molecules

MDPI

Article

The Integrative Method Based on the Module-Network for Identifying Driver Genes in Cancer Subtypes

Xinguo Lu [1,*], Xing Li [1], Ping Liu [2], Xin Qian [1], Qiumai Miao [1] and Shaoliang Peng [1,3,*]

1 College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; xingleo@hnu.edu.cn (X.L.); qianxin@hnu.edu.cn (X.Q.); qiumaimiao@hnu.edu.cn (Q.M.)
2 Hunan Want Want Hospital, Changsha 410006, China; lp-simple123@126.com
3 School of Computer Science, National University of Defense Technology, Changsha 410073, China
* Correspondence: hnluxinguo@hnu.edu.cn (X.L.); pengshaoliang@nudt.edu.cn (S.P.); Tel.: +86-731-88821907(X.L.)

Received: 7 November 2017; Accepted: 8 January 2018; Published: 24 January 2018

Abstract: With advances in next-generation sequencing(NGS) technologies, a large number of multiple types of high-throughput genomics data are available. A great challenge in exploring cancer progression is to identify the driver genes from the variant genes by analyzing and integrating multi-types genomics data. Breast cancer is known as a heterogeneous disease. The identification of subtype-specific driver genes is critical to guide the diagnosis, assessment of prognosis and treatment of breast cancer. We developed an integrated frame based on gene expression profiles and copy number variation (CNV) data to identify breast cancer subtype-specific driver genes. In this frame, we employed statistical machine-learning method to select gene subsets and utilized an module-network analysis method to identify potential candidate driver genes. The final subtype-specific driver genes were acquired by paired-wise comparison in subtypes. To validate specificity of the driver genes, the gene expression data of these genes were applied to classify the patient samples with 10-fold cross validation and the enrichment analysis were also conducted on the identified driver genes. The experimental results show that the proposed integrative method can identify the potential driver genes and the classifier with these genes acquired better performance than with genes identified by other methods.

Keywords: integrative analysis; module network; cancer subtypes; breast cancer; copy number variation; gene expression

1. Introduction

Breast cancer is one of the most common malignant tumors in women. The incidence rate is 7–10% of all kinds malignant tumors which is usually associated with genetic alterations [1]. Breast cancer has been categorized into five subtypes, including luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), basal-like (Basal), and normal-like (Normal) types. Previous studies have shown that each cancer subtype has its own gene imprint and tumor markers, and genetic variation will increase the risk of cancer. However, not all of the aberrations have the same impact on tumor progression. To understand the mechanism of cancer, identifying driver genes from genomic aberrations has become the focus of research. Meanwhile, the gene expression profiles play an important role in understanding the pathogenesis of disease. The gene expression profiles provide information about their activity level. Activation or deactivation of the functional parts of a genome, or genes, determines the pathological states and development of a disease. The over-expression of an oncogene or under-expression of a tumor suppressor gene also has a certain influence on cancer process. So it is reasonable to believe that an over/under-expressed gene has a footprint in a genome in the form of an aberration that can

be used as a biomarker [2–4]. Besides, with the availability of huge amounts of multiple genomics data, the comprehensive analysis across the different genomics data will improve the understanding of the role of biomarkers in breast cancer pathogenesis and procession [5].

With the development of high-throughput sequencing technologies, huge volumes of diseased-based histological data have been provided in life science research. These data are publicly available in databases, such as The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) in which several high throughput genomic data types for hundreds of sample on tens of cancer types have generated. Recently, many methods which integrated multi-types of genomic data have been developed to reveal combinatorial patterns and discover new biological mechanisms [6]. Zhang et al. developed an unbiased adaptive clustering approach to integrate and analyze the multi-types of genomics data for ovarian cancer, including genome-wide gene expression, DNA methylation, microRNA expression, and copy number alteration profiles. And they developed an algorithm to uncover molecular signatures that distinguish cancer subtypes [7]. Huang et al. proposed a multiple regression based method to construct an integrative network with gene expression, microRNA, methylation and copy number variation [8]. In addition, to explore the important role of genomics aberrations and gene expression profiles in disease progression, some studies have also committed to discover novel candidate driver genes by integrating gene expression and CNV data. Li et al. identified the breast cancer subtype-specific drivers by integrating and analyzing the copy number aberrations data and miRNA-mRNA dual expression profiling data [9]. However, these proposed methods are based on linear models, such as regression analysis and correlation analysis, which is not suitable for heterogeneous data that have extremely high within-group variations. As we know that there is a significant heterogeneity in breast cancer data. We expect this limitation of linear approaches can be solved for the heterogeneous data.

To this end, we present a novel computational framework using module network analysis [10] to identify the breast cancer subtype-specific driver genes by integrating the gene expression and CNV data. In this framework, we first selected the subsets of gene expression and CNV data with the differential analysis. Then, a module network is constructed as a form of Bayesian network in which the similarly behaving variables are clustered into modules and the same parents and parameters are learned for each module, instead of each variable. A module is defined as a set of co-expressed or co-regulated genes that share a common statistical model. Through the process of module network learning, we obtained the candidate drivers of each subtype. The final subtype-specific driver genes were acquired by comparing paired-wise subtypes. The classification algorithm is used to validate whether subtype-specific driver genes can distinguish subtypes as well. To better understand the underlying biological significance, the pathway impact analysis and functional analysis is applied to explicate regulation mechanism of these driver genes. The results have shown that the proposed method is able to detect highly mutated gene as subtype-specific driver genes and the identified driver genes can classify breast cancer subtypes as well.

2. Results

We downloaded the clinical records and five breast cancer subtypes of high-throughput data consisting of initial 825 patients from TCGA. To increase to the statistic power, we take a filtering strategy to ensure that each sample both has gene expression and CNV data for analysis. And this process resulted in 485 samples. For each sample, the gene expression data includes expression level of 17,268 genes and the CNV values of 20,871 genes is obtained.

Our experiment is constituted by three parts: (1) identifying the subtype-specific driver genes using the proposed integrative method; (2) comparing the classification performance. The classifiers are constructed by these driver genes. And the performance is compared with the information gain, Chi-squared and lemon-tree methods; (3) analyzing biological significance of the obtained driver genes, including topology-based pathway analysis, Gene ontology (GO) functional enrichment, KEGG pathway enrichment analysis.

For gene expression and CNV data, the difference between each subtype and other three subtypes was analyzed. In gene expression data,we selected genes with q-values < 0.1 which represented the gene was differentially expressed between subtypes. And for CNV data, we firstly selected genes with their q-values < 0.1. Then we calculated the frequency of amplification and deletion for each gene in each of the two subtypes samples and selected genes for which the difference of frequencies between two subtypes is more than 20%. Additionally, we used a threshold of log2 copy number variation ratio of $0.3/-0.3$ to call amplified/deleted genes. And the gene subsets are converged and the relevant subset of data is selected for each subtype. We selected the genes as candidate modulators with mutation frequency fre > 0.6. We compiled a list of 965 candidate modulators for the HER2, 336 candidate modulators for lumA, 1213 candidate modulators for lumB, 815 candidate modulators for basal.

Through data preprocessing above, module-network analysis is used to select candidate driver genes. The whole learning algorithm is performed 100 times. We filtered the set of candidate modulators and left only genes that appeared in at least one regulation program in at least 40% of the runs. These modulators are considered as candidate driver genes. Thus, the procedure resulted in 148 modules with 9 candidate drivers for HER2 samples, 550 modules with 28 candidate drivers for lumA samples, 258 modules with 12 candidate drivers for lumB samples, 201 modules with 14 candidate drivers for basal samples. Several modules shared the same modulators in each subtypes of samples. Interestingly, the result of module network analysis for HER2 and other three subtypes represented that they shared three genes. Finally, the candidate driver genes in this subtype, but not those in other subtypes, are the subtype-specific driver genes, including 8 Her2 drivers, 11 Basal drivers, 28 LumA drivers and 10 LumB drivers.

The effectiveness of our approach was compared against two other state-of-art model-based algorithm, Lemon-tree [11] and NetNorM [12], in which the breast cancer datasets is utilize to analyze. As shown in Figure 1, the 57 driver genes are identified by Module-network, while Lemon-tree and NetNorM identified 48 and 60 drivers respectively. There are two driver genes identified by three methods, ATP1A2 and IFI16 respectively. In addition, the 21 drivers were identified by methods of Module-network and Lemon-tree together, and only 4 driver genes are selected by methods of Module-network and NerNorM together. This may be because NetNorM method is based on the mutation data and gene-gene interaction data, while the other two methods integrated the mutation data and gene expression data. The results suggest that the integrative approach can recognize the driver genes for each subtypes.

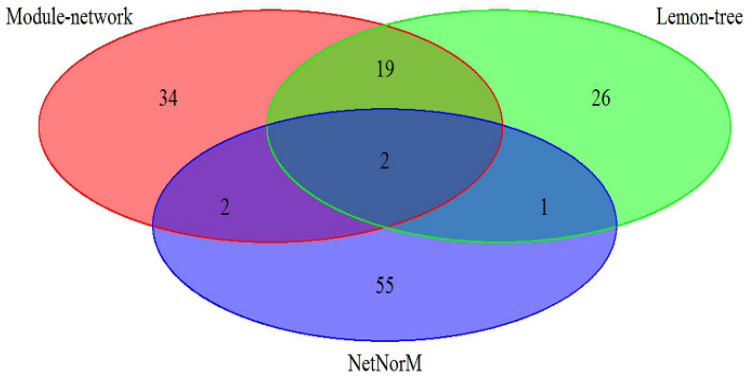

Figure 1. A comparison of identified cancer drivers between module-network and two state of the integrative methods on breast cancer subtypes datasets.

2.1. Identified Novel Subtype-Specific Driver Genes in LumA

In LumA , we identified 28 candidate driver genes which are frequently mutated. A heatmap of the 28 significantly differentially expressed genes of LumA is shown in Figure 2 which represented

that expression profiles can be clearly clustered into four subtypes by using the selected driver genes. In addition, we found that some samples were mixed between the LumA and the LumB in the process of clustering. We supposed that LumA and LumB belong to the luminal, and luminal subtype has the lowest overall mutation rate.

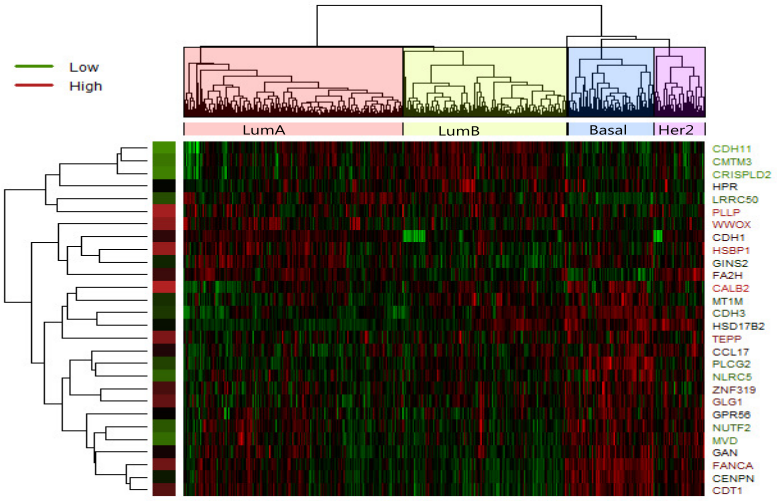

Figure 2. Heatmap of expression values of 28 most significant differently expressed gene in LumA-subtype. Clustering method on expression values was used to generate the heatmap. There are clear clusters of genes for the four tumor subtypes.

A novel significant driver of the subtype is a frequently mutated gene in specific subtype that has not been classified as a driver gene. Here we defined novel significant driver genes which satisfy the following requirements: (1) subtype-specific driver genes recurrently mutated in multiple subtypes. The mutation frequency of the driver genes is different from other three subtypes, and their difference is more than 0.2; (2) not previously classified as a driver by CGC database [13]. In LumA, we found 23 novel significant driver genes which recursively mutated. We sorted them according to mutation frequency and the mutation frequency of top 12 driver genes in LumA_subtype is shown in Figure 3a. The mutation frequency of 12 driver genes identified in LumA is more than 0.7. Moreover, the proportion of mutation samples in each breast cancer subtype was shown in Figure 3b. The proportion of LumA is obviously higher than the other three cancer subtypes.

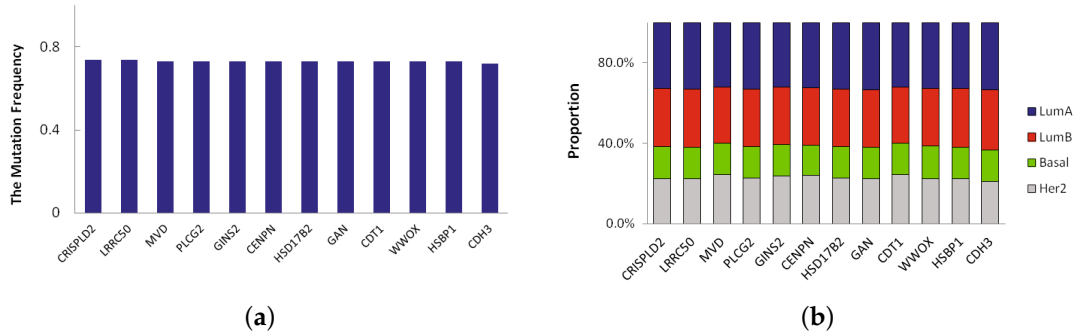

Figure 3. (**a**) The top 12 mutation frequency of driver genes in LumA; (**b**) The mutation proportion of the top 12 genes in each breast cancer subtype samples .

Of the 23 driver genes, CDH3, GLG1, CCL17 and PLCG2 are the most promising. These genes are significantly mutated and are involved in several cancer functions and pathways. CDH3 has the highest

mutation frequency and is mutated in 72% of LumA samples. CDH3 belongs to the family of classic cadherins that are engaged in various cellular activities including motility, invasion, and signaling of tumor cells, in addition to cell adhesion [14]. GLG1 is a key ligand-receptor in the early response of cells [15]. The mutation frequency of CCL17 in LumA samples is 0.7. CCL17 is a central regulator of Treg homeostasis, and CCL17 might be a target for vascular therapy [16]. A CCL17 gene is a candidate as one of the genetic factors in some allergic diseases [17]. PLCG2 encodes phospholipase $C_{\gamma 2}$($PLC_{\gamma 2}$), an enzyme with a critical regulatory role in various immune and inflammatory pathways, and plays a key role in the regulation of immune response [18]. PLAID-associated deletions of PLCG2 cause diminished receptor-mediated activity at physiologic temperatures in B cells and natural killer cells with enhanced spontaneous signaling in mast cells and B cells at subphysiologic temperatures [19].

2.2. Identified Novel Subtype-Specific Driver Genes in Basal/LumB/Her2

Similarly, we can also identify the subtype-specific driver genes of other three subtypes. In Basal, there are several novel significant driver genes and two of these are strong novel drivers: FDPS and ATP1A2. FDPS is a key enzyme in the isoprenoid pathway responsible for cholesterol biosynthesis, post-translational protein modifications and synthesis of steroid hormones, whose expression is regulated by phorbol esters and polyunsaturated fatty acids [20]. FDPS is necessary for osteoclast survival and activity and is considered as a major molecular target of aminobisphosphonates [21]. The ATP1A2 gene encodes the $\alpha 2$ subunit of Na^+-, K^+-ATPase, a plasma membrane enzyme that counter transports Na^+ and K^+ across cell membranes [22]. ATP1A2 gene mutation result in degeneration of the amygdala and pyriform cortex [23].

In LumB, two potential novel drivers is FASLG and RGS2. Alteration of FASLG pathway regulating cell death may lead to cancer development [24]. Studies have revealed that increased FASLG expression facilitate development and progression of tumors, including gastric cancer. These results suggest that variants of the FASLG gene is likely to be associated with the initiation and development of gastric cancer [25]. The mutation frequency of FASLG is about 80% in LumB. RGS2 is a member of a family of proteins that negatively modulate G-protein coupled receptor transmission. Variations in the RGS2 gene were found to be associated in humans with anxious and depressive phenotypes [26]. RGS2 is mutated in 78% Basal samples.

In Her2, there are fewer candidate driver genes than Luma, however, we detected two genes that have powerful potential to become novel drivers: ZFPM2 and EGLN1. ZFPM2 protein is an important cofactor for GATA family of transcription factors. In adult tissues, the ZFPM2 protein of 1151 amino acids is expressed predominantly in brain, heart, and testis. ZFPM2 may act as repressor or activator, depending on specific promoter and cell type [27]. EGLN1 is a key oxygen sensor gene that negatively regulates the activity of hypoxia-in-ducible factor (HIF-1A). Owing to its important function as an oxygen sensor, EGLN1 is relevant to the human hypoxic response, both at high altitude in hypoxic conditions or in cellular hypoxia [28]. EGLN1 is an important gene functioning at the upstream of the HIF pathway and showed consistent selective signals across multiple studies [29].

2.3. Validation of Classification between Subtypes

To validate whether our subtype-specific driver genes obtained from integration analysis are also applicable to distinguish subtypes, we classified the samples between subtypes using SVM [30]. SVM is a linear maximum-margin model for classification [31]. Here, we applied the gene expression profiles of 216 samples of LumA, 92 samples of Basal, 122 samples of LumB and 55 samples of Her2. We selected half of the samples of each subtype separately to train the classifier, and the rest of the samples were used to test the classification performance. In addition, the classification performance is compared with the other two key gene selection methods: Information gain and Chi-squared. These two methods are simply based on gene expression data but not considered the mutation data. Due to our method integrated the gene expression data and mutation data, we also compared the classification performance with another integrative method, called Lemon-tree.

To evaluate the performance of the model, we used accuracy measure. Accuracy can be computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

where *TP*, *TN*, *FP* and *FN* are true positive, true negative, false positive and false negative respectively. And *F-measure* uses both *Precision* and *Recall* measures to compute the score as follows:

$$F\text{-}measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2}$$

where

$$\begin{aligned} Precision &= \frac{TP}{TP + FP} \\ Recall &= \frac{TP}{TP + FN} \end{aligned} \tag{3}$$

In this study, we used the 10-fold cross-validation to assess the model. As shown in Tables 1–3, the accuracy, recall and F-measure of prediction of the our method is up to 98.82%, 0.976 and 0.964 respectively between Basal and other subtypes. For other three methods, the integrative method of Lemon-tree is prior to other two methods, with the accuracy of 98.03%, the recall of 0.928 and the F-measure of 0.939 between Basal and other subtypes. However, the selected driver genes can not successfully classify the subtypes between LumA and other subtypes. The accuracy, recall and F-measure of LumA is 83.13%, 0.801 and 0.812 respectively. It is possible that the selected driver genes have a similar regulation in gene expression.

Table 1. The accuracy of 10-fold cross validation between subtypes.

Subtypes	Module-Network	Information Gain	Chi-Squared	Lemon-Tree
LumA-others	83.13%	79.60%	80.39%	85.88%
LumB-others	84.70%	76.47%	76.86%	80.39%
Basal-others	98.82%	97.64%	98.82%	98.03%
Her2-others	92.94%	92.94%	94.90%	92.15%

Table 2. The recall of 10-fold cross validation between subtypes.

Subtypes	Module-Network	Information Gain	Chi-Squared	Lemon-Tree
LumA-others	0.801	0.844	0.853	0.827
LumB-others	0.847	0.402	0.291	0.5
Basal-others	0.988	0.952	0.976	0.928
Her2-others	0.929	0.56	0.6	0.44

Table 3. The F-measure of 10-fold cross validation between subtypes.

Subtypes	Module-Network	Information Gain	Chi-Squared	Lemon-Tree
LumA-others	0.812	0.790	0.798	0.842
LumB-others	0.719	0.491	0.415	0.590
Basal-others	0.964	0.930	0.964	0.939
Her2-others	0.590	0.608	0.697	0.523

2.4. Topology-Based Pathway Analysis of Identified Driver Genes

To better understand the underlying biological phenomenon, the pathway analysis is used to uncover pathway activity and influence of driver genes. Here, we applied Mirna enrIched paTHway Impact anaLysis (MITHrIL) [32], a technique that extends Draghici et al. [33] and Tarca et al. [34], by combining their effectiveness while improving the reliability of the results. Starting from expression

values of genes and/or microRNAs, MITHrIL returns a list of pathway stored according to the degree of their deregulation, together with the corresponding statistical significance (p-values), as well as predicted degree of alteration for each endpoint (a pathway node whose alteration, based on current knowledge, affects the phenotype in some way). All terms were first ranked by p-value and only the pathways with p-value less than 0.01 were selected. The top-18 pathways obtained for breast cancer are shown in Table 4.

Table 4. Top-18 pathways obtained for breast cancer subtypes after enrichment performed by MITHrIL.

LumA		**LumB**	
Pathway	***p*-Value**	**Pathway**	***p*-Value**
Chemokine signaling pathway	0	MAPK signaling pathway	0
HIF-1 signaling pathway	0	PI3K-Akt signaling pathway	0
VEGF signaling pathway	0	Apoptosis	0
Osteoclast differentiation	0	Neurotrophin signaling pathway	0
Hippo signaling pathway	0	Type I diabetes mellitus	0
Her2		**Basal**	
Pathway	***p*-Value**	**Pathway**	***p*-Value**
ErbB signaling pathway	0	Natural killer cell mediated cytotoxicity	0
Calcium signaling pathway	0	Chagas disease (American trypanosomiasis)	0
HIF-1 signaling pathway	0	HTLV-I infection	0
Focal adhesion	0		
Adherens junction	0		

As shown in the Table 4, for LumA, the important driver genes are mainly enriched in pathways in Chemokine signaling pathway, HIF-1 signaling pathway, VEGF signaling pathway, Osteoclast differentiation, Cell adhesion molecules (CAMs) and so on after pathway analysis. In Basal, pathways in Natural killer cell mediated cytotoxicity, Chagas disease (American trypanosomiasis), HTLV-I infection are enriched in pathways. In LumB, the pathways analysis are MAPK signaling pathway, PI3K-Akt signaling pathway, Apoptosis, Neurotrophin signaling pathway, Type I diabetes mellitus. In Her2, the main pathways are ErbB signaling pathway, Calcium signaling pathway, HIF-1 signaling pathway, Focal adhesion and Adherens junction after MITHrIL analysis.

2.5. Functional Analysis of Driver Genes for Each Subtype

We did Gene Ontology (GO) biological process (BP) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enriched among the subtype driver genes with R package clusterProfiler (http://www.bioconductor.org/packages/release/bioc/html/clusterProfiler.html) [35]. Only the enriched GO terms with p-value less than 0.01 and the enriched KEGG pathways with p-value less than 0.01 were selected to analyze.

For LumA, the important driver genes are mainly enriched in pathways in Cell adhesion molecules (CAMs), Terpenoid backbone biosynthesis, Thyroid cancer, African trypanosomiasis and so on after KEGG pathway enrichment. With respect to the biological process, nuclear DNA replication, steroid metabolic process, B cell receptor signaling pathway, organic hydroxy compound biosynthetic process are enriched via the GO functional enrichment.

In Basal, pathways in protein digestion and absorption, Lysosome, Natural killer cell mediated cytotoxicity, Apoptosis are enriched in KEGG pathways. In terms of biological process, cholesterol biosynthetic process, secondary alcohol biosynthetic process, multicellular organism catabolic process, multicellular organismal macromolecule metabolic process, steroid biosynthetic process are significantly enriched in GO functional enrichment.

In LumB, the pathways after KEGG enrichment are apoptosis, African trypanosomiasis, NOD-like receptor signaling pathway, Allograft rejection, Graft-versus-host disease. In terms of biological process

in GO functional enrichment, driver genes are enriched in response to positive regulation of peptidase activity, activation of innate immune response, positive regulation of innate immune response, cellular chloride ion homeostasis.

In Her2, the main pathways are the HIF-1 signaling pathway, MicroRNAs in cancer, Primary immunodeficiency, Bladder cancer and Pathways in cancer after KEGG enrichment analysis. In terms of biological process, blood vessel morphogenesis, regulation of T cell proliferation, regulation of microtubule-based process, T cell proliferation and regulation of mononuclear cell proliferation are enriched in GO functional enrichment.

3. Discussion

In this work, we introduced a module-based framework by integrating transcriptome and genomic data to identify significant driver genes in breast cancer subtypes. By virtue of the consideration the differential expression of genes, a subset of gene whose aberration/expression profile were significantly different between two subtypes may be selected. Also, we constructed a module network by integrating multi-genomic data to identify the specific driver genes.

Because breast cancer data is high heterogeneous and the conventional method for analyzing heterogeneous data perform poorly, we applied this method to the challenging problem of identifying driver genes of breast cancer subtypes. The final result demonstrated that the power of integrative analysis can generate a biological meaningful short list of genes as subtype-specific driver genes.

Furthermore, there are also some limitations of our method. Firstly, in computational and statistical analysis, the method is computationally expensive due to iteration in training the model and searching for the best model. In additional, because the method is based on the statistical machine-learning, it will work better if there are more samples. If there are a few samples in each condition, this method will not perform well. In a word, we developed a novel integrated method based on statistical machine-learning analysis to discover the drivers of breast cancer subtypes, and we showed that the method can generate a short list of biologically meaningful genes that can promote the process of biomarkers discovery.

4. Materials and Methods

4.1. Breast Cancer Patients Materials

We used processed and normalized breast cancer genomic data as given by TCGA. We downloaded gene expression data from the Agilent 244 K Custom Gene Expression platform, CNV data from the Affymatix Genome-Wide Human SNP 6.0 platform. In this data set, genomic data of 825 patients were obtained. We examined the clinical data from these patients to identify the subtype patients with gene expression data and CNV data. We selected 216 Luminal A, 122 Luminal B, 98 Basal-like and 58 HER2-enriched among the 825 patients, for which their gene expression and CNV data were available as well.

4.2. Differential Expression Analysis for Data with Intraclass Heterogeneity

The within group variations are extremely high for the heterogeneous data of breast cancer. Conventional methods to select gene subsets perform poorly when applied to these data with high within class heterogeneity. In this manuscript, the approach of *EMD* (Earth mover's distance) is applied to acquire gene subset [36]. *EMD* is a measure of distance between two distributions that reflects the minimum cost of transforming one distribution into the other. Whereas test statistics generated by standard differential expression approaches reflect the likelihood that the difference of mean expression between two groups is non-zero or reflect the significance of the association between abundance of short reads of the two groups, the *EMD* test statistic reflects the overall difference between two normalized distributions. These two distributions are represented by signatures.

For the differential expression analysis of genomics data, the signatures are data densities computed from gene expression values' histograms from each class of samples. Given two signatures

P and Q which are represented as: $P = \{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$, where p_i is the center of the ith histogram cell and w_{pi} is the weight of the cell, which represents the frequency of p_i; and $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$, where q_j is the center of the jth histogram cell and w_{qj} is the weight of the cell, which represents the frequency of q_j. Given P, Q, and d_{ij} (the Euclidean distance between p_i and q_j), f_{ij} (the optimal flow between p_i and q_j), the *EMD* scores are calculated as follows:

$$EMD(P, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}. \tag{4}$$

The q-value is the permutation-based estimate of the *FDR* (false discovery rate) which is the expected proportion of rejected null hypotheses [37]. To generate the null contribution, we permuted the sample labels and computed the EMD between the permuted classes for each iteration. We performed 1000 iterations and constructed a null distribution by the median of permuted EMDs for each gene to compute the FDRs.

Then, the FDRs are obtained from a range of significance thresholds. Given $M = [m_1, \ldots, m_N]$, a vector of median of permuted EMDs, and $EMD = [emd_1, \ldots, emd_N]$, a vector of observed EMDs, the mathematical representation of FDR for gene j and significance threshold i, t_i, is defined as follows:

$$FDR_{ji} = \begin{cases} \frac{\sum_{k=1}^{N} I(m_k, t_i)}{\sum_{k=1}^{N} I(emd_k, t_i)} & if\ emd_j \geq t_i \\ 1 & otherwise \end{cases} \tag{5}$$

where I is the indicator function:

$$I(K, t_i) = \begin{cases} 1 & if\ K \geq t_i \\ 0 & otherwise \end{cases} \tag{6}$$

N is the number of genes in the dataset, and $t_i's$ are in descending order from T to zero with step $\triangle : \{T,\ T - \triangle,\ T - 2\triangle,\ \ldots\ ,\ \triangle,\ 0\}$. We set $\triangle$ = 0.001 and T to the rounded maximum END minus 1 (T = 3). Then, the q-value for gene j is calculated as:

$$q\text{-}value_j = min\left(FDR_j\right). \tag{7}$$

4.3. The Selection of Candidate Modulators

The candidate modulators can regulate other genes in their module. We selected these genes as candidate modulators from the list of aberrant genes if their CNVs frequencies is high across the tumors. The mutation frequency of each gene is calculated as follows:

$$fre = \frac{\#Mutgenes}{N}. \tag{8}$$

where $\#Mutgenes$ represents the number of mutation genes in across sample, N is the total researched samples.

4.4. Initial Modules Construction Based on Normal Gamma Score

Innumerable functional studies suggest that driver mutation are expected to alter gene expression of their cognate proteins, their interacting partners, or genes that share the same biochemical pathway. This will lead to a correlated pattern of gene expression in a network of genes associated with a driver mutation. The initial modules construction step establishes an initial pairing between candidate modulators and gene expression modules by associating each target gene with the single modulator gene that fits it best.

First, for each single candidate modulator gene, we use the gene expression values of the amplified/deleted samples to guide the selection of threshold and consider the gene expression of the amplified/deleted samples to represent appropriate high/low expression levels. We use k-means clustering, using $k = 2$ and the normal and amplified/deleted samples as the two initial clusters to fit two normal distributions. The boundary between two clusters is the selected expression threshold level for this candidate modulator gene. The selected threshold is used to split the expression of each target gene.

Then, the expression of each target gene is split into two sets: those in the tumor samples in which the driver's expression is below the threshold, and those in the tumor samples in which the driver's expression is above the threshold. We adopted the possible models and used a statistically based likelihood score called the Bayesian score to evaluate a model's fitness to the data [38]. We use Normal Gamma distribution for our likelihood function. The Normal Gamma scoring function is used to compute the quality of this split, thus measuring a target gene's fit with a candidate modulator. Given $leaf$, a vector of the gene expression values of entire samples, and a vector $Leaf$ which represents the split gene expression values contained in the $leaf$; α and λ are the parameters and N is the size of $Leaf$. The Normal Gamma score is described below:

$$Score = -N * \ln\left(\sqrt{2\pi}\right) + \frac{\ln\left(\frac{\lambda}{\lambda+N}\right)}{2} + \ln\left(\Gamma(\alpha^+)\right) - ln\left(\Gamma(\alpha)\right) + \alpha * \ln(\beta) - \alpha^+ * \ln\left(\beta^+\right) \quad (9)$$

where,

$$\beta = Max(1, \frac{\lambda * (\alpha - 2)}{\lambda + 1}) \quad (10)$$

$$\beta^+ = \beta + \frac{Var(Leaf) * N}{2} + N * \lambda * \frac{\overline{Leaf^2}}{2 * (N + \lambda)} \quad (11)$$

$$\alpha^+ = \alpha + \frac{N}{2} \quad (12)$$

A split is scored by comparing the score of the split data to the score without split, along with a penalty for the split. After the score is computed for all pair-wise combinations of candidate modulators and target genes, each gene is assigned to the single highest scoring candidate modulator.

4.5. Module-Network Learning

The module-network learning step uses the modules generated by the initial modules step as a starting point and uses an iterative approach to improve the score of the modules and their regulation programs. In each iteration, the procedure learns a regulation program for each module and re-assigns each gene to the module whose program best predicts its behavior. The learning process is shown in Figure 4.

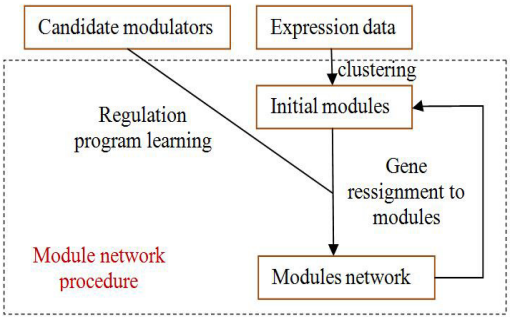

Figure 4. The process of module-network learning. It is an iterative procedure that determines both the partition of genes to modules and the regulation program for each module.

In our iterative learning procedure, the Expectation Maximization (EM) algorithm [39] is applied to search for the model with the highest score. Normal Gamma gives a higher score to data with lower variance and hence finds splits that create two different contexts that represent two distinct behaviors. In the procedure of learning, we also use Normal Gamma distribution for our likelihood function. We adopted the strategy in [38]. The likelihood function as follows:

$$L(M : D) = P(D|M) = \prod_{m=1}^{M} (P(x[m]|\tau, A)). \tag{13}$$

The M is a triple (C, τ, A), where C is a module set, τ is a module network template for C, and A is a module assignment function for C. Given the training set $D = \{x_1, ..., x_M\}$, which consisting of M instances drawn independently from an unknown distribution $P(X)$. We assume that the set of modules C is given, and we wish to estimate this distribution using a module network over C.

The procedure of EM algorithm consists two steps: an E-step and an M-step. These two steps are iterated until that fewer than 10% of the target genes have been re-assigned to a different module.

(i) In the M-step, the procedure is given a partition of the genes into modules and learns the best regulation program for each module. For computational efficiency, some M-step optimize only the parameters and leave the regulation program structure unchanged. We recursively learned the regulation choosing, at each point, the candidate modulator that best splits the gene expression of the module genes into two distinct behaviors. All candidate modulators and split values were evaluated and the modulator-split combination that achieves the highest improvement in score is selected.

(ii) In the E-step, given the inferred regulation programs, we determined the module whose associated regulation program best predicts each gene's behavior. Specifically, we iterated over all genes, one at a time, and moved each gene into the module which provides the highest improvement in the score. This step is guaranteed to improve the score, or leave it the same.

4.6. The Identification of Candidate Driver Genes

We proposed an integrative frame based on module-network to identify candidate driver genes. The framework process is shown in Figure 5. First the EMD difference analysis and frequency analysis were used to select subset of data. Then the initial modules were constructed by k-means clustering and Normal Gamma scoring function. And the module-network learning was used to obtain the final modules and candidate modulators. The module-network learning step was run N times. We calculated the frequency of appearance of each candidate modulators as follows:

$$fre\_app = \frac{\#times\ of\ appearance}{N}. \tag{14}$$

where $\#times\ of\ appearance$ is the appearance times of each modulators in N runs. we filtered the set of candidate modulators and left only genes that appeared in at least one regulation program in at least 40% of the runs. The final run of the module-network learning algorithm was run using the filtered set of candidate modulators and these modulators were considered as candidate driver genes. The whole method is shown as Algorithm 1.

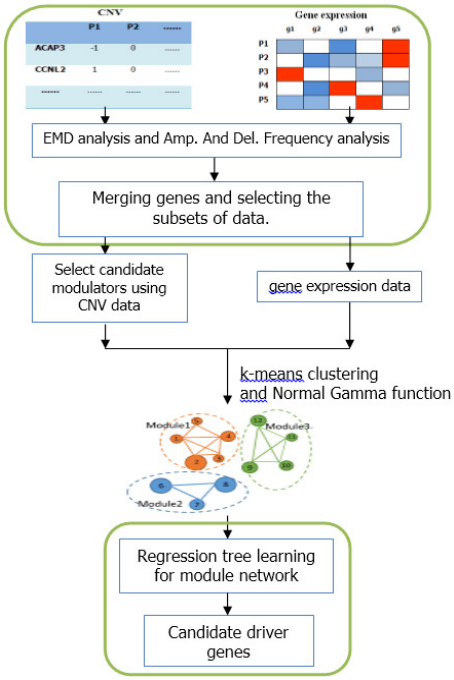

Figure 5. Schematic diagram of the integrative method based on module-network. The first part indicates the pre-processing steps based on the differential expression analysis. The middle part is the construction of the initial modules. The bottom part represents the process of module network learning.

Algorithm 1 Integrative model based on module-network for cancer subtypes

Input: CNV data and gene expression data of two subtypes

Output: A short list of gene sets

The 1th step: Difference analysis EMDSort (P, Q, f_{ij}, d_{ij})

(a) compute the $EMD(P, Q)$ using f_{ij} and d_{ij} according to the Formula (4). f_{ij} is the flow and the d_{ij} is the Euclidean distance.

(b) compute the FDR_{ji} according to the emd-values.

(c) compute the q-value according to the FDR_{ji}.

The 2th step: Initial modules construction

(a) fit two normal contributions by k-means clustering and select the threshold T for each modulator.

(b) split the expression of the target gene into two sets (A, B) according to the threshold T.

(c) Given a leaf vector $leaf$, the parameters α and λ, the size of $Leaf\ N$.

(d) compute the $Score(target\_gene, modulator)$ of the split using the Formula (9).

(e) assign the target gene into the single highest scoring candidate modulator.

The 3th step: Module network learning

repeat

(a) search for a regulation program for each module.

(b) reassign each gene to the module whose program best predicts its behavior.

(c) compute the proportion of re-assigned genes pro.

until $(pro < 0.1)$

The 4th step: The identification of candidate driver genes.

Supplementary Materials: The main code is available online.

Acknowledgments: The authors are grateful to Dana Pe'er Lab who provided the module networks learning analysis method with CONEXIC Toolkit and Guangchuang Yu's team who provided the functional enrichment analysis method with clusterProfilter package. This work was supported by the National Natural Science Foundation grant of China and China Scholarship Council grant.

Author Contributions: X.G.L. and X.L. conceived and designed the experiments; X.G.L., X.L. and X.Q. performed the experiments; Q.M.M., P.L. and S.L.P. analyzed the data; X.Q., P.L. and Q.M.M. contributed analysis tools; X.G.L., X.L. and Q.M.M. wrote the paper.

Conflicts of Interest: The authors declare that they do not have any conflicts of interest.

References

1. Mok, T.S.; Wu, Y.L.; Thongprasert, S.; Yang, C.H.; Chu, D.T.; Saijo, N.; Sunpaweravong, P.; Han, B.; Margono, B.; Ichinose, Y. Gefitinib or CarboplatinPaclitaxel in Pulmonary Adenocarcinoma. *New Engl. J. Med.* **2009**, *361*, 947–957.
2. Akavia, U.D.; Litvin, O.; Kim, J.; Sanchezgarcia, F.; Kotliar, D.; Causton, H.C.; Pochanard, P.; Mozes, E.; Garraway, L.A.; Pe'Er, D. An integrated approach to uncover drivers of cancer. *Cell* **2010**, *143*, 1005–1017.
3. Lahti, L.; Schafer, M.; Klein, H.U.; Bicciato, S.; Dugas, M. Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: A comparative review. *Brief. Bioinform.* **2013**, *14*, 27–35.
4. Lu, X.; Lu, J.; Liao, B.; Li, X.; Qian, X.; Li, K. Driver pattern identification over the gene co-expression of drug response in ovarian cancer by integrating high throughput genomics data. *Sci. Rep.* **2017**, doi:10.1038/s41598-017-16286-5.
5. Enerly, E.; Steinfeld, I.; Kleivi, K.; Leivonen, S.K.; Aure, M.R.; Russnes, H.G.; Ronneberg, J.A.; Johnsen, H.; Navon, R.; R?dland, E. miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors. *PLoS ONE* **2011**, *6*, e16915.
6. Chen, J.; Zhang, S. Integrative cancer genomics: models, algorithms and analysis. *Front. Comput. Sci.* **2017**, *11*, 1–15.
7. Zhang, W.; Liu, Y.; Sun, N.; Wang, D.; Boyd-Kirkup, J.; Dou, X.; Han, J.D. Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer. *Cell Rep.* **2013**, *4*, 542–553.
8. Hopf, T. The Integrative Network of Gene Expression, MicroRNA, Methylation and Copy Number Variation in Colon and Rectal Cancer. *Curr. Bioinform.* **2016**, *11*, 59–65.
9. Li, D.; Xia, H.; Li, Z.; Hua, L.; Li, L. Identification of Novel Breast Cancer Subtype-Specific Biomarkers by Integrating Genomics Analysis of DNA Copy Number Aberrations and miRNA-mRNA Dual Expression Profiling. *Biomed. Res. Int.* **2015**, *2015*, 1–17.
10. Segal, E.; Al, E. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* **2003**, *34*, 166–176.
11. Bonnet, E.; Calzone, L.; Michoel, T. Integrative Multi-omics Module Network Inference with Lemon-Tree. *PLoS Comput. Biol.* **2015**, *11*, e1003983.
12. Le, M.M.; Zinovyev, A.; Vert, J.P. NetNorM: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLoS Comput. Biol.* **2017**, *13*, e1005573.
13. Futreal, P.A.; Coin, L.; Marshall, M.; Down, T.; Hubbard, T.; Wooster, R.; Rahman, N.; Stratton, M.R. A census of human cancer genes. *Nat. Rev. Cancer* **2004**, *4*, 177–183.
14. Taniuchi, K.; Nakagawa, H.; Hosokawa, M.; Nakamura, T.; Eguchi, H.; Ohigashi, H.; Ishikawa, O.; Katagiri, T.; Nakamura, Y. Overexpressed P-cadherin/CDH3 promotes motility of pancreatic cancer cells by interacting with p120ctn and activating rho-family GTPases. *Cancer Res.* **2005**, *65*, 3092–3099.
15. Li, C.J.; Li, R.W.; Elsasser, T.H.; Kahl, S. Lipopolysaccharide-induced early response genes in bovine peripheral blood mononuclear cells implicate GLG1/E-selectin as a key ligand-receptor interaction. *Funct. Integr. Genom.* **2009**, *9*, 335–349.
16. Weber, C.; Meiler, S.; Doring, Y.; Koch, M.; Drechsler, M.; Megens, R.T.; Rowinska, Z.; Bidzhekov, K.; Fecher, C.; Ribechini, E. CCL17-expressing dendritic cells drive atherosclerosis by restraining regulatory T cell homeostasis in mice. *J. Clin. Investig.* **2011**, *121*, 2898–2910.

17. Saeki, H.; Tamaki, K. Thymus and activation regulated chemokine (TARC)/CCL17 and skin diseases. *J. Dermatol. Sci.* **2006**, *43*, 75–84.
18. Zhou, Q.; Lee, G.S.; Brady, J.; Datta, S.; Katan, M.; Sheikh, A.; Martins, M.S.; Bunney, T.D.; Santich, B.H.; Moir, S. A hypermorphic missense mutation in PLCG2, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency. *Am. J. Hum. Genet.* **2012**, *91*, 713–720.
19. Aderibigbe, O.M.; Priel, D.L.; Lee, C.C.; Ombrello, M.J.; Prajapati, V.H.; Liang, M.G.; Lyons, J.J.; Kuhns, D.B.; Cowen, E.W.; Milner, J.D. Distinct Cutaneous Manifestations and Cold-Induced Leukocyte Activation Associated With PLCG2 Mutations. *JAMA Dermatol.* **2015**, *151*, 627–634.
20. Romanelli, M.G.; Lorenzi, P.; Sangalli, A.; Diani, E.; Mottes, M. Characterization and functional analysis of cis-acting elements of the human farnesyl diphosphate synthetase (FDPS) gene 5′ flanking region. *Genomics* **2009**, *93*, 227–234.
21. Olmos, J.M.; Zarrabeitia, M.T.; Hernandez, J.L.; Sanudo, C.; Gonzalezmacias, J.; Riancho, J.A. Common allelic variants of the farnesyl diphosphate synthase gene influence the response of osteoporotic women to bisphosphonates. *Pharmacogenom. J.* **2012**, *12*, 227–232.
22. Fernandez, D.M.; Hand, C.K.; Sweeney, B.J.; Parfrey, N.A. A novel ATP1A2 gene mutation in an Irish familial hemiplegic migraine kindred. *Headache* **2008**, *48*, 101–108.
23. Harriott, A.M.; Nicole, D.; Cheng, Y.C.; Ryan, K.A.; O'Connell, J.R.; Colin, S.O.; Mcardle, P.F.; Wozniak, M.A.; Stern, B.J.; Mitchell, B.D. Polymorphisms in migraine-associated gene,atp1a2,and ischemic stroke risk in a biracial population: the genetics of early onset stroke study. *Springerplus* **2013**, *2*, 1–8.
24. Lei, D.; Sturgis, E.M.; Wang, L.E.; Liu, Z.; Zafereo, M.E.; Wei, Q.; Li, G. FAS and FASLG genetic variants and risk of second primary malignancy in patients with squamous cell carcinoma of the head and neck. *Cancer Epidemiol. Prev. Biomark.* **2010**, *19*, 1484–1491.
25. Wang, M.; Wu, D.; Tan, M.; Gong, W.; Xue, H.; Shen, H.; Zhang, Z. FAS and FAS ligand polymorphisms in the promoter regions and risk of gastric cancer in Southern China. *Biochem. Genet.* **2009**, *47*, 559–568.
26. Lifschytz, T.; Broner, E.C.; Zozulinsky, P.; Slonimsky, A.; Eitan, R.; Greenbaum, L.; Lerer, B. Relationship between Rgs2 gene expression level and anxiety and depression-like behaviour in a mutant mouse model: serotonergic involvement. *Int. J. Neuropsychopharmacol.* **2012**, *15*, 1307–1318.
27. Greenbaum, L.; Smith, R.C.; Lorberboym, M.; Alkelai, A.; Zozulinsky, P.; Lifschytz, T.; Kohn, Y.; Djaldetti, R.; Lerer, B. Erratum to: Association of the ZFPM2 gene with antipsychotic-induced parkinsonism in schizophrenia patients. *Psychopharmacology* **2012**, *220*, 519–528.
28. Aggarwal, S.; Negi, S.; Jha, P.; Singh, P.K.; Stobdan, T.; Pasha, M.A.; Ghosh, S.; Agrawal, A.; Prasher, B.; Mukerji, M. EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 18961–18966.
29. Xiang, K.; Ouzhuluobu.; Peng, Y.; Yang, Z.; Zhang, X.; Cui, C.; Zhang, H.; Li, M.; Zhang, Y.; Bianba. Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation. *Mol. Biol. Evol.* **2013**, *30*, 1889–1898.
30. Lu, X.; Peng, X.; Deng, Y.; Feng, B.; Liu, P.; Liao, B. A Novel Feature Selection Method Based on Correlation-Based Feature Selection in Cancer Recognition. *J. Comput. Theor. Nanosci.* **2014**, *11*, 427–433.
31. Adankon, M.M.; Cheriet, M. Support Vector Machine. *Comput. Sci.* **2002**, *1*, 1–28.
32. Alaimo, S.; Giugno, R.; Acunzo, M.; Veneziano, D.; Ferro, A.; Pulvirenti, A. Post-transcriptional knowledge in pathway analysis increases the accuracy of phenotypes classification. *Oncotarget* **2016**, *7*, 54572–54582.
33. Draghici, S.; Khatri, P.; Tarca, A.L.; Amin, K.; Done, A.; Voichita, C.; Georgescu, C.; Romero, R. A systems biology approach for pathway level analysis. *Genome Res.* **2007**, *17*, 1537–1545.
34. Tarca, A.L.; Draghici, S.; Khatri, P.; Hassan, S.S.; Mittal, P.; Kim, J.S.; Chong, J.K.; Kusanovic, J.P.; Romero, R. A novel signaling pathway impact analysis. *Bioinformatics* **2009**, *25*, 75–82.
35. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **2012**, *16*, 284–287.
36. Nabavi, S.; Schmolze, D.; Maitituoheti, M.; Malladi, S.; Beck, A.H. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. *Bioinformatics* **2016**, *32*, 533–541.
37. Storey, J.D. A Direct Approach to False Discovery Rates. *J. R. Stat. Soc.* **2002**, *64*, 479–498.

38. Segal, E.; Pe'er, D.; Regev, A.; Koller, D.; Friedman, N. Learning module networks. *J. Mach Learn. Res.* **2005**, *6*, 557–588.
39. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.

Sample Availability: Samples of the compounds are not available from the authors.