shows the result of this experiment for $M = 1$ to 50 top markers out of 674 pathways for both case studies: TCGA survival in *Figure 33A-D* and AD neuropathology in *Figure 33E-H*. We present a more detailed comparison of the actual AUC values for $M = 50$ in *Appendix Figure 30A* and *Appendix Figure 30B*. EMBARKER almost always outperforms the four alternative methods in identifying robust pathway-level markers. High consistency achieved by EMBARKER for both gene-level and pathway-level markers in both case studies demonstrates the power of our approach in identifying robust expression markers and its generalizability for different diseases and phenotypes.

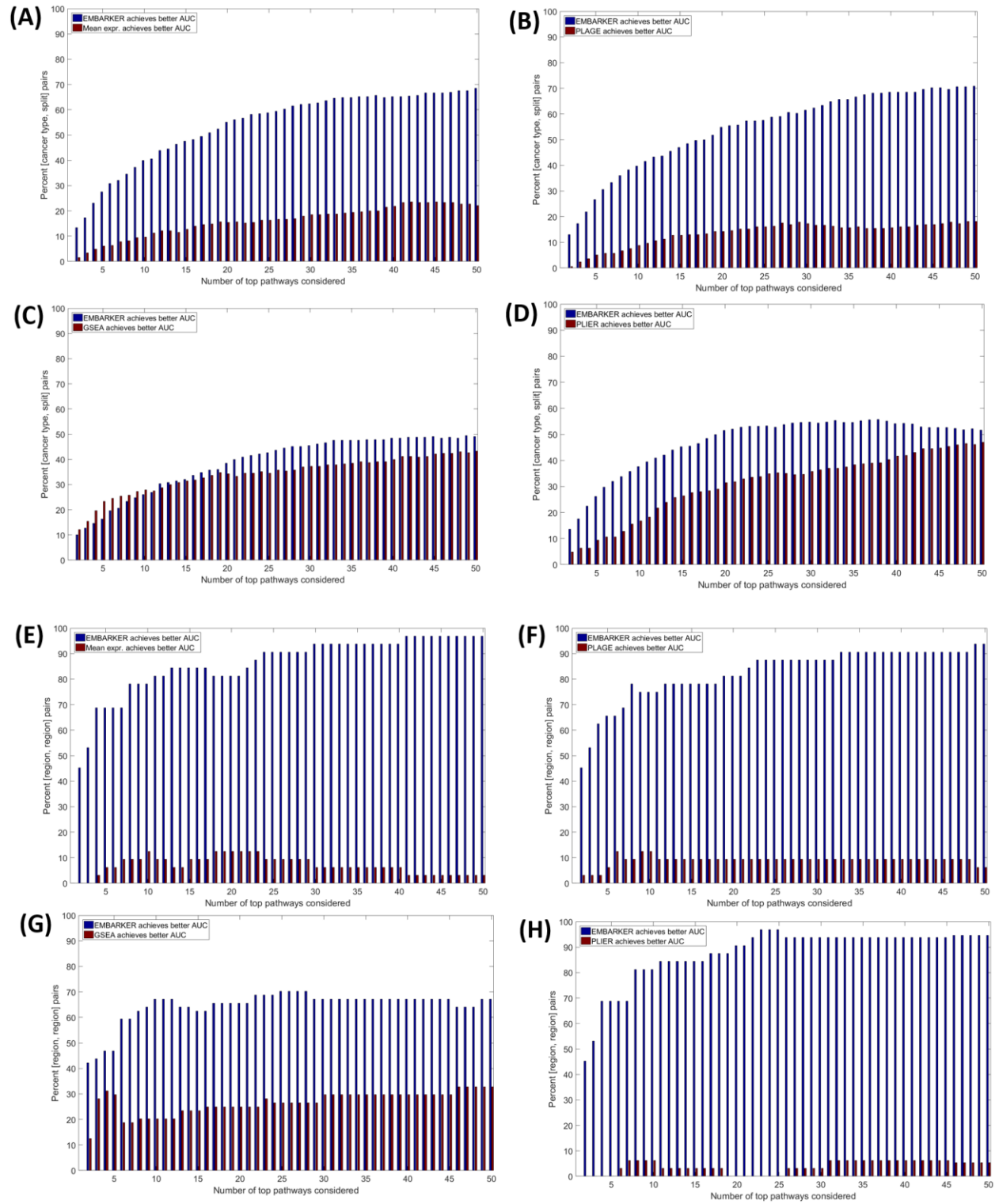


Figure 33. Statistical consistency of the pathway markers learned by EMBARKER vs. alternative methods.

*Consistency plots comparing pathway-level marker statistical consistency achieved by EMBARKER (blue bars) vs. the four alternative methods (red bars) in (A-D) TCGA survival case study, and (E-H) AD neuropathology case study. For varying number of top pathway markers considered (x-axis), the percent runs that achieved a higher consistency AUC (**Appendix Figure 27**) is reported for each method in comparison.*

5.3.3 EMBARKER incorporates the prior information based on the amount of relevant information available in the data.

As discussed earlier, a notable strength of EMBARKER is that it can choose how much of the pathway information (i.e., gene-pathway membership matrix **A**) it needs to incorporate with the observed data. This is done by determining the value of the hyperparameter γ selected through cross-validation (CV) experiments (details in APPENDIX C). A high γ value means that EMBARKER chose to use much of the pathway information, in which case the expression and phenotype data have relatively low level of gene-phenotype association signals; low γ value means that EMBARKER used relatively more from the observed gene-phenotype association signals. That is, γ is a hyperparameter that adjusts the importance of ‘prior’ vs. ‘data-likelihood’ in the probabilistic objective.

Not surprisingly, the γ values selected by our CV experiments were negatively correlated with the previously known relevance of the brain regions to AD in each cohort (**Appendix Figure 28A** and **B**; **Appendix Table 29**). Out of four ACT regions with similar sample size, hippocampus selected the smallest γ followed by the cortical regions temporal and parietal cortex (**Appendix Figure 28A**). This is consistent with our knowledge that AD starts near hippocampus and then

spreads to temporal and parietal lobes.¹⁴⁷ Expression and phenotype data collected from hippocampus likely contain more association signals than those from other regions. White matter, suggested to be less vulnerable to AD than gray matter,¹⁴⁸ selected the largest γ among conditions we tested. We observed a similar pattern for MSBB (**Appendix Figure 28B**), where Brodmann area 36 (ectorrhinal cortex) was assigned the lowest γ ; Brodmann area 36 coincides with the medial temporal lobe, the brain part where AD-related cellular and structural alterations begin and have a more severe effect than other regions.¹⁴⁹ On the other hand, Brodmann area 10 (frontopolar cortex), which is known to be affected in later stages of AD progression¹⁵⁰ and would lead to weak expression-phenotype association signals, was assigned a relatively large γ .

These observations demonstrate the ability of EMBARKER to adapt the learning process (i.e., adaptively incorporate prior pathway information) depending on the amount of available gene expression association signals relevant to the phenotype.

5.3.4 *EMBARKER's natural extension to a multi-dataset setting is applied to nine brain regions to learn expression markers for AD neuropathology.*

There are often many datasets available for a disease and making use of all samples on these datasets to identify expression markers has been an effective strategy to increase statistical power.¹⁵¹ However, this is a complex undertaking because available data are often from different studies involving study-specific confounders. For AD, incorporating samples from different datasets into a single framework is even more challenging because the available expression and neuropathology data are usually from different brain regions which might involve region-specific

biological processes. Moreover, different brain regions are not equally relevant to AD as discussed in the previous section (**Appendix Figure 28**).

Most meta-analysis methods explicitly (e.g., Olkin-Pratt method) or implicitly (e.g., Fisher's method) highly weight densely sampled datasets. For example, when we use these methods for a meta-analysis of nine brain regions (**Appendix Table 24**), ROSMAP-DPFCx would be highly weighted since it contains much more samples than the other regions. However, ROSMAP-DPFCx, containing measurements from dorsolateral prefrontal cortex which is usually affected later as AD progresses,¹⁵⁰ may not be as relevant to AD as many other regions in **Appendix Table 24**. This is consistent with the observation that γ value ROSMAP-DPFCx was assigned by EMBARKER was close to the γ value selected by ACT_HIP (hippocampus), while the sample size of ROSMAP-DPFCx is about seven times the sample size of ACT_HIP. EMBARKER, as demonstrated earlier, extracts signals from gene expression data relevant to a particular phenotype by effectively utilizing prior pathway information. Thus, it is a suitable method for extracting information in this noisy AD meta-analysis setting. Here we extend EMBARKER to a multi-dataset setting and name this extended version of our approach EMBARKER-M, where “M” stands for “multiple datasets.” In our experiments to test EMBARKER-M, we used the three AD studies that we used earlier to test EMBARKER (Case Study 2), containing a total of 1,742 post-mortem brain tissue samples from 9 brain regions (**Appendix Table 24**).

5.3.5 *EMBARKER-M learns more robust markers than the alternative methods to combine multiple regions.*

We performed a CV experiment to test whether gene expression markers identified by EMBARKER-M are statistically more robust than the markers from alternative methods that attempt to incorporate association statistics from different regions into a single model. These are either methods designed for meta-analysis or our extension of the commonly used PCC to multiple regions:

- (1) **Fisher’s method:**¹⁴² A widely used method that computes a summary p -value based on the average logarithm of the study-specific p -values.
- (2) **Stouffer’s method:**¹⁴³ Similar to Fisher’s method, but based on Z-scores rather than p -values where the Z-score for study i is computed by $Z_i = \Phi^{-1}(1 - p_i)$ where Φ is the standard normal cumulative distribution function. This method allows weighting studies differently.
- (3) **Stouffer’s method (no weights):** Simpler version of Stouffer’s method where each study is equally weighted like in Fisher’s method.
- (4) **Olkin-Pratt method:**¹⁴⁴ A fixed effect model based on correlation coefficients rather than the p -values or Z-scores. The standard errors of the study specific estimates are adjusted based on the sample size.
- (5) **Dersimonian-Laird method:**¹⁴⁵ Similar to Olkin-Pratt method, but a random effects model accounting for the potential heterogeneity across the studies.
- (6) **Concatenated data PCC:** We concatenated the samples from different regions after standardizing each gene and the phenotype in each region. We then computed Pearson’s correlation between each gene and the phenotype and used the resulting coefficients as the gene scores.

- (7) **Mean of region PCCs:** We computed Pearson’s correlation of a gene’s expression with the phenotype in each region and used the mean for each gene across the regions as the score for that gene.
- (8) **PC1 of region PCCs:** We computed Pearson’s correlation of a gene’s expression with the phenotype in each region and applied PCA on the $\#genes (14,912) \times \#regions (9)$ correlation matrix where we used each region as a variable. Then we used the first principal component (PC1) as the gene scores.

For one of the 9 regions used as a test fold, we used the other 8 regions for training. We then computed the consistency between the markers learned by each training run. We then averaged the results from each method over the test folds. **Figure 34** shows the statistical consistency curve (details in APPENDIX C), the exact same metric we used to generate results in **Figure 32** and **Figure 33**, while here we show the original curve instead of reporting the percent runs achieving a better curve. This is possible because here, we deal with only one phenotype association problem, which was 33 or 9 for our case studies from which we presented results in **Figure 32** and **Figure 33**. As demonstrated in **Figure 34**, gene markers identified by EMBARKER-M have a higher overlap with the markers identified on the test region than all alternative meta-analysis methods, which indicates a higher robustness of the markers identified by EMBARKER-M. EMBARKER-M enables leveraging the expression and phenotype data available in different datasets in a single framework by leveraging the prior pathway information. EMBARKER-M’s outperformance over the alternative approaches to combine multiple regions implies that our way to incorporate prior pathway information with the observed expression and phenotype data leads to retrieval of important signals out of the potential high amount of noise introduced by discrepancies across different datasets.

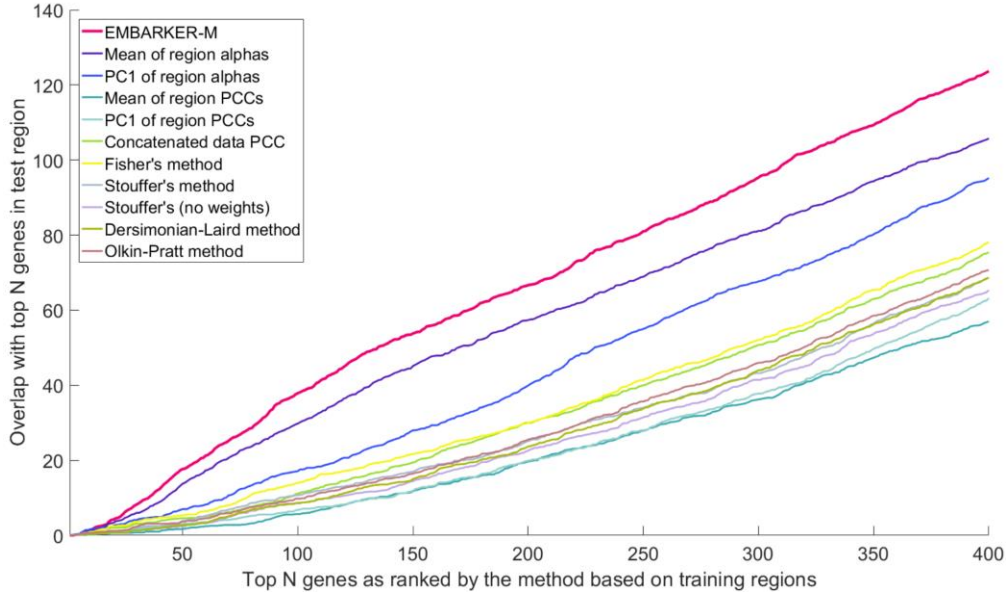


Figure 34. Statistical consistency of the gene markers learned by EMBARKER-M vs. alternative meta-analysis methods.

The average statistical consistency results from a cross-validation (CV) experiment where each of the 9 regions is used as a separate fold. EMBARKER-M applied on 9 brain regions is compared to the alternative methods that can combine the 9 regions.

5.3.6 Markers from EMBARKER-M are highly enriched for AD-related pathways.

We first computationally investigated the biological relevance of our findings. For this, we examined the functional enrichment of identified gene markers. There are 237 significant gene-neuropathology PCC on average over 9 regions after Bonferroni correction. Therefore, we chose the top 237 genes ranked by absolute value of the learned α vector as “identified gene markers”. For functional enrichment analysis, we considered 1,077 REACTOME, BioCarta, and KEGG canonical pathways from the C2 collection of the MSigDB.²⁶ We note that out of these 1,077 pathways, we only used 674 REACTOME pathways for training our EMBARKER-M model (i.e.,

for creating the **A** matrix). We computed the significance of the overlap between each pathway and the identified markers measured by using Fisher’s exact test (FET) p-value, and then applied false discovery rate (FDR) correction for multiple hypotheses testing.

As shown in **Table 2**, the pathways that are significantly enriched for the top gene markers identified by EMBARKER-M include pathways for neurodegenerative diseases AD, Parkinson’s disease and Huntington’s disease, and the electron transport chain and oxidative phosphorylation, implicated in a variety of neurodegenerative processes.^{152–154} **Appendix Table 25** lists the learned EMBARKER-M gene-level scores for all 14,912 genes used in our experiment. **Appendix Table 26** lists the learned EMBARKER-M pathway-level scores for 674 REACTOME pathways we used to train our EMBARKER-M model. We note that out of the 1,077 MSigDB C2 pathways we used for enrichment analysis, we only used 674 REACTOME pathways for training our model.

Table 2. FDR-corrected Fisher’s exact test (FET) p-values for the top pathways for which the identified markers from EMBARKER-M were enriched.

We cut the list at the 7th pathway because this was the elbow point — the next corrected p-value > 10^{-20} .

Pathway	FET p-value (FDR corrected) from EMBARKER-M
TCA Cycle and Respiratory Electron Transport (from REACTOME)	7.7e-112
Respiratory Electron Transport ATP Synthesis by Chemiosmotic Coupling and Heat Production by Uncoupling Proteins (from REACTOME)	2.4e-95
Oxidative Phosphorylation (from KEGG)	3.7e-80
Respiratory Electron Transport (from REACTOME)	5.7e-78
Parkinson’s Disease (from KEGG)	2.0e-73
Huntington’s Disease (from KEGG)	5.6e-66
Alzheimer’s Disease (from KEGG)	1.2e-65

5.3.7 *In vivo validation of EMBARKER-M identifies mitochondrial Complex I as a potential AD therapeutic target.*

Next, to gain insight into the biological relevance of markers identified by EMBARKER-M to A β toxicity, we utilized the nematode *C. elegans*, a well-established animal model of A β proteotoxicity. We conducted experiments with a transgenic worm line displaying an age-associated aggregation of human A β ₁₋₄₂ peptide resulting in rapid onset of age-associated paralysis.¹⁵⁵ A stringent reciprocal best hits (RBH) approach (BLAST *e*-value $\leq 10^{-30}$; details in APPENDIX C) was used to identify nematode orthologs for the human genes. The top gene from EMBARKER-M was *NDUFA9*, which encodes the alpha subcomplex subunit 9 of Complex I of the mitochondrial electron transport chain (mETC). The nematode ortholog of *NDUFA9* is Y53G8AL.2, a previously uncharacterized protein.

To first confirm that Y53G8AL.2 is a functional ortholog of *NDUFA9*, we performed mitochondrial respiration assays to measure oxygen consumption in whole animals. As expected for a subunit of Complex I, which is the largest enzyme complex in the mETC,¹⁵⁶ RNAi knockdown of Y53G8AL.2 resulted in a dramatic reduction in oxygen consumption (**Figure 35A**). Based on this result and the high sequence homology to *NDUFA9*, we conclude that Y53G8AL.2 is likely the true ortholog of *NDUFA9* and refer to Y53G8AL.2 hereafter as *nduf-9*, consistent with the *C. elegans* nomenclature.

We next used RNAi by bacterial feeding¹⁵⁷ to reduce expression of *nduf-9* in animals expressing toxic A β . This resulted in a striking delay in paralysis relative to control animals treated with empty vector (EV) RNAi (**Figure 35B**) (details in APPENDIX C). Suppression of paralysis by knockdown of *nduf-9* was comparable to RNAi knockdown of the insulin-like

receptor gene *daf-2* (**Figure 35B**), one of the strongest known suppressors of A β toxicity in worms.¹⁵⁸

In addition to *NDUFA9*, we observed that 5 of the top 10 markers identified by EMBARKER-M were Complex I genes. We therefore hypothesized that Complex I may act as a key regulator of A β toxicity. To test this hypothesis, we individually knocked down 12 additional *C. elegans* genes that encode homologs of human Complex I proteins (**Appendix Table 27**) in A β ₁₋₄₂ worms. Strikingly, knockdown of any of these 12 additional Complex I RNAi clones significantly delayed paralysis, with several exceeding the effect of *nduf-9* knockdown (**Figure 35C and D**).

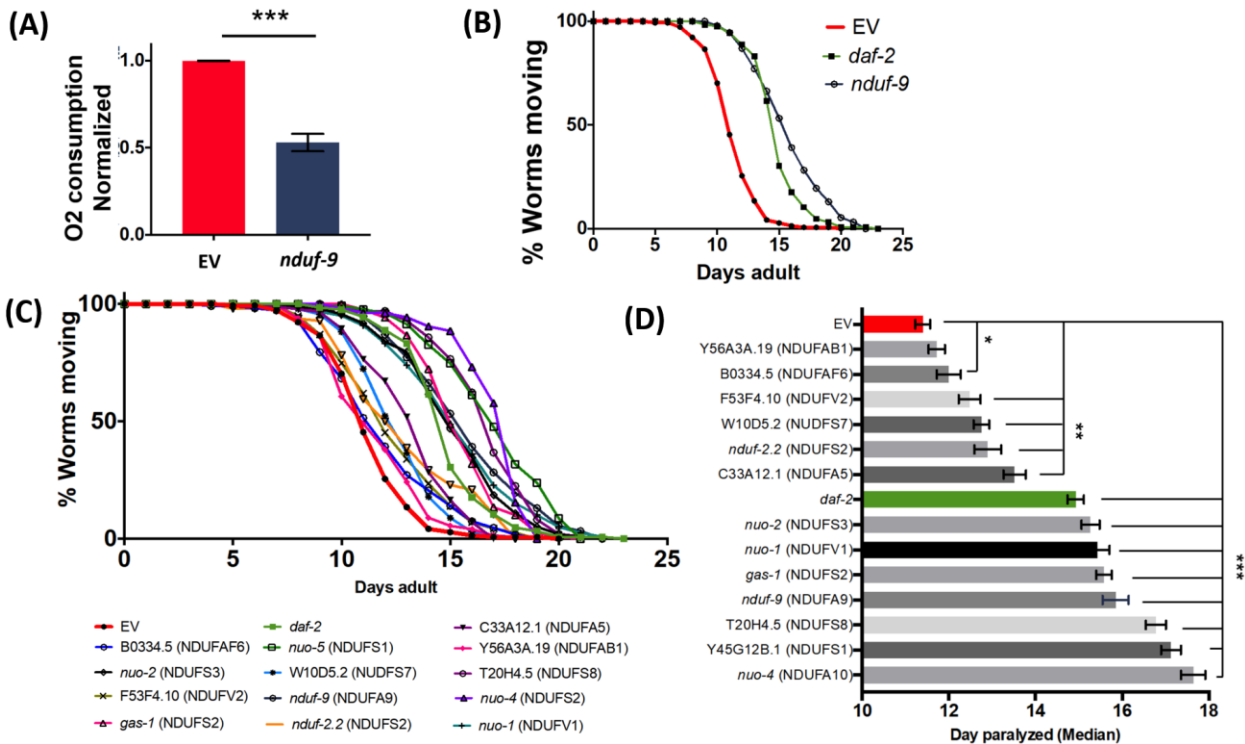


Figure 35. In-vivo experimental validation of NDUFA9 and the other Complex-I genes.

(A) RNAi knockdown of *NDUFA9* homolog *nduf-9* greatly reduces whole animal oxygen consumption. O₂ consumption is normalized to the O₂ consumption rate of one animal on control RNAi (EV). (B) Knockdown of the *NDUFA9* homolog *nduf-9* strongly suppresses A β -induced paralysis and this suppression is comparable to the potent proteotoxicity suppression

treatment with *daf-2* RNAi. **(C)** Paralysis curves for the reciprocal best orthologs of human Complex I genes. All tested RNAi conditions significantly suppress paralysis. **(D)** The same data as in (C) plotted as the median day of paralysis for each population of worms. Half of the conditions showed even stronger suppression than *daf-2* and *nduf-9* RNAi conditions. Error bars mean standard error of the mean across experiments. * $p\text{-value} < 0.05$. ** $p\text{-value} < 0.01$. *** $p\text{-value} < 0.001$.

5.4 DISCUSSION

Here, we present EMBARKER, a machine learning approach that utilizes prior knowledge on gene membership to pathways to identify gene-level and pathway-level expression markers for disease phenotypes or other general biological traits. We demonstrate that EMBARKER's way to incorporate pathway information to marker identification discovery dramatically increases the rate of true positive markers. Moreover, since it learns each pathway's phenotype association in addition to each gene's, EMBARKER enables better interpretation of the identified markers, which is essential for the translation of the markers to clinical practice.

Results from our computational and *in vivo* biological experiments pinpoint Complex I of the mETC as a critical mediator of A β proteostasis and mild inhibition of Complex I as a potential new paradigm for developing future AD therapeutics. This finding is supported by prior data indicating that mild inhibition of Complex I reduced A β and tau levels in mouse models of familial AD,¹⁵⁹ a diet rich of capsaicin — a natural product that can inhibit Complex I — was associated with lower serum A β levels in the elderly,¹⁶⁰ and capsaicin reduced AD-associated changes in the hippocampus of type 2 diabetes rats.¹⁶¹

Chapter 6. CONCLUSIONS

Gene expression measured in a tissue sample is an informative proxy to the protein levels in that tissue and known to drive several disease outcomes such as cancer survival, chemotherapy drug response, or lesions like plaques and tangles in Alzheimer’s disease (AD). Robust identification of genes whose expression has a driving role in disease outcome is an important and commonly studied problem, and drugs developed to target these novel genes might offer promising ways for treatment.

Our goal in this dissertation is to understand biological and disease mechanisms, and ultimately discover markers that would explain disease mechanisms and potentially be used as drug targets. We discussed three challenges (*high-dimensionality*, *study-specific confounders*, and *different molecular elements*) with applying machine learning on genomics data, specifically on gene expression data. Due to these challenges, the potential for overfitting is high when we solely rely on the samples provided to the learning algorithm. This implies that models that allow for arbitrarily rich dependencies among variables (such as those used in deep learning methods) cannot necessarily be applied without overfitting the data. A successful machine learning method needs to reduce the dimensionality of the problem and/or incorporate reasonable prior biological knowledge that allows generalization across datasets. We provided four different approaches we developed to resolve the aforementioned challenges using effective and biologically intuitive approaches.

MGL is an aggressive dimensionality reduction technique to learn an informative low-dimensional representation from a high-dimensional gene expression data. INSPIRE extends MGL to multiple expression datasets with only a partial overlap among their variables. Both MGL and

INSPIRE *assume* an existence of groups of genes responsible for a similar function and aim to *learn* them, while EMBARKER utilizes *known* gene sets to identify robust and reliable expression markers for disease phenotypes. We apply EMBARKER to integrating data from different brain regions to identify expression markers for Alzheimer’s disease to potentially be used as therapeutic targets. MERGE incorporates multi-omic prior information in terms of different cancer driver features to identify robust and reliable expression markers for cancer drug response mechanisms.

We anticipate that the methods we introduce will increase the reliability, generalizability and robustness of learned biomarkers, facilitate personalized treatment of complex and heterogeneous diseases like cancer or Alzheimer’s disease, and pave the way to precision medicine.

BIBLIOGRAPHY

1. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
2. Greene, C. S. The future is unsupervised. *Sci. Transl. Med.* **8**, 346ec108 (2016).
3. Celik, S., Logsdon, B. A. & Lee, S.-I. Efficient Dimensionality Reduction for High-Dimensional Network Estimation. in *International Conference on Machine Learning* (2014).
4. Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **61**, 611–622 (1999).
5. Toh, H. & Horimoto, K. Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics* **18**, 287–297 (2002).
6. He, Y., Qi, Y., Kavukcuoglu, K. & Park, H. Learning the dependency structure of latent factors. in *NIPS* (2012).
7. Chandrasekaran, V., Parrilo, P. A. & Willsky, A. S. Latent variable graphical model selection via convex optimization. *The Annals of Statistics* **40**, 1935–1967 (2012).
8. Ambroise, C., Chiquet, J. & Matias, C. Inferring sparse Gaussian graphical models with latent structure. *Electron. J. Stat.* **3**, 205–238 (2008).
9. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–38 (1977).
10. Guo, J. & Wang, S. Modularized gaussian graphical model. *Comput. Stat. Data Anal.* (2010).
11. Duchi, J. & Gould, S. Projected subgradient methods for learning sparse gaussians. *Twenty-*

- fourth Conf.* 145–152 (2008). at <<http://uai.sis.pitt.edu/papers/08/p153-duchi.pdf>>
12. Schmidt, M. W., Berg, E. Van Den, Friedlander, M. P. & Murphy, K. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. *Int. Conf. Artif. Intell. Stat.* **5**, 456–463 (2009).
 13. Marlin, B. M., Schmidt, M. & Murphy, K. P. Group sparse priors for covariance estimation. in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* 383–392 (2009). at <<http://dl.acm.org/citation.cfm?id=1795159%5Cnpapers2://publication/uuid/1FD0ECB6-5C32-47BC-A124-1ED550274FC6>>
 14. Marlin, B. M. & Murphy, K. Sparse gaussian graphical models with unknown block structure. in *ICML* (2009).
 15. Mardia, K. V, Kent, J. T. & Bibby, J. M. *Multivariate Analysis*. (Academic Press, 1979). at <<http://www.amazon.com/dp/0124712525>>
 16. Lauritzen, S. L. *Graphical Models*. (Oxford University Press, 1996).
 17. Yuan, M. & Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35 (2007).
 18. Banerjee, O., El Ghaoui, L. & d’Aspremont, A. Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data. *J. Mach. Learn. Res.* **9**, 485–516 (2008).
 19. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–441 (2008).
 20. Zhao, T., Liu, H. & Roeder, K. The huge package for high-dimensional undirected graph estimation in r. *J. Mach. ...* **13**, 1059–1062 (2012).

21. Bast, R. C., Hennessy, B. & Mills, G. B. The biology of ovarian cancer: new opportunities for translation. *Nat. Rev. Cancer* **9**, 415–28 (2009).
22. Akavia, U. D. *et al.* An integrated approach to uncover drivers of cancer. *Cell* **143**, 1005–1017 (2010).
23. Bell, D. *et al.* Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
24. Tothill, R. W. *et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin. Cancer Res.* **14**, 5198–5208 (2008).
25. Denkert, C. *et al.* A prognostic gene expression index in ovarian cancer - Validation across different independent data sets. *J. Pathol.* **218**, 273–280 (2009).
26. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).
27. Coussens, L. M. & Werb, Z. Inflammation and cancer. *Nature* **420**, 860–867 (2002).
28. Pietras, K., Sjoblom, T., Rubin, K., Heldin, C.-H. & Ostman, A. PDGF receptors as cancer drug targets. *Cancer Cell* **3**, 439–443 (2003).
29. Celik, S. *et al.* Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer. *Genome Med.* **8**, (2016).
30. Unsupervised Feature Learning and Deep Learning Tutorial.
31. Långkvist, M., Karlsson, L. & Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.* **42**, 11–24 (2014).
32. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

33. Szegedy, C., Toshev, A. & Erhan, D. Deep Neural Networks for Object Detection. in *Advances in Neural Information Processing Systems* 2553–2561 (2013).
34. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
35. Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Sci. Transl. Med.* **5**, 181re1 (2013).
36. Cheng, W.-Y., Yang, T.-H. O. & Anastassiou, D. Development of a Prognostic Model for Breast Cancer Survival in an Open Challenge Environment. *Sci. Transl. Med.* **5**, 181ra50 (2013).
37. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted gene co-expression network analysis. *BMC Bioinformatics* **9**, 559 (2008).
38. Sherlock, G. Analysis of large-scale gene expression data. *Brief. Bioinform.* **2**, 350–62 (2001).
39. Lee, S.-I. & Batzoglou, S. in *Advances in Neural Information Processing Systems 16* (eds. Thrun, S., Saul, L. & Schölkopf, B.) (MIT Press, 2004).
40. Koller, D. & Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. (The MIT Press, 2009).
41. Chuang, H.-Y., Lee, E., Liu, Y.-T., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**, 140 (2007).
42. Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T. & Lee, D. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.* **4**, (2008).
43. Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat. Biotechnol.* **27**, 199–204 (2009).

44. Ravasi, T. *et al.* An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell* **140**, 744–752 (2010).
45. Herschkowitz, J. I. *et al.* Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.* **8**, R76 (2007).
46. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
47. Lee, S.-I. & Batzoglou, S. Application of independent component analysis to microarrays. *Genome Biol.* **4**, R76 (2003).
48. Jiang, D., Tang, C. & Zhang, A. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering* **16**, 1370–1386 (2004).
49. INSPIRE web page: <http://inspire.cs.washington.edu>.
50. Tseng, P. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.* **109**, 475–494 (2001).
51. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning. Book 2nded*, (2001).
52. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
53. Konstantinopoulos, P. A. *et al.* Gene expression profile of BRCAness that correlates with responsiveness to chemotherapy and with outcome in patients with epithelial ovarian cancer. *J. Clin. Oncol.* **28**, 3555–3561 (2010).
54. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26**, 2438–44 (2010).

55. Vastrik, I. *et al.* Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8**, R39 (2007).
56. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277-80 (2004).
57. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* **41**, (2013).
58. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–15 (2013).
59. Segal, E. *et al.* Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **34**, 166–176 (2003).
60. Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* **45**, 1127–1133 (2013).
61. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
62. Way, G. P. *et al.* High-grade serous ovarian cancer subtypes are similar across populations. *Biorxiv* (<http://dx.doi.org/10.1101/030239>)
63. Wels, J., Kaplan, R. N., Rafii, S. & Lyden, D. Migratory neighbors and distant invaders: tumor-associated niche cells. *Genes Dev.* **22**, 559–74 (2008).
64. Le Blanc, K. & Mougiakakos, D. Multipotent mesenchymal stromal cells and the innate immune system. *Nat. Rev. Immunol.* **12**, 383–396 (2012).
65. Heuvers, M. E. *et al.* Patient-tailored modulation of the immune system may revolutionize future lung cancer treatment. *BMC Cancer* **12**, 580 (2012).
66. Liu, H. *et al.* Therapeutic potential of perineural invasion, hypoxia and desmoplasia in

- pancreatic cancer. *Curr. Pharm. Des.* **18**, 2395–403 (2012).
67. Riester, M. *et al.* Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. *J. Natl. Cancer Inst.* **106**, dju048- (2014).
68. Puisieux, A., Brabletz, T. & Caramel, J. Oncogenic roles of EMT-inducing transcription factors. *Nat. Cell Biol.* **16**, 488–94 (2014).
69. Ilić, D. *et al.* Reduced cell motility and enhanced focal adhesion contact formation in cells from FAK-deficient mice. *Nature* **377**, 539–44 (1995).
70. Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N. & Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* **39**, D235-40 (2011).
71. Barker, T. H. *et al.* SPARC regulates extracellular matrix organization through its modulation of integrin-linked kinase activity. *J. Biol. Chem.* **280**, 36483–93 (2005).
72. Dong, J. *et al.* VEGF-null cells require PDGFR alpha signaling-mediated stromal fibroblast recruitment for tumorigenesis. *EMBO J.* **23**, 2800–10 (2004).
73. Singh, A. & Settleman, J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. *Oncogene* **29**, 4741–51 (2010).
74. Seton-Rogers, S. Layers of regulation. *Nat. Rev. Cancer* **11**, 689 (2011).
75. Yang, J. & Weinberg, R. A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Dev. Cell* **14**, 818–29 (2008).
76. Gentles, A. J., Plevritis, S. K., Majeti, R. & Alizadeh, A. A. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *JAMA* **304**, 2706–15 (2010).
77. López-Novoa, J. M. & Nieto, M. A. Inflammation and EMT: an alliance towards organ

- fibrosis and cancer progression. *EMBO Mol. Med.* **1**, 303–14 (2009).
78. Steinmetz, R. *et al.* Mechanisms regulating the constitutive activation of the extracellular signal-regulated kinase (ERK) signaling pathway in ovarian cancer and the effect of ribonucleic acid interference for ERK1/2 on cancer cell proliferation. *Mol. Endocrinol.* **18**, 2570–82 (2004).
79. Tamura, R. E. *et al.* GADD45 proteins: central players in tumorigenesis. *Curr. Mol. Med.* **12**, 634–51 (2012).
80. Tetreault, M.-P., Yang, Y. & Katz, J. P. Krüppel-like factors in cancer. *Nat. Rev. Cancer* **13**, 701–13 (2013).
81. Gentles, A. J. *et al.* The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* (2015). doi:10.1038/nm.3909
82. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
83. Gravendeel, L. A. M. *et al.* Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res.* **69**, 9065–72 (2009).
84. Collisson, E. a. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–50 (2014).
85. Gentles, A. J. *et al.* A pluripotency signature predicts histologic transformation and influences survival in follicular lymphoma patients. *Blood* **114**, 3158–3166 (2009).
86. Chen, F. *et al.* Hop is an unusual homeobox gene that modulates cardiac development. *Cell* **110**, 713–723 (2002).
87. Kee, H. J. *et al.* Enhancer of polycomb1, a novel homeodomain only protein-binding partner, induces skeletal muscle differentiation. *J. Biol. Chem.* **282**, 7700–9 (2007).

88. Kook, H. *et al.* Cardiac hypertrophy and histone deacetylase-dependent transcriptional repression mediated by the atypical homeodomain protein Hop. *J. Clin. Invest.* **112**, 863–871 (2003).
89. Katoh, H. *et al.* Epigenetic Silencing of HOPX Promotes Cancer Progression in Colorectal Cancer. *Neoplasia* **14**, 559-IN6 (2012).
90. Waraya, M. *et al.* Cancer specific promoter CpG Islands hypermethylation of HOP homeobox (HOPX) gene and its potential tumor suppressive role in pancreatic carcinogenesis. *BMC Cancer* **12**, 397 (2012).
91. Chen, Y., Yang, L., Cui, T., Pacyna-Gengelbach, M. & Petersen, I. HOPX is methylated and exerts tumour suppressive function through Ras-induced senescence in human lung cancer. *J. Pathol.* (2014). doi:10.1002/path.4469
92. Jain, R. *et al.* Integration of Bmp and Wnt signaling by Hopx specifies commitment of cardiomyoblasts. *Science* (80-.). **348**, aaa6071-aaa6071 (2015).
93. Logsdon, B. A. *et al.* Sparse expression bases in cancer reveal tumor drivers. *Nucleic Acids Res.* **43**, 1332–1344 (2015).
94. Kim, J.-A., Choi, H.-K., Kim, T.-M., Leem, S.-H. & Oh, I.-H. Regulation of mesenchymal stromal cells through fine tuning of canonical Wnt signaling. *Stem Cell Res.* **14**, 356–368 (2015).
95. Macheda, M. L. & Stacker, S. a. Importance of Wnt signaling in the tumor stroma microenvironment. *Curr. Cancer Drug Targets* **8**, 454–65 (2008).
96. Savagner, P., Yamada, K. M. & Thiery, J. P. The zinc-finger protein slug causes desmosome dissociation, an initial and necessary step for growth factor-induced epithelial-mesenchymal transition. *J. Cell Biol.* **137**, 1403–19 (1997).

97. Mishra, P. J. *et al.* Carcinoma-associated fibroblast-like differentiation of human mesenchymal stem cells. *Cancer Res.* **68**, 4331–4339 (2008).
98. Li, N. *et al.* Single-Cell Analysis of Proxy Reporter Allele-Marked Epithelial Cells Establishes Intestinal Stem Cell Hierarchy. *Stem Cell Reports* **3**, 876–91 (2014).
99. Jain, R. *et al.* Plasticity of Hopx(+) type I alveolar cells to regenerate type II cells in the lung. *Nat. Commun.* **6**, 6727 (2015).
100. Ma, S. *et al.* Immunobiology of mesenchymal stem cells. *Cell Death Differ.* **21**, 216–25 (2014).
101. Karnoub, A. E. *et al.* Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. *Nature* **449**, 557–563 (2007).
102. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
103. Lee, S.-I. *et al.* A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.* **9**, (2018).
104. PhRMA. *Summer 2016 chart pack of the Pharmaceutical Research and Manufacturers of America (PhRMA)*. (2016).
105. Cheng, W.-Y., Yang, T.-H. O. & Anastassiou, D. Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput. Biol.* **9**, e1002920 (2013).
106. Haibe-Kains, B. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* **504**, 389–93 (2013).
107. The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia — NEJM. *N. Engl. J. Med.* **368**, 2059–2074 (2013).

108. Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–5 (2012).
109. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–7 (2012).
110. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **67**, 768–768 (2005).
111. Heiser, L. M. *et al.* Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2724–9 (2012).
112. Li, C., Li, M., Lange, E. M. & Watanabe, R. M. Prioritized subset analysis: improving power in genome-wide association studies. *Hum. Hered.* **65**, 129–41 (2008).
113. Roeder, K., Bacanu, S.-A., Wasserman, L. & Devlin, B. Using linkage genome scans to improve power of association in genome scans. *Am. J. Hum. Genet.* **78**, 243–52 (2006).
114. Genovese, C., Roeder, K. & Wasserman, L. False discovery control with p-value weighting. *Biometrika* **93**, 509–524 (2006).
115. Hildreth, C. A Quadratic Programming Procedure. *Nav. Res. Logist. Q.* **4**, 79–85 (1957).
116. Warga, J. Minimizing Certain Convex Functions. *SIAM J. Appl. Math.* **11**, 588–93 (1963).
117. Clark, S. J. & Melki, J. DNA methylation and gene silencing in cancer: which is the guilty party? *Oncogene* **21**, 5380–7 (2002).
118. Yuan, H., Paskov, I., Paskov, H., González, A. J. & Leslie, C. S. Multitask learning improves prediction of cancer drug sensitivity. *Sci. Rep.* **6**, 31619 (2016).
119. Pong, T. K., Tseng, P., Ji, S. & Ye, J. Trace Norm Regularization: Reformulations, Algorithms, and Multi-Task Learning. *SIAM J. Optim.* **20**, 3465–3489 (2010).
120. Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction

- algorithms. *Nat. Biotechnol.* **32**, 1–103 (2014).
121. Wierzbowska, A. *et al.* Cladribine combined with high doses of arabinoside cytosine, mitoxantrone, and G-CSF (CLAG-M) is a highly effective salvage regimen in patients with refractory and relapsed acute myeloid leukemia of the poor risk: a final report of the Polish Adult Leukem. *Eur. J. Haematol.* **80**, 115–26 (2008).
 122. Halpern, A. B. *et al.* Mitoxantrone, Etoposide, and Cytarabine (MEC) Following Epigenetic Priming with Decitabine in Adults with Relapsed/Refractory Acute Myeloid Leukemia (AML) or High-Risk Myelodysplastic Syndrome (MDS): Final Results from a Phase 1/2 Study. *Blood (Abstr)* **128**, 1068 (2016).
 123. Amadori, S. *et al.* Mitoxantrone, etoposide, and intermediate-dose cytarabine: an effective and tolerable regimen for the treatment of refractory acute myeloid leukemia. *J. Clin. Oncol.* **9**, 1210–4 (1991).
 124. Halpern, A. B. *et al.* Mitoxantrone, etoposide, and cytarabine (MEC) following epigenetic priming with decitabine in adults with relapsed/refractory acute myeloid leukemia or other high-grade myeloid neoplasms: A phase 1/2 study. *Leukemia* **31**, 2560–2567 (2017).
 125. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 267–288 (1996).
 126. Jin, L. *et al.* Pathway-based Analysis Tools for Complex Diseases: A Review. *Genomics. Proteomics Bioinformatics* (2014). doi:10.1016/j.gpb.2014.10.002
 127. Ridge, P. G., Mukherjee, S., Crane, P. K. & Kauwe, J. S. K. Alzheimer’s disease: Analyzing the missing heritability. *PLoS One* **8**, (2013).
 128. Ridge, P. G. *et al.* Assessment of the genetic variance of late-onset Alzheimer’s disease. *Neurobiol. Aging* **41**, (2016).

129. Zhang, B. *et al.* Integrated Systems Approach Identifies Genetic Nodes and Networks in Late-Onset Alzheimer's Disease. *Cell* **153**, 707–720 (2013).
130. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, 628–45 (2012).
131. Bennett, D. A. *et al.* Overview and findings from the rush Memory and Aging Project. *Curr. Alzheimer Res.* **9**, 646–63 (2012).
132. Wang, M. *et al.* Integrative network analysis of nineteen brain regions identifies molecular signatures and networks underlying selective regional vulnerability to Alzheimer's disease. *Genome Med.* **8**, 104 (2016).
133. Miller, J. A. *et al.* Neuropathological and transcriptomic characteristics of the aged brain. *Elife* **6**, (2017).
134. Yuan, L., Liu, J. & Ye, J. Efficient methods for overlapping group lasso. *IEEE Trans. Pattern Anal. Mach. Intell.* (2013). doi:10.1109/TPAMI.2013.17
135. Mao, W., Harmann, B., Sealfon, S. C., Zaslavsky, E. & Chikina, M. Pathway-Level Information ExtractoR (PLIER): a generative model for gene expression data. *Biorxiv* (<https://doi.org/10.1101/116061>) (2017).
136. Tomfohr, J., Lu, J. & Kepler, T. B. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* (2005). doi:10.1186/1471-2105-6-225
137. Voyle, N. *et al.* A Pathway Based Classification Method for Analyzing Gene Expression for Alzheimer's Disease Diagnosis. *J. Alzheimers. Dis.* (2015). doi:10.3233/JAD-150440
138. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btq182

139. Drier, Y., Sheffer, M. & Domany, E. Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1219651110
140. Ahn, T., Lee, E., Huh, N. & Park, T. Personalized identification of altered pathways in cancer using accumulated normal tissue data. in *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu449
141. Wang, H. *et al.* Individualized identification of disease-associated pathways with disrupted coordination of gene expression. *Brief. Bioinform.* (2016). doi:10.1093/bib/bbv030
142. Fisher, R. Statistical methods for research workers. *Biological monographs and manuals* (1925). doi:10.1056/NEJMc061160
143. Riley, J. W. *et al.* The American Soldier: Adjustment During Army Life. *Am. Sociol. Rev.* (1949). doi:10.2307/2087216
144. Olkin, I. & Pratt, J. W. Unbiased Estimation of Certain Correlation Coefficients. *Ann. Math. Stat.* (1958). doi:10.1214/aoms/1177706717
145. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control. Clin. Trials* (1986). doi:10.1016/0197-2456(86)90046-2
146. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1132
147. Masters, C. L. *et al.* Alzheimer's disease. *Nature Reviews Disease Primers* **1**, (2015).
148. Douaud, G. *et al.* Preventing Alzheimer's disease-related gray matter atrophy by B-vitamin treatment. *Proc. Natl. Acad. Sci.* **110**, 9523–9528 (2013).
149. Villain, N. *et al.* Relationships between Hippocampal Atrophy, White Matter Disruption, and Gray Matter Hypometabolism in Alzheimer's Disease. *J. Neurosci.* **28**, 6174–6181 (2008).

150. Masters, C. L. *et al.* Alzheimer's disease. *Nat. Rev. Dis. Prim.* 15056 (2015). at <http://dx.doi.org/10.1038/nrdp.2015.56>
151. Celik, S. *et al.* Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer. *Genome Med.* **8**, 66 (2016).
152. Federico, A. *et al.* Mitochondria, oxidative stress and neurodegeneration. *J. Neurol. Sci.* **322**, 254–262 (2012).
153. Patten, D. A., Germain, M., Kelly, M. A. & Slack, R. S. Reactive oxygen species: Stuck in the middle of neurodegeneration. *Journal of Alzheimer's Disease* **20**, (2010).
154. Morán, M. *et al.* Mitochondrial respiratory chain dysfunction: Implications in neurodegeneration. *Free Radical Biology and Medicine* **53**, 595–609 (2012).
155. McColl, G. *et al.* Utility of an improved model of amyloid-beta (A β 1-42) toxicity in *Caenorhabditis elegans* for drug screening for Alzheimer's disease. *Mol. Neurodegener.* 2012 71 **17**, 857–872 (2012).
156. Alberts B, Johnson A, Lewis J, *et al.* Electron transport chains and their proton pumps. *Mol. Biol. cell 4th Ed.* 1–10 (2002).
157. Kamath, R. S. & Ahringer, J. Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods* **30**, 313–321 (2003).
158. Steinkraus, K. A. *et al.* Dietary restriction suppresses proteotoxicity and enhances longevity by an hsf-1-dependent mechanism in *Caenorhabditis elegans*. *Aging Cell* **7**, 394–404 (2008).
159. Zhang, L. *et al.* Modulation of Mitochondrial Complex I Activity Averts Cognitive Decline in Multiple Animal Models of Familial Alzheimer's Disease. *EBioMedicine* **2**, 294–305

- (2015).
160. Liu, C. H. *et al.* The Associations between a Capsaicin-Rich Diet and Blood Amyloid- β Levels and Cognitive Function. *J. Alzheimer's Dis.* **52**, 1081–1088 (2016).
 161. Xu, W. *et al.* Capsaicin reduces Alzheimer-associated tau changes in the hippocampus of type 2 diabetes rats. *PLoS One* **12**, (2017).
 162. http://gdac.broadinstitute.org/runs/stddata__2014_03_16/data/OV/20140316/.
 163. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
 164. Bonome, T. *et al.* A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res.* **68**, 5478–5486 (2008).
 165. Hendrix, N. D. *et al.* Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas. *Cancer Res.* **66**, 1354–1362 (2006).
 166. Mok, S. C. *et al.* A Gene Signature Predictive for Outcome in Advanced Ovarian Cancer Identifies a Survival Factor: Microfibril-Associated Glycoprotein 2. *Cancer Cell* **16**, 521–532 (2009).
 167. Meyniel, J.-P. *et al.* A genomic and transcriptomic approach for a differential diagnosis between primary and secondary ovarian carcinomas in patients with a previous history of breast cancer. *BMC Cancer* **10**, 222 (2010).
 168. Ferriss, J. S. *et al.* Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: Predicting platinum resistance. *PLoS One* **7**, (2012).
 169. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. Affy - Analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307–315 (2004).

170. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez gene: Gene-centered information at NCBI. *Nucleic Acids Res.* **39**, (2011).
171. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, (2005).
172. Tibshirani, R. *et al.* Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74**, 245–266 (2012).
173. Marquardt, D. W. & Snee, R. D. Ridge Regression in Practice. *Source Am. Stat.* **29**, 3–20 (1975).
174. Sardy, S. On the practice of rescaling covariates. *Int. Stat. Rev.* **76**, 285–297 (2008).
175. de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D. & Calus, M. P. L. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345 (2013).
176. Schmidt, M., Niculescu-Mizil, A. & Murphy, K. P. Learning Graphical Model Structure using L1-Regularization Paths. *Proc. AAAI Conf. Artif. Intell.* **22**, 1278 (2007).
177. Mu, B. & How, J. P. Learning Sparse Gaussian Graphical Model with l 0 -regularization. *Tech. Rep.* 1–13 (2014).
178. Friedman, J., Hastie, T. & Tibshirani, R. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. *Tech. Rep.* 1–22 (2010).
179. Lee, S.-I., Pe’er, D., Dudley, A. M., Church, G. M. & Koller, D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 14062–14067 (2006).
180. Lee, S. I. *et al.* Learning a prior on regulatory potential from eQTL data. *PLoS Genet.* **5**, (2009).

181. cBio Cancer Genomics Portal: <http://cbioportal.org>.
182. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
183. Josse, J., Chavent, M., Liquet, B. & Husson, F. Handling Missing Values with Regularized Iterative Multiple Correspondence Analysis. *J. Classif.* **29**, 91–116 (2012).
184. Witten, D. M., Friedman, J. H. & Simon, N. New Insights and Faster Computations for the Graphical Lasso. *J. Comput. Graph. Stat.* **20**, 892–900 (2011).
185. Harrell, F. E., Lee, K. L. & Mark, D. B. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med.* **15**, 361–387 (1996).
186. ImageJ: <http://imagej.nih.gov/ij/>.
187. Hotchkiss, R. S. & Nicholson, D. W. Apoptosis and caspases regulate death and inflammation in sepsis. *Nat. Rev. Immunol.* **6**, 813–822 (2006).
188. Yiu, G. K. *et al.* SPARC (secreted protein acidic and rich in cysteine) induces apoptosis in ovarian cancer cells. *Am. J. Pathol.* **159**, 609–622 (2001).
189. Thiede, C. *et al.* Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: Association with FAB subtypes and identification of subgroups with poor prognosis. *Blood* **99**, 4326–4335 (2002).
190. Whitman, S. P. *et al.* Absence of the wild-type allele predicts poor prognosis in adult de novo acute myeloid leukemia with normal cytogenetics and the internal tandem duplication of FLT3: a cancer and leukemia group B study. *Cancer Res.* **61**, 7233–9 (2001).
191. Man, C. H. *et al.* Sorafenib treatment of FLT3-ITD + acute myeloid leukemia: Favorable initial outcome and mechanisms of subsequent nonresponsiveness associated with the

- emergence of a D835 mutation. *Blood* **119**, 5133–5143 (2012).
192. Kuchenbauer, F. *et al.* Detailed analysis of FLT3 expression levels in acute myeloid leukemia. *Haematologica* **90**, 1617–25 (2005).
193. Stone, R. M. *et al.* The Multi-Kinase Inhibitor Midostaurin (M) Prolongs Survival Compared with Placebo (P) in Combination with Daunorubicin (D)/Cytarabine (C) Induction (ind), High-Dose C Consolidation (consol), and As Maintenance (maint) Therapy in Newly Diagnosed Acute Mye. *Am. Soc. Hematol. Annu. Meet.* **126(23)**, A6 (2015).
194. Quentmeier, H., Reinhardt, J., Zaborski, M. & Drexler, H. G. FLT3 mutations in acute myeloid leukemia cell lines. *Leuk. Off. J. Leuk. Soc. Am. Leuk. Res. Fund, U.K* **17**, 120–124 (2003).
195. Cui, L. *et al.* Low expressions of ARS2 and CASP8AP2 predict relapse and poor prognosis in pediatric acute lymphoblastic leukemia patients treated on China CCLG-ALL 2008 protocol. *Leuk. Res.* **39**, 115–23 (2015).
196. Juárez-Velázquez, R. *et al.* Significance of CASP8AP2 and H2AFZ expression in survival and risk of relapse in children with acute lymphoblastic leukemia. *Leuk. Lymphoma* **55**, 2305–11 (2014).
197. Park, T. S. *et al.* CASP8AP2 is a novel partner gene of MLL rearrangement with t(6;11)(q15;q23) in acute myeloid leukemia. *Cancer Genet. Cytogenet.* **195**, 94–5 (2009).
198. DiNardo, C. D. *et al.* Serum 2-hydroxyglutarate levels predict isocitrate dehydrogenase mutations and clinical outcome in acute myeloid leukemia. *Blood* **121**, 4917–24 (2013).
199. Gross, S. *et al.* Cancer-associated metabolite 2-hydroxyglutarate accumulates in acute myelogenous leukemia with isocitrate dehydrogenase 1 and 2 mutations. *J. Exp. Med.* **207**, 339–44 (2010).

200. Xu, W. *et al.* Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α -ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17–30 (2011).
201. Chowdhury, R. *et al.* The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases. *EMBO Rep.* **12**, 463–9 (2011).
202. Chaturvedi, A. *et al.* Mutant IDH1 promotes leukemogenesis in vivo and can be specifically targeted in human AML. *Blood* **122**, 2877–87 (2013).
203. Kapoor, I. *et al.* Proteomic discovery of MNT as a novel interacting partner of E3 ubiquitin ligase E6AP and a key mediator of myeloid differentiation. *Oncotarget* **7**, 7640–56 (2016).
204. Popov, N., Wahlström, T., Hurlin, P. J. & Henriksson, M. Mnt transcriptional repressor is functionally regulated during cell cycle progression. *Oncogene* **24**, 8326–37 (2005).
205. Sanchez, R., Meslamani, J. & Zhou, M.-M. The bromodomain: from epigenome reader to druggable target. *Biochim. Biophys. Acta* **1839**, 676–85 (2014).
206. Asangani, I. A. *et al.* Therapeutic targeting of BET bromodomain proteins in castration-resistant prostate cancer. *Nature* **510**, 278–282 (2014).
207. Stahl, M., Gore, S. D., Vey, N. & Prebet, T. Lost in translation? Ten years of development of histone deacetylase inhibitors in acute myeloid leukemia and myelodysplastic syndromes. *Expert Opin. Investig. Drugs* **25**, 307–17 (2016).
208. Garcia-Manero, G. *et al.* Phase II trial of vorinostat with idarubicin and cytarabine for patients with newly diagnosed acute myelogenous leukemia or myelodysplastic syndrome. *J. Clin. Oncol.* **30**, 2204–10 (2012).
209. Ocio, E. M. *et al.* Panobinostat as part of induction and maintenance for elderly patients with newly diagnosed acute myeloid leukemia: phase Ib/II panobidara study. *Haematologica* **100**, 1294–300 (2015).

210. Odenike, O. M. *et al.* Histone deacetylase inhibitor romidepsin has differential activity in core binding factor acute myeloid leukemia. *Clin. Cancer Res.* **14**, 7095–101 (2008).
211. Kazi, J. U., Kabir, N. N. & Rönnstrand, L. Brain-Expressed X-linked (BEX) proteins in human cancers. *Biochim. Biophys. Acta* **1856**, 226–33 (2015).
212. Liu, S. *et al.* Sp1/NFkappaB/HDAC/miR-29b regulatory network in KIT-driven myeloid leukemia. *Cancer Cell* **17**, 333–47 (2010).
213. Volk, A. *et al.* Co-inhibition of NF- κ B and JNK is synergistic in TNF-expressing human AML. *J. Exp. Med.* **211**, 1093–108 (2014).
214. Röhrs, S. *et al.* Hypomethylation and expression of BEX2, IGSF4 and TIMP3 indicative of MLL translocations in acute myeloid leukemia. *Mol. Cancer* **8**, 86 (2009).
215. Wilson, B. G. *et al.* Residual complexes containing SMARCA2 (BRM) underlie the oncogenic drive of SMARCA4 (BRG1) mutation. *Mol. Cell. Biol.* **34**, 1136–44 (2014).
216. Witkowski, L. *et al.* Germline and somatic SMARCA4 mutations characterize small cell carcinoma of the ovary, hypercalcemic type. *Nat. Genet.* **46**, 438–43 (2014).
217. Dykhuizen, E. C. *et al.* BAF complexes facilitate decatenation of DNA by topoisomerase II α . *Nature* **497**, 624–7 (2013).
218. Buscarlet, M. *et al.* Essential role of BRG, the ATPase subunit of BAF chromatin remodeling complexes, in leukemia maintenance. *Blood* **123**, 1720–8 (2014).
219. Eguchi, T., Prince, T., Wegiel, B. & Calderwood, S. K. Role and Regulation of Myeloid Zinc Finger Protein 1 in Cancer. *J. Cell. Biochem.* **116**, 2146–54 (2015).
220. Hromas, R. *et al.* Forced over-expression of the myeloid zinc finger gene MZF-1 inhibits apoptosis and promotes oncogenesis in interleukin-3-dependent FDCP.1 cells. *Leukemia* **10**, 1049–50 (1996).

221. Gaboli, M. *et al.* Mzf1 controls cell proliferation and tumorigenesis. *Genes Dev.* **15**, 1625–30 (2001).
222. Noll, L. A., Peterson, F. C., Hayes, P. L., Volkman, B. F. & Sander, T. Heterodimer formation of the myeloid zinc finger 1 SCAN domain and association with promyelocytic leukemia nuclear bodies. *Leuk. Res.* **32**, 1582–1592 (2008).
223. Delleire, G. *et al.* Promyelocytic leukemia nuclear bodies behave as DNA damage sensors whose response to DNA double-strand breaks is regulated by NBS1 and the kinases ATM, Chk2, and ATR. *J. Cell Biol.* **175**, 55–66 (2006).
224. Kurz, E. U. & Lees-Miller, S. P. DNA damage-induced activation of ATM and ATM-dependent signaling pathways. *DNA Repair* **3**, 889–900 (2004).
225. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
226. Döhner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: Recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453–474 (2010).
227. Hande, K. R., Stein, R. S., McDonough, D. A., Greco, F. A. & Wolff, S. N. Effects of high-dose cytarabine. *Clin. Pharmacol. Ther.* **31**, 669–674 (1982).
228. Moarii, M., Boeva, V., Vert, J.-P. & Reyat, F. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* **16**, 873 (2015).
229. Tibshirani, R. Regression Selection and Shrinkage via the Lasso. *J. R. Stat. Soc. B* **58**, 267–288 (1994).
230. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat. Med.* **16**, 385–

- 395 (1997).
231. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
 232. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **33**, 1–22 (2010).
 233. Schemper, M., Kaider, A., Wakounig, S. & Heinze, G. Estimating the correlation of bivariate failure times under censoring. *Stat. Med.* **32**, 4781–4790 (2013).
 234. Montine, T. J., Sonnen, J. A., Montine, K. S., Crane, P. K. & Larson, E. B. Adult Changes in Thought study: dementia is an individually varying convergent syndrome with prevalent clinically silent diseases that may be modified by some commonly used therapeutics. *Curr. Alzheimer Res.* **9**, 718–23 (2012).
 235. Postupna, N. *et al.* Novel antibody capture assay for paraffin-embedded tissue detects wide-ranging amyloid beta and paired helical filament-tau accumulation in cognitively normal older adults. *Brain Pathol.* **22**, 472–484 (2012).
 236. AMP-AD Knowledge Portal. at
<<https://www.synapse.org/#!/Synapse:syn2580853/wiki/409840>>
 237. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
 238. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–10 (1990).
 239. Sutphin, G. L. & Kaeberlein, M. Measuring *Caenorhabditis elegans* life span on solid media. *J. Vis. Exp.* 1–7 (2009). doi:10.3791/1152

- 240. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
- 241. Han, S. K. *et al.* OASIS 2: online application for survival analysis 2 with features for the analysis of maximal lifespan and healthspan in aging research. *Oncotarget* (2016). doi:10.18632/oncotarget.11269
- 242. Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother. Rep.* **50**, 163–70 (1966).
- 243. Cox, D. R. Regression models and life tables. *J. R. Stat. Soc. Ser. B* **34**, 187–220 (1972).

APPENDIX A (APPENDIX TO INSPIRE)

1. Expression data preprocessing

We downloaded the gene level processed expression data (level 3) for TCGA ovarian cancer from the Firehose pipeline as of the March, 2014 analysis freeze¹⁶² for all three platforms available for ovarian cancer (Affymetrix U133A, Agilent g4502, Human Exon array). We first removed potential plate level batch effects with ComBat¹⁶³ for all expression datasets. As was done in the TCGA ovarian cancer study²³, we combined the three separate expression measurements for each of 11864 genes to produce a single estimate of gene expression level by performing a factor analysis across the three studies. All data are log transformed. For other datasets, we downloaded the raw cell intensity files (CEL) for Affymetrix U133 Plus 2.0 and U133A arrays (Affymetrix, Santa Clara, CA, USA) from the Gene Expression Omnibus⁵⁷ for accessions: GSE14764²⁵, GSE26712¹⁶⁴, GSE6008¹⁶⁵, GSE18520¹⁶⁶, GSE19829⁵³, GSE20565¹⁶⁷, GSE30161¹⁶⁸, GSE9899²⁴. Expression data were then processed using MAS5.0 normalization with the ‘Affy’ Bioconductor package¹⁶⁹ and mapped to Entrez gene annotations¹⁷⁰ using custom chip definition files (CDF)¹⁷¹ which was followed by natural log transformation of MAS5.0 normalized intensities. The expression data was then Z-transformed so that each gene has zero mean and unit variance across the samples within each dataset. As stated in Tibshirani (1996),¹²⁵ Z-transformation of expression data is a standard practice for any method that uses a sparsity tuning parameter so that the sparsity tuning parameter is invariant to the scale of the variables, particularly before applying a penalized regression technique such as lasso (L_1 penalty) or ridge (L_2 penalty).^{125,172–175} Since the graphical model likelihood is indeed equivalent to multiple coupled regression likelihoods, this is

generalized to the network estimation problem where we optimize a graphical model likelihood.^{3,22,59,176–180}

2. Copy number variation (CNV) data processing

We downloaded the CNV data from 488 ovarian cancer patients in the TCGA cohort from the cBio Cancer Genomics Portal web page¹⁸¹. We used R package *cgdsr* to download the data. The 16,597 CNV levels in the downloaded data were derived from the copy-number analysis algorithm GISTIC¹⁸², and indicate the copy-number level per gene. CNV level ‘-2’ is a deep loss, possibly a homozygous deletion, ‘-1’ is a shallow loss (possibly heterozygous deletion), ‘0’ is diploid, ‘1’ indicates a low-level gain, and ‘2’ is a high-level amplification.

3. Data imputation

To our knowledge, there are no published methods for learning modules and their dependencies from multiple datasets that contain different sets of genes (**Figure 6**). Thus, we adapted the state-of-the-art methods (which can run on a single dataset) by imputing the missing values on genes that are not presented in each of the datasets, and applied these methods to the imputed data. These are the ‘Imp--’ methods in Table 1. We employed the iterative PCA algorithm to generate the imputed data for all ‘Imp--’ methods and initializing INSPIRE. The results were robust to the imputation method; INSPIRE method consistently outperformed alternative approaches when other imputation methods were used. We used CRAN R package *missMDA*¹⁸³ to generate the imputed data.

4. Initialization of the INSPIRE latent variables

INSPIRE is an iterative learning algorithm that consists of three update steps, Equations (10)-(12), to learn the following sets of parameters: $\mathbf{L}$ – values on the latent variables, $\mathbf{Z}$ – gene-module assignments, and θ_L – the dependency network among the latent variables. So we need to have some starting point, i.e., initial values on any of these three sets of parameters. SLFA and MGL are also iterative learning algorithms that require a starting point. Therefore, for INSPIRE, SLFA and MGL, we used the same initial gene-module assignments obtained by running the k -means clustering algorithm on the imputed data (see above) because the imputed data contain all genes and all samples.

We initialized MGL with k -means centroids, and we followed that approach for the MGL variants (MGL1, ImpMGL, and InterMGL) in our experiments. Given that INSPIRE is an extension to MGL for multi-data setting, to directly test whether the INSPIRE outperforms MGL, we used the output of MGL as a starting point for INSPIRE. The authors of the SLFA algorithm did not specify any initialization method; so for a fair comparison among all these methods, we used the same initial gene-module assignments for SLFA and MGL - the centroids obtained by running the k -means clustering algorithm on the imputed data. The result of the k -means clustering algorithm also depends on the initial clusters which are randomly determined. So, to rule out the possibility to make a conclusion based on a particular set of initial parameters, for every experiment on comparison across methods, we performed 10 runs with different initial parameters (i.e., different random initial clusters in the k -means clustering algorithm) and presented the average results.

5. Runtime of INSPIRE on gene expression datasets

Running INSPIRE with the module count parameter $k = 90$ and the sparsity tuning parameter $\lambda = 0.1$ in our application on nine datasets (**Appendix Table 1**) with a total number of $p \cong 20,000$ genes and $n \cong 1,500$ samples took 13.7 minutes on a machine with an Intel(R) Xeon(R) E5645 2.40GHz CPU and 24GB RAM, once the latent variables are initialized. As mentioned above, for initialization of the latent variables, we used the module graphical lasso (MGL)³ method on the imputed data, which took 10.2 minutes on the same machine.

6. Synthetic data generation

We synthetically generated data based on the joint distribution in Equation (8). We first generated the sparse $k \times k$ inverse covariance matrix Σ_L^{-1} by creating a $k \times k$ matrix G as

$$\forall i, G_{ii} = 0,$$

$$G_{ij} \ (i > j) \sim \begin{cases} 0 & \text{w. prb. } (1 - d) \\ \text{Uniform distribution } (0, 0.5) & \text{w. prb. } \frac{d}{2} \\ \text{Uniform distribution } (0.5, 1) & \text{w. prb. } \frac{d}{2} \end{cases},$$

and letting $\Sigma_L^{-1} = G + G^T$ so that Σ_L^{-1} is symmetric. We set $\forall i, G_{ii} = \epsilon$ afterwards by selecting ϵ such that the resulting matrix Σ_L^{-1} is positive definite. $d \in [0,1]$ controls the density of Σ_L^{-1} and the results we reported from synthetic data experiments were generated using $k = 10$ and $d = 0.2$. The results were consistent for varying values of k and d .

Then, we generated the latent variables $L = \{L_1, \dots, L_k\}$ from $L \sim N(0, \Sigma_L)$ and we randomly generated a binary $p_T \times k$ matrix Z of module assignments which randomly assigns each of p_T genes to exactly one of the latent variables. Then we generated a high-dimensional data matrix X of p_T genes from the distribution $X | ZL, \sigma^2 \sim N(ZL, \sigma^2)$, and selected a portion of the samples

and genes in X to form a smaller dataset that we call ‘Dataset1’. Then we selected the remaining samples and a portion of the genes from X to form a second ‘Dataset2’.

We considered three simulated settings that correspond to different ratio of overlapping genes (**Appendix Figure 1a**). Each setting is characterized by $[OL, D1, D2]$ where OL denotes the number of genes that are present in both Dataset1 and Dataset2, $D1$ is the number of genes that are present only in Dataset1, and $D2$ means the number of genes that are present only in Dataset2. The settings we consider are $[150, 100, 0]$, $[200, 50, 0]$ and $[250, 0, 0]$, where the sample sizes of Dataset1 and Dataset2 are 20 and 30, respectively (**Appendix Figure 1a**). $[250, 0, 0]$ means that all genes are shared between the two datasets. We repeated the generation of data X 20 times in each of the three settings and presented the mean of the results for each method in (**Figure 8-Figure 10**). We show the p -values on the bars that represent the statistical significance of the difference between each method and INSPIRE across 20 different data instantiations.

Appendix Figure 1a illustrates the two datasets in each of these three settings. In each rectangle, each row represents a variable, and each column represents a sample. For simplicity in presentation of the evaluation results, we set $D2 = 0$. The results were consistent for varying $D2$. We note that $D2 \cong 0$ assumption holds in many real-world settings we are interested in, where the newer technology contains almost all of the genes in the older technology. We demonstrate this real-world situation in the second set of experiments on the ovarian cancer expression data (**Figure 12**).

7. Comparison of the scalability across all six methods in simulation experiment

We precisely measured the runtimes of six methods – GLasso, UGL, SLFA, WGCNA, MGL (Table 1) and INSPIRE – when running on the synthetic data with varying numbers of genes (p); $p = 300, p = 1,500, p = 3,000$. We generated the data exactly the same way as in the simulation experiments. We used 50 as sample size (20 samples in Dataset1 and 30 samples in Dataset2). We tested these methods on the ‘Imp---’ setting where we imputed the missing data before applying the algorithms, because 5 of these methods (except INSPIRE) cannot accommodate multiple datasets consisting of variables with discrepancies. We used varying sparsity tuning parameters in the interval of $(.5, .0001)$, exactly the same set of values that we used for choosing λ (via cross-validation tests) in our experiments. The runtimes of these methods are known to grow cubically or at least quadratically depending on the availability of a special efficient technique for the method¹⁸⁴ with increasing p (when gene-level dependencies are learned – GLasso and UGL), or k (when module level dependencies are learned – SLFA and MGL). Also, WGCNA grows quadratically with increasing p since it includes correlation computation and hierarchical clustering. Therefore, we determined that the methods whose runtime is >10 hours for $p = 3,000$ are not scalable enough to be useful on genome-wide analysis. Since the runtimes of the methods except MGL, WGCNA and INSPIRE already exceeded 10 hours at $p = 3,000$ (**Appendix Figure 2a**), it is clear that all methods other than MGL, WGCNA and INSPIRE are too slow to be used when $p > 3,000$ and >500 hours when p is near 20,000 (see the trend line in **Appendix Figure 2b**). We note that we increased the module count (k) with increasing p such that the average number of genes in a module is always 30, and SLFA was unable to run for $p > 1,500$ where the module count (k) exceeded the sample size (50). **Appendix Figure 2a-b** indicate that GLasso, UGL and SLFA are not practically useful to be used on genome-wide expression datasets, and furthermore, they do not perform well on smaller synthetic data on which we ran all six methods.

Thus, we excluded GLasso, UGL and SLFA for the evaluation on the genome-wide expression datasets. The runtime measurements were done on a very powerful machine with an Intel(R) Xeon(R) E7-8850 v2 @ 2.30GHz CPU and 528 GB RAM.

8. Computing the cross-validation test log-likelihood

We performed a 5-fold CV to choose λ for INSPIRE and each of the competing methods in our experiments to evaluate INSPIRE (synthetic data experiments and the experiments with two gene expression datasets). We measured the CV test log-likelihood on the test data portion of the first dataset (Dataset1 or OV1 which contains all or almost all genes) in each fold, which was common test data across all methods. For each of the 5 test folds, we computed the test data log-likelihood of the $p \times p$ gene-level dependency matrix that is computed using the dependencies among the latent variables (representing modules) inferred by each of the INSPIRE and its competitors, where p is the total number of genes in the two datasets. For the methods that optimize a non-convex objective function, we averaged the CV test log-likelihoods across multiple runs with different initial assignment of genes to modules. We tested a range of sparsity tuning parameter values (λ) and observed the “cup-shaped” underfitting/overfitting pattern in the λ (x-axis) vs. average CV test log-likelihood (y-axis) curves for all methods, as expected.

9. Evaluation of learned network in synthetic data experiments

In the synthetic data experiments, the correspondence between the modules in a learned model and the modules in the true model is not clear because each method can end up having different optimal number of modules, even if they started with the same number of initial modules. Therefore, we compared the methods in terms of the accuracy of the $p \times p$ gene-level dependency matrix that

is computed using the dependencies among the modules inferred by each of the INSPIRE and its competitors, where p is the total number of genes in the two datasets.

10. Measuring the significance of difference between INSPIRE and 13 competing methods

We repeated the synthetic data generation 20 times in each of the three settings, and presented the average results with the Wilcoxon signed rank test p -value measuring the significance of differences based on the Wilcoxon signed rank test. More specifically, it measures the probability that the corresponding method gave a better result in terms of mean rank than INSPIRE across 20 different data instantiations.

11. Comparison of the prediction performance with alternative methods

We compared INSPIRE with principal component analysis (PCA)⁴⁶ and subnetwork analysis method⁴¹ based on how well each method can predict each of the six phenotypes (resectability as defined by 0 cm of residual tumor vs. >0 cm of residual tumor after surgery, survival time, and four manually curated histologic phenotypes) from The Cancer Genome Atlas (TCGA) data. We used the *lasso*¹²⁵ (L₁ regularized linear regression) for predicting the continuous-valued phenotype (percent stroma), L₁ regularized logistic regression for predicting binary phenotypes (stroma type, vessel formation, invasion pattern, and residual tumor), and L₁ regularized Cox regression for predicting survival. The prediction performance was measured in left-out data via leave-one-out cross validation (LOOCV) tests for histologic phenotypes that have relatively less number of samples (~100), and 50-fold cross-validation for resectability and survival that have larger number of samples (~500). To evaluate a survival prediction model in a CV setting, we used two different

methods to summarize the prediction results across CV tests. This is because unlike other phenotypes, the prediction performance on survival time is measured by a ranking-based metric – the concordance index (CI) that measures the proportion of pairs of samples whose observed survival are concordant with the predicted survival in terms of which of the two samples experienced an event (death) before the other (or survived shorter)¹⁸⁵. First, we predicted the survival (i.e., hazard scores) of all ~500 samples (specifically, 550 samples) when each sample was treated as a test sample in one of the 50 folds. Then we computed one CI value based on predicted survival across all 550 samples, which leads to **Figure 16b** (middle). Second, we considered computing CIs within test samples in each fold, which would allow us to have multiple CIs (# of folds $\times$ # of CV rounds) and compute the confidence interval of the CIs for INSPIRE compared to the CIs for the alternative methods (**Figure 16b** right). Especially, in this analysis, we performed 50 rounds of 10-fold CV tests, and reported the average of a total of 500 CIs (i.e., y-axis of **Figure 16b** right) together with the associated Wilcoxon signed rank test p -value measuring the significance of the difference between INSPIRE and each of the alternative methods (PCA-based method⁴⁶, subnetwork analysis⁴¹ and individual genes). All p -values are smaller than .01 which means that the population mean of CIs from INSPIRE is statistically significantly higher than the population mean of CIs from each of the alternative methods. For each of the alternative methods, we also report the 95% confidence interval for the mean of a normal distribution fitted to the difference of the method's CIs from INSPIRE's CIs. Given that all of the three intervals cover the positive-valued ranges, we can say that INSPIRE predicts survival better than the alternative methods with 95% confidence.

The sparsity tuning parameter λ was chosen within training data by performing LOOCV tests, which is a standard way of choosing λ ¹²⁵. For a fair comparison with PCA⁴⁶ and the subnetwork method⁴¹, we used top 90 principal components, and 90 subnetworks that are most correlated with the phenotype, respectively. The subnetwork analysis method runs on binary phenotypes, but ‘percent stroma’ is continuous-valued; so, to make the subnetwork method work on this phenotype, we binarized the values by making >50% to be 1 and <50% be 0.

12. Learning subtypes based on the INSPIRE latent variables

We used the k -means clustering algorithm on the INSPIRE latent variables, each of which corresponds to a module, to cluster patients into four subtypes. We chose four as the number of subtypes to make it comparable to alternative subtyping methods (TCGA study²³ and the NBS method⁵⁸). Since k -means is non-deterministic, the resulting subtypes could depend on the starting point of the subtype assignments. In order to get the most coherent groups of patients, we ran k -means 10 times with different random initial assignments of the patients into subtypes, and chose the clustering which gives the lowest within cluster sum of squares.

13. Supervised model to predict tumor resectability

We trained supervised models of tumor resectability using different combinations of the POSTN expression and the latent variables corresponding to module 5 and module 6 in TCGA ovarian cancer data for 489 patients to predict 0 cm of residual tumor vs. >0 cm of residual tumor. The proportion of the sub-optimally debulked patients was 62% (= 139/223) in Tothill²⁴ and was 77% (= 378/489) in TCGA²³. Logistic regression was used to train the models. Five distinct models were constructed: 1) a model with only the POSTN expression, 2) a model with only the latent

variable corresponding to module 5, 3) a model with only the latent variable corresponding to module 6, 4) a model with POSTN expression and the latent variable corresponding to module 5, and 5) a model with the latent variables corresponding to module 5 and module 6. We trained each of those models along with (**Figure 19**) and without (**Appendix Figure 3**) the clinical covariates of age and stage. Performance was determined based on the results of each fitted model in the Tothill²⁴ data in terms of the Area Under the Curve (AUC) measure from a Receiver Operator Characteristic (ROC) curve (**Figure 19** and **Appendix Figure 3**).

14. Extraction of tumor histologic phenotypes from TCGA images

We manually curated multiple tumor histopathology features from image data on H-E staining of ovarian tumor section from TCGA. We primarily used 98 randomly sampled patients to test the association between tumor histopathology features and the latent variables learned by INSPIRE. Features were curated in a blinded fashion. Five histopathological features were evaluated including percent stroma, percent tumor, vessel formation, stroma type, and pattern of invasion. Percent tumor was defined as the percent area involved by viable neoplastic cells across the entire slide while percent stroma was the percent area of fibrous tissue (fibroblasts and collagen). Vessel formation was scored as minimal, moderate, or abundant based on the number of formed vessels identified at 100X magnification. Stroma type was defined as fibrous (dense collagen with relatively fewer fibroblasts) or desmoplastic (many fibroblasts embedded in a loose, myeloid extracellular matrix). Pattern of invasion related to how the neoplastic cells interacted with the surrounding stroma and was scored as expansile, infiltrative, papillary, or mixed. Expansile invasion was characterized by cohesive tumor cells growing in a cluster with relatively well-circumscribed borders with the surrounding stroma while infiltrative invasion included tumor cells

which grew in small nests or tentacles with abundant stroma surrounding the individual tumor cells. Tumors classified as having papillary invasion had abundant fibro-vascular cores upon which the neoplastic cells grew in arborizing branches. Mixed invasion patterns were identified and classified as such.

15. Immunohistochemistry

Ten patients were sampled for staining based on either having good tumor resection and survival (> 3 year survival, optimal debulking with residual tumor < 1cm) vs. poor tumor resection and survival (< 3 years survival, > 1 cm residual tumor). Tissue and clinical information were collected with patient consent by the University of Washington Gynecologic Oncology Tissue Bank under approval from the human subjects division (IRB 27077). Tumor tissue was collected at the time of primary surgery and flash frozen in liquid nitrogen, transported to the lab and stored at -80 C. The 17 frozen block was cryo-sectioned and one 8 mm section placed on a charged slide for IHC testing and H-E staining.

Frozen tissue slices fixed to glass slides were allowed to thaw at room temp for ten min. Slides were fixed in a Coplin jar in cold acetone for ten min at -20C. Slides were removed from acetone and placed tissue side up on a shaker. PBS was added to the slide (1mL, enough to cover tissue slice) for five min shaking. PBS wash was repeated for a total of two five min washes. After final wash, PBS was poured off the slide and tissues were blocked with 2% milk/PBS (Carnation Instant Nonfat Dry Milk dissolved in PBS) for one hour at room temperature, while shaking. Blocking solution was removed and primary antibody added, diluted in 1% milk. Antibody dilutions were per manufacturers recommendations. Slides were allowed to incubate overnight at 4°C while shaking with primary antibody. If primary antibody was conjugated to fluorescent molecule, slides

were also incubated in the dark overnight. Slides were washed three times with PBS at room temperature. Secondary antibody was diluted in 1% milk/PBS and incubated at room temperature for 30 min, shaking. Slides were then washed with PBS for 10 min, three times. Nuclear stain diluted in PBS was added to tissues. Either Dapi (300ng/mL, Sigma-Aldrich, catalog # D9542) or Sytox Green Nuclear Stain (Life Technologies, catalog # S7020) was used depending on the secondary antibodies used for staining. Last PBS wash was done at room temperature for 5min. Coverslips were mounted to slides using Fluoroshield (Sigma-Aldrich, catalog # F6182) and sealed with clear nail polish. Images were taken on a Nikon TiE Inverted Widefield Fluorescence High Resolution Microscope.

Primary Antibodies used: Anti-E Cadherin antibody conjugated to Allophycocyanin (Abcam, catalog # ab99885), Hop Antibody (Santa Cruz, catalog # sc-30216), Anti-CD73 antibody (Abcam, catalog # ab54217), GCS-a-1 Antibody (Santa Cruz, sc-23801)

Secondary Antibodies used: CD73 antibody was detected with Goat anti-mouse IgG-FITC (Santa Cruz, catalog # sc-2010). When co-stained with CD73, HOPX was detected with Donkey anti-rabbit IgG-CFL 647 (Santa Cruz, catalog # sc-362291). When co-stained with E Cadherin, HOPX antibody was detected with Chicken anti-rabbit IgG H&L FITC (Abcam, catalog # ab6825).

Fluorescence images were analyzed using ImageJ¹⁸⁶ and the plugin JACoP was used for co-localization analysis.

16. Genes whose CNV is associated with the subtypes

Stromal subtype:

Genes (association p-value threshold is 8×10^{-5} , gene count is 70):

ABCB4, ACN9, AKAP9, ANKIB1, ASNS, BHLHA15, BRD4, C7ORF62, C8ORF37, CALB1, CALCR, CCNE2, CNBD1, CNGB3, CPNE3, CPQ, CSMD3, CYP51A1, DCAF4L2, DECR1, DLX5, DLX6.AS1, EIF3H, EPHX3, FKBP6, GATAD1, GEM, HEPACAM2, KRIT1, LAPTM4B, LMTK2, MTDH, MTERF, MTERFD1, NBN, NDUFAF6, NECAB1, NPTX2, OSGIN2, OTUD6B, PCLO, PEX1, PLEKHF2, PTDSS1, RAD21, RBM48, RIPK2, RUNX1T1, SDC2, SEMA3A, SEMA3D, SHFM1, SLC25A40, SLC26A7, SRI, STEAP1, STEAP2, TAC1, TECPR1, TMEM55A, TMEM64, TP53INP1, TRHR, TRIM50, TRIQK, TRPS1, TSPYL5, UQCRB, UTP23, YWHAG

Functional enrichment of those genes (MSigDB C2 category, Bonferroni-corrected $-\log_{10}(p)$):

NIKOLSKY_BREAST_CANCER_8Q12_Q22_AMPLICON, 29.7339391253105

NIKOLSKY_BREAST_CANCER_7Q21_Q22_AMPLICON, 23.7934623738506

ONKEN_UVEAL_MELANOMA_UP, 2.59997036554253

Vascular subtype:

Genes (association p-value threshold is 10^{-12} , gene count is 99):

ADCK5, ADCY8, AGO2, ANXA13, ASAP1, ASAP1.IT1, BREA2, C8ORF31, CHRA1, COL22A1, COLEC10, CPSF1, CYC1, CYHR1, CYP11B2, DENND3, DGAT1, EEF1D, ENPP2, EPPK1, EXOSC4, FAM135B, FAM49B, FAM83H, FAM84B, FAM91A1, FBXL6, FOXH1, GLI4, GPAA1, GPIHBP1, GPR20, GPT, GRINA, GSDMC, GSDMD, HHLA1, HPYR1, HSF1, KCNK9, KCNQ3, KHDRBS3, KIAA0196, KIAA1875, KIFC2, LINC00964, LOC731779, LRRC6, LY6E, LY6H, MAF1, MAPK15, MFSD3, MROH1, MROH6, MTSS1, NAPRT1, NDUFB9, NOV, NRBP2, NSMCE2, OPLAH, PARP10, PHF20L1, PLEC, PPP1R16A, PTK2, PTP4A3, PUF60, PVT1, PYCRL, RECQL4, RHPN1, RHPN1.AS1, RNF139, SCRIB, SCRT1, SHARPIN, SLA, SLC39A4, SLC45A4, SLC52A2, SPATC1, SQLE, TIGD5, TMEM65, TMEM71, TNFRSF11B, TONSL, TOP1MT, TRAPPC9, TRIB1, TRMT12, TRPS1, TSTA3, VPS28, ZC3H3, ZFAT, ZFP41

Functional enrichment of those genes (MSigDB C2 category, Bonferroni-corrected $-\log_{10}(p)$):

NIKOLSKY_BREAST_CANCER_8Q23_Q24_AMPLICON, 139.483967848423

AGUIRRE_PANCREATIC_CANCER_COPY_NUMBER_UP, 13.9610410158971

ONKEN_UVEAL_MELANOMA_UP, 5.5595399649524

DAVICIONI_MOLECULAR_ARMS_VS_ERMS_DN, 3.16089982649642

DEURIG_T_CELL_PROLYMPHOCYTIC_LEUKEMIA_UP, 2.94719563791031

PATIL_LIVER_CANCER, 2.31938878043002

Immunoresponsive subtype:

Genes (association p -value threshold is 10^{-5} , gene count is 78):

ABLM1, ACTR1A, ADCY8, AFAP1L2, ASAP1, ASAP1.IT1, ATRNL1, BAG3, C10ORF118, C10ORF82, C10ORF95, C1ORF87, C8ORF31, CACUL1, CASC2, CCDC172, CUEDC2, CYP2J2, EFR3A, EIF3A, EMX2, EMX2OS, FAM135B, FAM160B1, FAM204A, FAM45B, FAM49B, FAM84B, FBXL15, GFRA1, GLI4, GPIHBP1, GRK5, GSDMC, HHLA1, HOOK1, HPYR1, HSPA12A, INA, INPP5F, KCNQ3, KIAA1598, LRRC6, LY6H, MCMBP, NANOS1, NFKB2, OMA1, PDZD8, PHF20L1, PNLIP, PNLIPRP1, PNLIPRP2, PNLIPRP3, PPAPDC1A, PRDX3, PRLHR, PSD, PTK2, PVT1, RAB11FIP2, RGS10, SEC23IP, SFXN2, SFXN4, SH3PXD2A, SLC18A2, SUFU, TACSTD2, TDRD1, TIAL1, TMEM180, TMEM71, TOP1MT, TRIB1, TRUB1, WDR11, ZFP41

Functional enrichment of those genes (MSigDB C2 category, Bonferroni-corrected $-\log_{10}(p)$):

NIKOLSKY_BREAST_CANCER_8Q23_Q24_AMPLICON, 7.80574660503181

Fibrous subtype:

Genes (association p -value threshold is 8×10^{-4} , gene count is 76):

ABCA17P, ABCA3, AMDHD2, ATP6V0C, C16ORF59, CCDC114, CCDC64B, CCNF, CEMP1, CLDN6, CLDN9, COX6B2, EMP3, FAM71E2, FLYWCH1, FLYWCH2, FPR1, FPR2, FPR3, HAS1, HCFC1R1, IL11, IL32, ISOC2, KCTD5, KDELRL1, MEFV, METTL22, MKLN1, MMP25, NAPSA,

NAT14, NFKBIA, NLRP13, NLRP4, NLRP5, NLRP8, NTN3, OR1F1, OR1F2P, PDPK1, PODXL, PRSS21, PRSS22, PRSS27, PRSS30P, PRSS33, PSMA6, PTPRH, RPL28, SIGLEC5, SIGLEC6, SPACA6P, SRRM2, SRRM2.AS1, SSC5D, SUV420H2, SYNGR4, SYT5, TCEB2, THOC6, TMEM143, TMEM190, TMEM86B, TNFRSF12A, TOP1, ZG16B, ZNF114, ZNF175, ZNF200, ZNF205, ZNF205.AS1, ZNF213, ZNF350, ZNF480, ZNF528

Functional enrichment of those genes (MSigDB C2 category, Bonferroni-corrected $-\log_{10}(p)$):

NIKOLSKY_BREAST_CANCER_16P13_AMPLICON, 20.8790876399383

17. Subtype characterizations based on INSPIRE

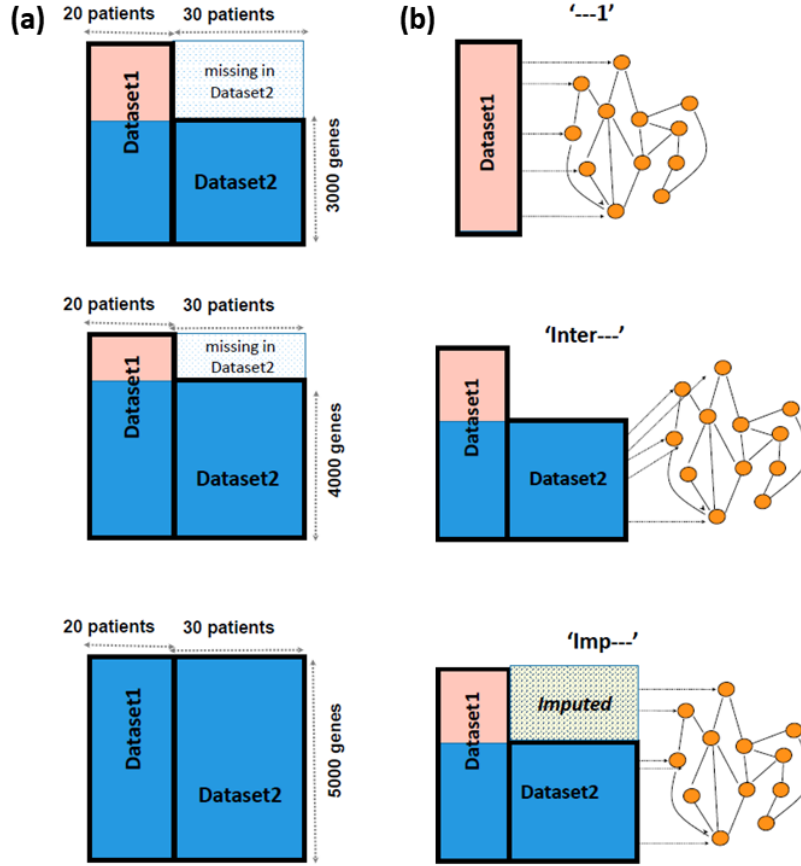
Specifications for the stromal subtype learned based on the INSPIRE latent variables was explained in the main text. Here we explain the other three INSPIRE subtypes – vascular, immunoresponsive and fibrous.

The vascular subtype is characterized by low expression of modules 70, 71, 78, 79, 81 (**Figure 18**), which indicates a strong association with the deactivation of immune system. Modules 70, 71, 78 and 79 are significantly enriched for the immune response, defense response and immune system process functions (**Appendix Table 10**). Module 81 is significantly enriched for caspase regulator activity, which modulates the activity of cysteine proteases involved in apoptosis (**Appendix Table 10**). Caspases are suggested to have many functions in the immune system besides acting as cell-death proteases¹⁸⁷. The vascular subtype is also characterized by moderate or abundant vessel formation (**Appendix Figure 11-iv**), which is potentially promoted by deactivation of immune system. Abundance of tumor vessels suggest fast tumor angiogenesis that facilitates the proliferation of tumor cells. The vascular subtype has a significant overlap (p -value: 6.81×10^{-82}) with the proliferative subtype revealed by the TCGA consortium²³ (**Appendix Table 11**).

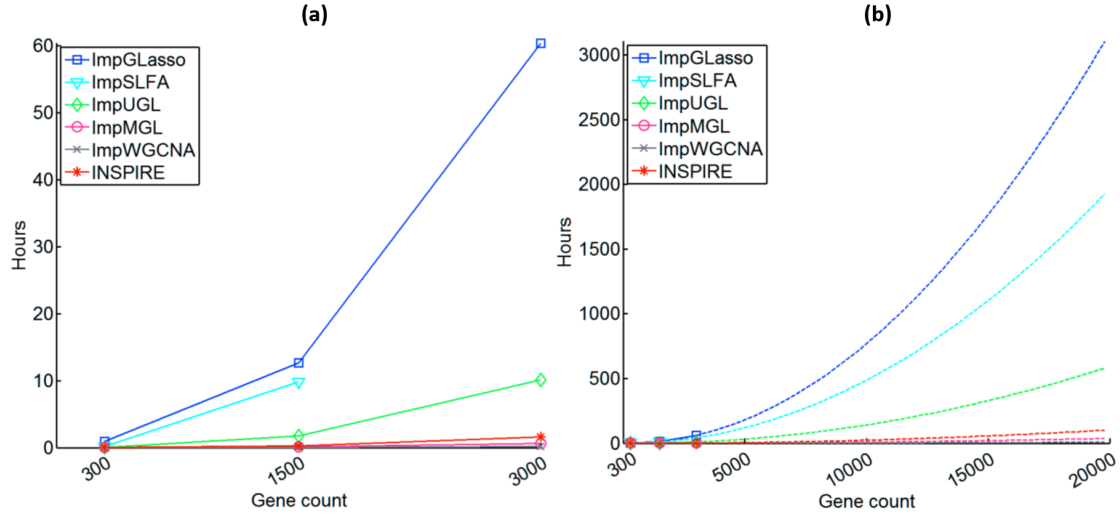
The immunoresponsive subtype is characterized by high expression of immune response, defense response and immune system process modules 70, 71 and 78 (**Appendix Table 10, Figure 18**) that are down-regulated in vascular subtype. The vessel formation amount is minimal (**Appendix Figure 11-iv**), which is potentially led by the high immune system response. The amount of stroma is low (**Appendix Figure 11-i**) and the stroma is mostly fibrous, suggesting a lower tumor-stroma interaction than the stromal subtype, while both immunoresponsive and stromal subtypes show high immune system response. Immunoresponsive subtype has a significant overlap (p -value: 6.54×10^{-13}) with the immunoreactive subtype discovered by the TCGA study²³ (**Appendix Table 11**).

The fibrous subtype is characterized by high expression of modules 87 and 89 (**Figure 18**), which are significantly enriched for G protein coupled receptor protein signaling pathway (**Appendix Table 10**), the series of molecular signals generated as a consequence of a G protein coupled receptor binding to its physiological ligand. The minimal desmoplasticity and minimal vessel formation in fibrous subtype (**Appendix Figure 11-ii,iv**) suggest antiproliferative effects, which potentially depends, in part, on signal transduction via a G protein coupled receptor¹⁸⁸. Fibrous subtype has a significant overlap (p -value: 2.17×10^{-17}) with the differentiated subtype revealed by the TCGA consortium²³ (**Appendix Table 11**).

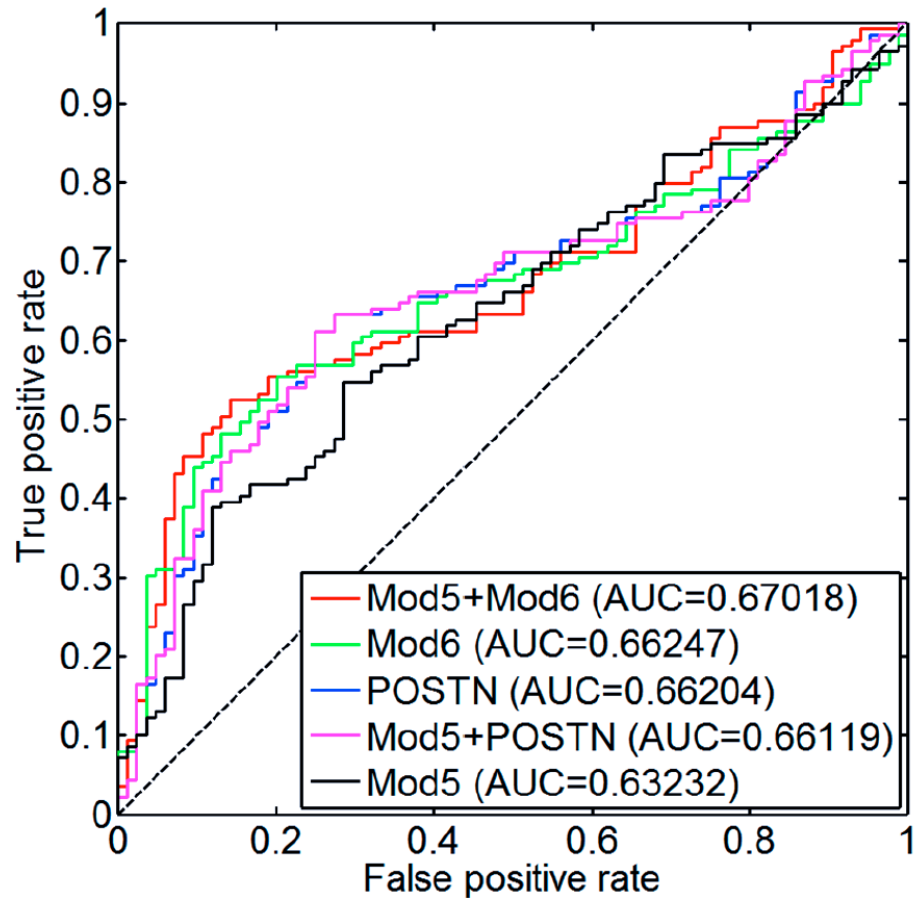
18. APPENDIX A Figures:



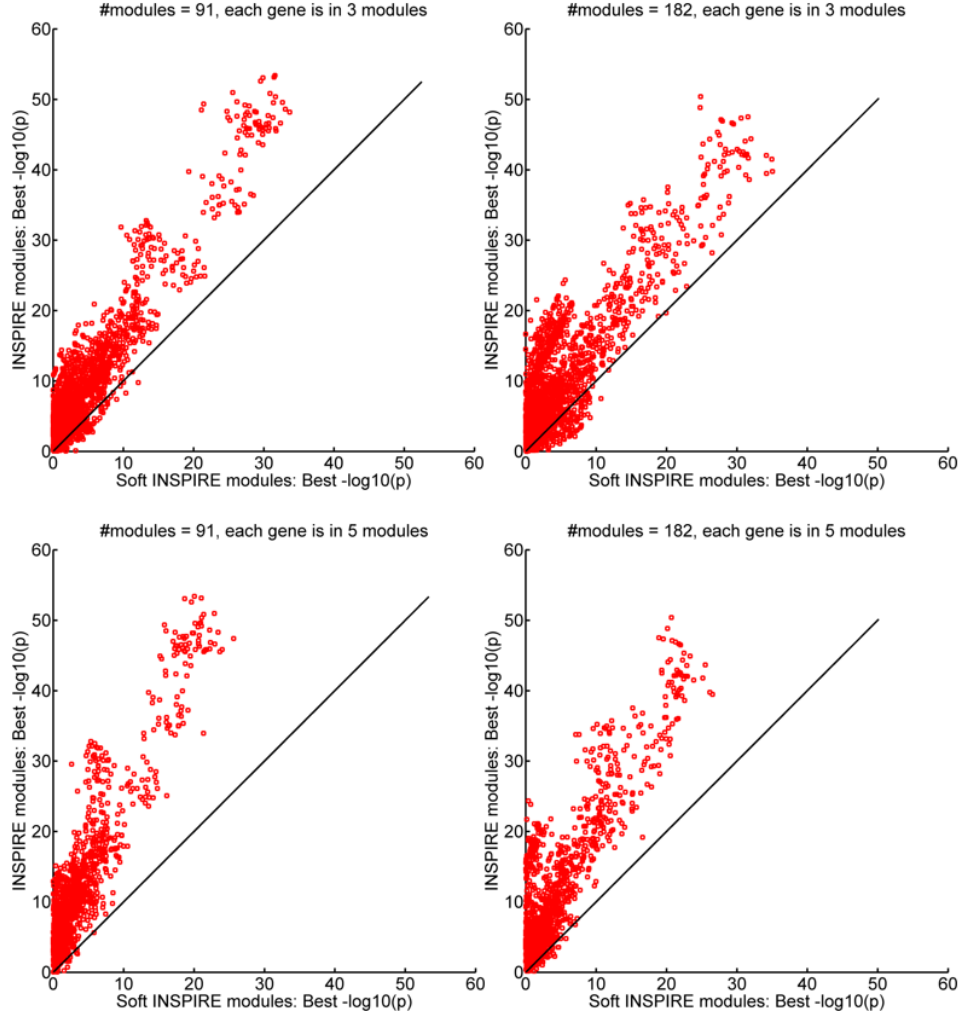
Appendix Figure 1. (A) The synthetic data for the three generated simulation settings are illustrated. Rows represent genes and columns represent samples. In each setting, different amount of genes overlap between datasets (60%, 80% and 100% respectively from top to bottom). **(B)** Adapted learning ways for the alternative methods as explained in Table 1 are illustrated ('---l', 'Inter---', and 'Imp---' respectively from top to bottom). The first illustration corresponds to '--l' which performs standard learning from a single dataset. The second illustration corresponds to 'Inter--' which learns the features using the overlapping genes, and map data-specific genes to the learned features. The third illustration corresponds to 'Imp--' which imputes the missing values and learns the features using the imputed data matrix.



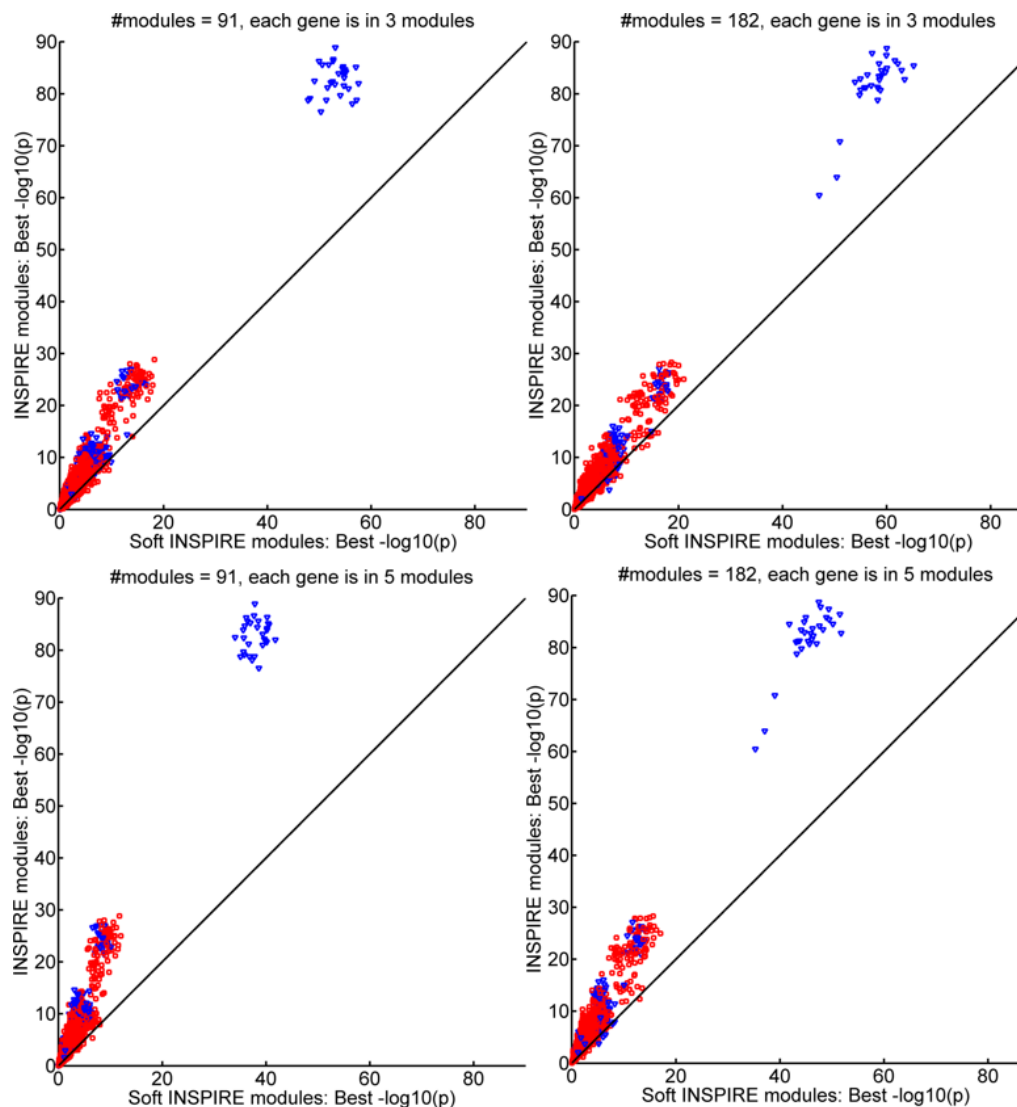
Appendix Figure 2. (A) Runtimes (in hours on the y-axis) of INSPIRE and five state-of-the-art methods that learn a network of modules or genes from a single dataset is compared for varying gene counts ($p = 300$, $p = 1500$ and $p = 3000$ as shown on the x-axis) where the imputed data are used for the methods that are unable make use of multiple datasets (all except INSPIRE). $p = 3000$ cannot be shown for ImpSLFA since it is unable to run for the cases where the module count (k) exceeds the sample size. **(B)** The trend lines from a quadratic fit are added to show the estimated runtimes (in hours on the y-axis) for bigger p and genome-wide data (on the x-axis).



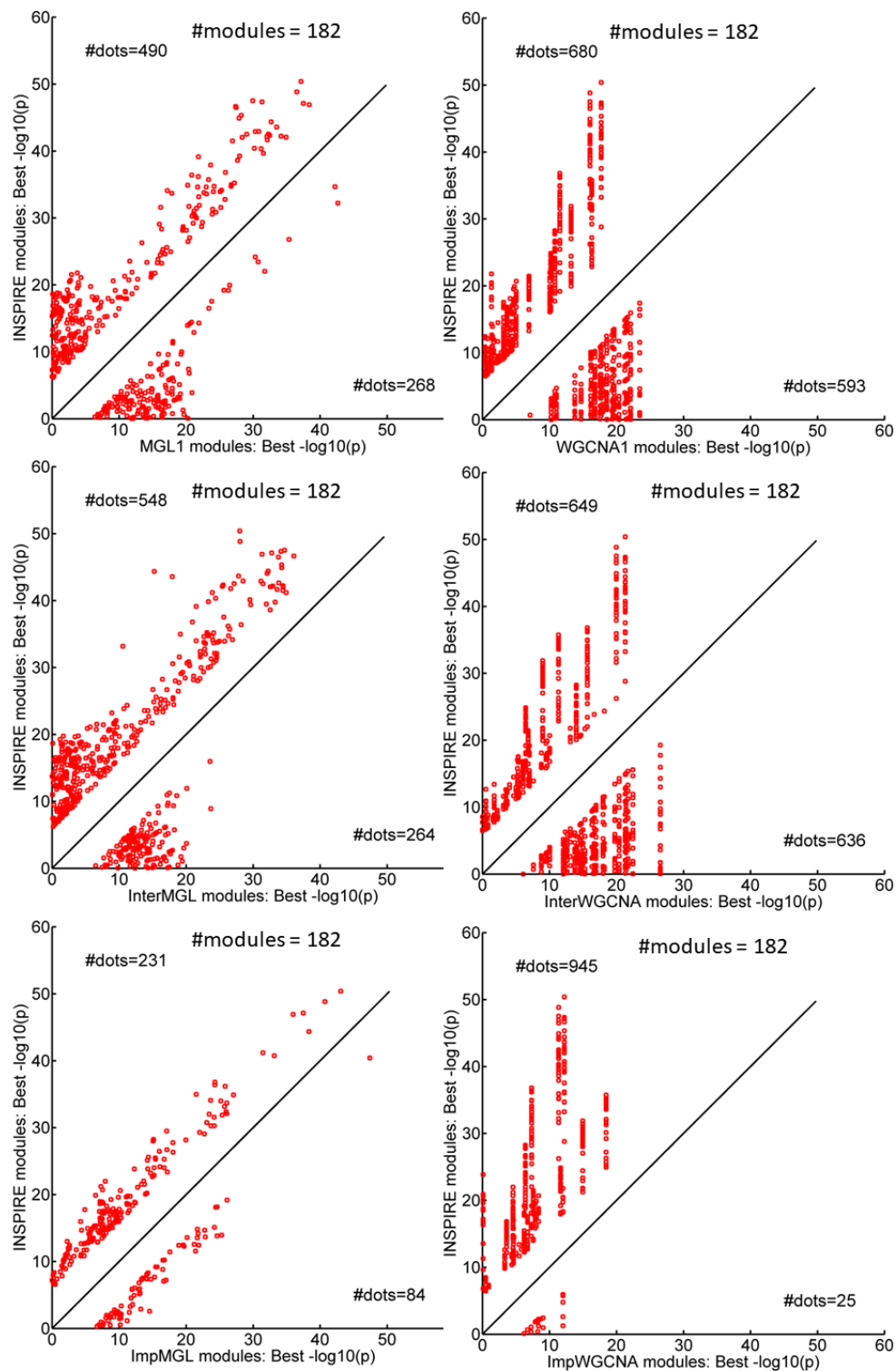
Appendix Figure 3. ROC curve of the supervised models for resectability prediction trained in TCGA and tested in Tothill data. Different combinations of POSTN and the INSPIRE features corresponding to modules 5 and 6 are used for training each model. The clinical covariates age and stage are not included in the models. AUC of each model is shown in the legend.



Appendix Figure 4. For $k = 91$ (left) and $k = 182$ (right), the best $-\log_{10} p$ from the functional enrichment of the modules learned by an INSPIRE extension that assigns each gene to more than one modules (on the x-axes) are compared to the best $-\log_{10} p$ from the functional enrichment of the modules learned by the proposed INSPIRE approach (on the y-axis). Each dot corresponds to a KEGG, Reactome or BioCarta GeneSet, and only the GeneSets with a Bonferroni corrected Fisher's exact test $p < .05$ in at least one of the compared two methods are shown on each plot. Two different versions of the INSPIRE extension was used in comparison; the one which assigns each gene to the top three modules with highest potential to contain that gene (Top), and the one which assigns each gene to the top five modules (Bottom). For both INSPIRE and alternative approach, the results from multiple runs are shown on each plot.

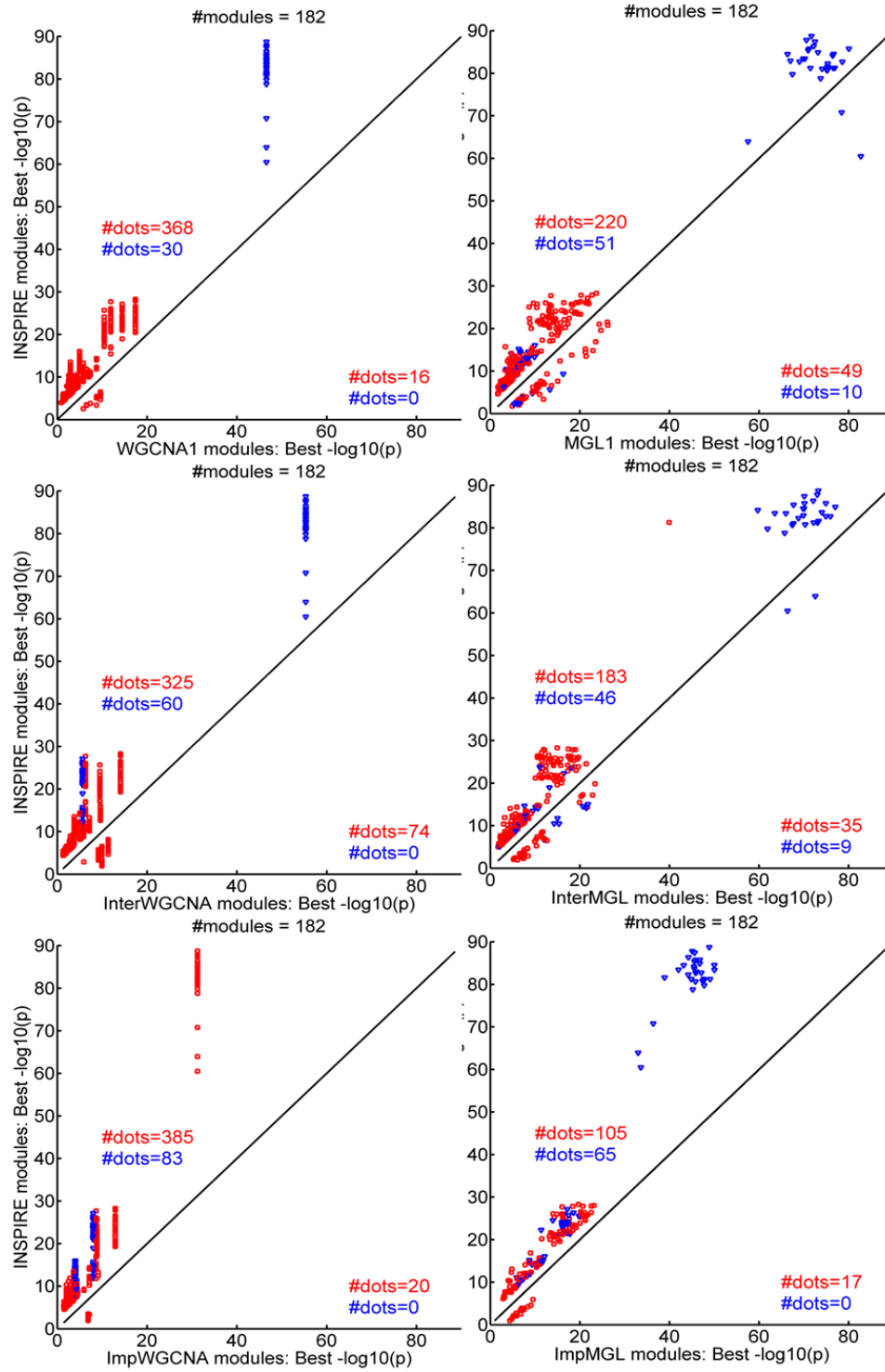


Appendix Figure 5. For $k = 91$ (left) and $k = 182$ (right), the best $-\log_{10} p$ from the ChEA enrichment of the modules learned by the INSPIRE extension (on the x-axes) are compared to the best $-\log_{10} p$ from the ChEA enrichment of the modules learned by INSPIRE (on the y-axis). Each dot is for a group of genes composed of a TF and its targets, and only the TFs with a Bonferroni corrected Fisher's exact test $p < .05$ in at least one of the compared two methods are shown on each plot. Two different versions of the INSPIRE extension was used in comparison; the one which assigns each gene to the top three modules with highest potential to contain that gene (Top), and the one which assigns each gene to the top five modules (Bottom). For both INSPIRE and alternative approach, the results from multiple runs are shown on each plot.



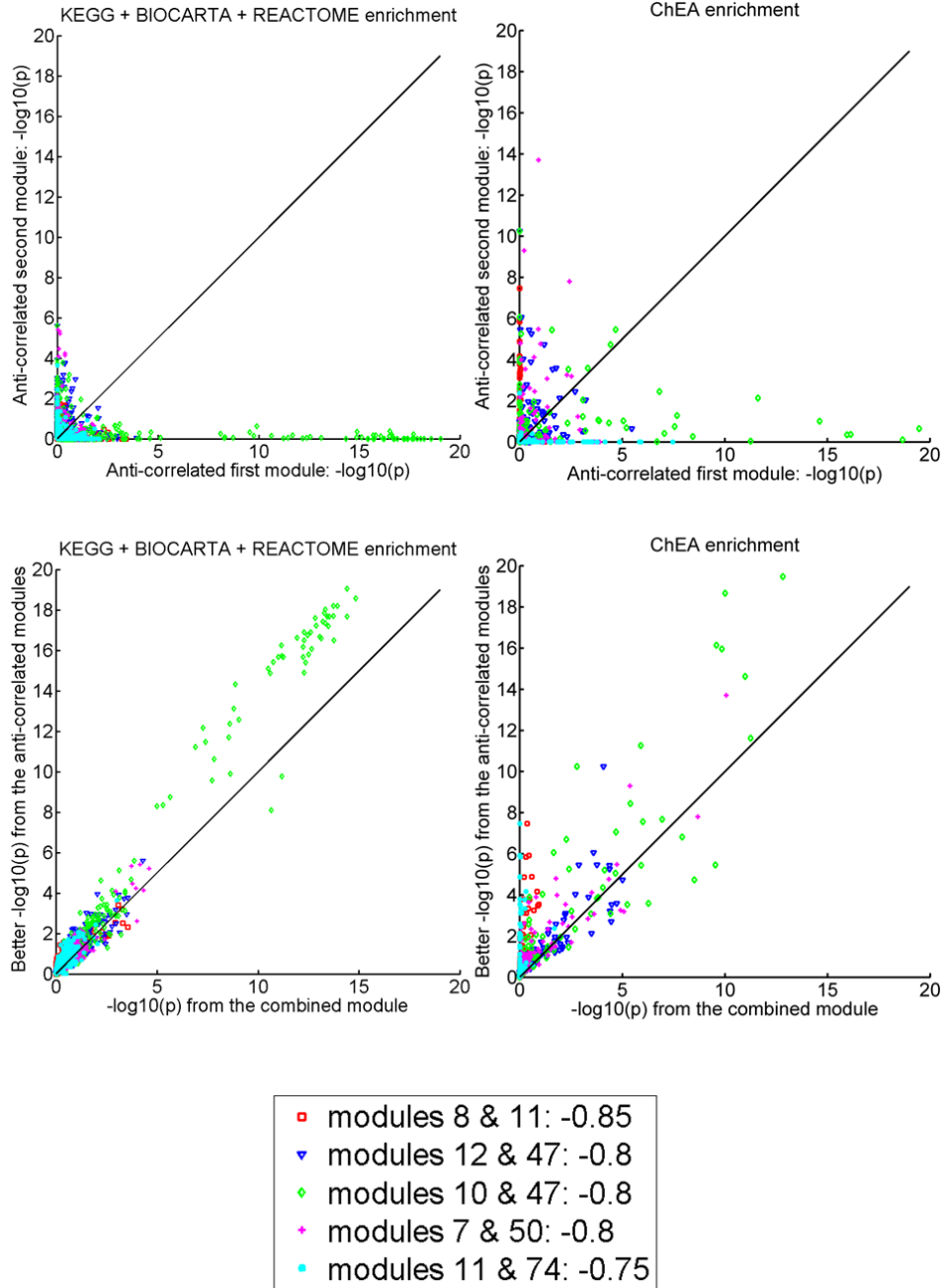
Appendix Figure 6. For $k = 182$, the best $-\log_{10} p$ from the functional enrichment of the modules learned by each of the six competing methods (on the x-axes) are compared to the best

$-\log_{10} p$ from the functional enrichment of the modules learned by INSPIRE (on the y-axis). Each dot corresponds to a KEGG, Reactome or BioCarta GeneSet, and only the GeneSets with a Bonferroni corrected Fisher's exact test $p < .05$ in at least one of the compared two methods are shown on each plot. For MGL variants and INSPIRE, the results from multiple runs are shown on each plot. We only considered the GeneSets that show sufficiently different levels of significance, i.e., $|\log_{10} p(i) - \log_{10} p(m)| \geq \delta$, where 'i' means INSPIRE and 'm' means the alternative method. $\delta = 6$ here and the results were consistent for varying δ .

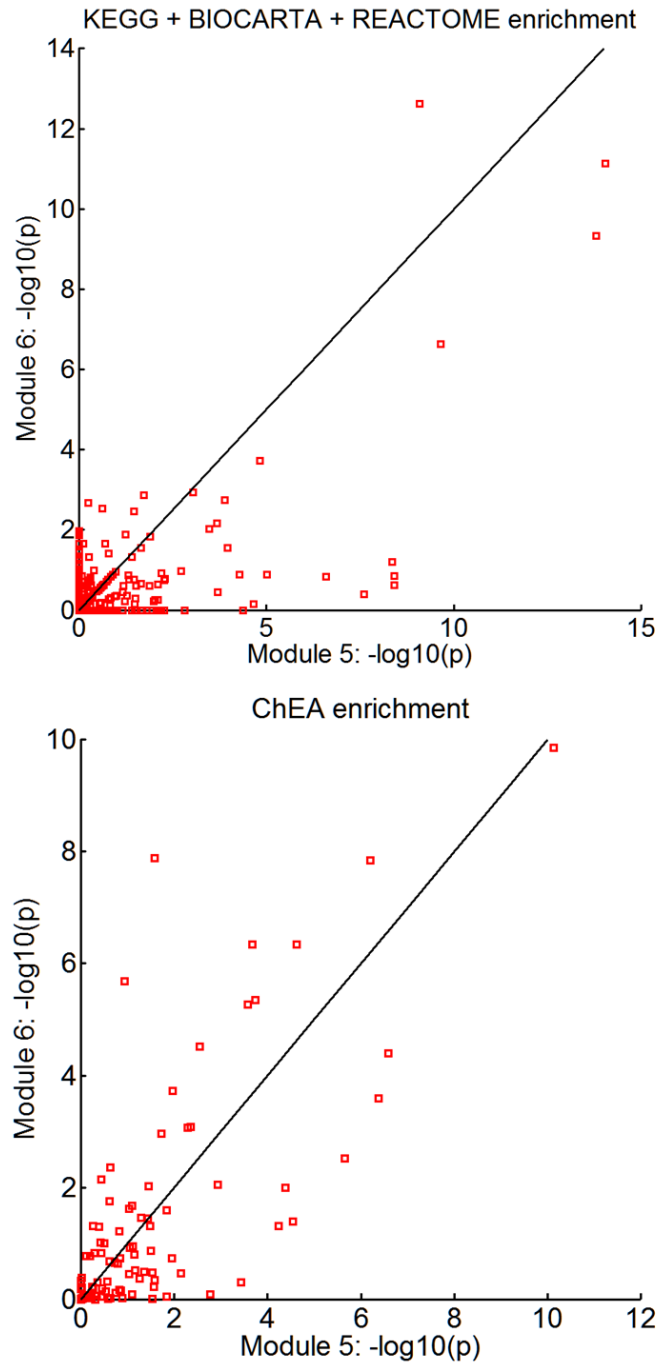


Appendix Figure 7. For $k = 182$, the best $-\log_{10} p$ from the ChEA enrichment of the modules learned by each of the six competing methods (on the x-axes) are compared to the best

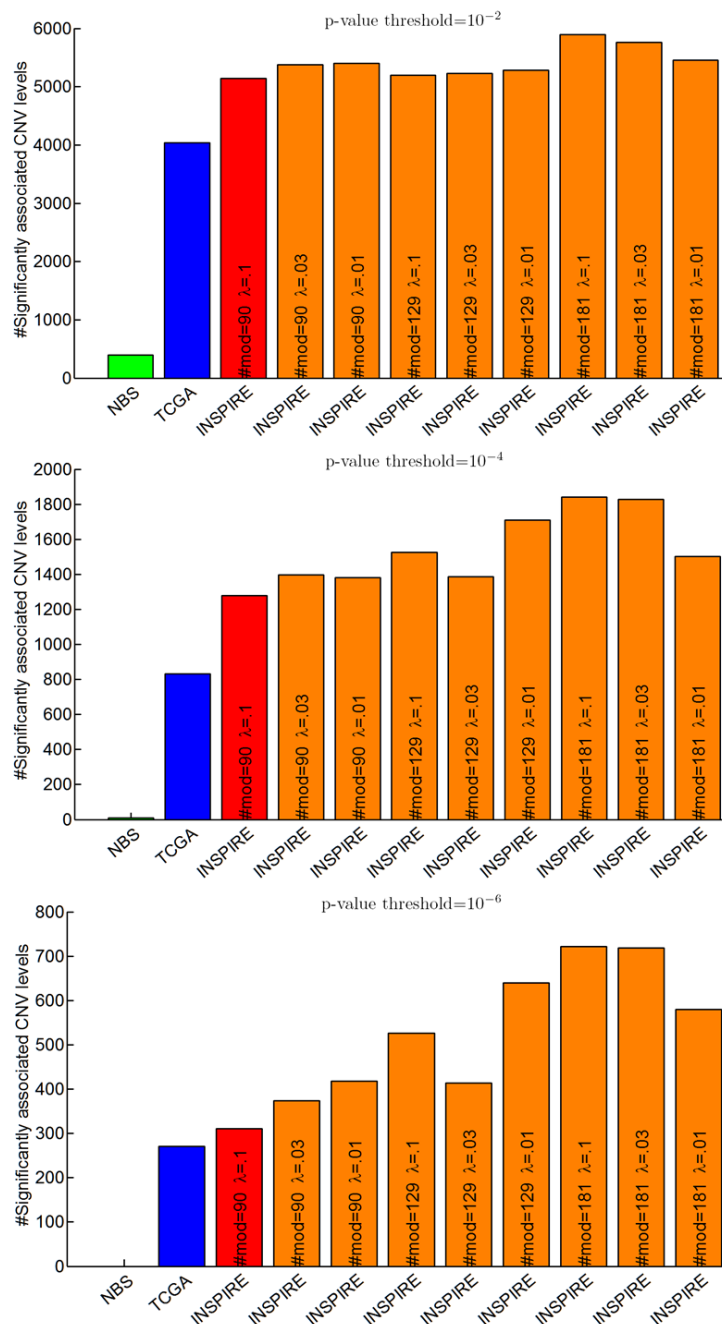
$-\log_{10} p$ from the ChEA enrichment of the modules learned by INSPIRE (on the y-axis). Each dot is for a group of genes composed of a TF and its targets, and only the TFs with a Bonferroni corrected Fisher's exact test $p < .05$ in at least one of the compared two methods are shown on each plot. For MGL variants and INSPIRE, the results from multiple runs are shown on each plot. We only considered the TFs that show sufficiently different levels of significance, i.e., $|\log_{10} p(i) - \log_{10} p(m)| \geq \delta$, where 'i' means INSPIRE and 'm' means the alternative method. We set $\delta = 3$ here and the results were consistent for varying δ . Blue dots represent the TFs which are contained by the INSPIRE module which was significantly enriched for the target genes of that TF, and red dots represent the TFs which are contained by an INSPIRE module different than the INSPIRE module that was significantly enriched for the target genes of that TF.



Appendix Figure 8. $-\log_{10} p$ from the KEGG, Reactome or BioCarta GeneSet enrichment (Left) and from the TF binding enrichment (Right) is compared for five negatively correlated module pairs with the biggest absolute correlation in the 9 datasets experiment. Each one of the two negatively correlated modules is shown on one of the x- or y-axis, and each dot corresponds to a KEGG, Reactome or BioCarta GeneSet (Left) or a group of genes composed of a TF and its targets (Right).

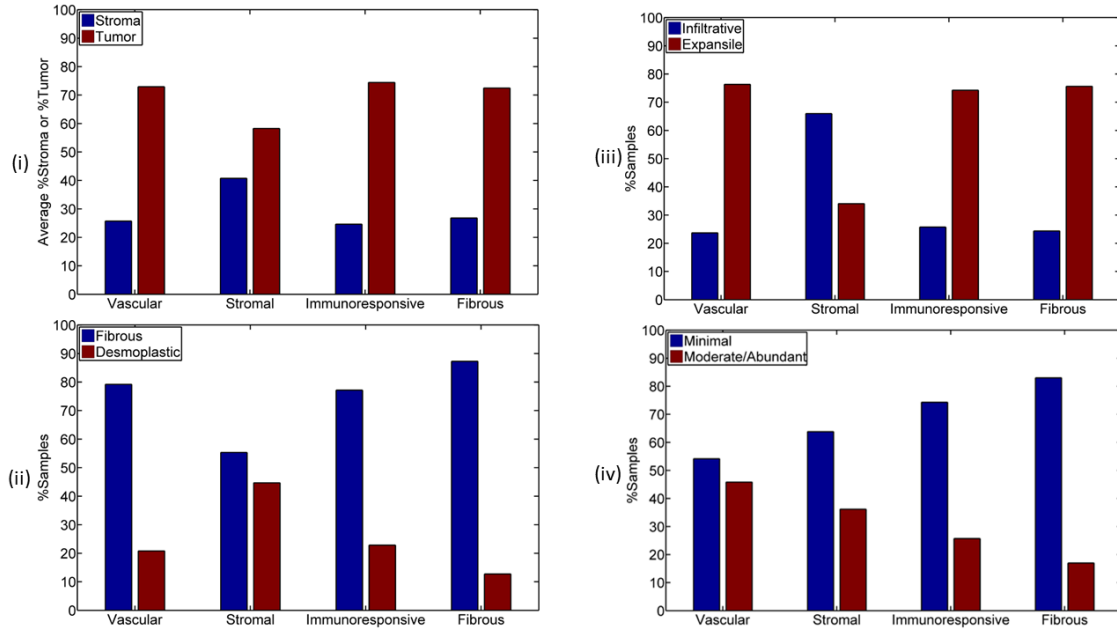


Appendix Figure 9. $-\log_{10} p$ from the KEGG, Reactome or BioCarta GeneSet enrichment (Top) and from the TF binding enrichment (Bottom) is compared for module 5 (on the x-axis) and 6 (on the y-axis). Each dot corresponds to a KEGG, Reactome or BioCarta GeneSet (Left) or a group of genes composed of a TF and its targets (Right).

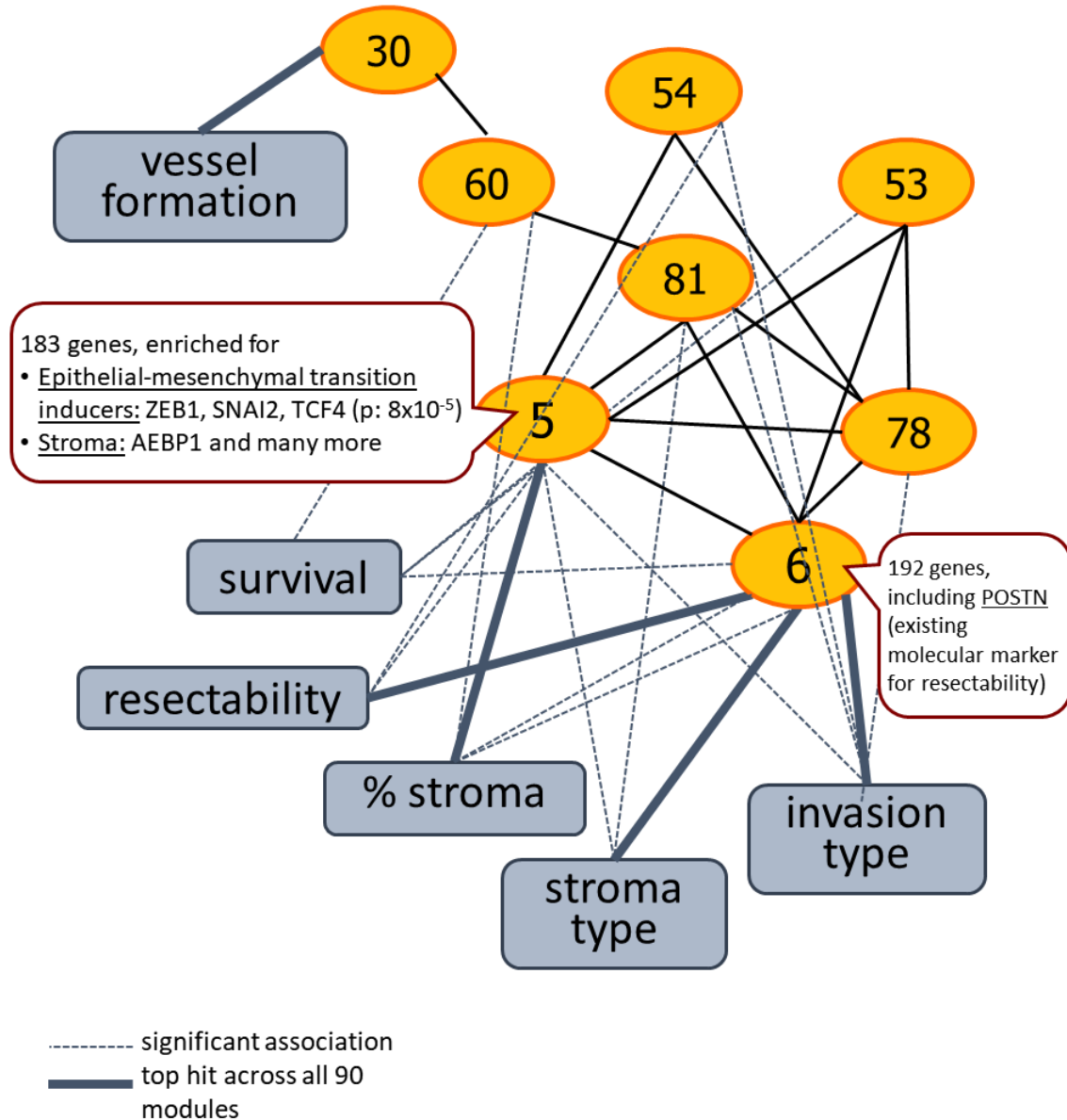


Appendix Figure 10. For three different Pearson's correlation p-value thresholds (10^{-2} , 10^{-4} , 10^{-6} respectively from top to bottom), the number of genes whose CNV levels are significantly associated with the learned ovarian cancer subtypes are shown for three methods: 1) the subtypes learned by a method that uses mutation profiles for the network-based stratification (NBS)⁵⁸ method (green), 2) the subtypes inferred from TCGA study²³ (blue), and 3) INSPIRE with varying sparsity tuning parameters (orange or red). Each bar for INSPIRE

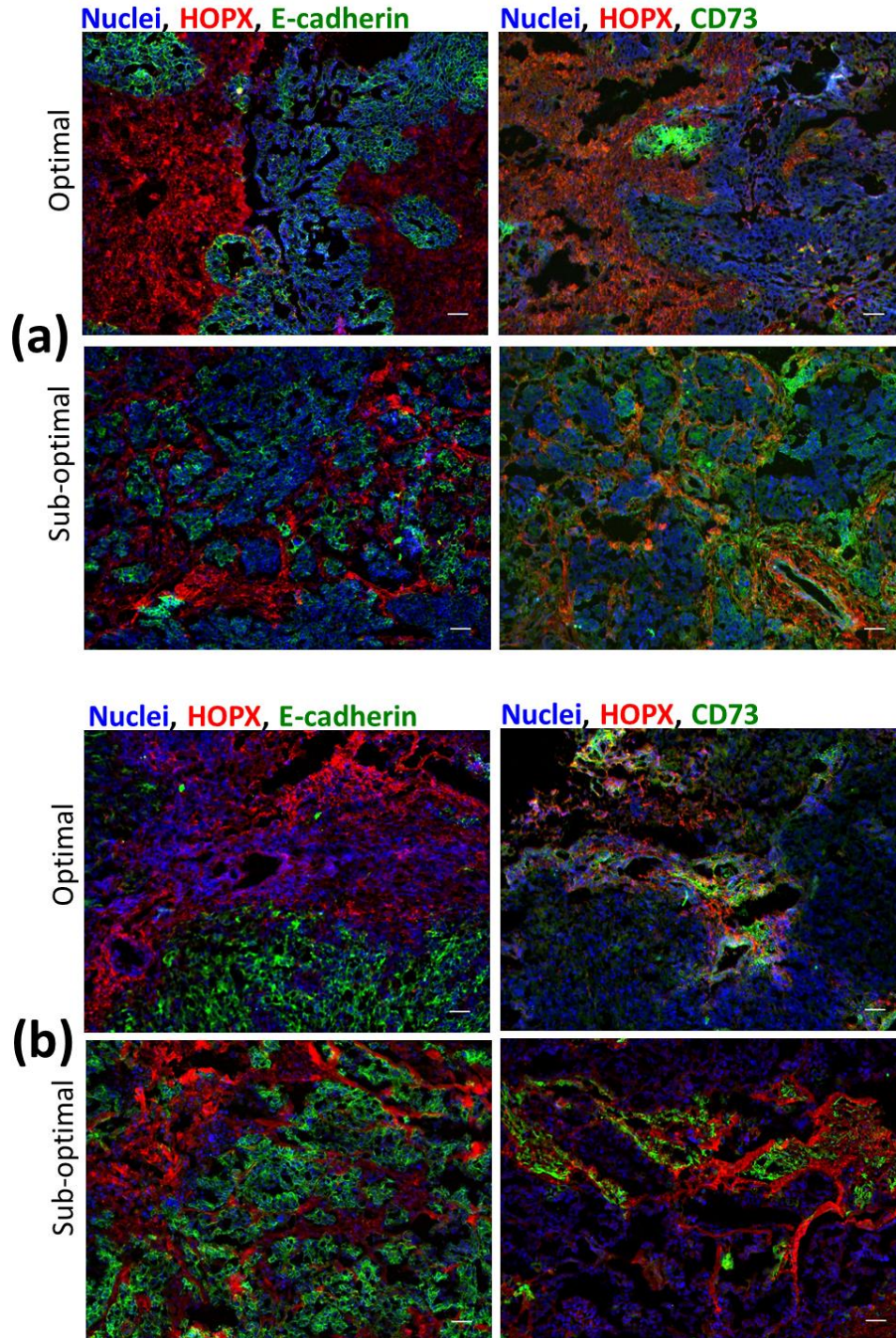
represents a setting with a different module count (k) and module network sparsity parameter (λ). The red bar for INSPIRE corresponds to the setting which our biological analyses are based on.



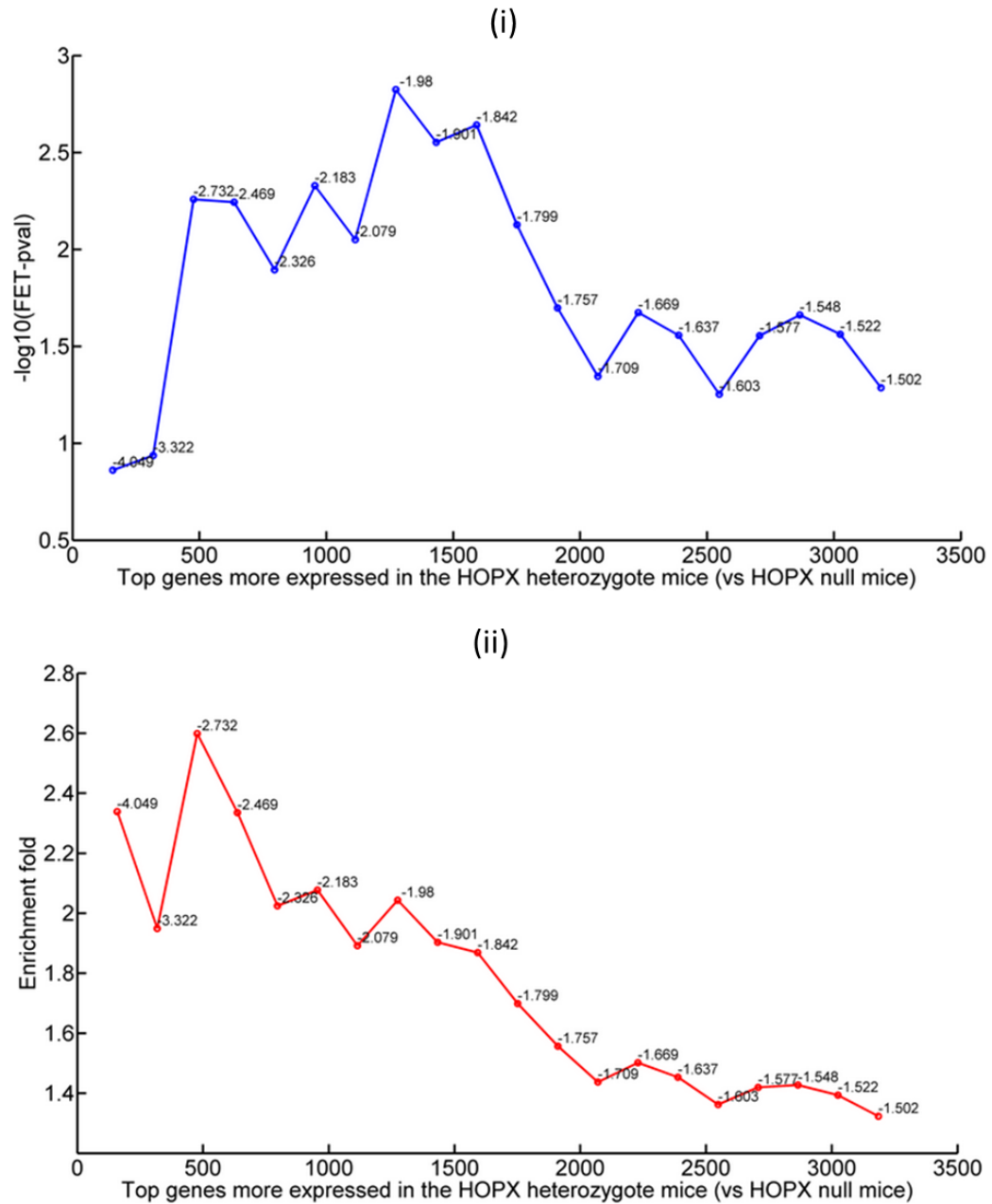
Appendix Figure 11. (i) For each of the INSPIRE subtypes, the percent stroma (blue bar) and the percent tumor (red bar) averaged over the patients in the subtype are shown. (ii) For each of the INSPIRE subtypes, the percentage of the patients in the subtype with fibrous stroma (blue bar) and desmoplastic stroma (red bar) are shown. (iii) For each of the INSPIRE subtypes, the percentage of the patients in the subtype with infiltrative invasion pattern (blue bar) and expansile invasion pattern (red bar) are shown. (iv) For each of the INSPIRE subtypes, the percentage of the patients in the subtype with minimal vessels (blue bar) and moderate or abundant vessels (red bar) are shown.



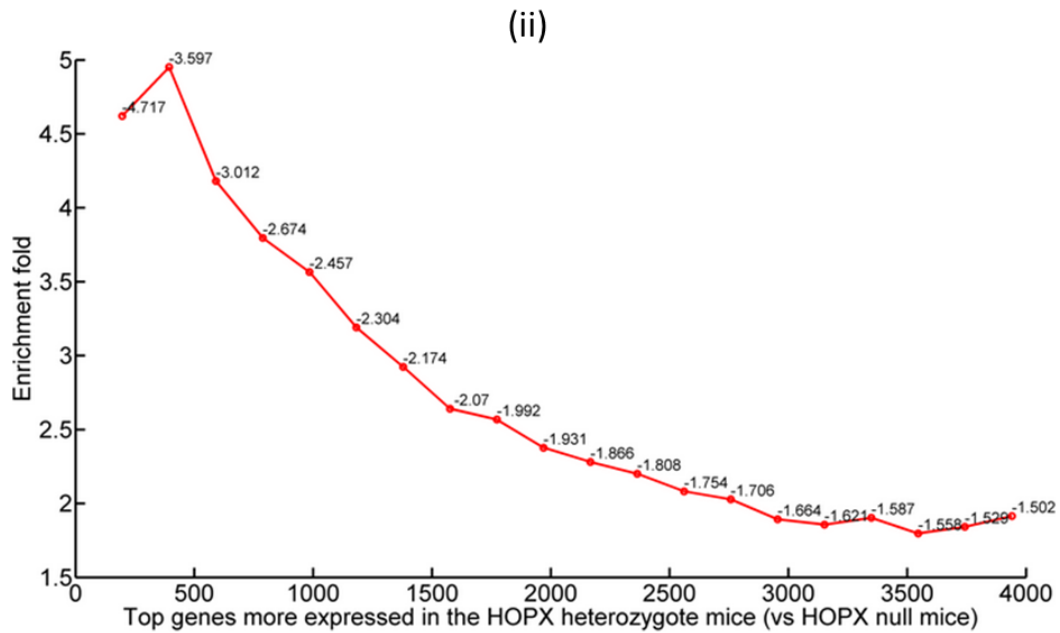
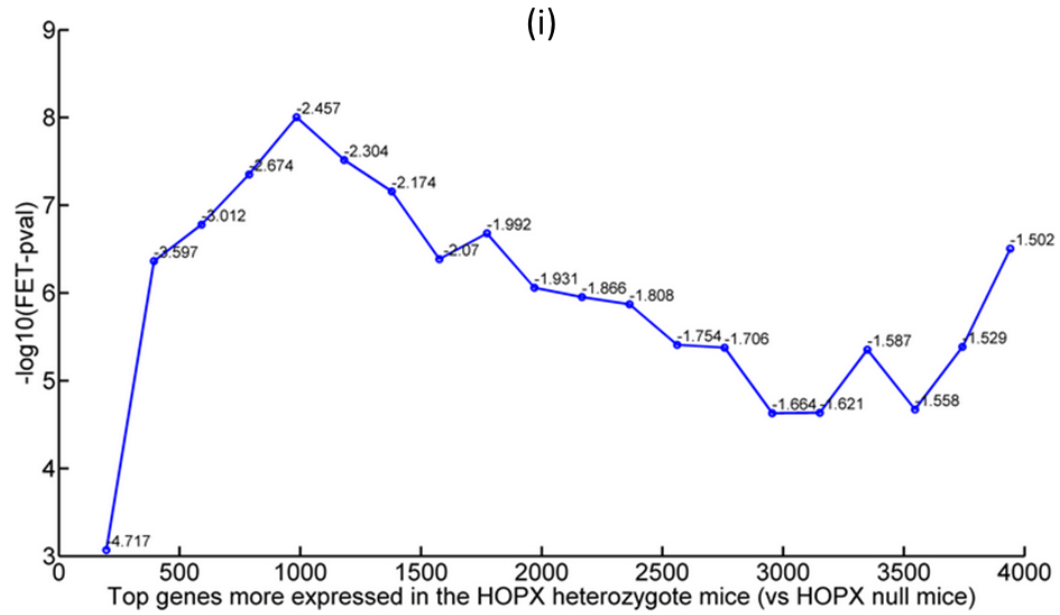
Appendix Figure 12. The interactions among the modules that show significant correlations with the important phenotypes in the TCGA ovarian cancer data, as shown by red bars in **Figure 15**. The edges are shown by black lines. Also, as a recap of **Figure 15** for those specific modules, the significant associations of each module with six important phenotypes in ovarian cancer are shown by dotted blue lines, and the associations that are the most significant among all modules are shown by solid blue lines. The details for the modules 5 and 6, which achieve top hits for a total of four phenotypes are given as well.



Appendix Figure 13. Additional fluorescent images of ovarian tumors from sub-optimally bulked and optimally debulked patients. Each row is a patient. As in **Figure 22**, HOPX is localized to the stroma and does not overlap with E Cadherin positive cancer cells. HOPX does however overlap with CD73, a mesenchymal stem cell marker.



Appendix Figure 14. The functional enrichment p-value (i) and the fold enrichment (ii) of module 5 genes for the genes down-regulated in HOPX-null mice for different thresholds of fold change in expression. The corresponding fold change threshold is displayed next to each dot on the curves. x-axes represent the number of down-regulated genes in HOPX-null mice whose expression fold change passes the fold change threshold displayed next to the corresponding value on the curve.



Appendix Figure 15. The functional enrichment p-value (i) and the fold enrichment (ii) of module 5 genes for the genes down-regulated in HOPX-null mice upon inhibition of Wnt signaling for different thresholds of fold change in expression. The corresponding fold change threshold is displayed next to each dot on the curves. x-axes represent the number of down-regulated genes in HOPX-null mice upon Wnt inhibition whose expression fold change passes the fold change threshold displayed next to the corresponding value on the curve.

19. APPENDIX A Tables:

Appendix Table 1. *The nine ovarian cancer gene expression datasets we used in the third set of experiments (biological application).*

Name	# Samples	# Genes	Platform
TCGA OV	560	11847	Affymetrix HGU133A + Agilent g4502 + Human Exon array
GSE14764	80	11942	Affymetrix HGU133A
GSE26712	185	11942	Affymetrix HGU133A
GSE6008	99	11942	Affymetrix HGU133A
GSE18520	53	18113	Affymetrix HGU133Plus2
GSE19829.GPL570	28	18113	Affymetrix HGU133Plus2
GSE20565	140	18113	Affymetrix HGU133Plus2
GSE30161	58	18113	Affymetrix HGU133Plus2
GSE9899	295	18113	Affymetrix HGU133Plus2

Appendix Table 2. *The two ovarian cancer gene expression datasets we used in the second set of experiments.*

Name	# Samples	# Genes	Platform
GSE19829.GPL570	28	18113	Affymetrix HGU133Plus2
GSE19829. GPL8300	42	8331	Affymetrix HGU95Av2

Appendix Table 3 (provided as a separate .xls file). *The properties of the modules learned by INSPIRE. Column A: ID of the module. Column B: The number of the neighbor modules in the learned module network. Column C: IDs of the neighbor modules. Column D: The number of genes the module contains. Column E: The names of the genes the module contains. Column F: The number of MSigDB C2 categories for which the module is significantly enriched (Bonferroni-corrected $p < 0.05$) based on a Fisher's exact test. Column G: The MSigDB C2 categories for which the module is significantly enriched and the corresponding Bonferroni-corrected $-\log_{10} p$ values. The enriched categories are ordered from most significant to least significant. Columns H-L: The p-value from Pearson's correlation of the feature corresponding to the module with six phenotypes: percent stroma (column H), percent tumor (column I), vessel*

formation and abundance (column J), stroma type (column K), invasion pattern (column L), residual tumor (column M), and survival (column N). As in **Figure 15**, we highlight modules 5, 6, 53, 54, 60, 78 and 81 that are significantly correlated with at least three of the six phenotypes by red. We also highlight module 30 since it is the only module that has a significant correlation with the vessel formation phenotype.

Appendix Table 4 (provided as a separate .xls file). The manually examined five histologic phenotypes for randomly selected 98 TCGA ovarian cancer patients with their TCGA patient IDs listed in Column A. The histologic phenotypes considered include percent stroma (Column B), percent tumor (Column C), vessel formation and abundance (Column D), stroma type (Column E), and invasion pattern (Column F).

Appendix Table 5 (provided as a separate .xls file). For five the most negatively correlated modules, the values of $-\log_{10} p$ from the pathway enrichment test (KEGG, Reactome, and BioCarta) (top) and from the TF binding enrichment test (bottom) are compared. For each of the five negatively correlated module pairs, we show five pathways and TFs that have the largest difference in the value of $-\log_{10} p$ between the two modules.

Appendix Table 6. The $-\log_{10} p$ from the pathway (KEGG, Reactome, and BioCarta) enrichment test (Top) and from the TF binding enrichment test (Bottom) are compared between module 5 and module 6. We show the five pathways or TFs that have the largest difference in the value of $-\log_{10} p$ between the two modules.

		Module 5	Module 6
Gene Set	REACTOME_SMOOTH_MUSCLE_CONTRACTION	8.4036	0.6308
	REACTOME_SIGNALING_BY_PDGF	8.41	0.8542
	REACTOME_MUSCLE_CONTRACTION	7.604	0.3971

TF	REACTOME_NCAM1_INTERACTIONS	8.358	1.1928
	REACTOME_NCAM_SIGNALING_FOR_NEURITE_OUT_GROWTH	6.5945	0.8311
	PAX3-FKHR	1.5796	7.8889
	POU3F2	0.9312	5.6912
	TP53	4.5343	1.3829
	AR	5.6507	2.5125
	TRIM28	3.4385	0.3065

Appendix Table 7. For module 5, module 6 and a hypothetical module containing all genes in module 5 and module 6, the prediction accuracy is compared in six prediction tasks via cross-validation tests. The best performance for each prediction task is highlighted by green.

	Module 5	Module 6	Module 5&6 merged
% stroma	0.8242	0.8063	0.818
Stroma type	0.8551	0.8634	0.86
Vessels	0.7903	0.7913	0.7892
Invasion	0.8382	0.847	0.8458
Residual tumor	0.8376	0.8397	0.8386
Survival	0.5465	0.5406	0.5457

Appendix Table 8. We remove 20%, 40%, 60%, and 80% of the genes in each module whose expression levels are least significantly associated with the respective phenotype, and regenerate latent variables from the rest of the genes in the modules. For each of those settings, the prediction performance is compared in six prediction tasks via cross-validation tests. The best performance(s) for each prediction task is highlighted by green.

	Original	20% removed	40% removed	60% removed	80% removed
% stroma	0.7951	0.7908	0.7712	0.7725	0.7591
Stroma type	0.8367	0.7857	0.7959	0.7959	0.8367
Vessels	0.7634	0.7419	0.7204	0.6989	0.6882
Invasion	0.7717	0.7609	0.75	0.7391	0.7174
Residual tumor	0.7912	0.7912	0.7842	0.7773	0.768

Survival	0.5689	0.5305	0.5302	0.5373	0.5459
----------	--------	--------	--------	--------	--------

Appendix Table 9 (provided as a separate .xls file). The assignment of the patients from TCGA ovarian cancer study²³ into four subtypes based on the INSPIRE latent variables.

Appendix Table 10 (provided as a separate .xls file). The enrichment of the marker modules that were selected by the Significance Analysis of Microarrays (SAM) procedure to be significantly differentially expressed in one of the four INSPIRE subtypes with MSigDB C5 categories of genes. Column A: ID of the module. Column B: The enriched MSigDB C5 category and $-\log_{10} p$ of the enrichment based on a Fisher's exact test.

Appendix Table 11. The confusion matrix representing the overlap between the INSPIRE subtypes and the subtypes revealed by the TCGA ovarian cancer study²³.

		TCGA subtypes			
		Proliferative	Mesenchymal	Immunoreactive	Differentiated
INSPIRE subtypes	Vas.	103 ($p = 6.81 \times 10^{-82}$)	8	2	10
	Str.	0	82 ($p = 1.03 \times 10^{-35}$)	12	10
	Imm.	17	9	82 ($p = 6.54 \times 10^{-13}$)	63
	Fib.	16	10	9	52 ($p = 2.17 \times 10^{-17}$)

Appendix Table 12. The six gene expression datasets we used in our pan-cancer survival analysis.

Name	Cancer type	# Samples	# Genes	Platform
Osloval	BRCA	184	17920	Illumina HTv3

GSE16011	GBM	159	18113	Affymetrix HGU133Plus2 + Agilent 244 k aCGH + Human Exon 1.0
TCGA-GBM	GBM	512	17814	Affymetrix HGU133A
Gentles	AML	784	16855	Affymetrix HGU133Plus2
GSE9899	OV	295	18113	Affymetrix HGU133Plus2
TCGA-LUAD	LUAD	395	16256	Illumina HiSeq

APPENDIX B (APPENDIX TO MERGE)

1. Interpretation of top expression markers identified by MERGE

FLT3 expression predicts response to Flt3 inhibitors

FLT3 (FMS like tyrosine kinase 3), the top-ranked gene in the Flt3 inhibitor class (**Appendix Table 14**), is a significant expression hub and regulator that is highly mutated (**Appendix Figure 19**). *FLT3* mutations are associated with a poor prognosis in AML^{189,190}. The principal type that occurs in about 25% of AML patients is internal tandem duplication (ITD). After treatment with some Flt3 inhibitors¹⁹¹, these patients often develop *FLT3* D835 (kinase domain) mutations, which is present in 8% of patients. One study shows a correlation between *FLT3* expression levels and prognosis among patients with wild-type *FLT3*¹⁹².

Our method led to a new finding: the high expression of *FLT3* is associated with increased sensitivity to the drugs sunitinib, ponatinib (previously AP24534), midostaurin (PKC412) and tandutinib (**Figure 29**). Sunitinib is a multi-targeted receptor tyrosine kinase inhibitor FDA-approved for renal cell carcinoma and imatinib-resistant gastrointestinal stromal tumor. Ponatinib, another multi-targeted receptor tyrosine kinase inhibitor, is FDA-approved for chronic myelogenous leukemia that has failed to respond to first-line inhibitors. Tandutinib is an inhibitor of the type III receptor tyrosine kinases *FLT3*, platelet-derived growth factor receptor (PDGFR), and KIT. Midostaurin was studied in a prospective, randomized trial in newly diagnosed patients with *FLT3* mutations undergoing induction, consolidation and maintenance therapy, and the group that received midostaurin exhibited prolonged event free (p=0.0044) and overall survival (p=0.007) as compared to placebo¹⁹³. Midostaurin recently received breakthrough therapy designation from the FDA for newly diagnosed *FLT3*-mutated AML.

Because *FLT3* mutation status is an important prognostic indicator in AML with the potential to guide therapy, we compared *FLT3* mRNA expression and *FLT3* mutation status in terms of significance of correlation with response to 53 drugs. For (a) patients and (b) cell lines, **Appendix Figure 20** shows the significance of association achieved by *FLT3* expression level (y-axis) vs. by *FLT3* mutation status (x-axis) for each drug. More dots appear above the diagonal in both **Figure 20a** and **b**, which implies that mRNA level achieved a more significant association than *FLT3* mutation status for a larger number of drugs (36 vs. 17 in patients; 31 vs. 22 in cell lines). For fair comparison, to generate **Figure 20a** for both *FLT3* mRNA and mutation status, we used only 27 patients for whom mutation status was known. Further, we used Quentmeier, Reinhardt, Zaborski, & Drexler (2003) as the source for the *FLT3* mutation status for the cell lines.

CASP8AP2 expression predicts sensitivity to Bcl2 inhibitors

CASP8AP2 (caspase 8 associated protein 2) is the top-ranked gene in the Bcl2 inhibitor class (**Appendix Table 14**). Overexpression of *BCL2* lets cancer cells evade apoptosis, the process of programmed cell death. The first Bcl2 inhibitor to gain approval for patients, venetoclax, was approved by the FDA on April 11, 2016 to treat the subset of relapsed patients with chronic lymphocytic leukemia that have the deletion of chromosome 17p, which contains the *TP53* gene. The drug is also undergoing evaluation in early phase AML clinical trials (e.g. NCT02203773, NCT02287233 on ClinicalTrials.gov). We found an association between an increased expression of *CASP8AP2* with an increased sensitivity to Bcl2 inhibitors (**Figure 29b**). Caspases mediate the activation of apoptosis, and low expression of *CASP8AP2* has been associated with a poor prognosis in pediatric acute lymphoblastic leukemia¹⁹⁵ and an increased risk of relapse¹⁹⁶.

Moreover, a single report notes a translocation of the poor risk *MLL* (mixed lineage leukemia) gene to the *CASP8AP2* gene in an AML patient with a t(6;11)(q15;q23) translocation¹⁹⁷. It is therefore not surprising that expression of a gene in the apoptotic pathways (e.g.-a caspase associated protein) would be associated with sensitivity to inhibiting the activity of one of the main antagonists of apoptosis. Restoring apoptosis by inhibiting *BCL2* may enhance cell death by increasing the level of expression of apoptotic pathway proteins such as *CASP8AP2*.

L2HGDH expression is the predictor for CDK inhibitors

Cyclin-dependent kinase (CDK) inhibitors prevent proliferation of cancer cells. *L2HGDH* (L-2-hydroxyglutarate dehydrogenase) is an enzyme that catalyzes conversion of L-2-hydroxyglutarate to alpha ketoglutarate. We found that the increased expression of *L2HGDH*, the only gene associated with the CDK inhibitor class (**Appendix Table 14**), was associated with increased sensitivity to CDK inhibitors. Elevated serum 2-hydroxyglutarate (2-HG) levels have been associated with isocitrate dehydrogenase (IDH1 and IDH2) mutations in AML^{198,199} as well as clinical outcomes¹⁹⁸. In addition, elevated 2-HG inhibits the function of *TET2*, resulting in DNA hypermethylation²⁰⁰. In addition, histone demethylases are competitively blocked by 2-HG²⁰¹. The elevated expression of *L2HGDH* would be expected to lower levels of 2-hydroxyglutarate and thus abrogate the deleterious secondary effects of DNA and histone methylation. CDK inhibitor genes were the most frequently downregulated genes in IDH1 mutants²⁰² with high 2-HG levels, thus linking the association we observed with susceptibility to CDK inhibitors with *L2HGDH*.

MNT and BAZ2B are expression markers for HDAC inhibitors

We found a correlation between susceptibility to HDAC inhibitors and increased expression levels of *MNT* and *BAZ2B* (**Figure 29b**), top two genes in the HDAC inhibitor class (**Appendix Table 14**). *MNT* encodes the protein Mnt, a member of the Myc/Max/Mad network of TFs that controls cell proliferation, differentiation, and death. It is known to be critical for normal myeloid differentiation in AML cell lines, and its loss leads to proliferation²⁰³. Mnt has a Sin3-interaction domain (SID) that lets it interact with mSin3A and mSin3B, which recruit histone deacetylases (HDACs), resulting in transcriptional repression²⁰⁴. Therefore, a high expression of *MNT* would be predictive of whether inhibiting HDACs has an effect.

BAZ2B (Bromodomain Adjacent to Zinc Finger Domain, 2B), is a bromodomain containing protein. The sole function of bromodomain is to recognize acetyl-lysine on histones and non-histone proteins²⁰⁵. Histone deacetylases remove acetyl groups from these sites. This suggests that high expression levels of such genes are associated with chemotherapy drug sensitivity for HDAC inhibitors, bringing a new insight into mechanisms that govern individual patient response to chemotherapy involving regulation of chromatin state. Additionally, a popular class of bromodomain inhibitors has shown potential for treating aggressive leukemia²⁰⁶.

Histone deacetylases impair myeloid differentiation. This is why HDAC inhibitors are potentially useful in therapy and have long been long investigated to treat myelodysplastic syndrome and AML²⁰⁷. Various publications address the efficacy of some HDAC inhibitors in clinical trials in AML, including vorinostat, panobinostat, and romidepsin^{208–210}.

BEX2 expression is a potential marker for NFkB inhibitors

NFkB (Nuclear factor kappa B), a TF, is constitutively active in many cancers and inhibits both apoptosis and drug resistance. BEX2, one of the top genes in the NFkB inhibitor class (**Appendix Table 14**), is overexpressed in breast cancer and glioma²¹¹. In KIT-driven AML, NFkB binds to Sp1 and transactivates KIT²¹². In addition, inhibiting both NFkB and JNK is effective in AML expressing the tumor necrosis factor²¹³. Lastly, *BEX2* expression occurs in AML with translocations involving the *MLL* gene²¹⁴.

SMARCA4 and MZF1 are markers for topoisomerase II inhibitors

SMARCA4 and *MZF1* are the top two genes in the anthracycline/topoisomerase inhibitor/DNA intercalator class (**Appendix Table 14**), and both genes are correlated with sensitivity to the drugs in that class (**Figure 29a**). The drugs *SMARCA4* is associated with are mitoxantrone, etoposide, and topotecan (**Figure 29b**). Mitoxantrone and etoposide are topoisomerase II inhibitors; topotecan is a topoisomerase I inhibitor. *SMARCA4* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4) is a component of one of the adenosine triphosphate (ATP)-dependent SWI/SNF-like Brg/Brm-associated factor (BAF) chromatin remodeling complexes. Mutations of *SMARCA4* (BRG1) occur in 10 to 35% of non-small cell lung carcinoma, 15% of Burkitt's lymphoma, 5 to 10% of childhood medulloblastoma, and less frequently in other cancers²¹⁵. Further, *SMARCA4* mutations characterize small cell carcinoma of the ovary of the hypercalcemic type²¹⁶. Furthermore, one recent study demonstrated a direct role for *SMARCA4* in facilitating decatenation of DNA by topoisomerase II²¹⁷. In leukemia, the core ATPase subunits are BRG/*SMARCA4* and BRM/*SMARCA2*, where BRG/*SMARCA4* is essential for the proliferation of both normal hematopoietic stem cells and leukemia stem cells²¹⁸.

MZF1 is a member of the SCAN-zinc finger family of TFs, frequently mutated in many types of cancer²¹⁹. Initially, it was believed to have a role in promoting myeloid malignancy²²⁰, but because knockout animals exhibited proliferation of hematopoietic progenitors, it was also thought to potentially suppress hematopoietic malignancy²²¹. *MZF1* localizes in the PML-NBs (promyelocytic leukemia nuclear bodies)²²², which are protein complexes involved in posttranslational modification of nuclear proteins and response to DNA damage²²³. It was shown that the topoisomerase II inhibitors etoposide and doxorubicin, known to create double stranded DNA breaks²²⁴, increased the number of PML-NBs by a fission mechanism, and then the number became regulated by cell cycle checkpoint control²²³. Thus, topoisomerase inhibitors directly affect the PML-NBs that incorporate *MZF1*.

2. Patient and cell line mRNA expression profiles

For all of the 30 samples, gene expression was analyzed using Affymetrix Human U133 Plus 2.0 arrays. The analysis was performed at the Department of Environmental and Occupational Health Science's Functional Genomics Core Laboratory at the University of Washington. RNA was isolated using the Qiagen RNeasy Mini kit (Qiagen Inc, Germantown MD). RNA quantity was assessed by measuring OD₂₆₀, and purity by OD_{260/280} and OD_{260/230} ratios with the NanoDrop spectrophotometer (Thermo Fisher Scientific Inc, Wilmington, DE). RNA integrity was assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Santa Clara, CA). RNA samples were required to meet these stringent quality control parameters and were then processed according to the manufacturer's recommendations. Briefly, 250 ng of total RNA was reverse transcribed. The resulting cDNA was converted to biotinylated cRNA. The biotinylated fragments were

hybridized to Affymetrix U133 Plus 2.0 arrays, washed and stained. The arrays were scanned with an Affymetrix GenChip® 3000 scanner. Image generation and feature extraction were performed using Affymetrix GeneChip Command Console software.

For the additional test set of 12 patient samples (from which we show the results in **Figure 26b**), RNA-seq data generation was performed at the Northwest Clinical Genomics Laboratory, UW Medicine Center for Precision Diagnostics using the TruSeq® stranded mRNA kit from Illumina (San Diego, CA) according to manufacturer's instructions. We had two RNA-seq replicates for each of the 12 samples. For MERGE experiments, we averaged the FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values from the Cufflinks²²⁵ output from the two replicates of each sample.

Gene expression for the 14 AML cell lines (EOL-1, MOLM/16, NB4, OCI-AML3, SKM1, HL60, KG1, MOLM/13, MV4.11, OCI-AML2, PL-21, U937, KASUMI1, THP-1) was retrieved from the Cancer Cell Line Encyclopedia (CCLE)¹⁰⁹.

3. Curve fitting to estimate drug sensitivity profiles

As was done in CCLE study¹⁰⁹, from each dose-response data for a combination of a drug and a patient sample, we extracted summary statistics, including area under the curve (AUC), half maximal inhibitory concentration (IC₅₀), half maximal effective concentration (EC₅₀), and maximal effect level (A_{max}). In brief, we fitted the dose-response data (averaged over duplicates) to the following 4 parameter sigmoid model:

$$y = A_{\text{inf}} + \left(\frac{A_0 - A_{\text{inf}}}{1 + \left(\frac{x}{\text{EC}_{50}} \right)^{\text{Hill}}} \right)$$

Here, A_0 and A_{inf} are the top and bottom asymptotes of the response (cell viability); EC_{50} is the inflection point of the curve; and Hill is the Hill slope, which describes the curve's steepness. Other key parameters derived from the models include IC_{50} (the concentration where the fitted curve crosses 50% in cell viability) and A_{max} (the maximal activity value reached within a model). We extracted these parameters (AUC, IC_{50} , EC_{50} , and A_{max}) after curve fitting by using MATLAB's 'nlinfit' function for nonlinear curve fitting. For inactive compounds, it is impossible to derive an IC_{50} ; in this instance, as was done in the CCLE study¹⁰⁹, we simply used the maximum tested concentration as the default value – which serves primarily as a placeholder to allow algorithms to work on all samples. Batch effects were corrected using ComBat¹⁶³ for each drug sensitivity summary based on the three batches of experiments for the 30 patients.

From these statistics, we chose AUC because it represents an average of drug sensitivity across a range of drug concentrations. Indeed, AUC showed by far the strongest association with gene expression levels; the number of significant associations (FDR corrected p -value < 0.1) between a gene and a drug sensitivity measure was 53,967 for AUC, 23,641 for IC_{50} , 15,112 for EC_{50} , and 7,132 for A_{max} . Considering additional drug sensitivity measures would increase the total number of hypotheses.

4. Summary of the clinical information and its consistency with our *in vitro* drug sensitivity data

Of the 30 AML patient samples, 24 were newly diagnosed, and 6 had relapsed. The majority of samples (19 of 30) were males. The median age was 54, and the age range was [19, 83]. Six patients had antecedent hematologic disorders. According to European LeukemiaNet criteria²²⁶,

seven samples were in favorable risk group, eleven were in intermediate-1, 2 were in intermediate-2, and ten were in the adverse risk cytogenetics group. Several different regimens were used to treat the patients, most of which included cytarabine, and many of which included an anthracycline. **Appendix Table 19** contains the detailed clinical information including usual evaluation (including risk group category and cytogenetic features), response to treatment, duration of remission, and the individual regimens for each of the samples.

We measured the standard clinical mutation status on *FLT3* and *NPM1* for most of the 30 patients (26 patients tested for *NPM1* mutation; 27 tested for *FLT3* mutation). Seven of 26 patients had an *NPM1* mutation, and seven of 27 patients had the *FLT3* ITD mutation. We observed a statistically significant association between *FLT3* mutation status and 12 drugs (FDR corrected p -value < 0.1): AS101, AT-7519, AZD7762, cladribine, mitomycin C, mitoxantrone, NVP-AUY-922, obatoclax, PIK-75, midostaurin (PKC412), sunitinib, and tandutinib. Three of these drugs (midostaurin, sunitinib and tandutinib) are known to have a *FLT3* inhibitory role. None of the drugs is associated with *NPM1* mutation status at the same significance level.

Additionally, we checked the statistical significance of the association between the complete remission (CR) status and the drug sensitivity measure (AUC) across all 53 drugs. Interestingly, 12 of 15 drugs (80%) (azacitidine, bortezomib, cladribine, clofarabine, daunorubicin, etoposide, fludarabine, mitoxantrone, panobinostat, PKC412, tretinoin, vorinostat) that have shown clinical efficacy in AML treatment were significantly associated with CR at FDR= 0.05, while only 21 of the other 38 drugs (55.3%) showed significant association with CR (**Appendix Table 18**). Some of the 38 drugs in the latter set have successor agents that have shown clinical efficacy in AML, such as the Bcl2 inhibitor ABT-199 (venetoclax), a successor to ABT-737 and ABT-263 that we

tested. Moreover, drugs might have additive or synergistic activity in combinations; however, in this study, the drugs were assayed as single agents. We note that cytarabine, commonly used to treat AML, is not listed in **Appendix Table 18** because it was not selected by the procedure that we followed to select the 53 drugs we focused on in our computational framework.

While cytarabine is commonly used to treat AML and a majority of the 30 AML patients received this drug, it was not one of the 53 drugs included in our MERGE computational framework: our criterion for selecting whether to include each of the 160 tested drugs in the study was exhibited activity (cell viability $\leq 50\%$) of the drug against at least half of the 30 patient samples; cytarabine did not satisfy this criterion. One possible reason is that the tested concentrations were too low for the 12 patients in the first batch (concentration range 2×10^{-10} M to 1×10^{-6} M), and cytarabine did not exhibit any activity for those patients until the highest concentration point. The concentrations were then increased for the next 18 patient samples (range 4×10^{-8} M to 1×10^{-4} M). For comparison, the peak mean plasma concentration after high dose cytarabine was 2×10^{-3} M²²⁷.

5. Extracting driver features from publicly available sources

The MERGE algorithm takes a set of *driver features* for each gene – namely, expression hubness, candidate regulators, mutation, copy number variation, and methylation – as input and estimates a MERGE score for each gene based on these features (**Figure 24b**). These features were extracted from publicly available sources, such as TCGA (The Cancer Genome Atlas) AML study¹⁰⁷, AML expression studies⁷⁶, and gene annotation databases, as described in detail below.

Expression hubness: In our prior study, we developed a novel computational method, named SPARROW (**SPAR**se selected exp**RESSi**On regulators identified **W**ith penalized regression), to estimate each gene's *hubness* purely based on expression data from cancer patients. SPARROW employs a sparse statistical model in which each gene's expression level is modeled as a linear combination of a small set of other genes (i.e., sparse basis), and determines the hubness of each gene based on how often it is chosen in the sparse basis for any other gene⁹³. To use gene hubness as a MERGE feature, we downloaded the R data object (.rda) containing the SPARROW results for AML from <http://sparrow-leelab.cs.washington.edu/data>. We then used 'sparrow1' scores (the number of downstream genes) from the 'basesFreq' object as the expression hubness feature.

Candidate regulators: The genes known to regulate other genes are more likely disease drivers and hence more reliable molecular markers for therapeutic response than those that are not. To incorporate this hypothesis, we used a list of genes known to have regulatory roles, including TFs, chromatin remodelers and signal transduction genes, constructed based on gene annotation databases⁷⁶. Based on this list of 3,052 genes, we generated a binary feature for each gene by assigning 1 if the gene was on the list and 0 otherwise.

Mutation: We downloaded significance measures of mutation frequencies for each gene measured in the MutSig2CV Analysis of the AML study from TCGA (<http://firebrowse.org/?cohort=LAML>). Each gene was given a *p*-value that measured the statistical significance that the gene had mutated more often than expected by chance given background mutation processes across patients; we used $-\log_{10}p$ -value as the feature value.

Copy number variation (CNV): We downloaded CNV measures (gdac.broadinstitute.org/LAML-TB.CopyNumber_Gistic2.Level_4.2015082100.0.0.tar.gz) from

http://gdac.broadinstitute.org/runs/analyses__2015_08_21/data/LAML/20150821/. Then, we used `all_data_by_genes.txt` in this tar.gz file to assign 1 (having CNV) or 0 (no CNV) to each gene. We set the CNV feature of a gene to 1 if the gene was amplified or deleted by at least .05 in at least 20 of 191 patients (~10%), and to 0 otherwise.

Methylation: We downloaded DNA methylation measures (gdac.broadinstitute.org_LAML.Methylation_Preprocess.Level_3.2015110100.0.0.tar.gz) from http://gdac.broadinstitute.org/runs/stddata__2015_11_01/data/LAML/20151101/. We then used `LAML.meth.by_mean.data.txt` in this tar.gz file to obtain the average methylation levels for each gene across all patients.

6. Averaging the methylation values across the CpG probes and across the samples

The AML study from TCGA includes preprocessed methylation profiles generated by four different pipelines that retrieve the gene-based methylation values from the CpG probes. These four include: (1) using the CpG probe with the highest anticorrelation with the gene's expression level, (2) using the CpG probe with the highest anticorrelation with the clinical data, (3) using the mean signal intensities across all CpG probes in proximity to the gene, and (4) using the CpG probe with the maximum standard deviation across all beta values.

A specific position of methylation (e.g., whether it occurs around the transcription start site or over a gene body) offers important information to understand the epigenetic cause of variation of downstream phenotypes, such as gene expression levels. Accordingly, the data generated by the first two preceding approaches were generally chosen when the goal was to explain a phenotype of interest (i.e., gene expression or clinical data) and to choose the probe likely to represent the

molecular basis for the phenotype. Since we sought to develop a general framework to utilize methylation data from the external AML sources, our study used the data generated by the approach (3), which is independent from any specific phenotype and provides a general summary of TCGA methylation data. We did not use the data from approach (4) because it discards probes with a standard deviation below a specified cutoff, so the dataset contained only ~2K genes, too small compared to the ~17K genes in the MERGE model.

After getting each sample's (patient's) methylation values for each gene as explained above, we used the sample mean across patients to compute each gene's methylation score. While other ways may provide a summary statistic across all patients, we considered only the sample mean in the current study: mean is the most commonly used measure of central tendency for continuous-valued variables, and several authors have averaged methylation values over samples (e.g., Moarii, Boeva, Vert, & Reyal, 2015).

7. Standardization of driver features, predictors and response variables

We standardized each driver feature before running the MERGE algorithm so that each feature had mean 0 and variance 1. We performed standardization so that despite different scales, driver features were treated fairly when estimating MERGE scores. We standardized binary features (candidate regulator and CNV) as we did other non-binary features. For any statistical model that combines data from several different kinds of variables (i.e., features) and aims to learn weights for the variables with regularization, one must provide all variables to the algorithm on the same scale to achieve fairness²²⁹. Binary variables are not exceptions²³⁰.

We also standardized predictor variables (gene expression levels) and response variables (drug sensitivity measures) for MERGE and all other methods compared to MERGE in our

experiments: ElasticNet, multi-task learning, Pearson's p -value, Spearman p -value, and Bayesian multi-task multiple kernel learning (MKL).

8. Measuring the significance of association between a gene expression level and a drug sensitivity measure

We computed the p -value of association between each gene and each drug using t statistics and associated p -values, measuring the correlation between the gene and drug in a univariate linear regression model. Then, we applied FDR correction for multiple hypotheses for all pairs of genes and drugs, using the Benjamini & Hochberg (1995) method. We considered the (genome-wide) FDR corrected p -value < 0.1 to be significant throughout the paper. When we measured the consistency of gene-drug associations in CCLE data, we used an uncorrected p -value of 0.1 as a cutoff to indicate that a gene-drug association discovered in patient data was replicated in validation data. We did not apply the FDR correction method in validation data because the FDR correction process is designed to maintain a certain FDR level when making a large number of hypotheses, and we do not make hypotheses during validation.

9. Clinical information on the validation data from additional 12 AML patient samples

The additional 12 patients we used for validation (**Figure 26c**) were enrolled in an open clinical trial (High Throughput Drug Sensitivity Assay and Genomics-Guided Treatment of Patients with Relapsed or Refractory Acute Leukemia NCT02551718). For those 12 samples, we measured gene expression levels using the newer RNA-seq technology while we had the microarray gene

expression from the initial 30 patient samples (details in APPENDIX B). The median age was 58, and 5 had antecedent hematologic disorders. According to European LeukemiaNet criteria²²⁶, 7 samples were in favorable risk group, 3 in intermediate-1, 1 in intermediate-2, and 8 were in the adverse risk cytogenetics group. Four samples were primary refractory, and 6 samples relapsed after allogeneic transplant. The average number of prior regimens used was 5.

10. Details on the implementations of the methods compared to MERGE

For ElasticNet, we used the R package *glmnet*²³² available on CRAN. For multi-task learning¹¹⁸, the MATLAB implementation was available from Pong, Tseng, Ji, & Ye (2010) (on <http://www.mypolyuweb.hk/~tkpong/>). For Bayesian multi-task MKL, we used the R code provided as supplementary to Costello et al. (2014).

As mentioned earlier, we standardized both the predictor and response variables before applying each method, as we did for MERGE. In addition, as with MERGE, we chose the tuning parameters for each method in the comparison using LOOCV. These include the mixing parameter α and regularization parameter λ for ElasticNet, the regularization parameter μ for multi-task learning, and α and β , which are, respectively, the shape and rate parameters of the Gamma priors, for Bayesian multi-task MKL.

Comparing MERGE to the other methods in **Figure 26** and **Figure 27c** requires ordering of the gene-drug associations by the other methods, as with MERGE. For Pearson's P-value or Spearman P-value methods, we sorted the gene-drug pairs based on decreasing significance of correlation (Pearson's or Spearman correlation, depending on the method). ElasticNet and multi-task learning are regression methods, and they learn coefficients for the predictor-response variable

pairs. For ElasticNet, we ran the method for each of the 53 drugs separately and concatenated the resulting gene coefficients to get a matrix of size $\#genes \times \#drugs$. The multi-task learning method, on the other hand, learns the coefficients for all drugs jointly; thus, a single run of this algorithm gives us a coefficient matrix of size $\#genes \times \#drugs$. After determining a coefficient matrix, we sorted gene-drug pairs, for each of ElasticNet and multi-task learning methods, based on decreasing absolute value of the coefficients (i.e., decreasing strength of gene-drug associations). After sorting gene-drug pairs for each method in comparison to MERGE, we incremented the number of considered top gene-drug pairs by 53 (the number of drugs per gene in our application) for each increment of 1 in the x-axis in **Figure 26** and **Figure 27c**.

To generate heat maps for the alternative methods (**Appendix Figure 18**), we computed gene scores (as with MERGE scores) based on each method. For this purpose, we first sorted the gene-drug pairs in *increasing* order of their importance for each method (i.e., in increasing significance of correlation for Pearson's and Spearman P-value methods, and in increasing absolute value of the coefficients for the ElasticNet and multi-task learning methods). For each gene G , the score was computed as the sum over all indices corresponding to that gene's associations in the sorted weight matrix. If a gene had many drug associations with high absolute weights, then those associations would be positioned towards the end of the sorted weight matrix, which would increase the score of G .

For the Bayesian multi-task MKL method, we used only gene-set views and the discretized view¹²⁰ computed on gene expression data since we had no epigenomic or proteomic profiling data from the 30 AML patient samples.

11. Computing the significance of the enrichment of the drug classes based on mechanism of action within drug clusters showing similar patient responses

To verify that drugs with similar mechanisms of action indeed showed similar response across patients, we used the dendrogram (**Figure 27a**) generated by agglomerative hierarchical clustering. For hierarchical clustering, we used Euclidean distance as the dissimilarity metric, and we used average linkage as the clustering method to cluster the drugs based on the response of 30 patient samples to those drugs.

In the dendrogram, for each internal node that had 3 to 8 drugs in its leaves, we first retrieved a unique set of drug classes that contained at least one drug from the node's downstream drugs. Then, we checked the enrichment of each of these drug classes in the set of the downstream drugs of that node using Fisher's exact test. Each of the enrichment p -values we present in **Figure 27a** is FDR-corrected for the number of drug classes tested for the corresponding node using the Benjamini & Hochberg (1995) method.

12. Computing each gene's drug class specificity (DCS) measure

First, for each pair of [gene A, class B], we computed the significance of the overlap between the drugs in B and the drugs gene A's expression level is significantly associated with, measured by Fisher's exact test p -value. Then, for each gene A, we computed a specificity measure that we named *drug class specificity* (DCS) score, by combining the p -values across all classes, as follows:

$$DCS_A = \frac{1}{\text{total \# of drug classes}} \sum_{B \in \{\text{drug classes}\}} -\log_{10}[\text{Fisher's exact test } p\text{-value}_{AB}].$$

Mathematically, Fisher's exact test is based on the following numbers for each pair of [gene A, class B]: (1) the total number of drugs with which A is significantly associated, (2) the number of drugs in class B with which A is significantly associated, (3) the number of drugs in class B, and (4) the total number of drugs (53 in our experiments). The p -value measures the significance of (2) based on the hypergeometric distribution. Therefore, it measures the specificity of the association between gene A and class B.

13. Summary description of the MERGE probabilistic model

MERGE is a probabilistic, model-based approach that uses MAP to estimate parameters. A probabilistic, model-based approach provides an expressive model to describe relationships among variables. Moreover, the probabilistic relationships can be read from the learned model and thus, often directly lead to a comprehensive biological interpretation. In MERGE, the prior variance of the gene-drug weights is interpreted as the corresponding gene's biomarker potential.

We modeled W and V in a Bayesian sense (i.e., they are parameters for which belief in their values is modeled), where W is modeled as a Gaussian random variable whose variance is modeled based on V , and V is modeled as a uniform random variable (i.e., $P(V)$ is constant).

Instead of using a traditional Bayesian approach (i.e., estimating $P(\theta | D)$, the full posterior distribution, over parameters θ where D represents the data), we employed *maximum a posteriori probability (MAP) estimation*, i.e., obtaining a point estimate of θ that maximizes $P(\theta | D)$. In other words, we optimized $\log P(X, Y, W, V)$ with respect to W and V . MAP estimation has two advantages over the traditional Bayesian approach. First, estimating specific parameter values makes biological interpretation straightforward. For example, specific parameter values of

v_k coefficients enable an efficient computation of MERGE scores and straightforward interpretation of how driver features affect the MERGE score (i.e., biomarker potential). Second, MAP estimation allows a much simpler parameter learning procedure, especially when $P(\theta | D)$ does not have a closed-form solution. Penalized linear regression models, such as LASSO (or Ridge), also employ MAP estimation for a probabilistic model (specifically, linear regression model) with a Laplacian (or Gaussian) prior for $P(\theta)$, where the parameter θ means the W values. MERGE extends the penalized linear regression models by explicitly modeling the variance of the W parameters based on V and the driver features of genes.

14. Initialization of MERGE parameters and identifiability of the MERGE model

In our application of MERGE, we initialized all five v_k values to zero, which provided an unbiased starting point (i.e., giving an equal prior variance to the weight values of all genes).

As noted previously, the MERGE objective function represented in Eq (2) is non-convex; thus, different initializations of the V vector (of v_k values, each for a different driver feature) may lead to different learned parameters, i.e., different local minima of the objective function. In practice, however, depending on the objective function and the input data, it is possible that a roughly unique solution can be empirically identified. One way to check is to try multiple runs with different parameter initializations and see whether these runs converge to roughly the same point.

We observed that when we tried different initializations of V (and correspondingly W), the learned parameters were very similar to each other. Below, we describe our results on the consistency between the zero and random initializations.

We performed 20 different runs of MERGE where we initialized v_k values so they could be generated randomly from a standard normal distribution. Then, we compared the resulting training objective function values and the MERGE scores from these 20 runs to those from the MERGE run used in our paper (i.e., where we initialized all five v_k values to zero). We performed this experiment with the same hyperparameter value selected by LOOCV and used for the final model ($\lambda = 20$) (**Appendix Figure 26a**) as well as with a different λ value ($\lambda = 50$) (**Appendix Figure 26b**). As shown in the top of **Appendix Figure 26a**, for $\lambda = 20$, only 4 of 20 runs resulted in a smaller objective function (i.e., better local optima) compared to our initialization with zero v_k values, and the difference was very small. All 20 runs with random initializations resulted in almost the same v_k values shown in **Figure 25**, and exactly the same gene rankings as those from the zero initialization of MERGE (**Appendix Figure 26a**, bottom). This indicates that different random initializations converged to roughly the same point. A different value of λ ($\lambda = 50$) showed consistent patterns (**Appendix Figure 26b**).

15. Cross-validation experiments and selection of the hyperparameter λ

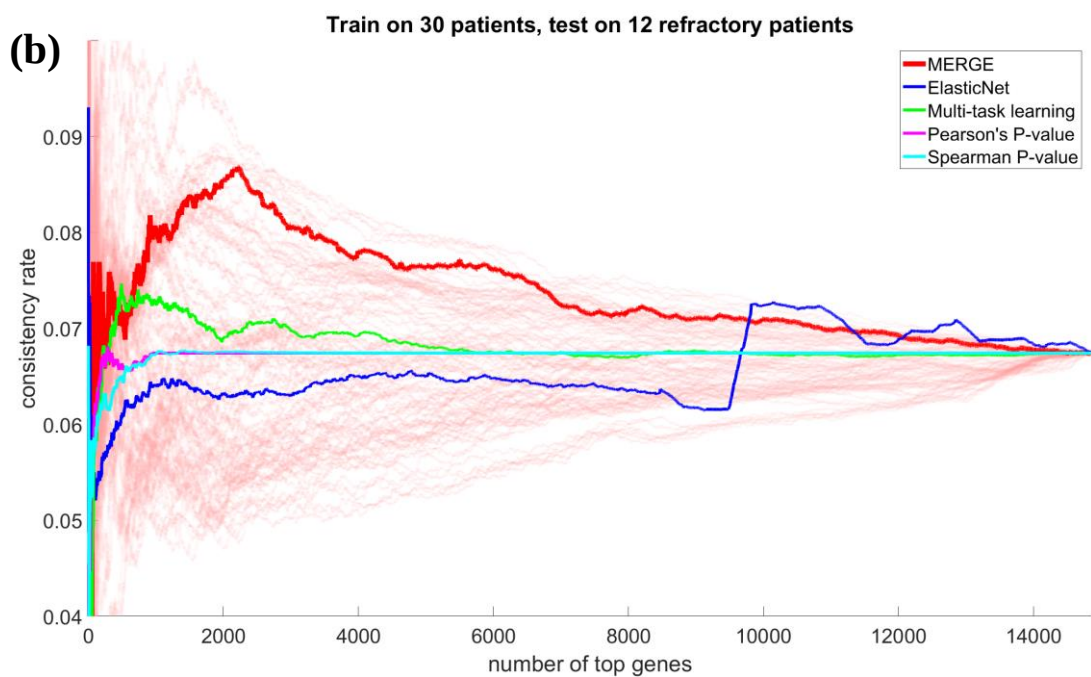
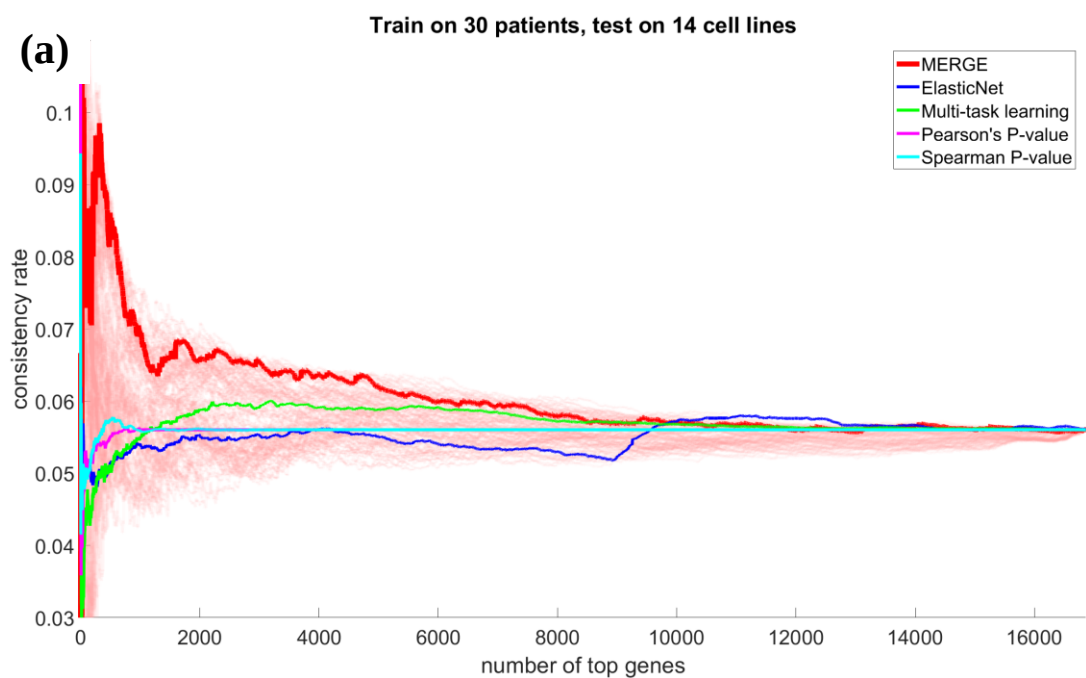
In each of our experiments, the regularization parameter λ was determined via cross-validation, a standard way to choose tuning parameters. Since our sample size was low, we employed LOOCV since it provides the maximum number of training samples for each fold. Cross-validation experiments were performed when measuring the prediction accuracy on left-out test data (**Figure 28**) besides when selecting the value of the hyperparameter (sparsity tuning parameter λ).

We performed LOOCV within the training data to choose one λ value from the 19 λ values in the wide range of [1,100]. We used LOOCV to choose the values of the tuning parameters for

other methods as well: elastic net regression, the multitask learning method, and the DREAM challenge winner Bayesian multi-task MKL method.

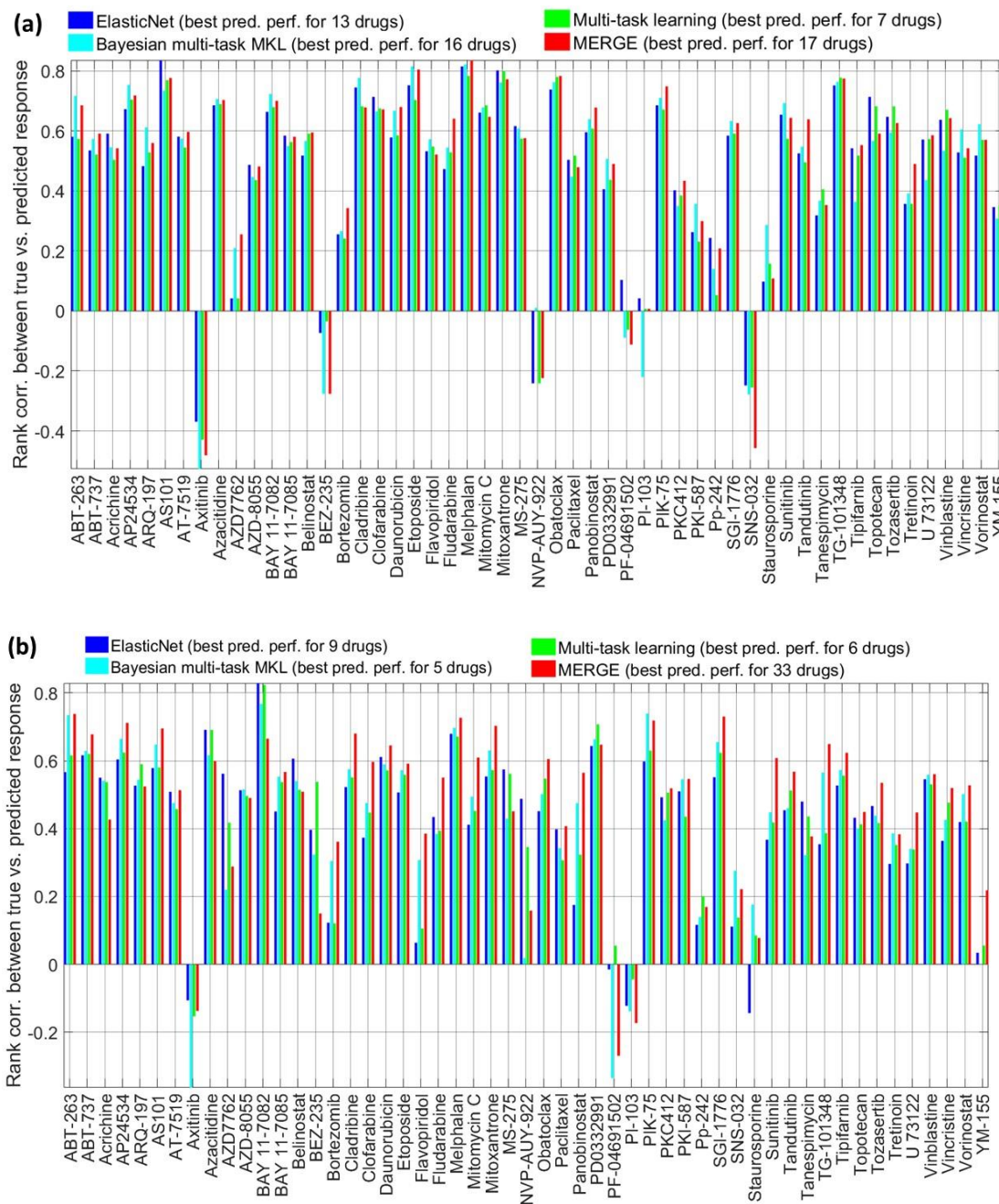
We used cross-validation tests in three settings to select the λ value: (1) *Training the model using all 30 samples (Figure 26)*. We first performed LOOCV on 30 samples to choose the λ value using mean squared error (MSE), and then trained the model using all 30 samples with the chosen λ value. (2) *Measuring the prediction accuracy by training the model using one 12-sample batch and testing on the other 12-sample batch (Figure 28a)*. We performed LOOCV tests within 12 samples in each batch to choose the value of λ . We used rank correlation between the actual and predicted responses as the evaluation metric. Then we trained the model using the selected λ value in each batch and tested the prediction performance on the other batch. (3) *Measuring the prediction accuracy via LOOCV (Figure 28b)*. In each fold for which we omitted one sample and used the remaining $(n-1)$ samples to train the model, we performed the “inner loop” LOOCV using those $(n-1)$ samples to choose the λ value and trained the model using $(n-1)$ samples with the chosen λ value. We used rank correlation between the actual and predicted responses as the evaluation metric.

16. APPENDIX B Figures



Appendix Figure 16. Feature consistency achieved by 100 MERGE runs (light red) each of which uses different permutations of the training samples is compared to the consistency achieved by the actual MERGE run that uses the original training data as well as to the alternative four methods as in **Figure 26**. We discovered gene-drug associations within the data

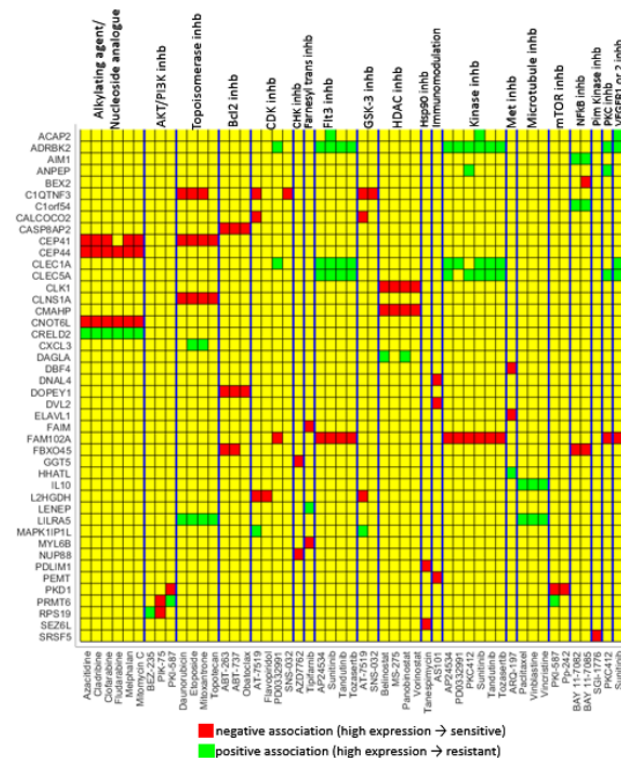
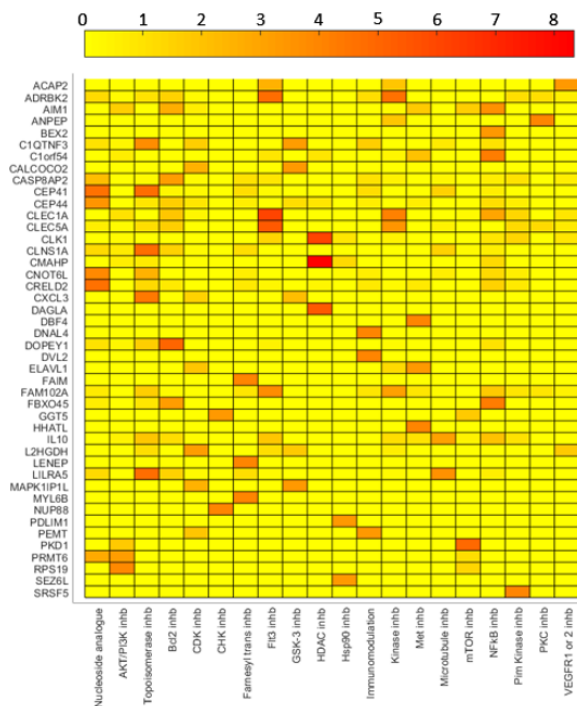
from 30 patient samples, and tested on **(a)** the 14 cell line samples, and **(b)** additional 12 refractory patient samples.



Appendix Figure 17. Detailed view of the prediction performance of MERGE and the three alternative methods in each of the two settings shown in **Figure 28**. The performance (y-axis) is measured by the rank-based (Spearman) correlation between the predicted response and actual

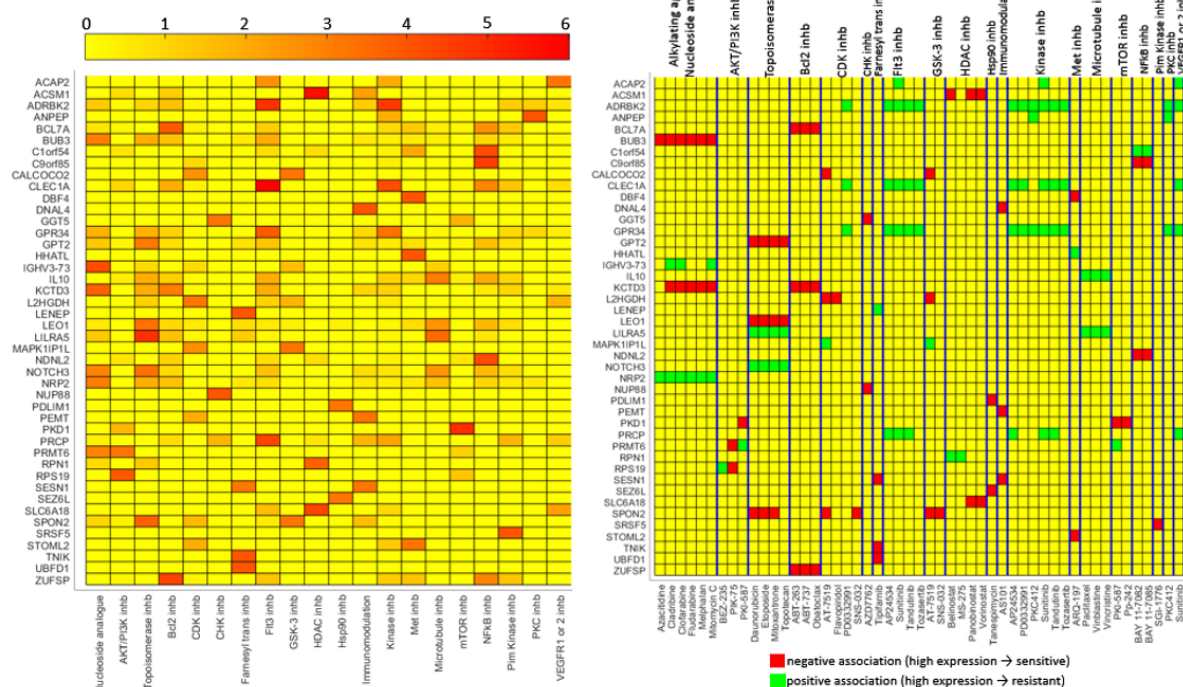
drug response in **(a)** split test, and **(b)** LOOCV. The numbers inside the parentheses in the figure's legend (bottom right) represent the number of drugs for which the corresponding method achieved the best prediction performance.

(a) Genes identified by ElasticNet

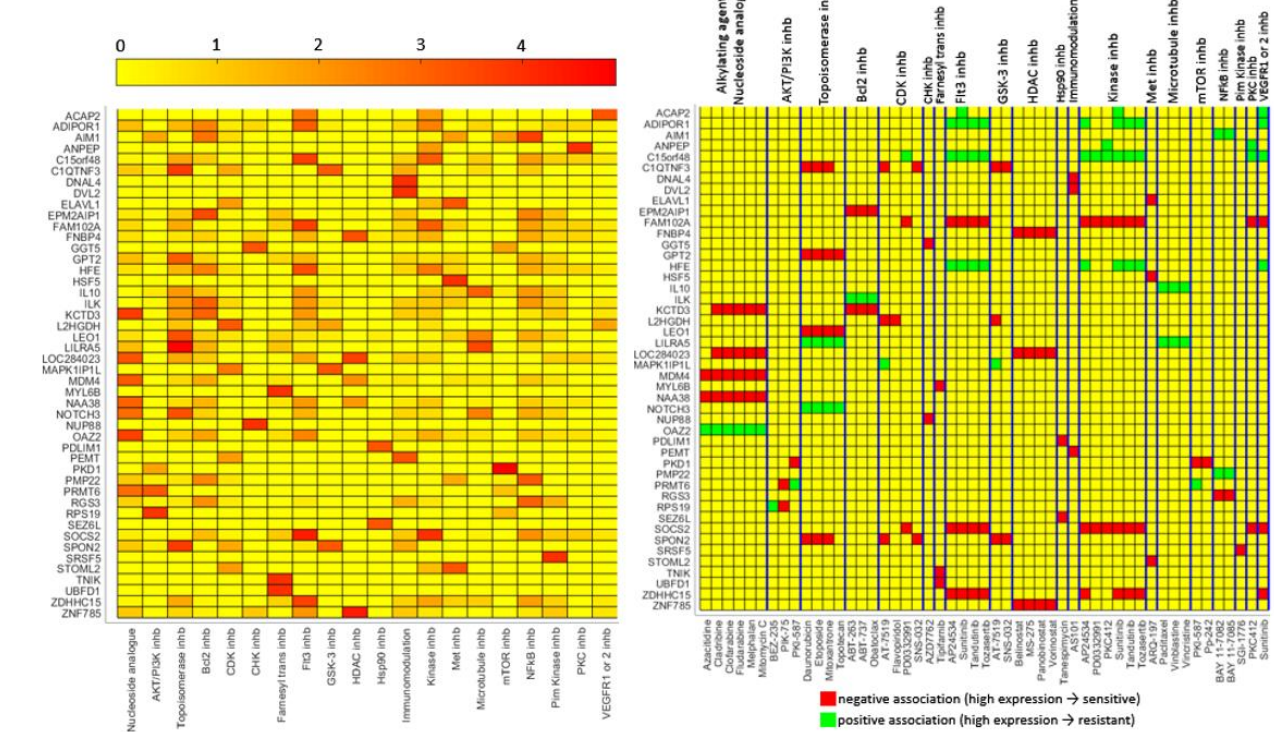


negative association (high expression → sensitive)
positive association (high expression → resistant)

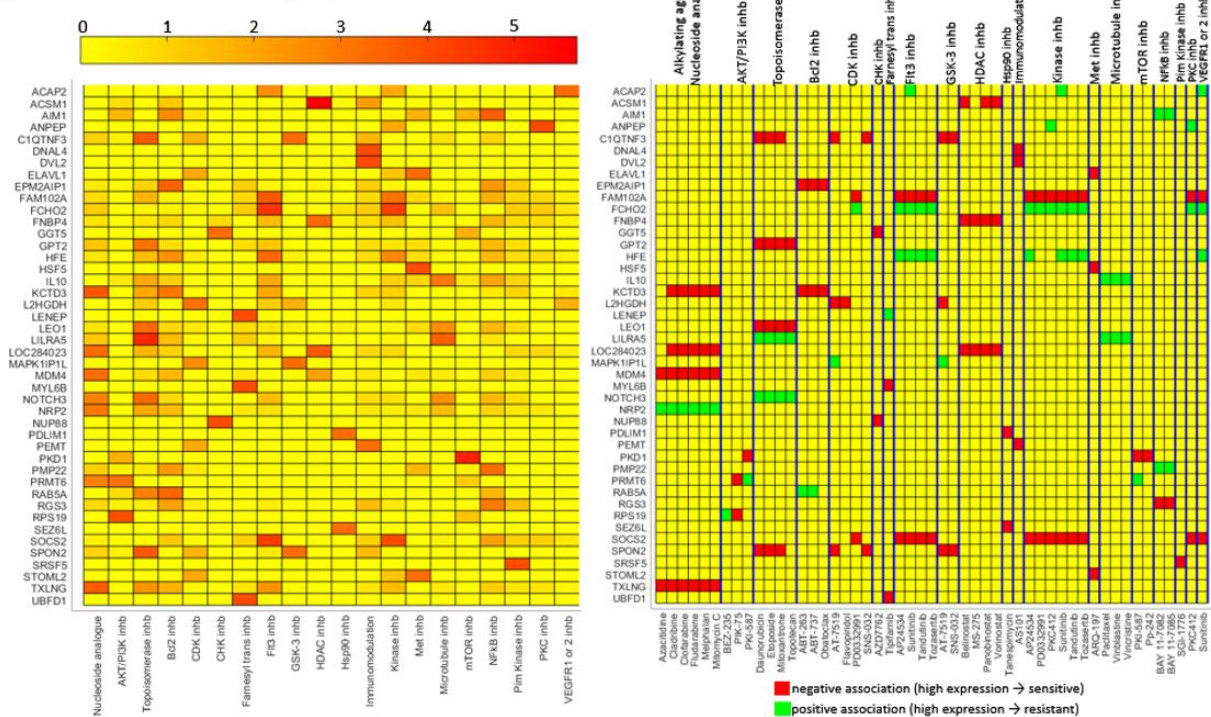
(b) Genes identified by multi-task learning



(c) Genes identified by Pearson's P-value

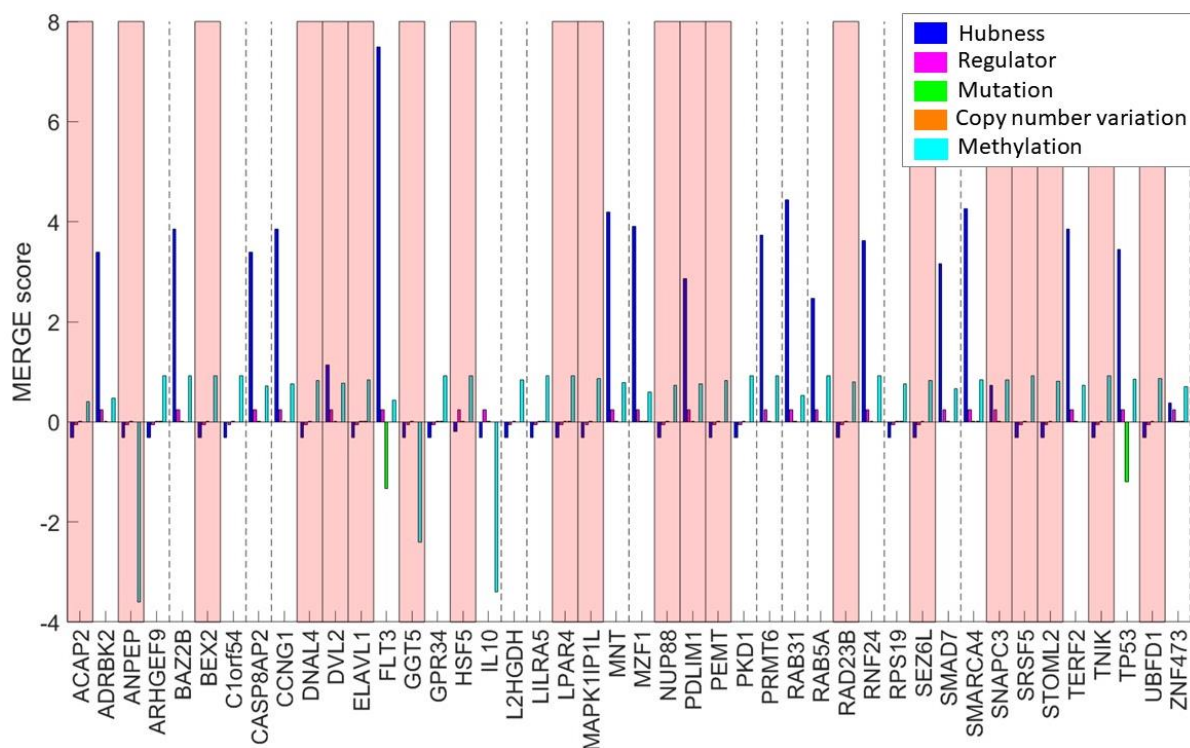


(d) Genes identified by Spearman P-value

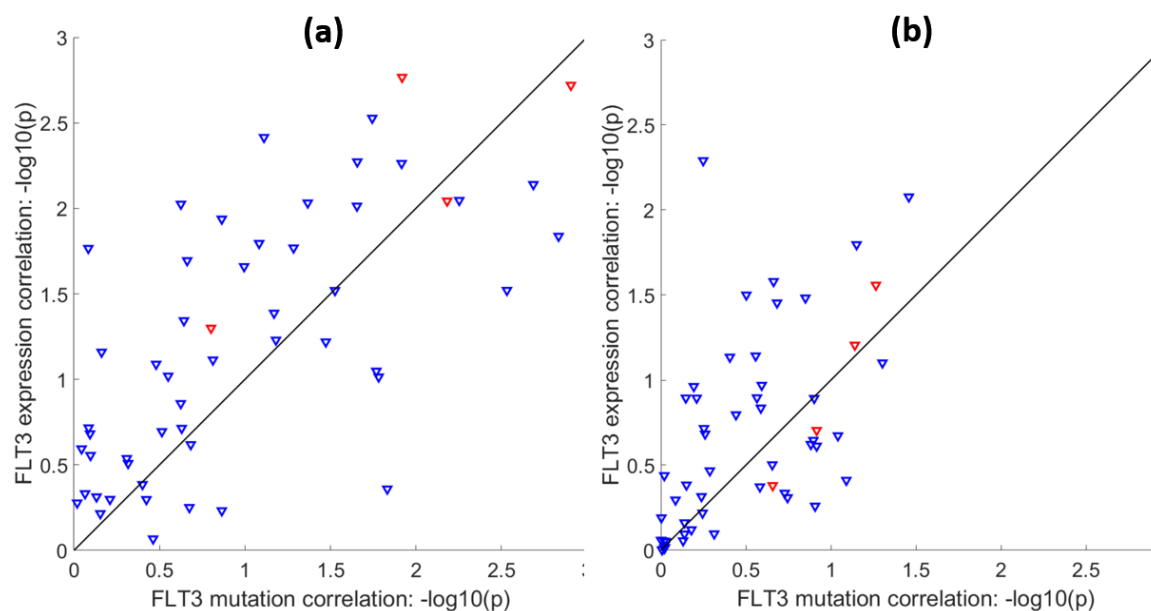


Appendix Figure 18. Heat maps that correspond to **Figure 29** for the four alternative methods (shown in **Figure 26** and **Figure 27c**). (a) ElasticNet, (b) Multi-task learning, (c)

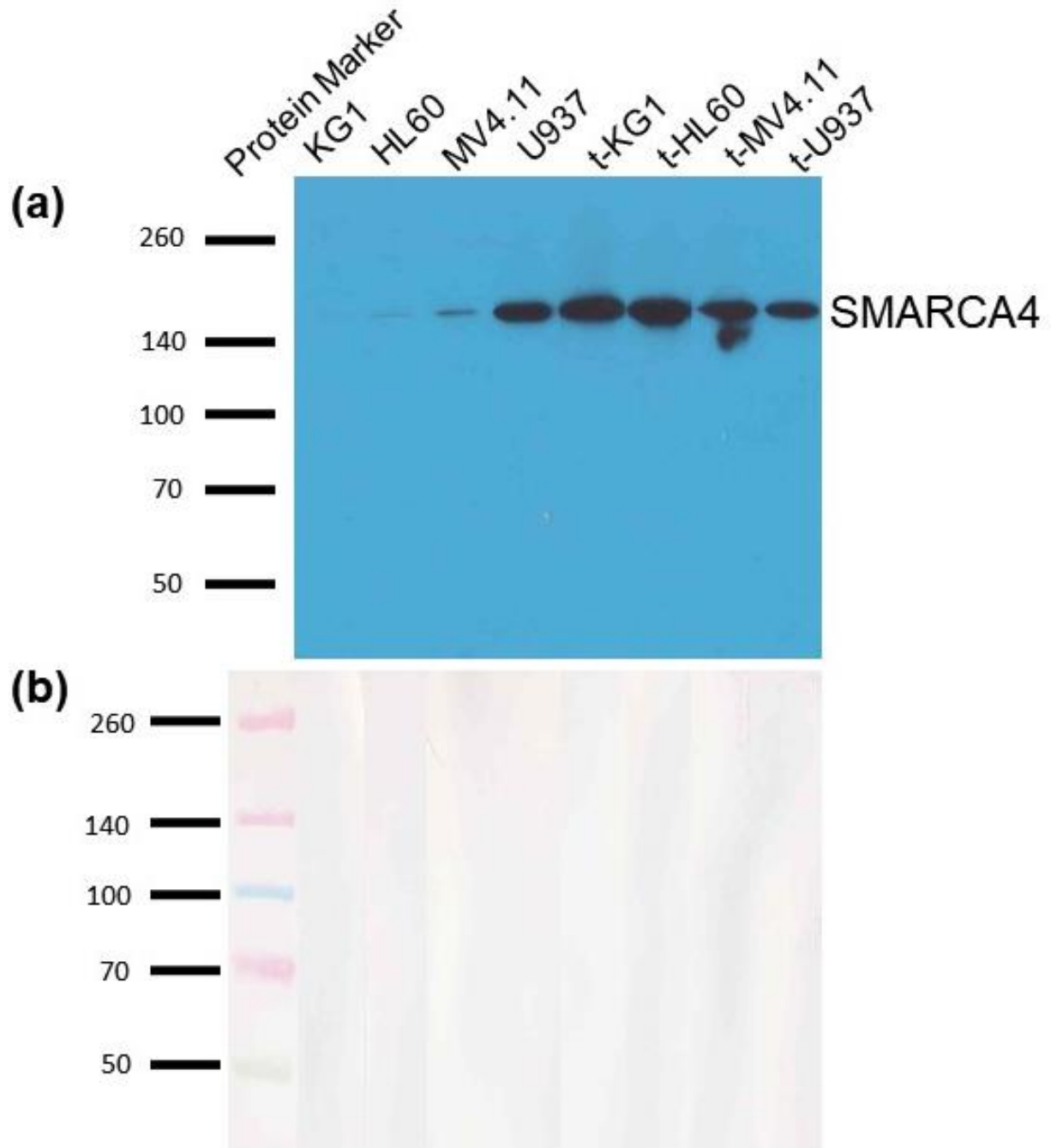
Pearson's P -value, and (d) Spearman P -value. In (a)-(d), the heat map on the left shows the level of specificity of each gene to each drug class, measured by $-\log_{10}[\text{Fisher's exact test } p\text{-value}]$ for the top 3 genes selected using the corresponding method; the right heat map shows the gene-drug association for the genes and drug classes shown on the left heat map.



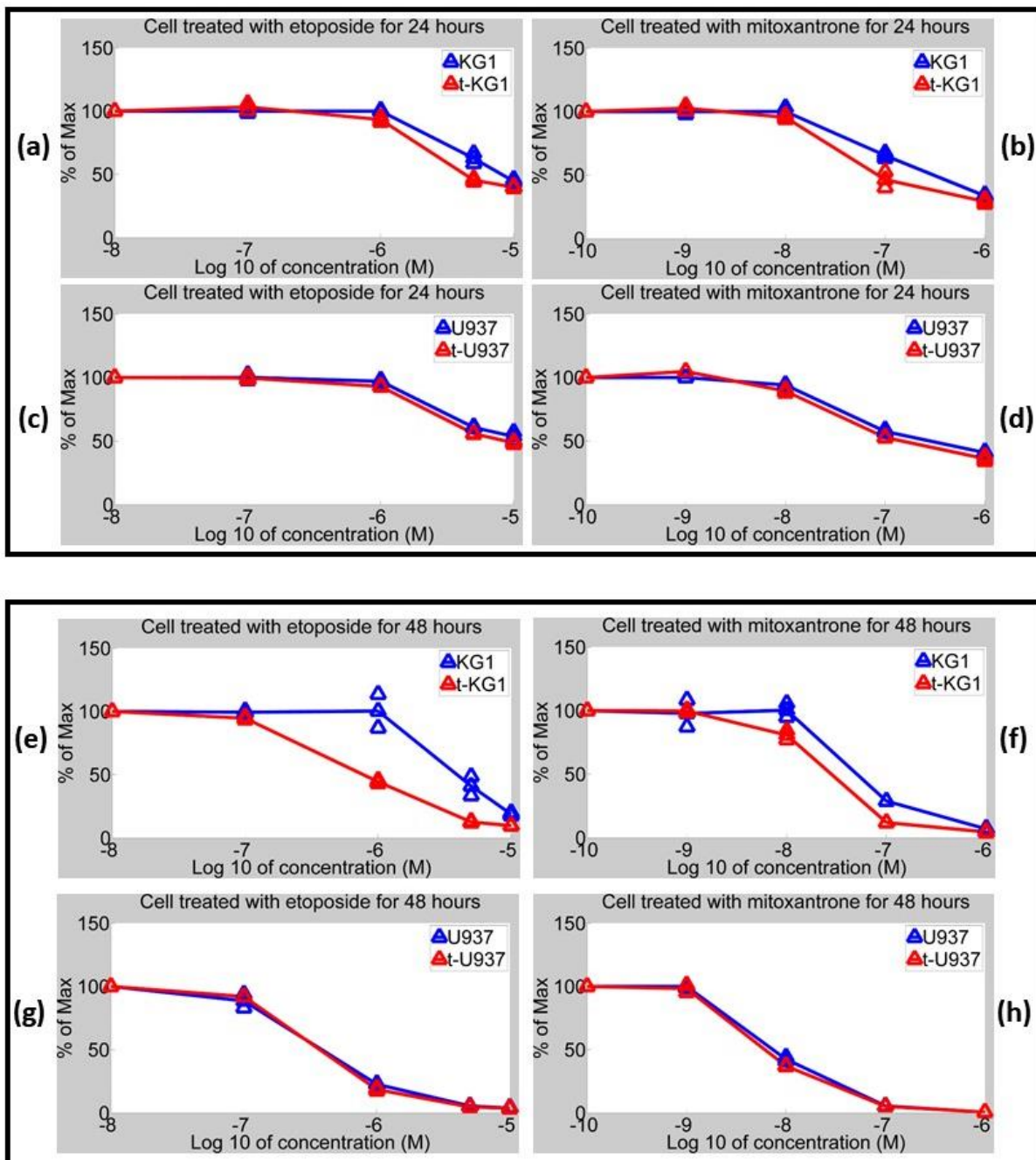
Appendix Figure 19. For each of the genes in **Figure 29a** and **b**, the amount of contribution of each driver feature on the MERGE score, measured by $[\text{driver feature weight} \times \text{driver feature value}]$ for a driver feature. The genes highlighted in red are associated with a single drug (i.e., the genes corresponding to the rows in **Figure 29b** that have a single unique red or green square). We note that the genes with a high hubness contribution (blue bar > 3) tend to have significant associations with more than one drug (Fisher's exact test $p\text{-value} = 1.7 \times 10^{-5}$).



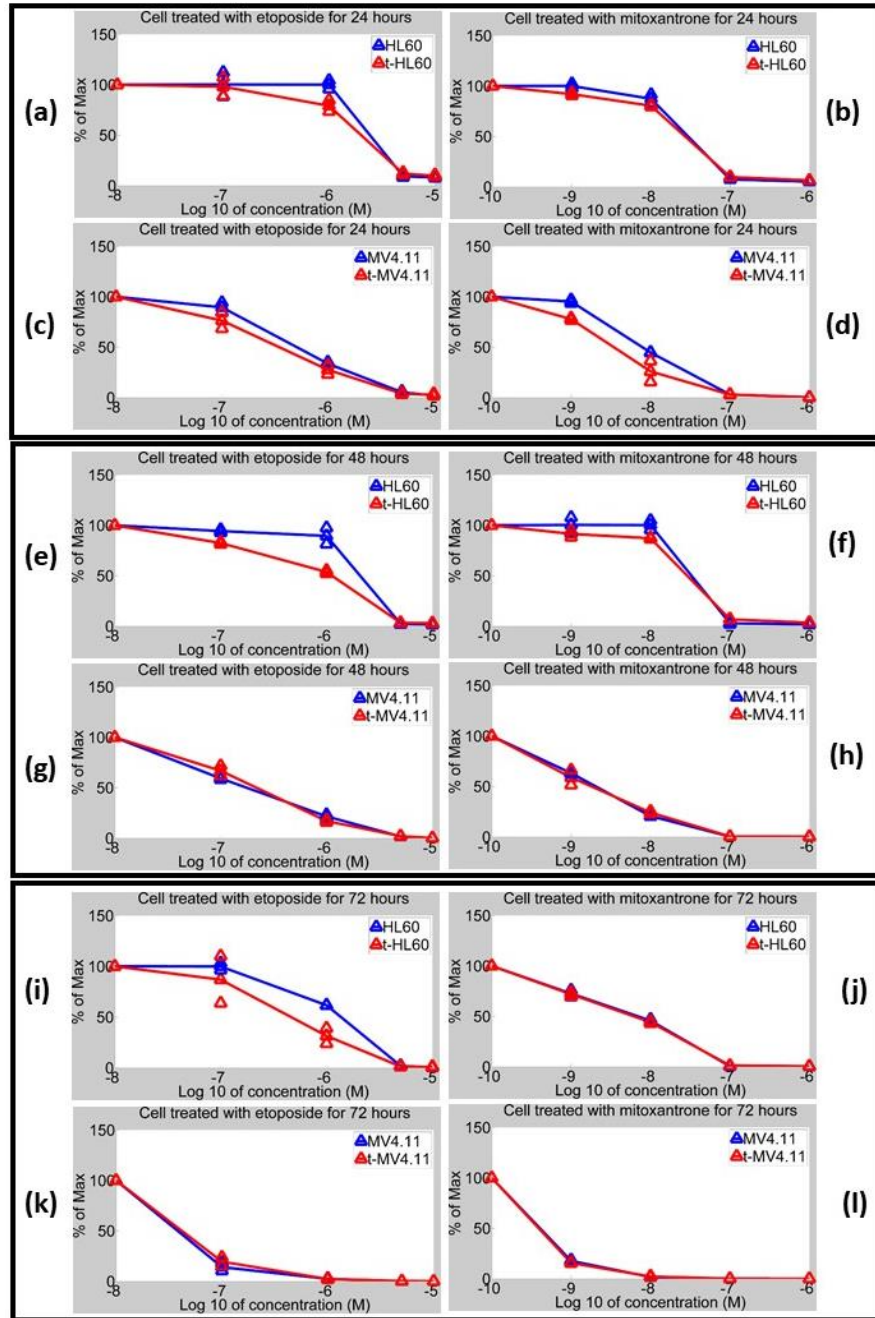
Appendix Figure 20. Comparison of *FLT3* expression level to *FLT3* mutation status in terms of significance of correlation with drug response. For each of the 53 drugs, the $-\log_{10}p$ -value representing the degrees of association between the drug response and *FLT3* expression level (y-axis) vs. *FLT3* mutation status (x-axis) is shown for **(a)** patients and **(b)** cell lines. The drugs highlighted in red on each scatter plot correspond to the four *FLT3* inhibitors (AP24534, Sunitinib, Tandutinib, Tozasertib) in the set of 53 drugs that we studied.



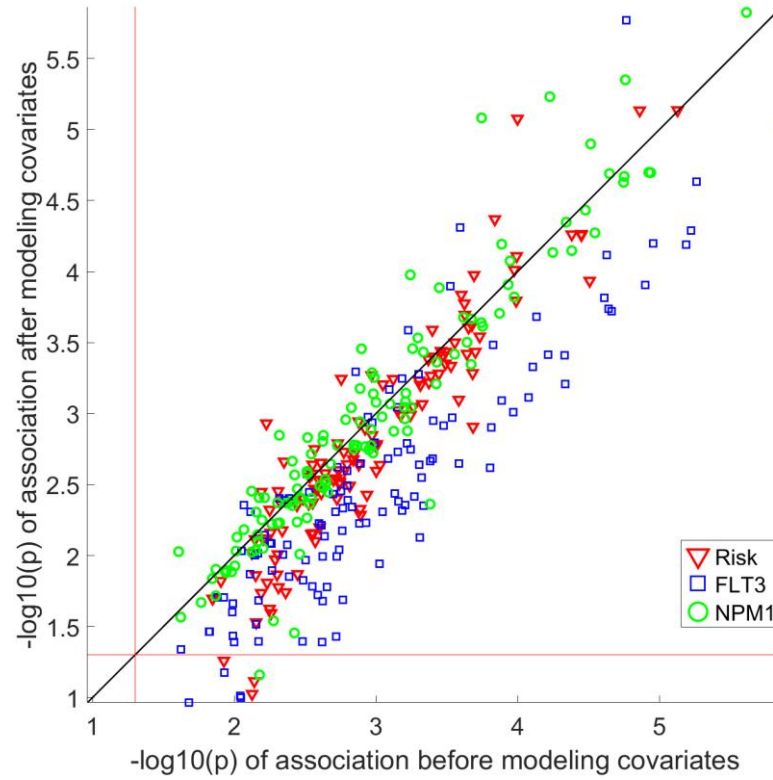
Appendix Figure 21. Uncropped image of the Western blot of control and SMARCA4 plasmid transfected AML cell lines: KG1, U937, HL60, and MV4.11. (a) Film. (b) Nitrocellulose membrane.



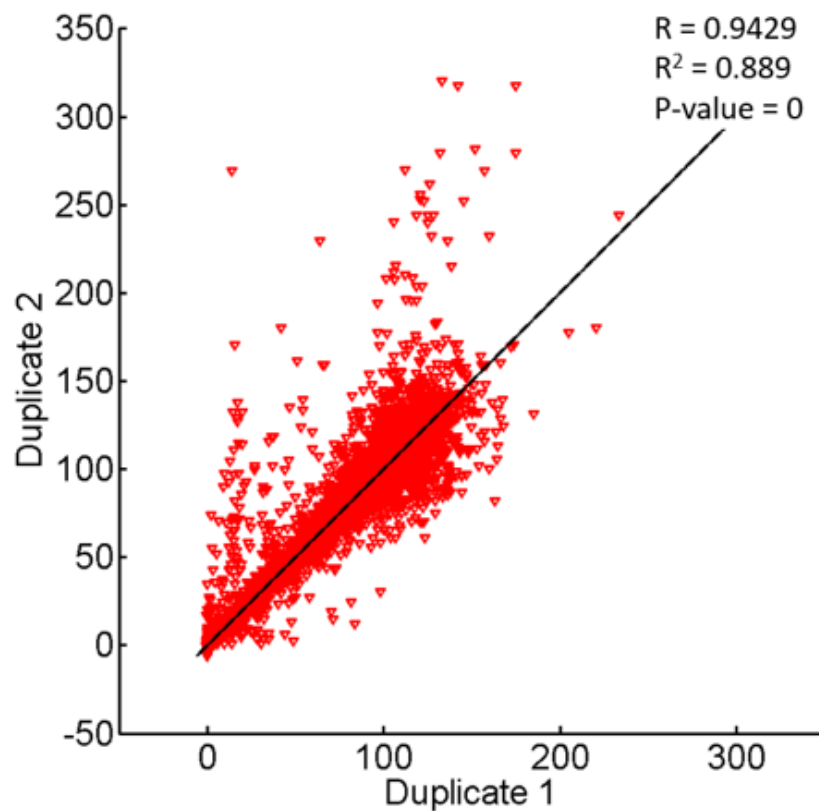
Appendix Figure 22. The dose-response curves for cell lines treated with etoposide (left panel) and mitoxantrone (right panel) after 24 hours (a)-(d) and 48 hours (e)-(h). Each plot compares KG1 with transfected KG1 in (a)-(b) and (e)-(f), and U937 and transfected U937 in (c)-(d) and (g)-(h). Triangular marks indicate individual data points in duplicates and the average among them. Lines connect averages of duplicates in each concentration measured.



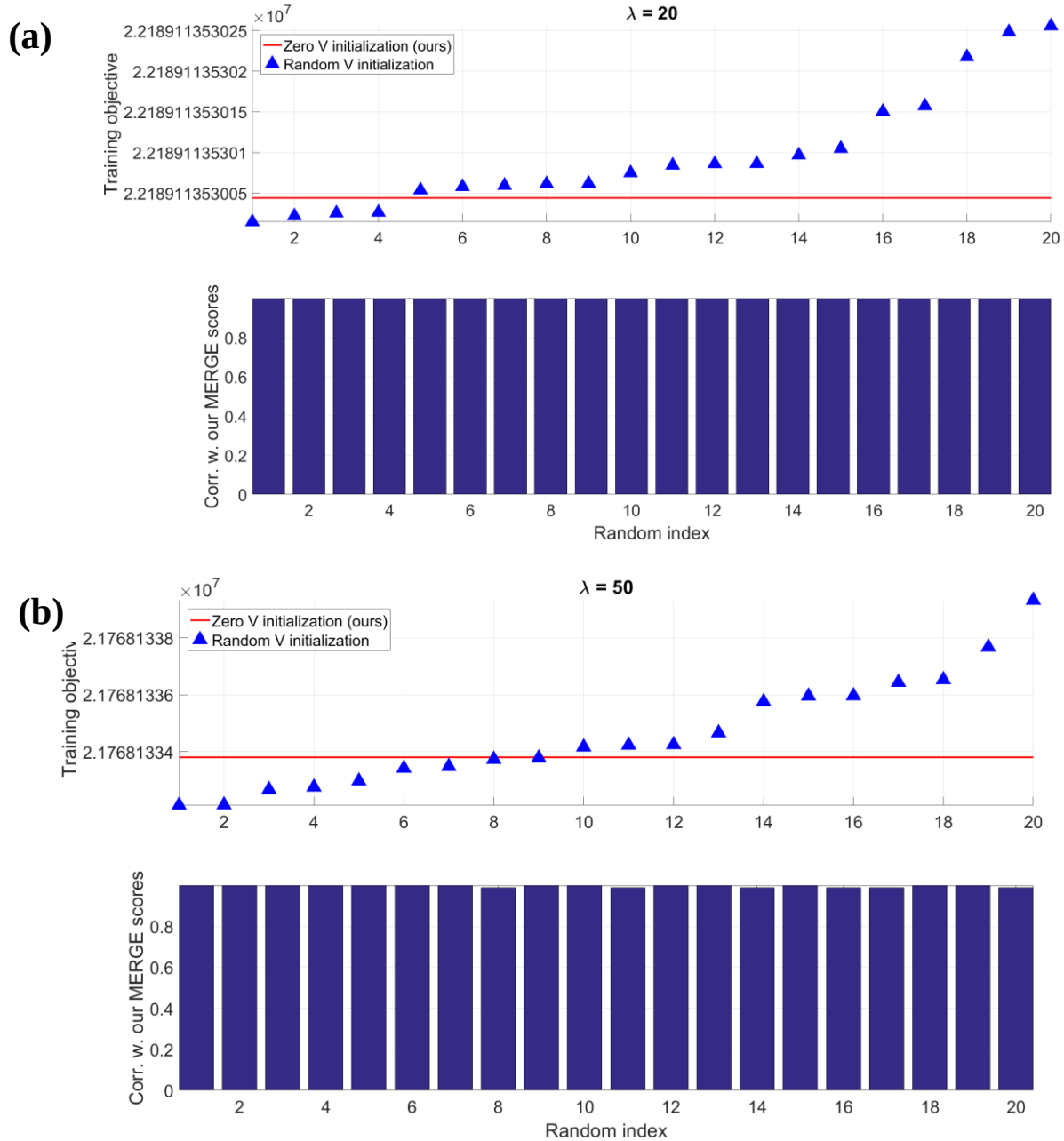
Appendix Figure 23. The dose-response curves for cell lines treated with etoposide (left panel) and mitoxantrone (right panel) after 24 hours (a)-(d), 48 hours (e)-(h), and 72 hours (i)-(l). Each plot compares HL60 with transfected HL60 in (a)-(b), (e)-(f), and (i)-(j) and MV4.11 and transfected MV4.11 in (c)-(d), (g)-(h), and (k)-(l). Triangular marks indicate individual data points in duplicates and the average over them. Lines connect averages over duplicates in each concentration measured.



Appendix Figure 24. Comparison of the significance of the gene-drug associations ($-\log_{10} p\text{-value}$) with (y-axis) vs. without (x-axis) adding the available risk group/cytogenetic features as covariates. Red corresponds to cytogenetic risk covariate, blue corresponds to *FLT3* mutation status, and green corresponds to *NPM1* mutation status. The red vertical and horizontal lines correspond to $-\log_{10} 0.05$. Each dot corresponds to one of the 119 unique gene-drug associations shown as red or green on the heat map in **Figure 29b**.



Appendix Figure 25. Scatter plot demonstrating the reproducibility of high-throughput drug sensitivity assay. Each dot corresponds to a combination of a drug, a sample, and a specific concentration. The correlation coefficient, R^2 (squared correlation coefficient), and p -value are computed based on Pearson's correlation.



Appendix Figure 26. For 20 random initializations of the learning parameters (x -axis), the training objective function values (top) and the Spearman correlation of the learned MERGE scores from each random initialization with those from zero initialization (used for our paper) (bottom). **(a)** We used the same λ value as the final MERGE model (i.e., trained based on 30 samples) that was selected by LOOCV ($\lambda = 20$). **(b)** We used an additional value for the hyperparameter $\lambda = 50$.

17. APPENDIX B Tables

Appendix Table 13. Some properties for each of the 53 drugs we used in application of our MERGE algorithm. *p*-values listed in the 4th column are FDR-corrected for the number of drugs.

Since some drugs are in multiple classes, the last column shows the number of genes significantly associated with the drug for each class the drug is in. This column is empty for the 3 drugs (Acrichine, U 73122 and YM-155) assigned only to the “Other” class, which contains the drugs that do not belong to any of the 24 drug mechanism classes based on their mechanisms of action. The footnotes (a)-(k) present details of the drugs that have a “1” in the 3rd column or have a “-” in the last column. We note that for the drugs ABT-263 and ABT-737, the “1” in the 3rd column reports a successor drug that is effective in AML.

Drug Name	Reported Clinical Efficacy in AML Treatment	Whether Any Undergoing Clinical Evaluations in AML	Significance of Association of AUC with CR	#Significantly Associated Genes	#Significantly Associated Genes after Cell Line Consistency Filter	#Significantly Associated Genes after Cell Line Consistency and Drug Class Specificity Filters
ABT-263	0	1 ^(a)	0.0023	1825	142	[Bcl2 inh: 10]
ABT-737	0	1 ^(b)	0.0039	1727	122	[Bcl2 inh: 11]
Acrichine	0	0	0.0267	1169	119	- ⁽ⁱ⁾
AP24534 (=Ponatinib)	0	0	0.0153	1911	93	[Flt3 inh: 11]; [Kinase inh: 5]
ARQ-197 (=Tivantinib)	0	0	0.003	1103	73	[Met inh: 6]
AS101	0	0	0.0086	2011	123	[Immunomodulation: 12]
AT-7519	0	0	0.4349	653	61	[CDK inh: 1]; [GSK-3 inh: 14]
Axitinib	0	0	0.9724	35	2	[Kinase inh: 0]; [VEGFR1 or 2 inh: 0]
Azacitidine	1	0	0.0379	1477	76	[Nucleoside analogue: 22]
AZD7762	0	0	0.8083	171	15	[CHK inh: 5]
AZD-8055	0	0	0.0039	649	32	[mTOR inh: 1]
BAY 11-7082	0	0	0.003	2568	141	[NFkB inh: 25]
BAY 11-7085	0	0	0.0021	1554	66	[NFkB inh: 11]
Belinostat	0	1 ^(c)	0.0159	696	42	[HDAC inh: 16]
BEZ-235	0	0	0.2664	138	11	[AKT/PI3K inh: 4]

Drug Name	Reported Clinical Efficacy in AML Treatment	Whether Any Undergoing Clinical Evaluations in AML	Significance of Association of AUC with CR	#Significantly Associated Genes	#Significantly Associated Genes after Cell Line Consistency Filter	#Significantly Associated Genes after Cell Line Consistency and Drug Class Specificity Filters
Bortezomib	1	0	0.0094	263	4	[Proteasome inh: 0]
Cladribine	1	0	0.0497	1420	169	[Nucleoside analogue: 69]
Clofarabine	1	0	0.0267	908	93	[Nucleoside analogue: 54]
Daunorubicin	1	0	0.0224	1833	73	[Topoisomerase inh: 12]
Etoposide	1	0	0.0317	1699	78	[Topoisomerase inh: 12]
Flavopiridol	1	0	0.6414	346	13	[CDK inh: 2]
Fludarabine	1	0	0.0086	1169	95	[Nucleoside analogue: 42]
Melphalan	0	1 ^(d)	0.0098	2323	151	[Nucleoside analogue: 45]
Mitomycin C	0	0	0.073	1428	80	[Nucleoside analogue: 33]
Mitoxantrone	1	0	0.0457	2106	90	[Topoisomerase inh: 14]
MS-275	0	0	0.1161	679	26	[HDAC inh: 7]
NVP-AUY-922	0	0	0.5051	121	4	[Hsp90 inh: 2]
Obatoclax	0	1 ^(e)	0.0276	1177	48	[Bcl2 inh: 6]
Paclitaxel	0	0	0.0684	590	27	[Microtubule inh: 4]
Panobinostat	1	0	0.0159	680	41	[HDAC inh: 18]
PD0332991 (=Palbociclib)	0	1 ^(f)	0.0023	1739	194	[CDK inh: 0]; [Kinase inh: 3]
PF-04691502	0	0	0.9764	63	1	[AKT/PI3K inh: 0]; [mTOR inh: 0]
PI-103	0	0	0.4311	47	0	[AKT/PI3K inh: 0]; [mTOR inh: 0]
PIK-75	0	0	0.0171	2151	103	[AKT/PI3K inh: 1]
PKC412 (=Midostaurin)	1	0	0.0344	794	38	[Kinase inh: 4]; [PKC inh: 1]
PKI-587 (=Gedatolisib)	0	1 ^(g)	0.0385	677	86	[AKT/PI3K inh: 1]; [mTOR inh: 1]
Pp-242	0	0	0.868	179	11	[mTOR inh: 0]
SGI-1776	0	0	0.0039	1758	95	[Pim Kinase inh: 2]
SNS-032	0	0	0.3771	178	15	[CDK inh: 1]; [GSK-3 inh: 4]

Drug Name	Reported Clinical Efficacy in AML Treatment	Whether Any Undergoing Clinical Evaluations in AML	Significance of Association of AUC with CR	#Significantly Associated Genes	#Significantly Associated Genes after Cell Line Consistency Filter	#Significantly Associated Genes after Cell Line Consistency and Drug Class Specificity Filters
Staurosporine	0	0	0.3771	83	11	[Kinase inh: 1]
Sunitinib	0	1 ^(h)	0.0344	1045	48	[Flt3 inh: 9]; [Kinase inh: 4]; [VEGFR1 or 2 inh: 1]
Tandutinib	0	0	0.0053	961	43	[Flt3 inh: 11]; [Kinase inh: 8]
Tanespimycin	0	0	0.2049	459	28	[Hsp90 inh: 6]
TG-101348	0	0	0.0344	1602	90	[JAK/STAT inh: 1]
Tipifarnib	1	0	0.1329	1592	67	[Farnesyl trans inh: 9]
Topotecan	1	0	0.1088	856	38	[Topoisomerase inh: 10]
Tozasertib	0	0	0.0239	984	46	[Aurora kinase inh: 0]; [Flt3 inh: 9]; [Kinase inh: 6]
Tretinoin	1	0	0.0039	534	69	[Retinoid: 4]
U 73122	0	0	0.0344	679	36	-(i)
Vinblastine	0	0	0.07	1469	180	[Microtubule inh: 10]
Vincristine	0	0	0.0874	1154	102	[Microtubule inh: 8]
Vorinostat	1	0	0.0171	621	29	[HDAC inh: 15]
YM-155	0	0	0.3771	249	9	-(k)

- (a) Its successor Bcl2 inhibitor ABT199 is effective in AML.
- (b) Its successor Bcl2 inhibitor ABT199 is effective in AML.
- (c) It has been tested in phase II trial in AML as a single agent.
- (d) It is one of the 2 drugs of the fludarabine-melphalan preparative regimen for allogeneic transplant that has been used frequently in AML.
- (e) It has been studied in a phase I/II clinical trial as a single agent in AML.
- (f) It is undergoing clinical trial in AML.
- (g) It is undergoing clinical trial in AML.
- (h) It has been studied in a phase I/II clinical trial in AML with chemotherapy.
- (i) It is an antimalarial drug in the “Other” class.
- (j) It is a phospholipase inhibitor in the “Other” class.
- (k) It is a survivin suppressant in the “Other” class.

Appendix Table 14. Top 10 MERGE-scoring genes in each drug class shown in **Figure 29a**. Some drug classes have less than 10 genes that are specifically associated with that class and whose associations are conserved in CCLE data. Column 2 (#Drugs) indicates the number of drugs in the corresponding drug class. Column 3 (#Genes) indicates the number of top MERGE genes specifically associated with the corresponding class (the full list is in **Appendix Table 22**). The classes highlighted in red contain only 1 drug, and the genes highlighted in yellow are those discussed in detail in APPENDIX B.

Drug Class	#Drugs	#Genes	Top 10 Genes
Alkylating agent/ Nucleoside analogue	6	58	RAB31, CCNG1, TERF2, ZNRD1, RNF113A, SMARCC1, ZNF697, MDM4, CNOT6L, ADAM28
AKT/PI3K inhibitor	5	2	PRMT6, RPS19
Topoisomerase inhibitor	4	24	SMARCA4, MZF1, TP53, LEO1, BCL2, ILF3, PRMT7, CEP41, CLNS1A, GPT2
Bcl2 inhibitor	3	13	CASP8AP2, SMAD7, RAB5A, ZNF420, FBXO45, ILK, BCL7A, ZUFSP, SHQ1, KCTD3
CDK inhibitor	4	1	L2HGDH
CHK Inhibitor	1	2	NUP88, GGT5
Farnesyl transferase inhibitor	1	6	SNAPC3, TNIK, UBFD1, MYL6B, FAIM, LENEP
Flt3 inhibitor	4	16	FLT3, ADRBK2, ZNF473, ZDHHC15, GPR34, ATAT1, FAM102A, FCHO2, MYH3, PITRM1
GSK-3 inhibitor	2	10	LPAR4, MAPK1IP1L, RAD23B, NXF3, RNASE2, TNFRSF18, CALCOCO2, RNASE3, SPON2, C1QTNF3
HDAC inhibitor	4	25	MNT, BAZ2B, RNF24, MZF1, ZNF785, ZNF763, RAB3D, PSMC5, CCDC167, CMAHP
Hsp90 inhibitor	2	2	PDLIM1, SEZ6L
Immunomodulation, anti-inflammatory	1	4	DVL2, PEMT, DNAL4, SESN1
Kinase inhibitor	8	11	FLT3, ADRBK2, GPR34, FAM102A, FCHO2, PITRM1, SRSF3, C15orf48, SOCS2, CLEC5A
Met inhibitor	1	5	HSF5, ELAVL1, STOML2, DBF4, HHATL
Microtubule inhibitor	3	2	LILRA5, IL10
mTOR inhibitor	5	1	PKD1
NFkB inhibitor	2	10	ARHGEF9, BEX2, C1orf54, AIM1, YY1AP1, NDNL2, C9orf85, GGA2, PMP22, RGS3
Pim Kinase inhibitor	1	1	SRSF5
PKC inhibitor	1	1	ANPEP
VEGFR1 or 2 inhibitor	2	1	ACAP2

Appendix Table 15. Whether each of the 8 expression markers identified by MERGE (rows) would have been identified by the alternative methods in **Figure 26** and **Figure 27a** (columns). A checkmark means the marker in the corresponding row was identified by the method in the corresponding column. The rows of heat maps in **Figure 27** show the entire list of genes identified by each method.

Gene	ElasticNet	Multi-task learning	Pearson's P-value	Spearman P-value
SMARCA4				
CASP8AP2	✓			
L2HGDH	✓	✓	✓	✓
FLT3				
MNT				
BAZ2B				
MZF1				
BEX2	✓			

Appendix Table 16. IC50 and AUC values in the KG1, transfected KG1, U937 and transfected U937 cell lines.

AUC (24 hr)	KG1	t-KG1	U937	t-U937
Etoposide	594.318	575.157	583.107	575.956
Mitoxantrone	479.329	457.108	466.419	453.636
AUC (48 hr)	KG1	t-KG1	U937	t-U937
Etoposide	603.722	503.202	463.981	464.517
Mitoxantrone	417.081	358.726	312.344	308.787
AUC (72 hr)	KG1	t-KG1	U937	t-U937
Etoposide	602.589	462	446.642	429.982
Mitoxantrone	427.033	306.129	293.408	292.742
IC ₅₀ (24 hr)	KG1	t-KG1	U937	t-U937
Etoposide	-5.252	-5.384	-5	-5.09
Mitoxantrone	-6.95	-7.034	-6.782	-6.932
IC ₅₀ (48 hr)	KG1	t-KG1	U937	t-U937
Etoposide	-5.318	-5.948	-6.374	-6.38
Mitoxantrone	-7.052	-7.496	-8.018	-8.03
IC ₅₀ (72 hr)	KG1	t-KG1	U937	t-U937
Etoposide	-5.354	-6.308	-6.548	-6.71
Mitoxantrone	-7.088	-8	-8.102	-8.108

Appendix Table 17 (provided as a separate .xls file). Expression data and drug sensitivity measures (AUC, IC50, EC50 and Amax) from the 30 AML patient samples and 14 leukemia cell lines.

Appendix Table 18 (provided as a separate .xls file). The 160 drugs in our customized drug panel and their functional classes.

Appendix Table 19 (provided as a separate .xls file). Clinical information of the 30 AML patient samples.

Appendix Table 20 (provided as a separate .xls file). MERGE score, five driver features, and the significant p-value count for each gene

Appendix Table 21 (provided as a separate .xls file). Processed RNA-seq data and drug sensitivity AUC measures from the 12 refractory AML patient samples used for validation.

Appendix Table 22 (provided as a separate .xls file). For each drug class shown in the columns of Fig. 6a, the list of significant genes in the order of the MERGE score, and the corresponding gene properties identified by the MERGE algorithm.

APPENDIX C (APPENDIX TO EMBARKER)

1. Data collection and preprocessing

For experiments with TCGA, we used clinical data and RNA-Seq data provided by the Broad Institute (gdac.broadinstitute.org). We used normalized and log-transformed RNA-Seq read counts from 15,097 protein-coding genes that are present in all cancers. We used *SurvCorr* R package²³³ available on CRAN for imputing the censored survival times. The TCGA data we processed and used as input to EMBARKER models is available on our project website suinlee.cs.washington.edu/projects/embarker.

For experiments with AD, we used data from the ROSMAP,^{130,131} ACT,^{133,234} and MSBB studies. Around half of the people in each cohort had been diagnosed with dementia by the time of death. MSBB neuropathology data was made available by the AMP-AD Knowledge Portal of Sage Bionetworks through synapse.org with Synapse ID syn6101474. We accessed neuropathology data from the ROSMAP and ACT studies through data use agreements. ACT neuropathology data includes four types of A β quantifications: (1) immunohistochemistry (IHC) measured on fresh-frozen brain tissue, (2) IHC measured on formalin-fixed, paraffin-embedded (FFPE) brain tissue, (3) histelide²³⁵ quantification from FFPE slides, and (4) Luminex. We chose histelide since it led to the strongest average association with gene expression levels. ROSMAP RNA-Seq data and MSBB RNA-Seq data were made available by Sage Bionetworks on the AMP-AD Knowledge Portal²³⁶ with Synapse IDs syn3505732 and syn7391833, respectively. The ACT RNA-Seq data¹³³ was collected by the Allen Institute for Brain Science, Kaiser Permanente Washington Health Research Institute (KPWHRI), and the University of Washington (UW), and

it was made available on aging.brain-map.org. In each study, we used the protein-coding genes that have a nonzero RNA-Seq read count in at least one-third of the samples. Overlapping these genes across the three studies resulted in 14,912 genes which we used in our experiments. We used normalized and log-transformed RNA-Seq read counts for all datasets. The AD data we processed and used as input to EMBARKER models is available on our project website suinlee.cs.washington.edu/projects/embarker.

ACT RNA-Seq data are part of the larger ACT project — a collaboration between KPWHRI and UW — that collected clinical and genetic data from thousands of individuals. The approximately 100 ACT individuals from whom we have RNA-Seq measurements were specifically selected for a traumatic brain injury (TBI) study; half of the selected individuals sustained a TBI with loss of consciousness during their lifetime. As a result, the RNA-Seq cohort does not necessarily reflect the demographics of the entire set of thousands of ACT individuals. Therefore, the ACT study assigned each RNA-Seq sample a weight based on the demographic information of each sample in the TBI study. Details of this weighting scheme and the weights themselves are provided on aging.brain-map.org. As recommended by the ACT study, we took the weights into account when computing gene expression-neuropathology associations for the ACT data so that the samples reflect the original ACT cohort demographics. These weights can be found on our project website suinlee.cs.washington.edu/projects/embarker.

2. Details on the computational experimental settings

For each run of EMBARKER on any gene expression-phenotype data, we chose the hyperparameter γ by a 10-fold CV, and we tested 100 different γ values in [10, 1000] range.

Appendix Table 29 lists γ values chosen by the CV experiments. We followed the same approach to choose the hyperparameters for the alternative methods, if applicable.

For TCGA survival case study, for each of the 33 cancer types, we randomly split the samples into two, and used one half for training and the other half for testing to check the consistency between the top markers identified in each half. We repeated this with 10 random splits and averaged over the results for each cancer type. For AD neuropathology case study, since each of the 9 regions are from brain tissue samples, rather than a CV in each region, we used each of the 72 region pair permutations as the training and test pairs.

As the evaluation metric used for comparing between methods in Figures 2 and 3 (in which we presented results for gene-level and pathway-level marker consistency, respectively), we used the area under the statistical consistency curves (AUC) for each cancer type-data split pair or for each region pair. Details of the statistical consistency curve are described in the next subsection.

3. Details on the comparisons to alternative methods

For OGL method to which we compared EMBARKER in Figure 2, we used MATLAB SLEP GitHub package and for PLIER method we used the PLIER GitHub package. For GSEA method to which we compared EMBARKER in Figure 3, we used the fgsea CRAN R package. For Fisher's method and Stouffer's method to which we compared EMBARKER in Figure 5B we used the metap CRAN R package, and for Olkin-Pratt and Dersimonian-Laird methods we used the metacor CRAN R package. We have listed these tools on suinlee.cs.washington.edu/projects/embarker.

Since elastic net and OGL are sparse regression methods, while computing the statistical consistency AUC for any run of these methods, the number of resulting nonzero coefficients (say,

N) could be less than the number of top genes we considered from each method ($M_{max} = 200$ in Figure 2). Similarly, for PLIER, the resulting U matrix may contain pathways which were not used at all to create any latent variable, which leads to an undefined correlation between each of these pathways and the phenotype. In this case, let N be the number of pathways that have a valid correlation with the phenotype (i.e., that were used to create at least one latent variable). In each of these cases, we set $M_{max} = N$ for both the alternative method and EMBARKER for that specific run, to ensure an entirely fair evaluation. **Appendix Table 28** lists the number of valid gene markers for elastic net and OGL as well as the number of valid pathway markers for PLIER in each run of these methods for Case Study 1 or 2.

4. Statistical consistency curve

Statistical consistency curve (**Appendix Figure 27**) shows how many of the top ranked markers identified based on the training samples overlap with those identified based on the test samples. The x-axis represents the sorted list of the markers prioritized by the method based on the *training set* samples, and the y-axis value for $x = M$ represents the number of overlapping markers between the top M markers from the application of the method to the *training set* and the top M markers from the application of the method to the *test set*. Thus, the area under the statistical consistency curve indicates how well the method identifies robust (i.e., replicated in test data) markers.

The x-axis represents the sorted list of the top M markers identified by a method when it is applied to the training set. The y-axis means the number of overlapping markers between the top M markers identified by the method when it is applied to the training set and the top M markers

identified by the method when it is applied to the test set. For all experiments, we used varying M values to make sure that we do not make a conclusion based on a specific M value and show that our conclusions are consistent for varying values of M .

For the results plotted in Figure 2 and 3 (in which we presented results for gene-level and pathway-level marker consistency, respectively), we computed the area under the statistical consistency curve (AUC) to enable a clear presentation of the comparative results across many gene-phenotype association problems: 33 problems (cancer survival phenotype for 33 cancers) for Case Study 1, and 9 problems (AD neuropathology phenotype in 9 brain regions) for Case Study 2. Figure 5A-B shows the entire curve for one phenotype association problem, i.e., identifying expression markers for AD neuropathology levels using all expression datasets in a meta-analysis setting.

5. Identification of *C. elegans* orthologs

To enable biological testing of the human genes identified using our computational analysis, we obtained the RBHs between human and *C. elegans*. We first identified all unique protein sequences for each potential marker gene using the *biomaRt* R package²³⁷ available on CRAN. Then, we used the latest version of the NCBI BLAST tool²³⁸, which we downloaded from <ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST>, version 2.6.0 as of 8/15/17, to identify the *C. elegans* orthologs for each complete human protein query sequence. We downloaded the *C. elegans* protein sequences from wormbase.org/species/c_elegans. We took into account only the protein pairs mapped from human to *C. elegans* with a BLAST e -value smaller than 10^{-30} . For each *C. elegans* isoform, we identified the corresponding human genes, again using the NCBI

BLAST tool, and used only the orthologs that achieved a BLAST *e*-value smaller than 10^{-30} . This process resulted in high-confidence RBHs for us to test in *C. elegans*.

6. *C. elegans* strain, cultivation, and RNAi treatment

Standard procedures for *C. elegans* strain maintenance and manipulation were used, as previously described.^{158,239} All experiments were performed using the GMC101 strain expressing the human A β ₁₋₄₂ peptide under the *unc-54* promoter.¹⁵⁵ Experimental worm populations of GMC101 animals were obtained from the Caenorhabditis Genetics Center (CGC) and cultivated on NGM plates with OP50 at 15C.^{155,240} Care was taken to ensure that the animals were never starved and the plates remained free of contamination.

The gene-specific RNAi clones were obtained from the commercial Ahringer or Vidal *C. elegans* RNAi-feeding libraries (BioScience, Nottingham, UK). Each bacterial clone was struck-out onto LB plates containing carbenicillin (50 ug/ml) and tetracycline (10 ug/ml). Single colonies were then seeded into 5 ml LB + carbenicillin (50 ug/ml) and tetracycline (10 ug/ml) for growth overnight on a 37C rotator. 100 ul of each overnight culture was then inoculated into 10 ml of LB containing carbenicillin (50 ug/ml) and tetracycline (10 ug/ml) and IPTG (5mM) and incubated on a 37C rotator for 4 hours. Each bacterial growth was then centrifuged at 3500 X G for 25 minutes, decanted, and the pellet resuspended in 0.5 ml of LB containing carbenicillin (50 ug/ml), tetracycline (10 ug/ml), and IPTG (5 mM). To verify RNAi conditions' plasmid DNA, each RNAi clone was purified and assessed through PCR (polymerase chain reaction) with sequence-specific primers or through Sanger sequencing.

7. Nematode paralysis assays

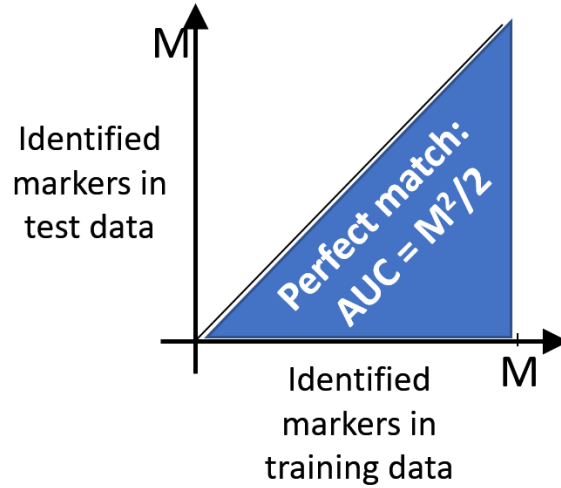
Paralysis assays were performed by visually inspecting animals daily to determine if they were capable of spontaneous movement or if they were paralyzed. A robotic system equipped with a digital camera was used to obtain images of individual wells of a 12 well-plate at 5-minute intervals over the entire course of the experiment. Each well contained 30-40 individual animals expressing A β . Through analysis of serial images from each plate, the age at which each animal stopped moving could be easily determined. We applied this system to the transgenic A β model line GMC101 to determine the time of paralysis onset for each individual animal. Paralysis data was plotted using Oasis2.²⁴¹ Statistical significance of mean paralysis time-points between RNAi conditions was determined by a weighted log-rank^{242,243} test.

Prior to loading on the experimental plates, animal populations were amplified on high-growth plates seeded with NA22 bacteria. Worm populations were developmentally synchronized by hypochlorite treatment, and the remaining eggs were deposited on unseeded plates overnight. Synchronized larval stage 1 animals were washed off unseeded plates and moved onto standard *C. elegans* RNAi plates containing carbenicillin (50 mg/ml), tetracycline (10 mg/ml), and IPTG (5 mM) 48h at 20C. These developmentally synchronized, late larval stage 4-populations were then washed and transferred to their respective RNAi conditions on 12-well plates. We used standard RNAi conditions plus FuDR (100 ug/ml) to prevent progeny and neomycin (200 mg/ml) to prevent fungal growth.²³⁹ Each RNAi condition was tested in 2-3 wells as technical replicates. At least three biological replicates, each started on different weeks, were conducted for each RNAi clone. Animals were maintained at ambient room temperature (generally 22-24C) over the course of the paralysis assay.

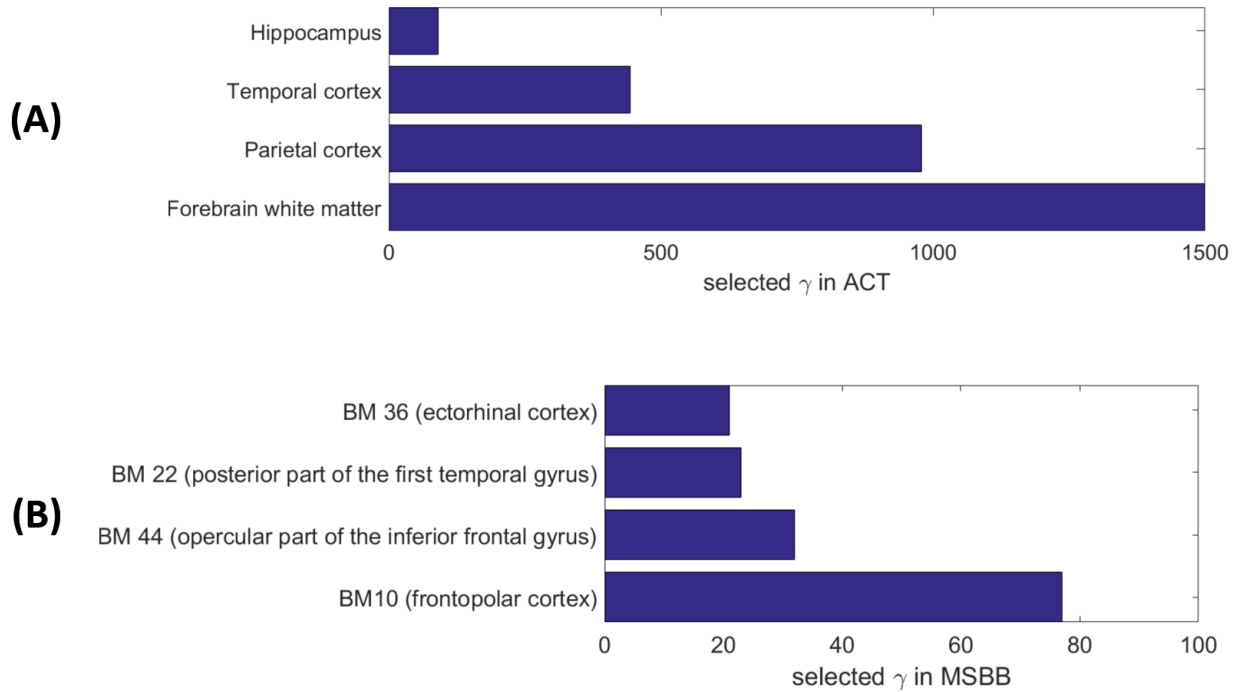
8. Oxygen consumption assay

We quantified the oxygen consumption rates of day-two adult worms cultivated from L1 on *nduf-9* RNAi or the EV control RNAi at 20°C. Oxygen consumption was measured utilizing the Seahorse X24 Bioanalyzer (Seahorse Biosciences, MA, USA).⁵⁴ Briefly, young adult worms were washed from NGM plates and rinsed from RNAi bacteria with M9 buffer. Approximately 40-50 worms were pipetted into each well in Seahorse XF24 Cell Culture Microplates, with the final volume of 500µL M9. The number of worms in each well was quantified. Basal respiration was analyzed using the average respiration of five technical replicates per condition, and eight recordings over the course of an hour were made of each condition. Three independent biological replicates were conducted on different weeks. For each experiment, we compared the relative rate of oxygen consumption between the paired conditions. In each plate, nematode-free wells were used as control. Oxygen consumption rates were then normalized by the number of worms. All three replicates showed that the animals cultivated on the *nduf-9* RNAi consumed about half as much oxygen as those grown on the control RNAi bacteria (**Figure 35A**).

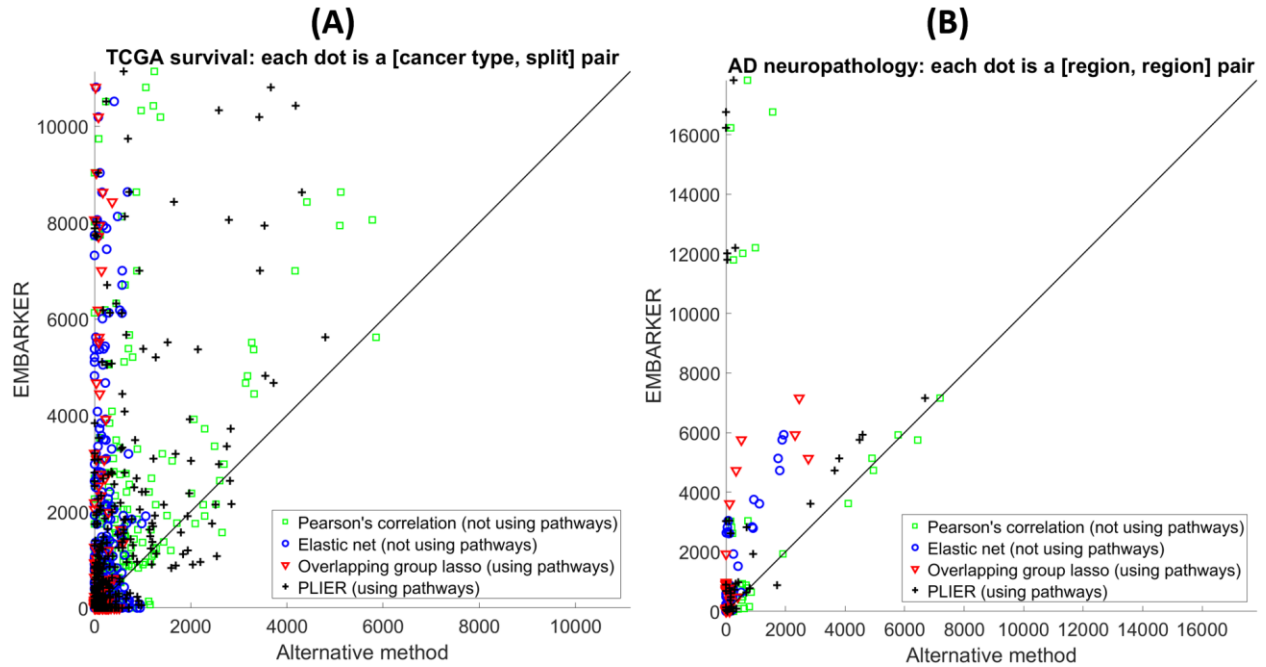
9. APPENDIX C Figures



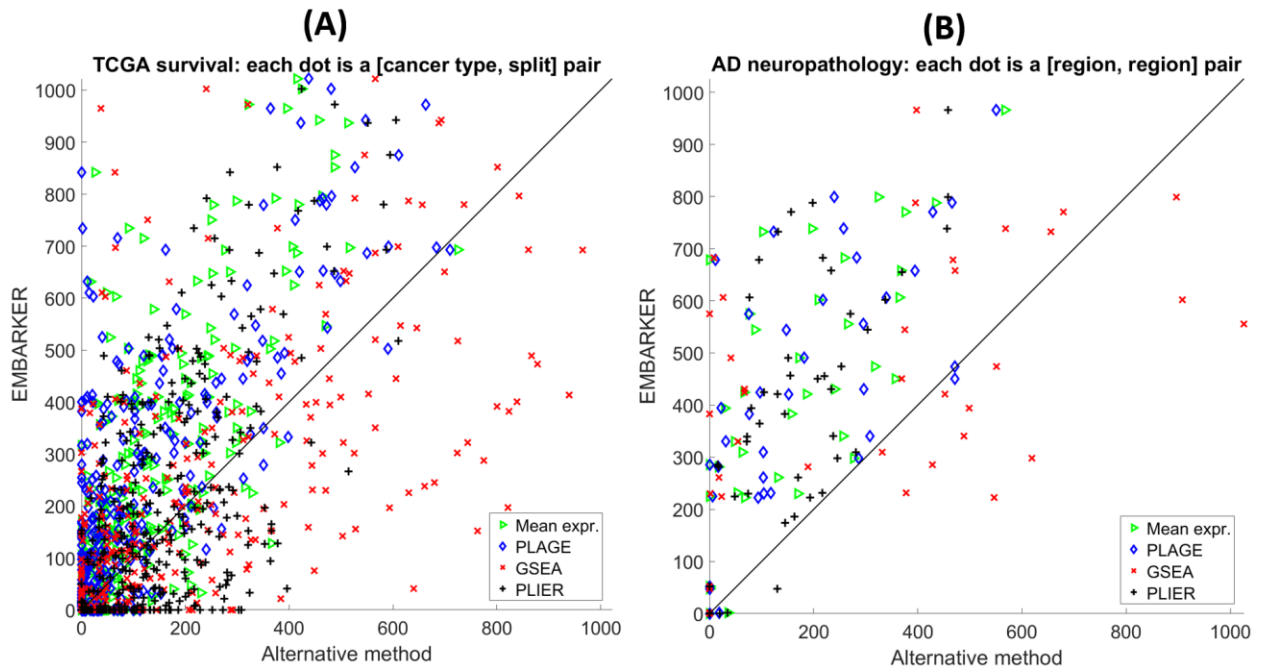
Appendix Figure 27. A plot showing our AUC metric for statistical consistency for the top M markers. The blue-highlighted area represents the perfect match between training and test which is the best statistical consistency that can be achieved for M markers.



Appendix Figure 28. The γ hyperparameter values selected for **(A)** four MSBB study regions and **(B)** four ACT study regions.



Appendix Figure 29. Scatter plots comparing gene-level marker statistical consistency achieved by EMBARKER (y-axis) vs. the four alternative methods (x-axis) in **(A)** TCGA survival case study, and **(B)** AD neuropathology case study for the top 200 identified gene-level markers.



Appendix Figure 30. Scatter plots comparing pathway-level marker statistical consistency achieved by EMBARKER (y-axis) vs. the four alternative methods (x-axis) in **(A)** TCGA survival

case study, and **(B)** AD neuropathology case study for the top 50 identified pathway-level markers.

10. APPENDIX C Tables

Appendix Table 23. Abbreviations of 33 cancer types we used in our Case Study 1 and the corresponding sample sizes.

Abbreviation	Cancer	Sample size	Uncensored (dead) sample size
ACC	Adrenocortical carcinoma	77	27
BLCA	Bladder urothelial carcinoma	381	171
BRCA	Breast invasive carcinoma	1073	150
CESC	Cervical and endocervical cancers	288	71
CHOL	Cholangiocarcinoma	34	17
COAD	Colon adenocarcinoma	226	55
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma	48	9
ESCA	Esophageal carcinoma	170	72
GBM	Glioblastoma multiforme	157	127
HNSC	Head and neck squamous cell carcinoma	508	215
KICH	Kidney chromophobe	64	8
KIRC	Kidney renal clear cell carcinoma	515	169
KIRP	Kidney renal papillary cell carcinoma	278	40
LAML	Acute myeloid leukemia	161	105
LGG	Brain lower grade glioma	511	125
LIHC	Liver hepatocellular carcinoma	358	126
LUAD	Lung adenocarcinoma	421	154
LUSC	Lung squamous cell carcinoma	466	198
MESO	Mesothelioma	86	73
OV	Ovarian serous cystadenocarcinoma	301	183
PAAD	Pancreatic adenocarcinoma	150	82
PCPG	Pheochromocytoma and paraganglioma	178	5

Abbreviation	Cancer	Sample size	Uncensored (dead) sample size
PRAD	Prostate adenocarcinoma	476	10
READ	Rectum adenocarcinoma	85	15
SARC	Sarcoma	222	85
SKCM	Skin cutaneous melanoma	416	196
STAD	Stomach adenocarcinoma	327	130
TGCT	Testicular germ cell tumors	133	4
THCA	Thyroid carcinoma	499	16
THYM	Thymoma	118	8
UCEC	Uterine corpus endometrial carcinoma	156	33
UCS	Uterine carcinosarcoma	55	34
UVM	Uveal melanoma	80	23

Appendix Table 24. The 9 brain regions from 3 studies we used in our Case Study 2 and meta-analysis setting, and the corresponding number of samples with gene expression and neuropathology ($A\beta$) quantifications.

Region abbreviation (Study-brain region)	Brain region	Sample size
ACT-TCx	Temporal cortex	85
ACT-PCx	Parietal cortex	79
ACT-HIP	Hippocampus	80
ACT-FWM	Forebrain white matter	81
ROSMAP-DPFCx	Dorsolateral prefrontal cortex	538
MSBB-BM10	Brodman area 10 (frontopolar cortex)	244
MSBB-BM36	Brodman area 36 (ectothal cortex)	200
MSBB-BM44	Brodman area 44 (opercular part of the inferior frontal gyrus)	208
MSBB-BM22	Brodman area 22 (posterior part of the first temporal gyrus)	227

Appendix Table 25 (provided as a separate .xlsx file). EMBARKER-M scores for all genes ordered by decreasing absolute value.

Appendix Table 26 (provided as a separate .xlsx file). EMBARKER-M scores for all REACTOME pathways used for training the model ordered by decreasing absolute value.

Appendix Table 27 (provided as a separate .xlsx file). Human Complex I genes (1st column), the strong ortholog of the NDUF gene in *C. elegans* (2nd column), BLAST e-values of human-to-*C. elegans* mapping (3rd column), and the ranking of the gene from EMBARKER-M applied on nine brain regions (4th column).

Appendix Table 28 (provided as a separate .xlsx file). The number of nonzero gene-level scores from elastic net and overlapping group lasso, and the number of valid pathway-level scores from PLIER. Each sheet contains the counts from a case study (TCGA survival or AD neuropathology).

Appendix Table 29 (provided as a separate .xlsx file). The γ values selected by cross-validation experiments. Each sheet contains the counts from one of the two case studies (TCGA survival or AD neuropathology).