Genome Medicine

METHOD Open Access

Extracting a low-dimensional description of multiple gene expression datasets reveals a potential driver for tumor-associated stroma in ovarian cancer

Safiye Celik[1], Benjamin A. Logsdon[2], Stephanie Battle[3], Charles W. Drescher[4], Mara Rendi[5], R. David Hawkins[3,6] and Su-In Lee[1,3*]

Abstract

Patterns in expression data conserved across multiple independent disease studies are likely to represent important molecular events underlying the disease. We present the INSPIRE method to infer modules of co-expressed genes and the dependencies among the modules from multiple expression datasets that may contain different sets of genes. We show that INSPIRE infers more accurate models than existing methods to extract low-dimensional representation of expression data. We demonstrate that applying INSPIRE to nine ovarian cancer datasets leads to a new marker and potential driver of tumor-associated stroma, *HOPX*, followed by experimental validation. The implementation of INSPIRE is available at http://inspire.cs.washington.edu.

Keywords: Gene expression, Variable discrepancy, Low-dimensional representation, Module, Conditional dependence, Latent variable, *HOPX*, Tumor-associated stroma

Background

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 [1–3]. Specific machine learning domains to see noted success recently include the development of deep learning algorithms [3], where authors demonstrate enormous increases in performance on difficult tasks such as image and text classification [4, 5]. 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 challenge [6]. 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 margin [7] 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 WGCNA [8] or simple clustering approaches [9, 10] will not necessarily capture complex dependence structures among the modules. Appropriately accounting for

* Correspondence: suinlee@uw.edu
[1]Department of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[3]Department of Genome Sciences, University of Washington, Seattle, WA, USA
Full list of author information is available at the end of the article

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 data [11]. 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 (INferring Shared modules from multiPle gene expREssion datasets), to infer highly coherent and robust modules of genes and their dependencies on the basis of gene expression datasets from multiple independent studies (Fig. 1). 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 model [12], where the latent variables are modeled as a Gaussian graphical model, which is the most commonly used probabilistic graphical model for continuous-valued variables (Fig. 1). Each observed gene is treated as a noisy and independent observation of these underlying latent variables. 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 (Fig. 1). 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 study [11], 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

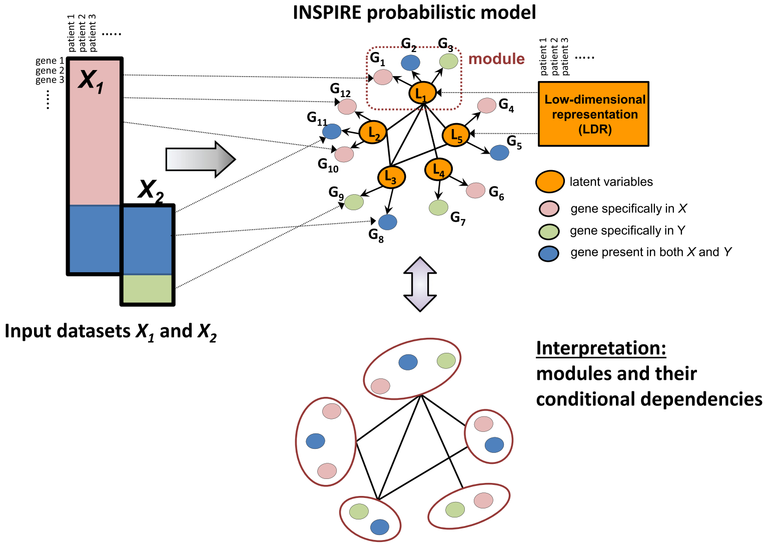

Fig. 1 Overview of the INSPIRE framework. INSPIRE takes as input multiple expression datasets that potentially contain different sets of genes and learns a network of expression modules (i.e. co-expressed sets of genes) conserved across these datasets. INSPIRE is a general framework that can take any 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 an LDR of the expression datasets. *Top middle*: As an example, five INSPIRE features $L_1, \ldots, L_5$ (*orange-shaded circles*), 12 genes $G_1, \ldots, G_{12}$ 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*: 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

and have better power to learn both the modules and their dependencies, even when the number of genes is much greater than the sample size. Through extensive simulated and real data analysis (Fig. 2), we demonstrate that our approach is a great practical trade-off between model complexity and model parsimony when understanding biological pathways characterizing the cancer transcriptome across ovarian cancer patients.

Previous approaches to extract LDR from expression data can be divided into two categories; (1) supervised

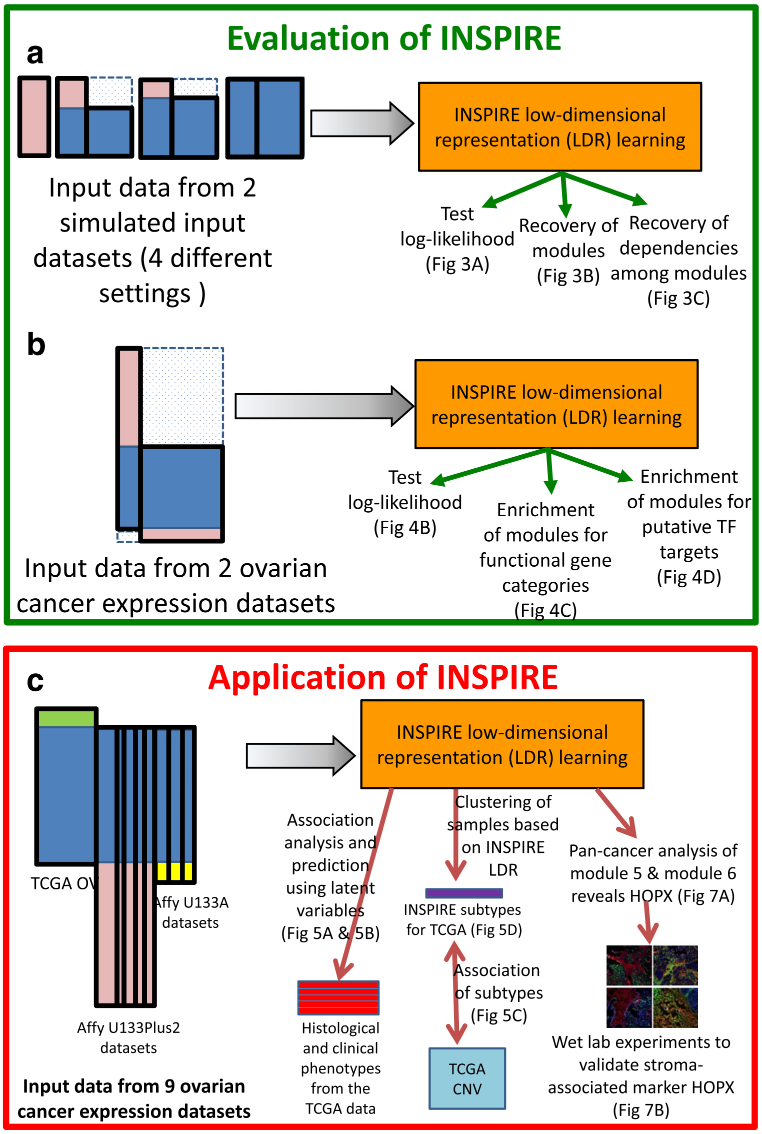
Evaluation of INSPIRE
a
INSPIRE low-dimensional representation (LDR) learning
Input data from 2 simulated input datasets (4 different settings)
Test log-likelihood (Fig 3A)
Recovery of modules (Fig 3B)
Recovery of dependencies among modules (Fig 3C)
b
INSPIRE low-dimensional representation (LDR) learning
Input data from 2 ovarian cancer expression datasets
Test log-likelihood (Fig 4B)
Enrichment of modules for functional gene categories (Fig 4C)
Enrichment of modules for putative TF targets (Fig 4D)
Application of INSPIRE
c
INSPIRE low-dimensional representation (LDR) learning
TCGA OV
Affy U133A datasets
Affy U133Plus2 datasets
Input data from 9 ovarian cancer expression datasets
Association analysis and prediction using latent variables (Fig 5A & 5B)
Histological and clinical phenotypes from the TCGA data
Clustering of samples based on INSPIRE LDR
INSPIRE subtypes for TCGA (Fig 5D)
Association of subtypes (Fig 5C)
TCGA CNV
Pan-cancer analysis of module 5 & module 6 reveals HOPX (Fig 7A)
Wet lab experiments to validate stroma-associated marker HOPX (Fig 7B)

Fig. 2 Overview of the evaluation and application of INSPIRE procedure. INSPIRE takes as input $K \geq 2$ datasets, and the method is an iterative procedure that determines 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 copy number variation (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

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.

A supervised method aims to extract an LDR that is discriminative between class labels in a particular prediction problem. 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 [13–16]. Chuang et al. [13] propose 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. [14] developed a similar algorithm to select subsets of genes from MSigDB (Molecular Signatures Database) C2 (curated) pathways that give the optimal discriminative power for the classification of leukemia/breast cancer phenotypes. Both Chuang et al. [13] and Lee et al. [14] determine LDR as the average expression levels of genes in each subnetwork and pathway, respectively. Taylor et al. [15] propose 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. [16] 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. [17] used 106 genes selected by the intrinsic analysis for a classification problem (122 mouse breast tumors/232 human breast tumors). The intrinsic analysis aims to select genes that are relevant to tumor classification by identifying genes whose expression show relatively low within-group variation and high between-group variation for known groups of tumors in each of human and mouse datasets [17]. Although supervised methods would be useful to infer an LDR relevant to a particular prediction problem, they have several disadvantages over unsupervised methods. First, we need to have a particular prediction problem with class labels, which may not be available. Second, they 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 LDR learning methods extract an LDR without knowing about the class labels, while the learned LDR can be used for classification purposes later. One of the most commonly used methods is the principal component analysis (PCA) [18] which sequentially extracts most of the variance (variability) of the data. Another is independent component analysis (ICA) [10, 19], a statistical technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. However, each principal component (PC - 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 algorithms [20], 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) [8], the adjacencies retrieved from Pearson's correlation of the expression levels of the gene pairs is transformed into topological overlap measure (TOM), namely network interconnectivity that takes into account the shared neighbors of each gene pair, which is then used in a hierarchical clustering to define modules. While WGCNA [8] 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 [11, 21, 22]. Chandrasekaran et al. [21] incorporated latent variables into a Gaussian graphical model among individual genes, while Celik et al. [11] divided variables into modules and learned module-level dependencies (module graphical lasso (MGL)). He et al. [22] 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. [11] in modeling a higher-level dependency structure, He et al. [22] does not form explicit clusters. Finally, Cheng et al. [7] 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 model [6].

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 (Fig. 1). We demonstrate that INSPIRE outperforms the imputation-based approaches (methods named "Imp–" in Figs. 3 and 4). 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 (Fig. 1). 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 (Fig. 2c), 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 study [23]. 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 website [24].

Methods

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 freeze (http://gdac.broadinstitute.org/runs/stddata__2014_03_16/data/OV/20140316/) for all three platforms available for ovarian cancer (Affymetrix U133A, Agilent g4502, Human Exon array). We first removed potential plate level batch effects with ComBat [25] for all expression datasets. As was done in the TCGA ovarian cancer study [23], we combined the three separate expression measurements for each of 11,864 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 Omnibus [26] for accessions: GSE14764 [27], GSE26712 [28], GSE6008 [29], GSE18520 [30], GSE19829 [31], GSE20565 [32], GSE30161 [33], GSE9899 [34]. Expression data were then processed using MAS5.0 normalization with the "Affy" Bioconductor package [35] and mapped to Entrez gene annotations [36] using custom chip definition files (CDF) [37] which was followed by natural log transformation of MAS5.0 normalized intensities. The expression data were then Z-transformed so that each gene has zero mean and unit variance across the samples within each dataset. As stated in Tibshirani [38], 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) [38–42]. 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 [11, 43–49].

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 page [50]. 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 GISTIC [51] 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.

INSPIRE learning algorithm

We present the INSPIRE 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

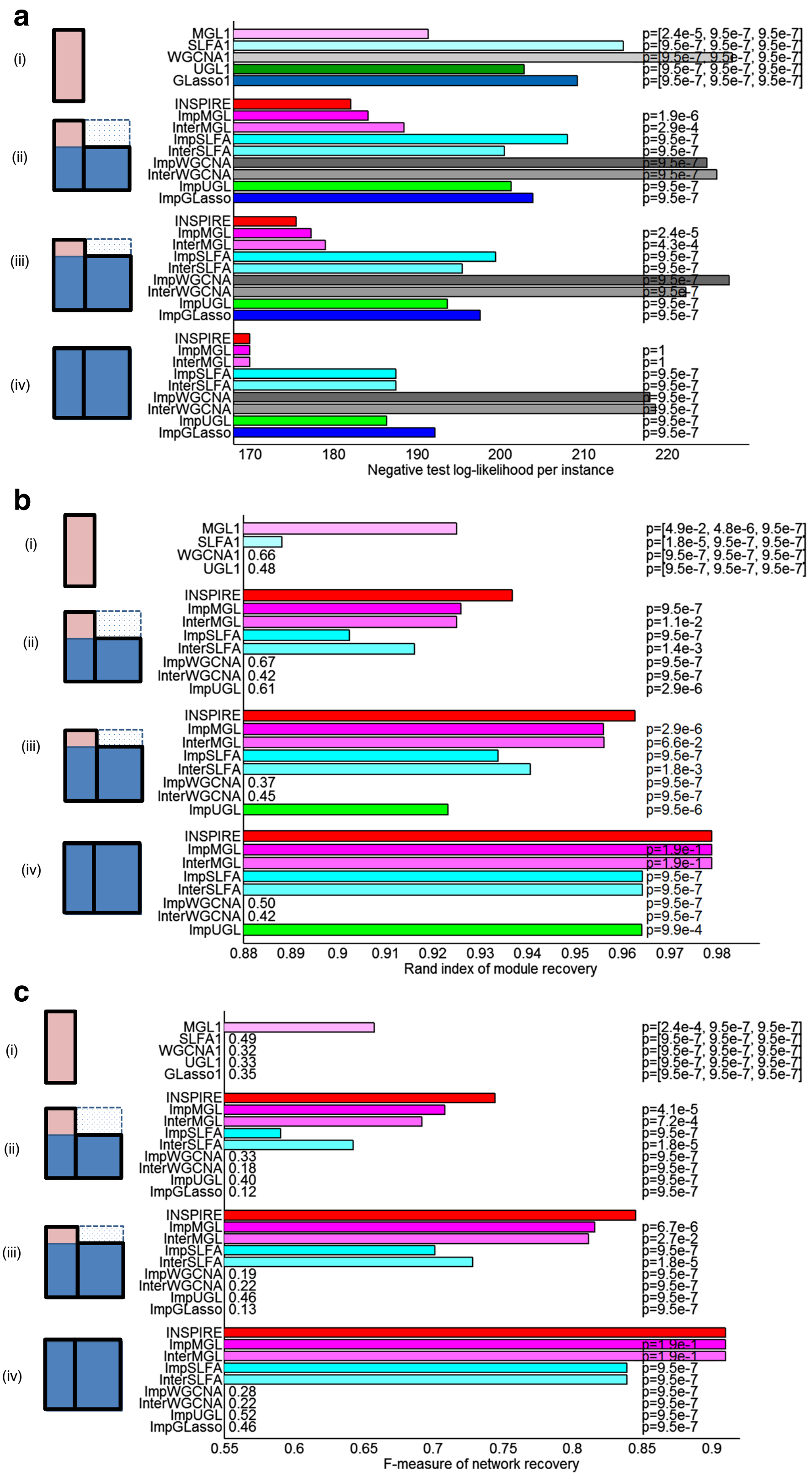

Fig. 3 (See legend on next page.)

(See figure on previous page.)
Fig. 3 *Illustration* of the synthetic data, aligned with four groups of bars in each of (**a**)–(**c**). *Rows* represent genes and columns represent samples. **a** Negative test log-likelihood per instance averaged over 20 different instantiations of the synthetic data (lower is better). **b** Rand index for module recovery averaged over 20 different instantiations of the synthetic data. **c** F-measure for feature dependency recovery averaged over 20 different instantiations of the synthetic data. 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

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 low-dimensional features for prediction tasks (e.g. predicting histopathological phenotypes) or clustering (e.g. patient stratification) (Fig. 1).

INSPIRE uses a formal probabilistic graphical model, specifically the Gaussian graphical model (GGM), to model the 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 [52, 53]. For example, in a simple latent network shown in Fig. 1, five latent variables (L_1, …, L_5) have mutual dependencies. So, let L = {L_1, …, L_5} ~ $N(0, \Sigma_L)$, then nonzero 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, and G_3 in Fig. 1 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, …, X^Q be a set of Q expression datasets where the qth dataset $X^q = \left\{X_1^q, \ldots, X_{p_q}^q\right\}$ 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, …, L^Q be a set of matrices where each L^q is associated with a dataset and consists of k latent variables. $L^q = \{L_1{}^q, \ldots, L_k{}^q\} \sim N(0, \Sigma_L)$, where Σ_L is a $k \times k$ covariance matrix. These latent variables can be viewed as a 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 Fig. 1. 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 ($p_q \le p_T$). 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 $\forall\ i,\ \sum_{c=1}^{c=k} Z_{ic} = 1$. Each observed dataset X^q is generated by the multivariate Gaussian distribution $X^q \mid Z^q L^q, \sigma^2 \sim N(Z^q L^q, \sigma^2)$, 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 jth module Mj can be defined as $M_j = \cup_{\{q\ =\ 1\}}^{Q} \{X_i{}^q \mid Z_{ij}{}^q = 1\}$. 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 $\forall\ i, j\ 0 \le Z_{ij} \le 1$.

INSPIRE jointly learns the latent variables $L = [L^1, \ldots, L^Q]$ each corresponding to a module; the module assignment indicator Z; and the feature dependence network Σ_L^{-1}. Given Q datasets X^1, …, X^Q, where $X^q \left(\in \mathbb{R}^{\{p_q \times n_q\}}\right)$ contains n_q observations on p_q genes and $n_T = \sum_{q=1}^{q=Q} n_{q,}$ INSPIRE aims to learn the following:

- $L^q \in \mathbb{R}^{\{k \times n_q\}}$ for each q ($\in \{1, \ldots, Q\}$) containing the values on k features in n_q samples in X^q
- $Z \mid \Sigma Z_i = 1$, a binary vector for each $i(\in \{1, \ldots, p_T\})$ specifying the module membership of the ith 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 $\Theta_L(\succ 0)$. 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:

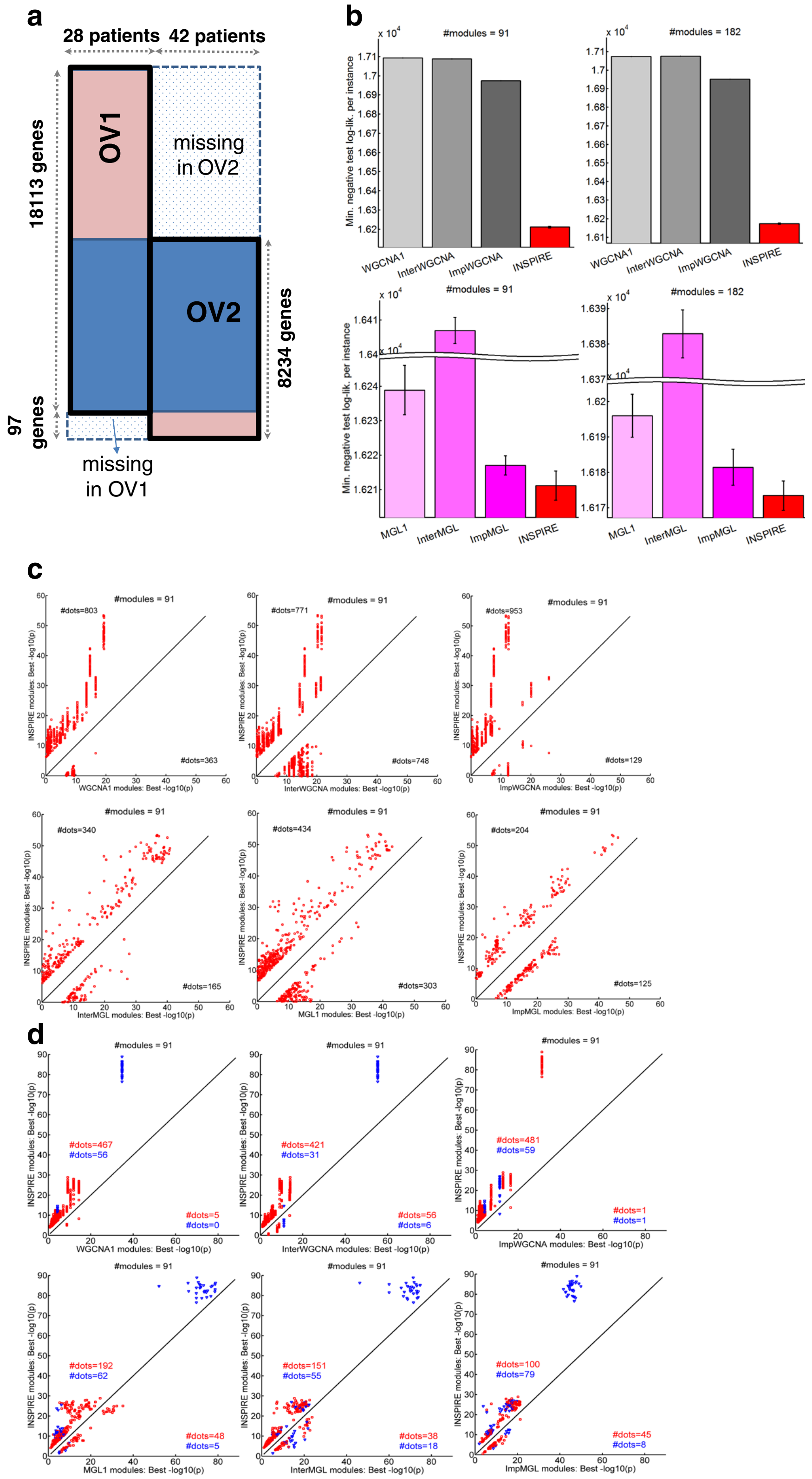

Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 a *Illustration* of the two OV datasets used for evaluating INSPIRE. *Rows* represent genes and *columns* represent samples. **b** For $k = 91$ (left) and $k = 82$ (right), INSPIRE is compared to WGCNA variants (*top*) and MGL variants (*bottom*) in terms of the best cross-validation (CV) negative test log-likelihood (lower is better) across all tested sparsity tuning parameters (λ). **c** 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 <0.05 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. $|\log_{10}p(\mathit{INSPIRE}) - \log_{10}p(\mathit{ALTERNATIVE\_METHOD})| \geq \delta$. $\delta = 6$ here and the results were consistent for varying δ. **d** 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 <0.05 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. $|\log_{10}p(\mathit{INSPIRE}) - \log_{10}p(\mathit{ALTERNATIVE\_METHOD})| \geq \delta$. $\delta = 3$ 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

$$\begin{aligned}\log\mathrm{P}\left(\mathrm{X}^1, \dots, \mathrm{X}^Q, \mathrm{L}^1, \dots, L^Q, \mathrm{Z}, \Theta_L; \lambda, \sigma\right) &= \sum_{q=1}^{Q} \log P(X^q|L^q, Z^q) + \sum_{q=1}^{Q} \log\mathrm{P}(L^q|\Theta_L) \\ &\quad + \log\mathrm{P}(\Theta_L) + \log P(Z) \\ &= \frac{n_T}{2}\{\log\ \det\ \Theta_L - tr\ (S_L\Theta_L)\} - \lambda\sum\nolimits_{j\neq j'}\left|(\Theta_L)_{jj'}\right| \\ &\quad - \frac{1}{2}\sum\nolimits_{q=1}^{Q}\frac{\|X^q - Z^qL^q\|_2^2}{\sigma^2} + const,\end{aligned} \tag{1}$$

where $S_L = \frac{1}{n_T}\sum\nolimits_{q=1}^{q=Q} L^q {L^q}^T$ 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 Eq. (1) given Z and Θ_L, we solve the following problem:

$$\max_{L^1, \dots, L^Q}\left\{-tr\left(L^q {L^q}^T \Theta_L\right) - \frac{\|X^q - Z^q L^q\|_2^2}{\sigma^2}\right\} \tag{2}$$

Setting the derivative of the objective function in Eq. (2) to zero with respect to L_c^q for $q \in \{1, \dots, Q\}$ and $c \in \{1, \dots, k\}$ leads to:

$$L^q{}_c = \frac{{Z^q}^T{}_c X^q - \sigma^2\sum\nolimits_{i\neq c}(\Theta_L)_{ic} L^q{}_i}{\|{Z^q}^T{}_c\|_2^2 + \sigma^2(\Theta_L)_{cc}}. \tag{3}$$

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

$$\min_{Z_1 \dots Z_{p_T}} \sum_{q=1}^{Q}\|X^q - Z^q L^q\|_2^2 \tag{4}$$

In the hard assignment paradigm that we follow throughout this paper, Eq. (4) 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 $\boldsymbol{\Theta_L}$ given $L^1, \dots, L^Q$, we solve the following optimization problem:

$$\max_{\Theta_L \succ 0}\left\{\log\det\Theta_L - tr(S_L\Theta_L) - \lambda\sum\nolimits_{j\neq j'}\left|(\Theta_L)_{jj'}\right|\right\}, \tag{5}$$

where the constraint $\Theta_L \succ 0$ restricts the solution to the space of positive definite matrices of size $k \times k$, and $S_L = \frac{1}{n_T}\sum\nolimits_{q=1}^{q=Q} L^q {L^q}^T$ is the empirical covariance matrix of L. Based on the estimated value of L, Eq. (5) can be solved by the graphical lasso [54], a well-known algorithm for learning the structure of a 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) [55], 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.

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 (Fig. 1). 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 [56] to generate the imputed data.

Initialization of the INSPIRE latent variables

INSPIRE is an iterative learning algorithm that consists of three update steps, Eqs. (3)–(5), to learn the following sets of parameters: L, values on the latent variables, 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.

To be more specific, the authors of the MGL algorithm suggested initializing 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.

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 (Additional file 1: Table S1) with a total number of $p \cong 20{,}000$ genes and $n \cong 1500$ samples took 13.7 min 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) [11] method on the imputed data, which took 10.2 min on the same machine.

Table 1 Methods we compared with the 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 lasso [54], UGL - unknown group L_1 regularization [62], SLFA - the structured latent factor analysis [22], WGCNA - weighted gene co-expression network analysis [8], and MGL - module graphical lasso [11] (see "Methods" for details). We adapted the input datasets such that we can apply these methods to datasets with variable discrepancy (Additional file 2: Figure S1B): "—1", learning a model from only Dataset1 that contains all genes; "Inter—", learning a model from the data on the overlapping genes (blue-shaded region in Fig. 1) 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 the expression of each of the modules); and "Imp—", imputing missing values in Dataset2 and learning a model from the imputed data (see "Methods" for details on imputation) (Additional file 2: Figure S1B). 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 Additional file 3: Figure S2) These methods are 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 Additional file 3: Figure S2)
		—1	Inter—	Imp—	
GLasso	Standard graphical lasso [54]	GLasso1	X	ImpGLasso	No
UGL	Unknown group L_1 regularization [62]	UGL1	X	ImpUGL	No
SLFA	Structured latent factor analysis [22]	SLFA1	InterSLFA	ImpSLFA	No
WGCNA	Weighted gene co-expression network analysis [8]	WGCNA1	InterWGCNA	ImpWGCNA	Yes
MGL	Module graphical lasso [11]	MGL1	InterMGL	ImpMGL	Yes

Synthetic data generation

We synthetically generated data based on the joint distribution in Eq. (1). We first generated the sparse $k \times k$ inverse covariance matrix λ by creating a $k \times k$ matrix G as

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

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

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

Then, we generated the latent variables $L = \{L_1, \ldots, L_k\}$ 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 $X \mid ZL, \sigma^2 \sim N(ZL, \sigma^2)$ 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 amount of overlapping genes (Additional file 2: Figure S1A). 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 (Additional file 2: Figure S1A). [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 (Fig. 3a–c). 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.

Additional file 2: Figure S1A 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 (Fig. 4b).

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); $p = 300$, $p = 1500$, $p = 3000$. 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 five of these methods (except INSPIRE) cannot accommodate multiple datasets. We used varying sparsity tuning parameters in the interval of (0.5, 0.0001), exactly the same set of values that we used for choosing λ (via cross-validation (CV) 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 method [57] 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 h for $p = 3000$ 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 h at $p = 3000$ (Additional file 3: Figure S2A), it is clear that all methods other than MGL, WGCNA, and INSPIRE are too slow to be used when p >3000 and >500 h when p is near 20,000 (see the trend line in Additional file 3: Figure S2B). 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 >1500 where the module count (k) exceeded the sample size (50). Additional file 3: Figure S2A and 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 (Fig. 3). Thus, we excluded GLasso, UGL, and SLFA from 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.

Computing the cross-validation test log-likelihood

We performed a fivefold 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 five 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) versus average CV test log-likelihood (y-axis) curves for all methods, as expected.

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.

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.

Comparison of the prediction performance with alternative methods

We compared INSPIRE with PCA [18] and the subnetwork analysis method [13] based on how well each method can predict each of the six phenotypes (resectability as defined by 0 cm of residual tumor versus >0 cm of residual tumor after surgery, survival time, and four manually curated histologic phenotypes) from TCGA data. We used the lasso [58] (L_1 regularized linear regression) for predicting the continuous-valued phenotype (percent stroma), L_1 regularized logistic regression for predicting binary phenotypes (stroma type, vessel formation, invasion pattern, and residual tumor), and L_1 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 CV 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) [59]. 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 these predicted survival across all 550 samples, which leads to Fig. 5b (middle). Second, we considered computing CIs within test samples in each fold, which would allow us to have multiple CIs (# of folds × # of CV rounds) and compute the confidence interval of the CIs for INSPIRE compared to the CIs for the alternative methods (Fig. 5b, right). Especially, in this analysis, we performed 50 rounds of tenfold CV tests, and reported the average of a total of 500 CIs (i.e. y-axis of Fig. 5b, 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 method [18], subnetwork analysis [13], and individual genes). All p values are smaller than 0.01 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 [58]. For a fair comparison with PCA [18] and the subnetwork method [13], we used the top 90 PCs and 90 subnetworks that are most

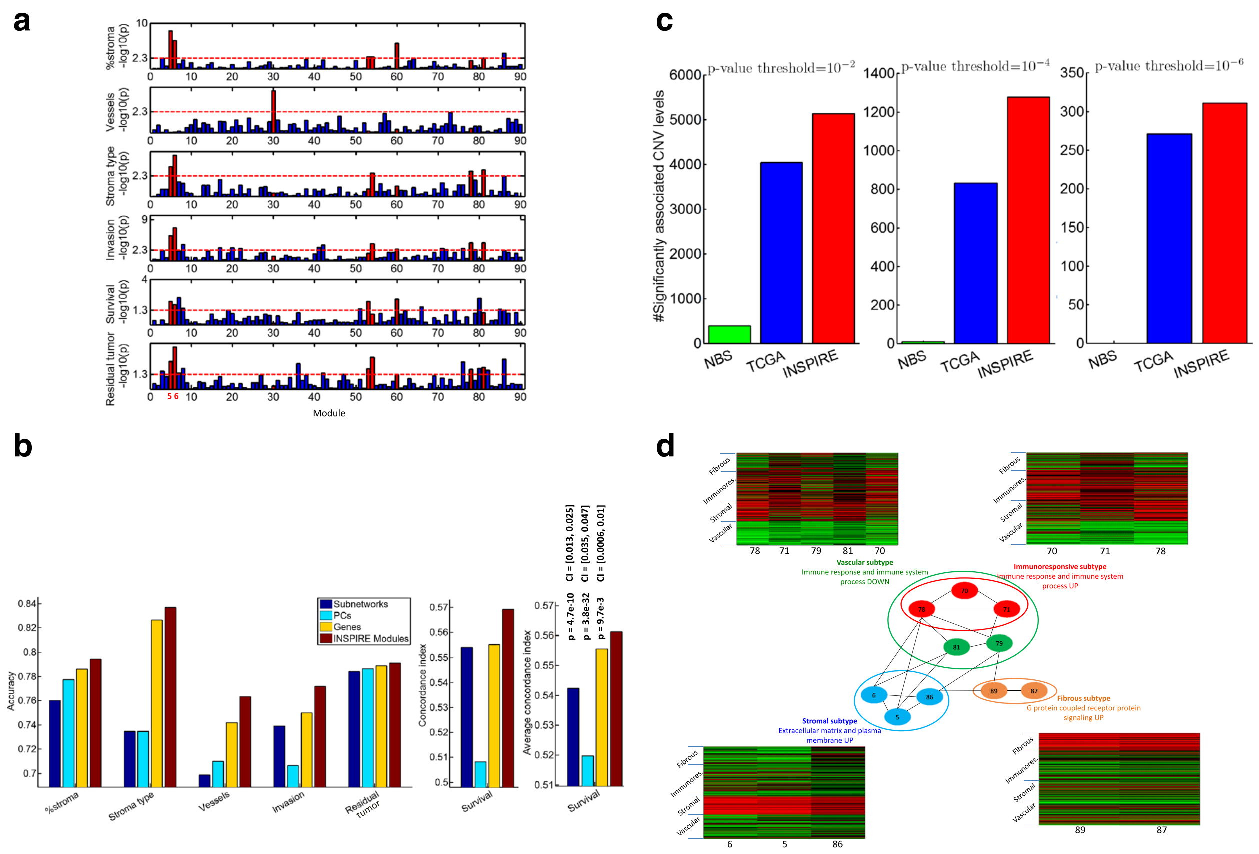

Fig. 5 **a** For each of 90 INSPIRE modules (*x-axis*), the $-\log_{10}p$ from the Pearson's correlation is shown (*y-axis*) for six different histological and clinical phenotypes. The *p* value threshold (shown by *red dotted horizontal lines*) is 5×10^{-3} for histological phenotypes and 5×10^{-2} 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 in *red*. We also highlight module 30 in *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. **b** For four different methods (the subnetwork markers, PCs, all genes, and INSPIRE latent variables), the prediction performance is compared for six prediction tasks in CV 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 the survival phenotype, an additional analysis is presented where the mean concordance index is reported across all 500 folds in 50 rounds of tenfold 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 each of the alternative three methods (individual genes, PCs, and subnetworks). **c** For three different Pearson's correlation *p* value thresholds (10^{-2}, 10^{-4}, and 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. **d** The modules that differentiate the subtypes that are learned using INSPIRE features and the interactions among those 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

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.

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 study [23] and the NBS method [60]). 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 ten times with different random initial assignments of the patients into subtypes and chose the clustering which gives the lowest within cluster sum of squares.

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 versus >0 cm of residual tumor. The proportion of the sub-optimally debulked patients was 62 % (=139/223) in Tothill [34] and was 77 % (=378/489) in TCGA [23]. 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 (Fig. 6) and without (Additional file 4: Figure S3) the clinical covariates of age and stage. Performance was determined based on the results of each fitted model in the Tothill [34] data in terms of the area under the curve (AUC) measure from a receiver operator characteristic (ROC) curve (Fig. 6 and Additional file 4: Figure S3).

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-

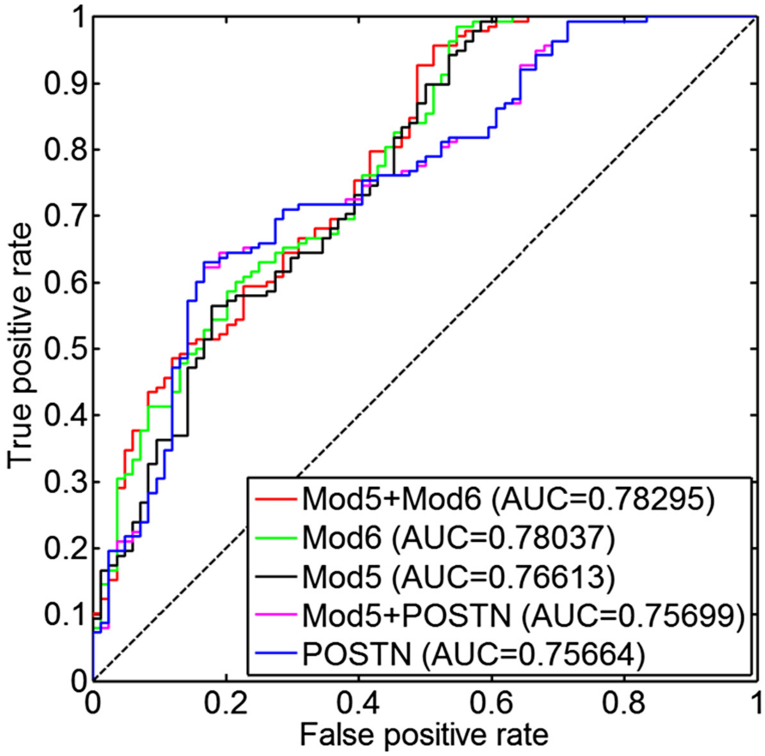

Fig. 6 *ROC curve* of the supervised models for resectability prediction trained in TCGA and tested in Tothill data. Different combinations of *POSTN* and the INSPIRE features corresponding to modules 5 and 6 are used for training each model. The clinical covariates age and stage are also included in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

vascular cores upon which the neoplastic cells grew in arborizing branches. Mixed invasion patterns were identified and classified as such.

Immunohistochemistry

Ten patients were sampled for staining based on either having good tumor resection and survival (>3-year survival, optimal debulking with residual tumor <1 cm) versus poor tumor resection and survival (<3-year 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 10 min. Slides were fixed in a Coplin jar in cold acetone for 10 min at –20 °C. Slides were removed from acetone and placed tissue side up on a shaker. Phosphate buffered saline (PBS) was added to the slide (1 mL, enough to cover tissue slice) for 5 min shaking. PBS wash was repeated for a total of two 5 min washes. After the 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 1 h at room temperature, while shaking. Blocking solution was removed and primary antibody added, diluted in 1 % milk. Antibody dilutions were per manufacturer's recommendations. Slides were allowed to incubate overnight at 4 °C while shaking with the primary antibody. If the 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. The 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 (300 ng/mL, Sigma-Aldrich, catalog # D9542) or Sytox Green Nuclear Stain (Life Technologies, catalog # S7020) was used depending on the secondary antibodies used for staining. The last PBS wash was done at room temperature for 5 min. 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 were: Anti-E Cadherin antibody conjugated to Allophycocyanin (Abcam, catalog no. ab99885); Hop Antibody (Santa Cruz, catalog no. sc-30216); Anti-CD73 antibody (Abcam, catalog no. ab54217); and GCS-a-1 Antibody (Santa Cruz, catalog no. sc-23801)

Secondary antibodies used were: CD73 antibody was detected with Goat anti-mouse IgG-FITC (Santa Cruz, catalog no. sc-2010); when co-stained with CD73, HOPX was detected with Donkey anti-rabbit IgG-CFL 647 (Santa Cruz, catalog no. sc-362291); when co-stained with E Cadherin, HOPX antibody was detected with Chicken anti-rabbit IgG H&L FITC (Abcam, catalog no. ab6825).

Analysis of immunohistochemistry

Fluorescence images were analyzed using ImageJ [61] and the plugin JACoP was used for co-localization analysis.

Results

Overview of the INSPIRE framework

INSPIRE extracts a 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 (Fig. 1). INSPIRE uses a standard iterative learning algorithm to optimize the joint log-likelihood objective function, Eq. (1), by iteratively updating its model parameters until convergence (see "Methods" for details). INSPIRE iterates the following three steps until convergence: (1) inferring the values of latent variables with all the other parameters held fixed, as described in Eq. (3); (2) assigning genes into latent variables as described in Eq. (4); and (3) learning a network of latent variables as described in Eq. (5). 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 Eq. (3). 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 (Fig. 1). 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 (Additional file 5: Figure S4). 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. 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) (see "Methods"). 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 would provide improved insights into the underlying disease process, as discussed below.

After evaluating INSPIRE by comparing with alternative methods on simulated data and a small set of genome-wide expression datasets (Fig. 2a, b), we applied INSPIRE to many ovarian cancer expression datasets, which lead to a novel marker and potential driver of tumor-associated stroma (Fig. 2c).

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 (Fig. 1). 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 (Additional file 2: Figure S1A, Fig. 3a (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 "Methods" for details of synthetic data generation).

We compare INSPIRE with the following five state-of-the-art methods: (1) GLasso, standard graphical lasso [54] that learns a gene-level conditional dependence network with no LDR or module assumption; (2) UGL, unknown group L_1 regularization [62] that learn sparse block-structured inverse covariance matrices with unknown block structure; (3) SLFA, structured latent factor analysis [22] that learn an LDR of the data as well as the relationship between the latent factors; (4) WGCNA, weighted gene co-expression network analysis [8] that allows to define modules based on a special metric derived from the correlations of the gene pairs; and (5) MGL, module graphical lasso [11] which simultaneously learns a 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 (Additional file 2: Figure S1B): (1) using only Dataset1 that contains all genes; (2) using data on the genes that are present in both datasets (blue-shaded region in Fig. 1), 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 run each method on 20 different instantiations of the synthetic data and present the average results with p values of significance of the difference with INSPIRE (see "Methods"; Fig. 3). 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. In order to make comparisons 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 rule out the possibility of making a conclusion based on a particular set of initial parameters, we performed the variants of those algorithms multiple times with different starting points (see "Methods" for details on initialization).

Test log-likelihood

The test log-likelihood that measures how well the learned models fit unseen data is a widely used evaluation metric on probabilistic models [11, 62, 63]. We generated test data Y containing 100 samples, which was created in the same way as the training data X (see "Methods"). The 13 learned models are tested based on the same unseen data Y. Each method selects its own regularization parameter using the standard CV test [64] selecting λ with the best average CV test log-likelihood measured on Dataset1 in X (see "Methods"). We used the test set of Dataset1 to compute the test log-likelihoods for all methods since Dataset1 contains all genes. Figure 3a 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 (Additional file 2: Figure S1A). Each bar (except INSPIRE) displays a p value from the Wilcoxon signed rank test that measures how significantly INSPIRE is better than the corresponding method across 20 instantiations of the data (see "Methods"). 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 eight 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 great potential to increase the chance to infer the true 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-likelihood of most methods increase with the increasing number of overlapping variables, from (i) to (iv).

Module recovery

We then evaluated how well important aspects of the true 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 index [65] 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 Fig. 3b, the module recovery performance of INSPIRE is significantly better than its 13 competitors. INSPIRE has a 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 ($p \leq 6.6 \times 10^{-2}$).

Module dependencies

Then, we evaluated 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 method [11]. 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 Fig. 3c, INSPIRE has the highest average F-measure that measures the accuracy of the dependencies learned by each method in (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 a lower average rand index (for modules) and F-measure (for 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 underlying modules and their dependencies better; INSPIRE is better than 13 alternative approaches in revealing the underlying model.

Evaluation on two genome-wide ovarian cancer expression datasets

Next, we evaluated INSPIRE based on the statistical robustness and biological relevance of the learned modules on two publicly available ovarian cancer gene expression datasets [31] (Fig. 4a): (1) OV1 that contains 18,113 genes and 28 patients (Affy U133 Plus 2.0 platform); and (2) OV2 that contains 8331 genes in a total of 42 patients (Affy U95Av2 platform) (see "Methods"; Additional file 6: Table S2).

We compared INSPIRE with six alternative methods that are scalable to genome-wide data (Table 1; Additional file 3: Figure S2). The runtime of all the other methods when $p = 3000$ is >10 h, which means that running these methods on genome-wide data would be too slow to be used. A total of 8234 genes are presented in both datasets (rows in the blue-shared region in Fig. 4a). As a preprocessing step, we standardized each dataset so that each gene has zero mean and unit variance across the samples within each dataset (See "Methods"). 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 (Additional file 7: Figure S5).

In the next three subsections, we show the results of the following evaluations (Fig. 2b): (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 database [66].

INSPIRE learns a statistically more robust LDR model than alternative approaches

We first evaluated the learned 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 fivefold CV tests (see "Methods"). 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 Fig. 4b, the best average test log-likelihood per sample across the tested λ values is plotted for each method. As can be seen in Fig. 4b, 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 ten runs of those methods with different initializations are represented by the error bars on the bottom panel in Fig. 4b.

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

INSPIRE uses a biologically intuitive 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 the 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 MSigDB [67]. 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 technique [64].

We considered 1077 GeneSets (pathways) from the C2 collection (curated gene sets from online pathway databases) of the current version of the MSigDB [67] based on Reactome [68], BioCarta, and KEGG [69]. 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 the Fisher's exact test p value, followed by the Bonferroni multiple hypothesis correction. Figure 4c and Additional file 7: Figure S5A show the 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 based on prior knowledge more accurately than the alternative methods.

INSPIRE modules are more significantly enriched for putative targets of the same TF than 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 database [66] 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 database [66], we computed the significance of the overlap between each module and each TF's putative targets from ChEA database measured by the Fisher's exact test p value followed by the Bonferroni correction. Figure 4d and Additional file 7: Figure S5B 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 Fig. 4d and Additional file 7: Figure S5B, 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 (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 messenger RNA (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.

Application to nine genome-wide ovarian cancer expression datasets

Encouraged by the in-depth evaluation described above, we applied INSPIRE to nine expression datasets comprising 1498 ovarian cancer patient samples downloaded from the TCGA project website and the Gene Expression Omnibus (GEO) [26] (Fig. 2c). This corpus of data consists of publically available transcriptomic characterizations of ovarian cancer across nine distinct studies where gene expression data collected in different studies come from distinct platforms. These data are therefore 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 study [23] and we anticipate that our approach will not only re-identify these subtypes based on the expression of our inferred modules, but will also further resolve potential molecular drivers of these subtypes through ancillary analyses of the INSPIRE inferred modules. These ancillary analyses are described below. We repeated our analyses for this application using varying module counts 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 the interpretation of a high-dimensional conditional dependence network.

We evaluated the learned LDR consisting of 90 modules and the corresponding latent variables based using three evaluation metrics:

(1) We performed gene set enrichment analysis to characterize each module based on its associated genes (see Additional file 8: Table S3 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 Additional file 8: Table S3), and we used inferred INSPIRE latent variables as features for predicting those phenotypes. Figure 5a shows the association between the learned latent variables with the six important phenotypes and Fig. 5b compares INSPIRE to the following based on the prediction of those phenotypes: (1) PCA [18], an unsupervised LDR method; (2) subnetwork analysis [13], a supervised LDR method; and (3) all genes when no LDR is learned. The histopathological phenotypes are provided as a resource for this paper (Additional file 9: Table S4) 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: (1) the subtypes recently described by the TCGA ovarian cancer study [23]; and (2) the subtypes learned by a method that uses mutation profiles for the network-based stratification of cancer patients (NBS) [60], 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 Additional file 1: Table S1 and the processing of the expression data is described in "Methods".

(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.

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 modular [49] 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 Additional file 10: Figure S6, we show scatter plots in which each dot corresponds to a GeneSet (from the pathway databases or TF binding information) and we plot the maximum −log10(p) obtained by each model (axis).

Additional file 10: Figure S6A (top) demonstrates that the modules that are strongly negatively correlated with each other show very distinct pathway (left) enrichment as well as TF binding enrichment (right). In Additional file 11: Table S5, the significance of enrichment from five negatively correlated module pairs with the biggest absolute correlation listed for five pathways or TFs for which the highest enrichment difference between the negatively correlated modules is observed.

We also compared between following two models in terms of functional enrichment of the modules: (1) two negatively correlated modules are defined as two separate modules as in the original work; and (2) instead of the two negatively correlated module, there is one hypothetical module that contains all genes in the two negatively correlated modules. Additional file 10: Figure S6A (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.

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 LDR of transcriptomic profiles across patients (Fig. 2c) 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 "Methods"; Additional file 9: Table S4). For each pair of a histologic phenotype and a latent variable from the INSPIRE model, we performed the Pearson's correlation test that produces a correlation coefficient and a p value. Additional file 8: Table S3 lists the p values from these association tests of INSPIRE latent variables, with each of the five histologic phenotypes. Figure 5a 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 Fig. 5a. 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 the Pearson's correlation test for tumor resectability.

Modules 5 and 6 show high correlations with the histopathological phenotypes, such as percent stroma, stroma type, and invasion pattern. As shown in Fig. 5a, 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 modules 5 and 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. Additional file 10: Figure S6B compares modules 5 and 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 modules 5 and 6 exhibit several unique biological properties. In Additional file 12: Table S6, 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 difference between modules 5 and 6 in terms of phenotypes associated with them, we compared 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): (1) modules 5 and 6 exist as two separate modules as in the original work; and (2) instead of modules 5 and 6, there is one hypothetical module that contains all genes in modules 5 and 6. As shown in Additional file 13: Table S7, 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 1 is a better predictor of all six phenotypes than model 2. 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.

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: (1) PCA [18], the most widely used unsupervised LDR method; (2) subnetwork analysis [13], a powerful supervised LDR method that extracts network markers; and (3) all genes when no LDR is learned. The subnetwork analysis method [13] learns small subnetworks 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 subnetworks of genes in a given PPI network such that the average expression level of each subnetwork 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, PCs, 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 “Methods” for details). The result of the comparison shows that the features learned by INSPIRE show the best prediction performance measured among all methods considered (Fig. 5b). This result strengthens our claim that the INSPIRE latent variables provide informative lower-dimensional features for prediction tasks.

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. In order 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. Additional file 14: Table S8 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.

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 study [23] (560 samples) into four subtypes by using the latent variables learned by the INSPIRE method as features for clustering patients (details in “Methods”). Additional file 15: Table S9 lists the assignment of the patients in the TCGA ovarian cancer data to the 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 with CNV of genes, an important genomic abnormality that can drive cancer (Fig. 5c and Additional file 16: Figure S7A). 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 mutations [70]. 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 Fig. 5c) 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 patients [60]; and (2) the subtypes inferred from a recent TCGA ovarian cancer study [23]. Figure 5c shows that INSPIRE results in subtypes that are more associated with CNV-based genomic abnormalities than alternative approaches. In Additional file 16: Figure S7A, 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 hyper-parameters. Figure 5c and Additional file 16: Figure S7A indicate that INSPIRE further resolves subtypes as defined by the potential genomic drivers of ovarian cancer when compared to alternative approaches. In Additional file 17: Supplementary Note 1, we list the CNV levels that are significantly correlated with each of the four subtypes. The enrichments of those CNV levels with the MSigDB [67] C2 (curated gene sets) categories and the corresponding $-\log_{10}p$ are also listed for each subtype.

Subtypes revealed by INSPIRE and their relationships with the TCGA subtypes

Figure 5d reveals a subnetwork learned by modules from an INSPIRE model using parameters $\lambda = 0.1$ 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 Fig. 5d. 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) algorithm [71] implemented in the R package *siggenes*. Additional file 18: Table S10 lists the enrichment of the marker modules with the MSigDB [67] C5 (GO gene sets) and the corresponding $-\log_{10}p$. We observed that the set of marker modules (Additional file 18: Table S10) 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 Fig. 5a and Additional file 8: Table S3 except module 30). Not surprisingly, the INSPIRE subtypes show diverse histologic features across subtypes, and we accordingly termed the INSPIRE subtypes "vascular," "stromal," "immunoresponsive," and "fibrous." See Fig. 5d and Additional file 18: Table S10, where the marker modules for the vascular, stromal, immunoresponsive, and fibrous subtypes are colored in green, blue, red, and orange, respectively.

Additional file 19: Table S11 shows a confusion matrix that describes the overlap between the INSPIRE subtype assignments and the TCGA subtype assignments [23] 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 findings of Way et al. [72]. 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 Fig. 5c). We further include the description of the stromal subtype here since it is characterized by the high expression of modules 5 and 6, which are strongly associated with the six important phenotypes in cancer (Fig. 5a). See Additional file 17: Supplementary Note 2 for the characterization of the other three ("vascular," "immunoresponsive," and "fibrous") subtypes.

The stromal subtype is characterized by high expression of modules 5, 6, and 86 (Fig. 5d) and associated increased percent stroma, infiltrative growth pattern, and desmoplastic stroma (Additional file 16: Figure S7B (i), (ii), (iii)). Modules 5 and 6 are significantly enriched for proteinaceous extracellular matrix gene sets (Additional file 18: Table S10), which is likely due to increased percent stroma. In Fig. 5d, 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 factors release from tumor stroma along with the vasculature itself [73]. This is supported by multipotent mesenchymal stromal cells having unique immunoregulatory and regenerative properties [74]. 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 potential [75]. We note that the immune system modules 78 and 81 that are connected to extracellular matrix modules 5 and 6 are also upregulated in the stromal subtype (Fig. 5d). Stromal subtype is a significant predictor of poor patient survival (Cox proportional hazards model logrank ($p = 8.8 \times 10^{-2}$) with a median survival of 914 days. Cancers associated with a reactive stroma is typically diagnostic of poor prognosis [76] 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 TCGA [23] (Additional file 19: Table S11).

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 cancer [77], 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 LOOCV across 1061 samples in eight datasets including the TCGA [23] and Tothill [34] datasets. *POSTN* is a member of module 6 that shows the most significant association with resectability among all 90 modules (Fig. 5a). 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 the *POSTN* expression. We observed that when training on TCGA data [23], 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 Tothill [34] dataset (see AUC values in Fig. 6). TCGA data [23] were used for training because of its large sample size. Tothill [34] was used for testing, because it has the largest sample size except TCGA data (Additional file 1: Table S1) 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 versus >0 cm) (see "Methods").

Since module 6 contains *POSTN*, outperformance of (1)–(3) means that the 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 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 Additional file 4: Figure S3). Modules 5 and 6, with strong stromal and mesenchymal properties (see below), provide potential novel molecular basis for tumor resectability.

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 Fig. 5a and Additional file 8: Table S3).

Module 5 contains known EMT inducers *ZEB1*, *SNAI2*, and *TCF4* (*E2.2*) [78], as well as multiple other genes known to be important in focal adhesion [79], extracellular matrix interaction [80], extracellular matrix organization [81], and markers of cancer-associated fibroblasts (*PDGFRB*, *PDGFRA*) [82] (see Additional file 8: Table S3). Similarly, module 6 contains EMT inducer *TWIST1* [78], many extracellular matrix genes, as well as genes associated with senescence and autophagy, collagen genes, and the well validated predictor of tumor resectability, *POSTN* [77] (see Additional file 8: Table S3). 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, they are likely to represent fairly different biological processes (Additional file 10: Figure S6B). When we combined these two modules and used one latent variable that represents the two modules, the overall prediction results became worse (Additional file 13: Table S7).

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 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) [83–86]. 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 *ZEB2* [78]. This indicates that module 78 may capture aspects of tumor associated inflammation, a known contributing factor to EMT [87]. Module 81 includes genes that regulate the MAPK and ERK cascades, signal transduction pathways that are known to be upstream of multiple oncogenic process [88]. Module 54 represents genes involved in pro-apoptotic and cell cycle regulation. *GADD45* genes, known to be upstream of JNK signaling [89], 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 inhibitors [90]. Modules 30 and 54 are indicative of the likely metabolic shift that cancers cells undergo as these modules are enriched in metabolic and biosynthesis pathways. When considering these modules jointly, we get a picture of multiple processes (Additional file 20: Figure S8) and potential tumor cell subpopulations that populate the tumor microenvironment and perpetuate aggressive tumor states in subpopulations 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.

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 a prognostic signature that generalizes 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 chance [91]. 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 [6, 34, 86, 92–94] from five cancer types: ovarian cancer, breast cancer, acute myeloid leukemia, glioblastoma, and lung cancer (see Additional file 21: Table S12 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 modules 5 and 6 are significantly associated with survival in at least three of the six datasets (Fisher's test statistic p value $= 1.68 \times 10^{-}$ for module 5 and $= 4.44 \times 10^{-8}$ 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 study [6] 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 module 5 or module 6 (Fig. 7a). As shown in the previous sections, module 5 (containing 183 genes) is highly associated with percent stroma (Fig. 5a), 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 11,119 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" [95]. Gentles et al. have defined a list of about 3000 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 impact [95]. This implies that *HOPX* could be a regulator in the stroma-associated processes.

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 [96–98]. It has been suggested to have tumor suppressive function in other cancer types [99–101], 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 state [102]. In our further analysis of *HOPX*, we observed that *HOPX* is one of the top candidate expression regulators for ovarian cancer [95, 103]. To understand how *HOPX* is associated with the genes in the tumor-associated stroma/mesenchymal module (module 5), we compared these genes with those downregulated in *Hopx*$^{-/-}$ mice compared to *Hopx*$^{+/-}$ control mice [102] 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. Additional file 22: Figure S9C shows the enrichment p value and the fold enrichment of the genes in the tumor-associated stroma/mesenchymal module with the downregulated genes in *Hopx*$^{-/-}$ mice for varying fold change of expression of the downregulated genes (x-axis).

Furthermore, genes downregulated in *Hopx*$^{-/-}$ mice after addition of XAV939, a potent inhibitor of Wnt

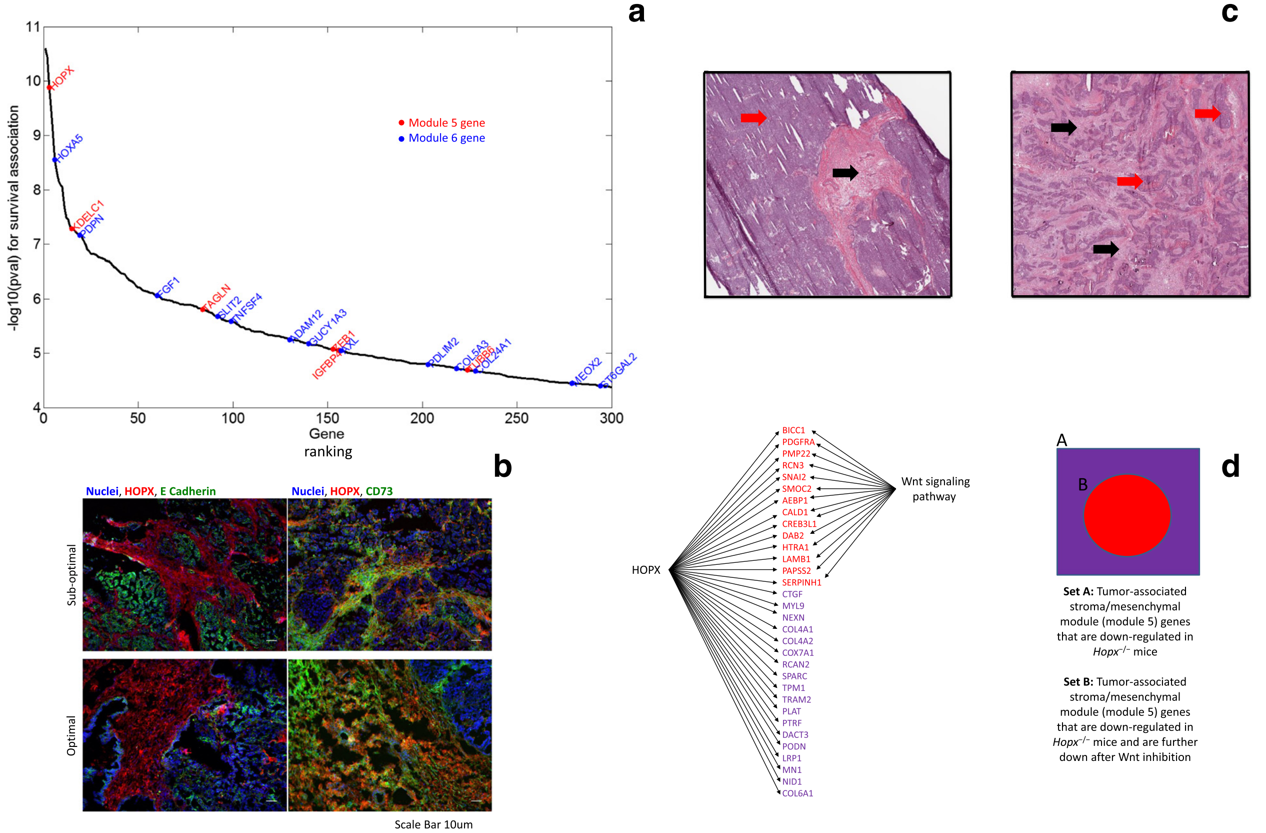

Fig. 7 **a** Fisher's combined *p* values for survival (*y-axis*) are shown for the top 300 genes (*x-axis*) which achieve the most significant survival association in the pan-cancer survival analysis. Module 5 genes are shown in *red* and module 6 genes are shown in *blue*. **b** Fluorescent staining of ovarian tumors from sub-optimally debulked 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 (**c**) *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 (H&E), 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*). H&E, 100X. **d** A total of 32 genes that are potential targets of *HOPX* are shown. The *purple-colored genes* are the potential targets whose expression does not depend on Wnt signaling and the *red-colored genes* are the potential targets which are downregulated in *Hopx*$^{-/-}$ mice and further down upon Wnt inhibition in *Hopx*$^{-/-}$ mice. It is highly likely that the expression of the *red-colored genes* is driven by both *HOPX* and Wnt signaling pathway

signaling to *Hopx*$^{-/-}$ mice [102], 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 inhibitor [102], 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 downregulated 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 mesenchymal stem cells (MSCs) [104, 105]. Additional file 22: Figure S9D shows the enrichment *p* value and the fold enrichment of the stroma/mesenchymal module genes that are downregulated in Wnt-inhibited *Hopx*$^{-/-}$ mice for varying fold change of expression of the downregulated genes (x-axis).

These results suggest that the genes in the tumor-associated stroma/mesenchymal module which are downregulated in both *Hopx*$^{-/-}$ mice and Wnt-inhibited *Hopx*$^{-/-}$ mice are good candidates as downstream targets of *HOPX*. Figure 7d shows those 32 potential targets of *HOPX*. The purple-colored genes are the potential targets that are downregulated in *Hopx*$^{-/-}$mice 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 potential targets of *HOPX* which are downregulated in *Hopx*$^{-/-}$mice and they are downregulated further upon Wnt inhibition ($|FC \text{ change}| \leq 0.93$). It is highly likely that the expression of the red-colored genes in Fig. 7d are driven by both *HOPX* and Wnt signaling pathway. We note that *HOPX* is, therefore, a potential driver for *SNAI2*, which is involved in EMT [106] and *AEBP1*, which is a stromal adipocyte enhancer-binding protein.

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 "Methods" for details). As shown in Fig. 7b, 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 MSC marker, as MSCs play an important role in the generation of cancer-associated fibroblasts and stroma [107]. 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 Additional file 22: Figures S9A and S9B.

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 [102, 108, 109]. Nonetheless, these results indicate *HOPX* as a putative novel marker for tumor-associated MSCs. In the patients with poor tumor resectability and prognosis, CD73 and *HOPX* expression is riddled throughout the tumor tissue (Fig. 7b). A typical patient with optimal resectability and low *HOPX* expression is shown on the left in Fig. 7c, 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.

Discussion

We propose the INSPIRE framework for learning a 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, comprising 1498 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 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 [96–98]. *HOPX* has recently emerged as a marker of numerous stem cell types [102, 108, 109]. 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 growth [110]. The tumor MSCs are known contributors to tumor-associated stroma via differentiation to cancer-associated fibroblasts (CAFs) [107] and may also promote metastasis [111]. *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 state [102], 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 MSCs.

INSPIRE is a general computational framework and can be applied to various diseases and different types of molecular data. For example, such as we applied it 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 apply INSPIRE to integrate microarray data. Since RNA-sequencing (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 method [112] 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 like 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 data [112].

INSPIRE provides a great, effective starting point to learn complex dependencies between genes, because we can learn a gene-level conditional dependence network by using for example the graphical lasso [54] algorithm within each module. There are several other potential next steps to improve technically on the proposed INSPIRE framework. One of those is to extend INSPIRE to the case where the latent network is not perfectly conserved across the 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 the mRNA expression levels of genes and the other dataset measures the protein levels. In this case, we will need to develop a novel method for discovering the 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.

Conclusions

In this work, 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 MSCs 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 the 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.

Additional files

Additional file 1: Table S1. The nine ovarian cancer gene expression datasets we used in the third set of experiments (biological application). (DOC 32 kb)

Additional file 2: Figure S1. 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 (“—1,” “Inter—,” and “Imp—,” respectively, from *top* to *bottom*). The first *illustration* corresponds to “–1” 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. (PDF 292 kb)

Additional file 3: Figure S2. 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*). (PDF 50 kb)

Additional file 4: Figure S3. *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. (PDF 139 kb)

Additional file 5: Figure S4. A 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-axis*) 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 $p < 0.05$ 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 top three modules with highest potential to contain that gene (*top*) and the one which assigns each gene to top five modules (*bottom*). For both INSPIRE and alternative approach, the results from multiple runs are shown on each plot. **B** 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-axis*) 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 <0.05 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 top three modules with highest potential to contain that gene (*top*) and the one which assigns each gene to top five modules (*bottom*). For both INSPIRE and alternative approach, the results from multiple runs are shown on each plot. (PDF 467 kb)

Additional file 6: Table S2. The two ovarian cancer gene expression datasets we used in the second set of experiments. (DOC 27 kb)

Additional file 7: Figure S5. A 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-axis*) 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 <0.05 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. **B** 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-axis*) 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 <0.05 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; *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. (PDF 738 kb)

Additional file 8: Table S3. 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 the 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 Fig. 5a, we highlight modules 5, 6, 53, 54, 60, 78, and 81 that are significantly correlated with at least three of the six phenotypes in red. We also highlight module 30 since it is the only module that has a significant correlation with the vessel formation phenotype. (XLSX 258 kb)

Additional file 9: Table S4. 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). (XLSX 12 kb)

Additional file 10: Figure S6. A $-\log_{10}p$ 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 nine-dataset experiment. Each one of the two negatively correlated modules is shown on one of the *x-axis* 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*). **B** $-\log_{10}p$ 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*). (PDF 194 kb)

Additional file 11: Table S5. 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. (XLSX 10 kb)

Additional file 12: Table S6. 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 modules 5 and 6. We show the five pathways or TFs that have the largest difference in the value of $-\log_{10}p$ between the two modules. (DOC 48 kb)

Additional file 13: Table S7. For module 5, module 6, and a hypothetical module containing all genes in modules 5 and 6, the prediction accuracy is compared in six prediction tasks via CV tests. The best performance for each prediction task is highlighted in green. (DOC 31 kb)

Additional file 14: Table S8. 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 CV tests. The best performance(s) for each prediction task is highlighted in green. (DOC 33 kb)

Additional file 15: Table S9. The assignment of the patients from TCGA ovarian cancer study [2] into four subtypes based on the INSPIRE latent variables. (XLSX 16 kb)

Additional file 16: Figure S7. A 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) [60] method (*green*); (2) the subtypes inferred from TCGA study [23] (*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 on which our biological analysis is based. **B** (1) 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; (2) 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; (3) 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; (4) 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. (PDF 379 kb)

Additional file 17: INSPIRE-SupplementaryInformation. Supplementary notes. (DOC 94 kb)

Additional file 18: Table S10. 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 Fisher's exact test. (XLSX 21 kb)

Additional file 19: Table S11. The confusion matrix representing the overlap between the INSPIRE subtypes and the subtypes revealed by the TCGA ovarian cancer study [23]. (DOC 99 kb)

Additional file 20: Figure S8. 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 Fig. 5a. The edges are shown by *black lines*. Also, as a recap of Fig. 5a 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 modules 5 and 6, which achieve top hits for a total of four phenotypes, are given as well. (PDF 248 kb)

Additional file 21: Table S12. The six gene expression datasets we used in our pan-cancer survival analysis. (DOC 30 kb)

Additional file 22: Figure S9. A, **B** Additional fluorescent images of ovarian tumors from sub-optimally debulked and optimally debulked patients. Each *row* is a patient. As in Fig. 6b, HOPX is localized to the stroma and does not overlap with E Cadherin positive cancer cells. HOPX does however overlap with CD73, a MSC marker. **C** The functional enrichment *p* value (i) and the fold enrichment (ii) of module 5 genes for the genes downregulated 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*. The *x-axes* represent the number of downregulated genes in *Hopx*-null mice whose expression fold-change passes the fold-change threshold displayed next to the corresponding value on the *curve*. **D** The functional enrichment *p* value (i) and the fold enrichment (ii) of module 5 genes for the genes downregulated 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*. The *x-axes* represent the number of downregulated 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*. (PDF 1484 kb)

Abbreviations

AUC, area under the curve; BIC, Bayesian information criterion; CAF, cancer-associated fibroblasts; ChEA, ChIP enrichment analysis; CI, concordance index; CNV, copy number variation; CV, cross-validation; EMT, epithelial-mesenchymal transition; GEO, gene expression omnibus; GGM, Gaussian graphical model; GISTIC, genomic identification of significant targets in cancer; GLasso, graphical lasso; INSPIRE, INferring Shared modules from multiPle gene expREssion datasets; LDR, low-dimensional representation; LOOCV, leave-one-out cross-validation; MAP, maximum a posteriori; MGL, module graphical lasso; MSC, mesenchymal stem cell; MSigDB, molecular signatures database; NBS, network-based stratification; OV, ovarian cancer; PC, principal component; PCA, principal component analysis; PPI, protein-protein interaction; ROC, receiver operator characteristic; SAM, significance analysis of microarrays; SLFA, structured latent factor analysis; TCGA, the cancer genome atlas; TF, transcription factor; TOM, topological overlap measure; UGL, unknown group L1 regularization; WGCNA, weighted gene co-expression network analysis

Funding

This work was supported by the National Institutes of Health (T32 HL 007312); National Science Foundation (DBI-1355899); American Cancer Society (127332-RSG-15-097-01-TBG); Solid Tumor Translational Research; American Association of University Women (AAUW) International Fellowship; and the Ford Foundation Fellowship.

Availability of data and materials

INSPIRE is freely available as an R package in the CRAN repository. The processed expression data used in the study, the inferred INSPIRE models, histopathologic features of 100 TCGA H&E stained images, and the results of our immunohistochemistry staining experiments are available on our website [24].

Authors' contributions

SC, BAL, and SIL designed the statistical methods and various analyses. SC developed the algorithm and performed computational experiments. SC, BAL, SB, MR, RDH, and SIL wrote the manuscript. RDH and SB designed and performed the immunohistochemistry experiment. MR extracted histopathologic features from the 100 H&E stained ovarian tumor section images obtained through TCGA. CD interpreted the results on the molecular basis for resectability. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

All input data of INSPIRE are publicly available through the Gene Expression Omnibus (GEO) web page [113] and The Cancer Genome Atlas (TCGA). For experimental validation on *HOPX*, we used frozen tissue slices fixed to glass slides from ten patients with ovarian cancer for immunohistochemistry staining. The tumor tissue and associated clinical variables were obtained from an institutional tumor bank, which prospectively collected specimens and clinical information for subjects who provided informed consent under an IRB-approved protocol (University of Washington IRB 27077).

Author details

[1]Department of Computer Science & Engineering, University of Washington, Seattle, WA, USA. [2]Sage Bionetworks, Seattle, WA, USA. [3]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [4]Translational Research Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [5]Department of Anatomic Pathology, University of Washington, Seattle, WA, USA. [6]Medical Genetics, Department of Medicine, University of Washington, Seattle, WA, USA.

Received: 13 January 2016 Accepted: 18 May 2016
Published online: 10 June 2016

References

1. Unsupervised Feature Learning and Deep Learning Tutorial: http://deeplearning.stanford.edu/tutorial/.
2. Längkvist M, Karlsson L, Loutfi A. A review of unsupervised feature learning and deep learning for time-series modeling. Pattern Recognit Lett. 2014;42:11–24.
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
4. Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection. Adv Neural Inf Process Syst. 2013;26:2553–61.
5. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86:2278–324.
6. Margolin AA, Bilal E, Huang E, Norman TC, Ottestad L, Mecham BH, et al. Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. Sci Transl Med. 2013;5:181re1.
7. Cheng WY, Yang THO, Anastassiou D. Development of a prognostic model for breast cancer survival in an open challenge environment. Sci Transl Med. 2013;5:181ra50.
8. Langfelder P, Horvath S. WGCNA: an R package for weighted gene co-expression network analysis. BMC Bioinformatics. 2008;9:559.
9. Sherlock G. Analysis of large-scale gene expression data. Brief Bioinform. 2001;2:350–62.
10. Lee SI, Batzoglou S. ICA-based clustering of genes from microarray expression data. Adv Neural Inf Process Syst. 2004;16:675–82.
11. Celik S, Logsdon BA, Lee S-I. Efficient dimensionality reduction for high-dimensional network estimation. Proc of the 31st International Conference on Machine Learning. 2014;31:1953–61.
12. Koller D, Friedman N. Probabilistic graphical models: principles and techniques. Harvard, MA: MIT Press; 2009.
13. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol. 2007;3:140.
14. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. PLoS Comput Biol. 2008;4, e1000217.
15. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, et al. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27:199–204.

16. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010; 140:744–52.
17. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. 2007;8:R76.
18. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:417–41.
19. Lee S-I, Batzoglou S. Application of independent component analysis to microarrays. Genome Biol. 2003;4:R76.
20. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: A survey. IEEE Trans Knowl Data Eng. 2004;16:1370–86.
21. Chandrasekaran V, Parrilo PA, Willsky AS. Latent variable graphical model selection via convex optimization. Ann Stat. 2012;40:1935–67.
22. He Y, Qi Y, Kavukcuoglu K, Park H. Learning the dependency structure of latent factors. Adv Neural Inf Process Syst. 2012; 25:2366–74.
23. Bell D, Berchuck A, Birrer M, Chien J, Cramer DW, Dao F, et al. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474:609–15.
24. INSPIRE web page: http://inspire.cs.washington.edu.
25. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: Archive for functional genomics data sets - Update. Nucleic Acids Res. 2013;41:D991–5.
27. Denkert C, Budczies J, Darb-Esfahani S, Györffy B, Sehouli J, Könsgen D, et al. A prognostic gene expression index in ovarian cancer - Validation across different independent data sets. J Pathol. 2009;218:273–80.
28. Bonome T, Levine DA, Shih J, Randonovich M, Pise-Masison CA, Bogomolniy F, et al. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. Cancer Res. 2008;68:5478–86.
29. Hendrix ND, Wu R, Kuick R, Schwartz DR, Fearon ER, Cho KR. Fibroblast growth factor 9 has oncogenic activity and is a downstream target of Wnt signaling in ovarian endometrioid adenocarcinomas. Cancer Res. 2006;66:1354–62.
30. Mok SC, Bonome T, Vathipadiekal V, Bell A, Johnson ME, Wong KK, et al. A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2. Cancer Cell. 2009;16:521–32.
31. Konstantinopoulos PA, Spentzos D, Karlan BY, Taniguchi T, Fountzilas E, Francoeur N, 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. 2010;28:3555–61.
32. Meyniel J-P, Cottu PH, Decraene C, Stern M-H, Couturier J, Lebigot I, 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. 2010;10:222.
33. Ferriss JS, Kim Y, Duska L, Birrer M, Levine DA, Moskaluk C, Theodorescu D, Lee JK. Multi-gene expression predictors of single drug responses to adjuvant chemotherapy in ovarian carcinoma: Predicting platinum resistance. PLoS One. 2012;7, e30550.
34. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, et al. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res. 2008;14:5198–208.
35. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy - Analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20:307–15.
36. Maglott D, Ostell J, Pruitt KD, Tatusova T. Entrez gene: Gene-centered information at NCBI. Nucleic Acids Res. 2011;39:D52–7.
37. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res. 2005;33, e175.
38. Tibshirani R. Regression selection and shrinkage via the Lasso. J R Stat Soc B. 1994;58:267–88.
39. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, et al. Strong rules for discarding predictors in lasso-type problems. J R Stat Soc Ser B Stat Methodol. 2012;74:245–66.
40. Marquardt DW, Snee RD. Ridge regression in practice. Source Am Stat. 1975;29:3–20.
41. Sardy S. On the practice of rescaling covariates. Int Stat Rev. 2008;76:285–97.
42. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MPL. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193:327–45.
43. Schmidt M, Niculescu-Mizil A, Murphy KP. Learning graphical model structure using L1-regularization paths. Proc AAAI Conf Artif Intell. 2007;22:1278.
44. Mu B, How JP. Learning Sparse Gaussian Graphical Model with l0 -regularization. Tech Rep. 2014; 1–13.
45. Friedman J, Hastie T, Tibshirani R. Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Tech. Rep. 2010; 1–22
46. Lee S-I, Pe'er D, Dudley AM, Church GM, Koller D. Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. Proc Natl Acad Sci U S A. 2006;103:14062–7.
47. Lee SI, Dudley AM, Drubin D, Silver PA, Krogan NJ, Pe'er D, et al. Learning a prior on regulatory potential from eQTL data. PLoS Genet. 2009;5, e1000358.
48. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, et al. An integrated approach to uncover drivers of cancer. Cell. 2010;143:1005–17.
49. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, et al. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003;34:166–76.
50. cBio Cancer Genomics Portal: http://cbioportal.org.
51. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011;12:R41.
52. Mardia KV, Kent JT, Bibby JM. Multivariate analysis. Washington, DC: Academic Press; 1979.
53. Lauritzen SL. Graphical models. Oxford: Oxford University Press; 1996.
54. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–41.
55. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. J Optim Theory Appl. 2001;109:475–94.
56. Josse J, Chavent M, Liquet B, Husson F. Handling missing values with regularized iterative multiple correspondence analysis. J Classif. 2012;29:91–116.
57. Witten DM, Friedman JH, Simon N. New insights and faster computations for the graphical lasso. J Comput Graph Stat. 2011;20:892–900.
58. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B. 1996;58:267–88.
59. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
60. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. Nat Methods. 2013;10:1108–15.
61. ImageJ: http://imagej.nih.gov/ij/.
62. Marlin BM, Murphy K. Sparse gaussian graphical models with unknown block structure. Proc of the 26th International Conference on Machine Learning. 2009;26:705–712.
63. Duchi J, Gould S. Projected subgradient methods for learning sparse gaussians. Proc of the Twenty-Fourth Conf on Uncertainty in Artificial Intelligence. 2008; 153–60.
64. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer New York Inc; 2001.
65. Rand WM. Objective criteria for the evaluation of clustering methods. J Am Stat Assoc. 1971;66:846–50.
66. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics. 2010;26:2438–44.
67. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
68. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007;8:R39.
69. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32:D277–80.
70. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nat Genet. 2013;45:1127–33.
71. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc. 2001;96:1151–60.
72. Way GP, Rudd J, Wang C, Hamidi H, Fridley BL, Konecny G, et al. High-grade serous ovarian cancer subtypes are similar across populations. Biorxiv. http://dx.doi.org/10.1101/030239.

73. Wels J, Kaplan RN, Rafii S, Lyden D. Migratory neighbors and distant invaders: tumor-associated niche cells. Genes Dev. 2008;22:559–74.
74. Le Blanc K, Mougiakakos D. Multipotent mesenchymal stromal cells and the innate immune system. Nat Rev Immunol. 2012;12:383–96.
75. Heuvers ME, Aerts JG, Cornelissen R, Groen H, Hoogsteden HC, Hegmans JP. Patient-tailored modulation of the immune system may revolutionize future lung cancer treatment. BMC Cancer. 2012;12:580.
76. Liu H, Ma Q, Xu Q, Lei J, Li X, Wang Z, et al. Therapeutic potential of perineural invasion, hypoxia and desmoplasia in pancreatic cancer. Curr Pharm Des. 2012;18:2395–403.
77. Riester M, Wei W, Waldron L, Culhane AC, Trippa L, Oliva E, et al. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. J Natl Cancer Inst. 2014;106:dju048.
78. Puisieux A, Brabletz T, Caramel J. Oncogenic roles of EMT-inducing transcription factors. Nat Cell Biol. 2014;16:488–94.
79. Ilić D, Furuta Y, Kanazawa S, Takeda N, Sobue K, Nakatsuji N, et al. Reduced cell motility and enhanced focal adhesion contact formation in cells from FAK-deficient mice. Nature. 1995;377:539–44.
80. Chautard E, Fatoux-Ardore M, Ballut L, Thierry-Mieg N, Ricard-Blum S. MatrixDB, the extracellular matrix interaction database. Nucleic Acids Res. 2011;39:D235–40.
81. Barker TH, Baneyx G, Cardó-Vila M, Workman GA, Weaver M, Menon PM, et al. SPARC regulates extracellular matrix organization through its modulation of integrin-linked kinase activity. J Biol Chem. 2005;280:36483–93.
82. Dong J, Grunstein J, Tejada M, Peale F, Frantz G, Liang W-C, et al. VEGF-null cells require PDGFR alpha signaling-mediated stromal fibroblast recruitment for tumorigenesis. EMBO J. 2004;23:2800–10.
83. Singh A, Settleman J. EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer. Oncogene. 2010;29:4741–51.
84. Seton-Rogers S. Layers of regulation. Nat Rev Cancer. 2011;11:689.
85. Yang J, Weinberg RA. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. Dev Cell. 2008;14:818–29.
86. Gentles AJ, Plevritis SK, Majeti R, Alizadeh AA. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. JAMA. 2010;304:2706–15.
87. López-Novoa JM, Nieto MA. Inflammation and EMT: an alliance towards organ fibrosis and cancer progression. EMBO Mol Med. 2009;1:303–14.
88. Steinmetz R, Wagoner HA, Zeng P, Hammond JR, Hannon TS, Meyers JL, 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. 2004;18:2570–82.
89. Tamura RE, de Vasconcellos JF, Sarkar D, Libermann TA, Fisher PB, Zerbini LF. GADD45 proteins: central players in tumorigenesis. Curr Mol Med. 2012;12:634–51.
90. Tetreault M-P, Yang Y, Katz JP. Krüppel-like factors in cancer. Nat Rev Cancer. 2013;13:701–13.
91. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015;21:938–45.
92. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8.
93. Gravendeel LAM, Kouwenhoven MCM, Gevaert O, de Rooi JJ, Stubbs AP, Duijm JE, et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. Cancer Res. 2009;69:9065–72.
94. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. Nature. 2014;511:543–50.
95. Gentles AJ, Alizadeh AA, Lee SI, Myklebust JH, Shachaf CM, Shahbaba B, et al. A pluripotency signature predicts histologic transformation and influences survival in follicular lymphoma patients. Blood. 2009;114:3158–66.
96. Chen F, Kook H, Milewski R, Gitler AD, Lu MM, Li J, et al. Hop is an unusual homeobox gene that modulates cardiac development. Cell. 2002;110:713–23.
97. Kee HJ, Kim J-R, Nam K-I, Park HY, Shin S, Kim JC, et al. Enhancer of polycomb1, a novel homeodomain only protein-binding partner, induces skeletal muscle differentiation. J Biol Chem. 2007;282:7700–9.
98. Kook H, Lepore JJ, Gitler AD, Lu MM, Yung WWM, Mackay J, et al. Cardiac hypertrophy and histone deacetylase-dependent transcriptional repression mediated by the atypical homeodomain protein Hop. J Clin Invest. 2003;112:863–71.
99. Katoh H, Yamashita K, Waraya M, Margalit O, Ooki A, Tamaki H, et al. Epigenetic silencing of HOPX promotes cancer progression in colorectal cancer. Neoplasia. 2012;14:559–IN6.
100. Waraya M, Yamashita K, Katoh H, Ooki A, Kawamata H, Nishimiya H, et al. Cancer specific promoter CpG Islands hypermethylation of HOP homeobox (HOPX) gene and its potential tumor suppressive role in pancreatic carcinogenesis. BMC Cancer. 2012;12:397.
101. 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. 2015;235:397–407.
102. Jain R, Li D, Gupta M, Manderfield LJ, Ifkovits JL, Wang Q, et al. Integration of Bmp and Wnt signaling by Hopx specifies commitment of cardiomyoblasts. Science. 2015;348:aaa6071–1.
103. Logsdon BA, Gentles AJ, Miller CP, Blau CA, Becker PS, Lee SI. Sparse expression bases in cancer reveal tumor drivers. Nucleic Acids Res. 2015;43:1332–44.
104. 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. 2015;14:356–68.
105. Macheda ML, Stacker SA. Importance of Wnt signaling in the tumor stroma microenvironment. Curr Cancer Drug Targets. 2008;8:454–65.
106. Savagner P, Yamada KM, Thiery JP. The zinc-finger protein slug causes desmosome dissociation, an initial and necessary step for growth factor-induced epithelial-mesenchymal transition. J Cell Biol. 1997;137:1403–19.
107. Mishra PJ, Mishra PJ, Humeniuk R, Medina DJ, Alexe G, Mesirov JP, et al. Carcinoma-associated fibroblast-like differentiation of human mesenchymal stem cells. Cancer Res. 2008;68:4331–9.
108. Li N, Yousefi M, Nakauka-Ddamba A, Jain R, Tobias J, Epstein JA, et al. Single-cell analysis of proxy reporter allele-marked epithelial cells establishes intestinal stem cell hierarchy. Stem Cell Rep. 2014;3:876–91.
109. Jain R, Barkauskas CE, Takeda N, Bowie EJ, Aghajanian H, Wang Q, et al. Plasticity of Hopx(+) type I alveolar cells to regenerate type II cells in the lung. Nat Commun. 2015;6:6727.
110. Ma S, Xie N, Li W, Yuan B, Shi Y, Wang Y. Immunobiology of mesenchymal stem cells. Cell Death Differ. 2014;21:216–25.
111. Karnoub AE, Dash AB, Vo AP, Sullivan A, Brooks MW, Bell GW, et al. Mesenchymal stem cells within tumour stroma promote breast cancer metastasis. Nature. 2007;449:557–63.
112. Law CW, Chen Y, Shi W, voom Smyth GK. Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
113. Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo/.