Open Access

OBM Genetics

Original Research

MSIGNET: A Bayesian Approach for Disease-associated Gene Network Identification

Xi Chen *, Jianhua Xuan

Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203, USA; E-Mails: xichen86@vt.edu; xuan@vt.edu

* Correspondence: Xi Chen; E-Mail: xichen86@vt.edu

Academic Editor: Lunawati L Bennett

Special Issue: Cancer Genetics and Epigenetics Alterations

OBM Genetics
2020, volume 4, issue 2
doi:10.21926/obm.genet.2002107

Received: October 19, 2019
Accepted: April 01, 2020
Published: April 07, 2020

Abstract

The analysis of gene networks and signalling pathways plays a key role in understanding gene functions, i.e., their effects on the development of a particular disease. Yet, for many heterogeneous diseases, the number of known disease-associated genes is limited. Identifying disease-associated genes is still an open challenge. To understand the functions of genes associated with a disease, we develop a Metropolis-Hastings sampling based SIGnificant NETwork (MSIGNET) identification approach. MSIGNET integrates disease gene expression data and human protein-protein interactions in a Bayesian network, and identifies interactions of genes specifically expressed under the disease condition. We applied MSIGNET to simulation and benchmark data. Results demonstrated its superior performance over conventional network identification tools on disease-associated gene network identification when multiple local gene modules existed. To learn genes and functional signalling pathways associated with ovarian cancer recurrence, we identified a gene network using TCGA ovarian cancer gene expression data and further validated results using an independent gene expression data set. Genes in the identified network were significantly enriched with cellular processes relevant to ovarian cancer development, and as

© 2020 by the author. This is an open access article distributed under the conditions of the Creative Commons by Attribution License, which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is correctly cited.

features, they demonstrated predictive power on ovarian cancer recurrence. MSIGNET can be accessed at https://sourceforge.net/projects/msignet/.

Keywords
Bayesian network; differential gene expression; protein-protein interaction; ovarian cancer

1. Introduction

With comprehensive cancer-specific gene expression data acquired from the TCGA project [1], for many cancer types we are able to reliably identify cancer-specific genes from patient samples. Although some genes were previously individually studied for their functional roles in cancer development, genes indeed exhibited emergent properties when functioning together [2], like revealing synthetic lethality [3]. Gene signalling pathway analysis (including both genes and their interactions) plays a key role in understanding biological signal transduction in cancer. For example, the deficiency of DNA repair is one of the hallmarks of the high grade serous ovarian cancer [4]. The chemotherapy drug like olaparib that targets DNA damage response taking advantage of clinical synthetic lethality has already shown therapeutic benefit in ovarian cancer [5]. There were many cellular processes or signalling pathways but only a few were associated with disease development under a specific context [6]. Thus, the analysis of gene interactions and their coherent functions for a specific cancer type can facilitate the identification of novel disease-associated genes and signalling pathways [7].

For disease-associated gene network identification, a number of methods were developed to reflect significant disease-specific gene expression from the network aspect, where a gene expression dataset (with both disease and control samples) and protein-protein interactions (PPI) [8] were used as input [9-11]. In general, these methods proposed objective functions and identified networks in which gene features would maximize the objectives. As limited knowledge about disease-gene association is available [12], especially for cancers that are highly heterogeneous, given a specific disease, the number of genes or their interactions associated with that disease is largely unknown. In most cases, identifying an optimal disease-associated gene network is NP-hard [13]. Another challenge of studying gene PPI networks [8] is that, among interactions between tens of thousands of genes, there exist many local gene modules. Network searching tools could easily stick within local modules and miss important genes associated with disease. Therefore, computational tools that can globally and efficiently search large-scale networks and identify interactions of genes with specific and coherent expression under a disease condition are still needed.

In this paper, we developed a network identification tool, MSIGNET (Metropolis–Hastings sampling based SIGnificant NETwork identification), by integrating disease-specific gene expression data with human PPIs. MSIGNET uses Markov chain Monte Carlo (MCMC) to search, sample and combine local gene modules into a global network. During the MCMC process, MSIGNET samples a subset of genes and their interactions from the large-scale human PPI network, evaluates the significance of gene differential expression between disease and control samples and co-expression of PPI connected genes, and then determines if the sampled network

meets the criterion of acceptance. To improve the sampling efficiency, MSIGNET uses an informative proposal function to make the Markov Chain more likely to jump to a disease relevant state (with more differentially expressed and co-expressed genes). In the end, all sampled subnetworks are summed together as a weighted global network where the sampling frequency on each gene or gene interaction denotes the probability of its association with the disease.

To demonstrate the accuracy of MSIGNET on disease-associated gene networks identification, we applied MSIGNET to simulation data and compared its performance to competing network identification approaches. Results revealed a superior performance of MSIGNET on identifying disease-associated genes and their interactions, especially in networks with multiple local modules. We further applied MSIGNET to studying Parkinson's disease for which disease-associated gene networks were previously identified [14]. Not only did we reproduce Parkinson's disease-associated gene networks but identified additional biologically meaningful and interactive genes associated with Parkinson's disease.

We finally used MSIGNET to identify genes and signalling pathways associated with ovarian cancer recurrence. Ovarian cancer is one of the most aggressive cancers in women in the United States. Generally, 70% of advanced stage ovarian cancer relapses; and even in stage I or II patients, the relapse rate is 20%-25% [15]. Using MSIGNET we identified a network for genes differentially expressed between two groups (aggressive versus less aggressive) of TCGA ovarian cancer patients with short and long survival time. Validating the results network on another independent data set we demonstrated that the MSIGNET-learned gene network includes key ovarian cancer genes as well as functionally relevant genes that provide prediction power on cancer recurrence.

2. Methods

Given a gene expression dataset $\mathbf{Y}=[y_{n,1},...,y_{n,M_D},y_{n,M_D+1},...,y_{n,M_D+M_C}]$ including N genes ($n\in\{1,2,...,N\}$) for M_D disease and M_C control subjects, to identify disease-associated genes and their interactions, we define a gene vector $\mathbf{G}=[g_n]$ and a network matrix $\mathbf{V}=[v_{n_1,n_2}]$ and infer the states of each elements in them using Bayesian inference as follows:

$$P(\mathbf{V},\mathbf{G}\,|\,\mathbf{Y})\propto P(\mathbf{Y}\,|\,\mathbf{G})P(\mathbf{Y}\,|\,\mathbf{V})P(\mathbf{V},\mathbf{G}) \quad (1)$$

The prior information of $\mathbf{G}$ and $\mathbf{V}$ is obtained from the input PPI data as follows:

$$\begin{cases} v_{n_1,n_2}=1 \text{ or } 0, \text{ if there is a protein-protein interaction} \\ v_{n_1,n_2}=0, \text{ else} \end{cases}, \quad (2)$$

$$\begin{cases} g_n=1, \text{ if } (\sum_{n'} v_{n,n'})>0 \\ g_n=0, \text{ else} \end{cases}. \quad (3)$$

As the number of genes and interactions that are truly associated with a particular disease is unknown, we assume a uniform (noninformative) prior on $\mathbf{G}$ and $\mathbf{V}$. The conditional probability $P(\mathbf{Y}\,|\,\mathbf{G})$ represents the likelihood of genes in $\mathbf{G}$ that are differentially expressed between disease subjects and control subjects in the gene expression dataset $\mathbf{Y}$. The conditional probability $P(\mathbf{Y}\,|\,\mathbf{V})$ represents the relatedness (co-expression) for genes interacting in PPI network $\mathbf{V}$. For each gene $g_n=1$, we calculate its differential expression p-value using t-statistics and then transfer the p-value to a z-score z_n based on the inverse cumulative distribution of Gaussian.

$P(\mathbf{Y}|\mathbf{G})$ is calculated as a probability of the sum of z_n for all $g_n = 1$, and is assumed to follow Gaussian distribution. For each edge $v_{n_1,n_2} = 1$, we calculate the Pearson correlation coefficient of gene expression for genes n_1 and n_2, and then Fisher-transform the coefficient to z-score e_{n_1,n_2}. $P(\mathbf{Y}|\mathbf{V})$ is calculated as a probability of the sum of e_{n_1,n_2} for all $v_{n_1,n_2} = 1$, and is assumed to follow Gaussian distribution. $P(\mathbf{Y}|\mathbf{G})$ and $P(\mathbf{Y}|\mathbf{V})$ can be approximated as follows:

$$P(\mathbf{Y}|\mathbf{G}) = \mathbf{P}\left(\frac{1}{\sqrt{N_i}}\sum_{n=1}^{N}(z_n g_n);\ \mu_{Z,N_i}, \sigma^2_{Z,N_i}\right), \quad (4)$$

$$P(\mathbf{Y}|\mathbf{V}) = \mathbf{P}\left(\frac{1}{\sqrt{L_i}}\sum_{n_1=1}^{N-1}\sum_{n_2>n_1}^{N}(e_{n_1,n_2} v_{n_1,n_2});\ \mu_{E,L_i}, \sigma^2_{E,L_i}\right), \quad (5)$$

where $N_i = \sum_{n=1}^{N} g_n$ and $L_i = \sum_{n_1=1}^{N-1}\sum_{n_2>n_1}^{N} v_{n_1,n_2}$. The mean and variance parameters μ_{Z,N_i}, σ^2_{Z,N_i}, μ_{E,L_i} and σ^2_{E,L_i} are pre-estimated by randomly selecting N_i genes and L_i interactions and calculating the mean and variance of the score sums. The specific values will change when we select different networks. In MSIGNET, we control N_i to vary from 20 to 50 so that each subnetwork captures a local module without increasing much of the computational complexity. We also control the number of edges L_i under 100.

To initiate the Markov chain, we randomly select a small number of genes (e.g., 30) and their PPI interactions. Between two consecutive rounds of sampling, the network jumps from one state to the next by adding or deleting (with an equal prior) one node and one edge, controlled by a proposal function h.

Given $\mathbf{G}^i$ and $\mathbf{V}^i$, in the $(i+1)$-th round of sampling, in an adding action we first randomly select a seed gene n in $\mathbf{G}^i$ (the dark node in Figure 1B). Then we sample a new neighbour gene n' (the red node in Figure 1B) among genes connected to the seed gene in the global PPI but not included by $\mathbf{V}^i$. The gene differential expression score $z_{n'}$ of gene n' and its relatedness (co-expression) $e_{n,n'}$ to the seed gene jointly reflect the positive contribution of this adding action to the disease-associated network. Thus, we design the proposal function $h(\mathbf{G}', \mathbf{V}'|\mathbf{G}^i, \mathbf{V}^i)$ (a conditional probability) proportional to $P(z_{n'})P(e_{n,n'})$. For a deleting action, to avoid breaking the network connectivity, we only delete a leaf node n' from the network $\mathbf{V}^i$ with $\sum_n v_{n,n'} = 1$ (the cyan node in Figure 1B). Specifically, in the network $\mathbf{V}^i$, we delete a node that is less differentially expressed and less intensely related to its connected node. Therefore, the proposal function $h(\mathbf{G}', \mathbf{V}'|\mathbf{G}^i, \mathbf{V}^i)$ is proportional to $(1-P(z_{n'}))(1-P(e_{n,n'}))$.

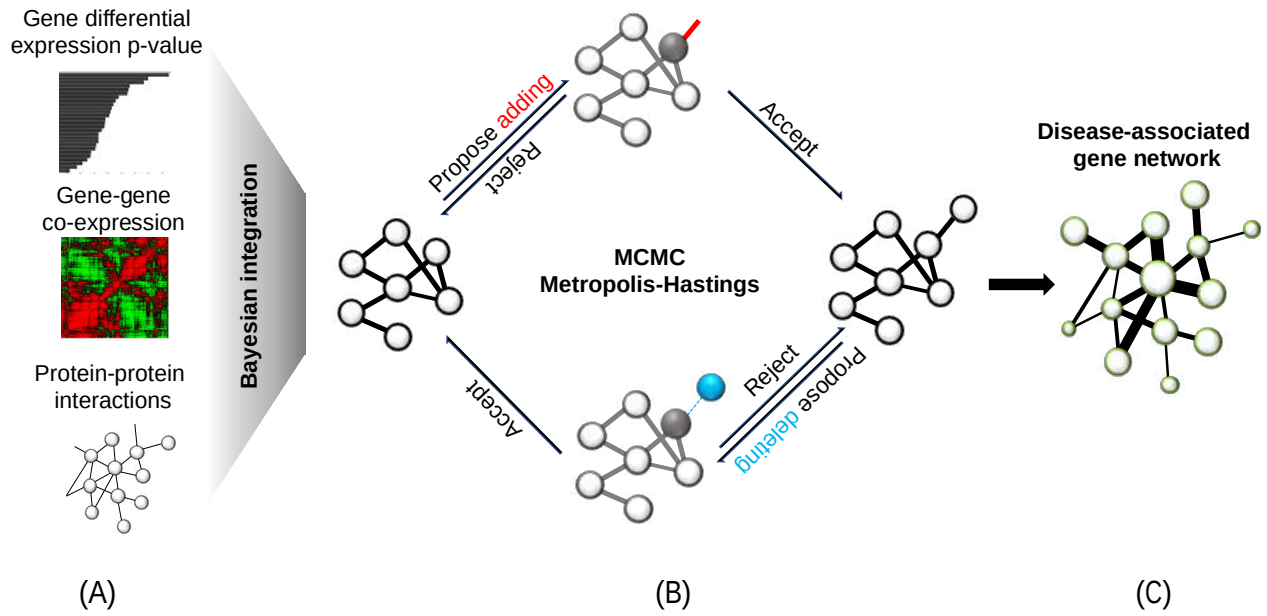

Figure 1 Workflow of the MSIGNET approach. For a specific disease, given a gene expression dataset, (A) MSIGNET calculated gene differential expression between disease samples and control samples for each gene and also calculated gene co-expression for pairs of genes connected by PPIs. (B) Under a Bayesian framework, MSIGNET used Metropolis-Hastings sampling to sample genes and their interactions, evaluated gene differential expression and co-expression in each sampled subnetwork and accepted or rejected samples. (C) In the end, summing all sampled subnetworks together, MSIGNET returned a global network with weighted nodes (genes) and edges.

After the above adding/deleting action, we proposed a new gene vector $\mathbf{G}'$ and a new network matrix $\mathbf{V}'$ to reflect the changed gene and edge states. Then, the network jumps to a new proposed state. According to the Metropolis-Hastings rule (defined in Eq. (6)), we evaluate the posterior probability improvement of the proposed state over the previous state and decide if we accept or reject it. If we accept the proposed state, we assign the network state for the $(i+1)$-th round of sampling as $\mathbf{G}^{i+1}=\mathbf{G}'$ and $\mathbf{V}^{i+1}=\mathbf{V}'$; otherwise we hold the network state as $\mathbf{G}^{i+1}=\mathbf{G}^{i}$ and $\mathbf{V}^{i+1}=\mathbf{V}^{i}$.

$$\alpha=\min\left(1,\ \frac{P(\mathbf{G}',\mathbf{V}')h(\mathbf{G}^{i},\mathbf{V}^{i}\,|\,\mathbf{G}',\mathbf{V}')}{P(\mathbf{G}^{i},\mathbf{V}^{i})h(\mathbf{G}',\mathbf{V}'\,|\,\mathbf{G}^{i},\mathbf{V}^{i})}\right) \quad (6)$$

We monitor the sampling convergence by running multiple MCMC sequences initiating from different network states [16]. Once the sampler appears to converge to the stationary distribution, we start to record $\mathbf{G}$ and $\mathbf{V}$. In real data analysis, especially when the highly heterogeneous cancer data is used, the number of patient samples is usually small and the data noise is high, which significantly impacts the disease-associated gene network identification. To obtain robust network results, we bootstrap samples of the gene expression dataset 100 times and run MSIGNET with each, to combat the effects of any outlier samples and data noise. We sum all sampled network states together and obtain the final estimation as $\bar{\mathbf{G}}$ and $\bar{\mathbf{V}}$, where the sampling frequency of each unit in gene vector $\bar{\mathbf{G}}$ or in matrix $\bar{\mathbf{V}}$ represents the posterior probability of disease association for each gene or edge.

3. Results

3.1 MSIGNET Performance on Simulated Networks

To evaluate the accuracy of MSIGNET on network identification, especially when there exit multiple local modules in a large gene network, we used real human PPIs from the HPRD database [8] (http://www.hprd.org/download, release 9) for network simulation and also simulate gene expression data with both disease-associated genes (differentially expressed (DE)) and background genes (equally expressed (EE)).

Using human PPIs, we constructed three networks with different levels of connectivity complexity between DE and EE genes (Supplementary Material, S1; Table S1; Figures S1-S3). Network 1 had two DE modules with dense interactions in between, including 127 DE genes and 927 DE-DE interactions, with 133 EE genes and 736 DE-EE or EE-EE interactions as background (Supplementary Figure S1). Network 2 also had two DE modules including 125 DE genes and 360 DE-DE interactions. Between the two DE modules there were very few direction interactions but 62 EE genes and 650 DE-EE or EE-EE edges (Supplementary Figure S2). In Network 3, there were multiple DE modules (220 DE genes and 1,513 DE-DE edges) in a large network (totally 2518 genes and 15,266 edges) (Supplementary Figure S3). The number of EE genes is ten times higher than that of DE genes and those EE genes were densely around DE modules.

Given each constructed network and gene states (DE or EE), we further used a two-step model to simulate gene expression data under two conditions (i.e., disease versus control). In the first step, we modelled the probabilities of the DE or EE state for each network gene using a Markov random field (MRF) model [17], where the network connectivity and neighbor gene states jointly determined how likely a DE gene can be 'true'. Second, we simulated gene expression data using a Gamma-Gamma (GG) model [18] with the MRF probabilities as input, and generated 50 samples under each condition, with Gaussian noise added. More details about gene expression data simulation were introduced in Supplementary Material, S2.

We compared MSIGNET with several network identification approaches including RegMOD [9], COSINE [11], MATISSE [19], Heinz [20], jActive [13], GiGA [21] and DeGAS [14] on network gene and edge identification. Most tools (other than RegMOD) reported the finally identified network as binary. Therefore, for performance compare, in Figure 2 we presented receiver operating characteristic (ROC) curves for MSIGNET and RedMOD using their weighted network outputs and in Table 1 we presented F-measures (2*precision*recall/(precision + recall)) of the other methods. Both results demonstrated that MISGNET exhibits a superior performance over existing approaches, especially when the network consists of multiple local modules. For Network 1 that had simple network connectivity, MSIGNET was comparable and better than the other methods. While for Network 3 where most methods were stuck in partial local modules, only able to capture partial DE genes, MSIGNET achieved a much better performance on capturing a high proportion of DE genes (F-measure = 0.88), with 0.2 F-measure improvement over the other methods. To connect those DE genes, each method had to frequently pass through EE genes and their interactions so that many 'bridges' between DE and EE genes were frequently selected, leading to a relatively low accuracy on edge identification. As MSIGNET was designed to reject samples of such 'bridge' edges as much as possible without breaking the network connectivity, its

performance on Network 3 DE-DE edge identification (F-measure of 0.55) was two times higher than that of any other method.

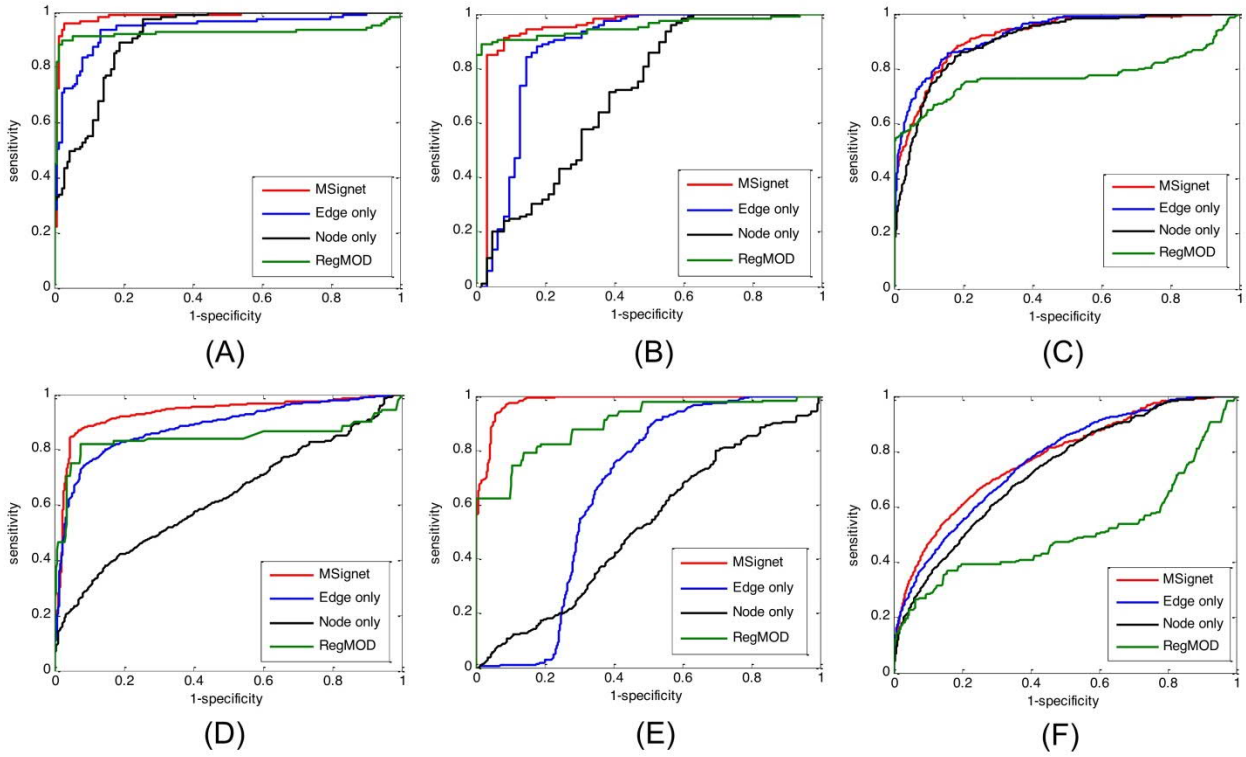

Figure 2 MSIGNET receiver operating characteristic curves for network identification. Performances of MSIGNET, its simpler versions modelling on either node (using differential expression) or edge (using co-expression) and a competing method RegMOD on identifying DE genes (A, B and C) and DE-DE edges (D, E, F) in Networks 1 - 3.

Table 1 F-measure performances of competing methods on node and edge identification.

Networks		MSIGNET	COSINE	MATISSE	Heinz	jActive	GiGA	DeGAS
Network 1	Gene	0.96	0.49	0.92	0.92	0.94	0.94	0.70
	Edge	0.90	0.31	0.54	0.90	0.89	0.89	0.51
Network 2	Gene	0.95	0.52	0.92	0.91	0.97	0.84	0.87
	Edge	0.82	0.34	0.54	0.72	0.81	0.66	0.72
Network 3	Gene	0.88	0.15	0.67	0.67	0.49	0.42	0.56
	Edge	0.55	0.17	0.23	0.26	0.27	0.16	0.21

3.2 Parkinson's Disease associated Gene Networks

To benchmark the performance of MSIGNET on disease-associated network identification, we applied it to Parkinson's disease gene expression data [22] and compared the results with existing Parkinson's disease-associated gene networks [14]: one up-regulated network including 73 genes over-expressed in Parkinson's disease patient samples and one down-regulated network including 67 down-regulated genes. In the PPI database, most up-regulated genes were within two-jump-

neighbour genes around gene YWHAB and most down-regulated genes were within two-jump-neighbour genes around gene HD. Therefore, MSIGNET integrated the Parkinson's disease gene expression data [22] with PPI inputs of 693 neighbour genes around gene YWHAB and identified an up-regulated gene network. Meanwhile, MSIGNET identified a down-regulated network using gene expression and PPI inputs of 347 genes around gene HD. Among the top 100 genes mostly sampled in each network, MSIGNET has successfully captured 80% up-regulated genes (red genes in Figure 3A; hypergeometric p-value = 5.6e-34) and 77% down-regulated genes (green genes in Figure 3B; hypergeometric p-value = 4.9e-37).

Further, focusing on the top 100 up-regulated network genes identified by MSIGNET, gene functional enrichment analysis using DAVID [23] returned significantly enriched cellular processes like regulation of apoptosis, cellular response to stress and regulation of cellular protein metabolic process (Supplementary Material, S3; Supplementary Table S2), consistent to the earlier study [14]. While among the top ranked 100 down-regulated genes, enrichment analysis revealed 11 Parkinson's disease-associated genes (adjusted p-value 3.7e-5), with 5 more genes identified if compared to the earlier approach (6 out of 11 genes were reported) [14]. In addition, we found that many genes could be associated with neurological diseases like Alzheimer's disease, Huntington's disease, neurological system process, regulation of neurogenesis and learning, memory and behavior (Supplementary Table S3), all of which were highly relevant to the development of Parkinson's disease.

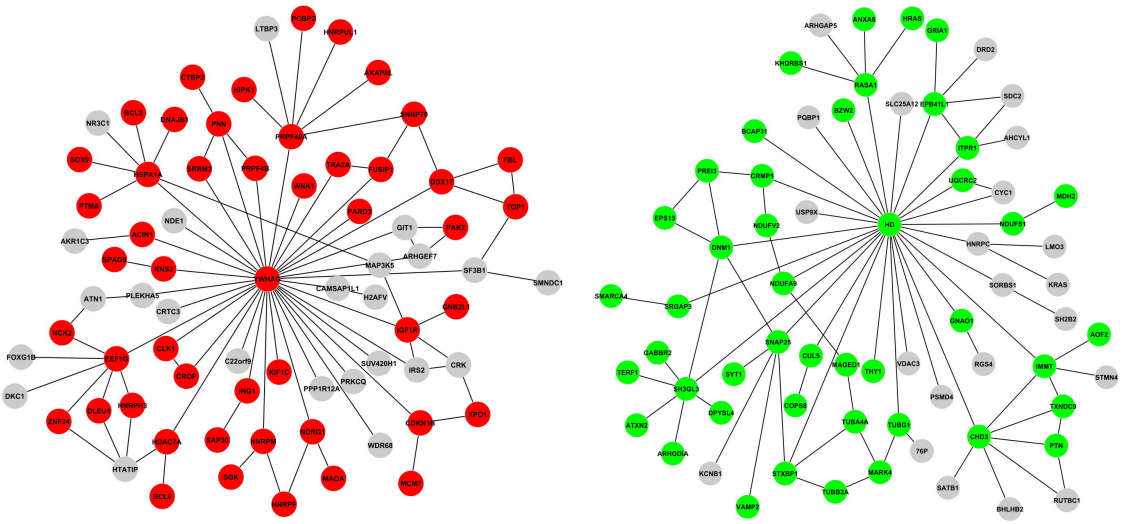

Figure 3 MSIGNET identified Parkinson's disease-associated gene networks. (A) A network with up-regulated genes, where genes previously reported as associating with Parkinson's disease in an independent network study [14] are labelled in red. (B) A network with down-regulated genes, where genes previously reported as associating with Parkinson's disease [14] are labelled in green.

3.3 Ovarian Cancer-associated Gene Network

Ovarian cancer is one of the most aggressive women cancer types. Generally, 70% of advanced stage ovarian cancer relapses; and even in stage I or II patients, the relapse rate can be as high as

25% [15]. Ovarian cancer recurrence can occur as soon as six months with a median interval of 18 to 24 months. But it is also noted that there are still long-term survivors of recurrent ovarian cancer with survival time longer than five years. Clinical features at the time of primary therapy like optimal surgical cytoreduction and platinum-sensitivity [24], and genomic features like a lower proportion of somatic copy number alterations and a lower average ploidy [25] have been associated with long survival. Yet, there is still much unknown about the biomarkers associated with short/long survival in women with ovarian cancer. Therefore, identifying predictive and interpretable features for ovarian cancer patient survival is important in understanding the mechanism of ovarian cancer development and helpful for drug selection or new drug development.

We investigated gene networks to be associated with ovarian cancer recurrence. Here, we specifically focused on the genes that were differentially expressed between samples of patients with different survival time distributions (Figure 4A and 4B). We downloaded gene expression data of ovarian cancer patients (with chemotherapy) from the TCGA project (https://portal.gdc.cancer.gov/) and another similar ovarian cancer gene expression data set from the GEO database (http://www.ncbi.nlm.nih.gov/geo/, GSE3149). As mentioned above, for aggressive ovarian cancer, the recurrence can occur as soon as six months with a median interval of 1.5 to 2 years while there are still long-term survivors with survival >5 years [26]. As shown in Fig.4, based on the overall survival time distribution of patients in TCGA or in GSE3149, three groups of patients can be found: survival <2 years, 2-4 years and > 4 years. Therefore, to identify genes networks associated with ovarian cancer recurrence, for a group of patients that had very aggressive ovarian cancer (the cancer comes back within a short period after treatment), we selected samples with survival time < 2 years as the 'early' recurrent group. We then selected samples with survival time > 4 years as the 'late' recurrent group, for patients that were more sensitive to the chemotherapy and had a longer survival time, with a relatively lower risk of cancer recurrence. The TCGA dataset included 72 'early' samples and 131 'late' samples (Figure 4A). Gene TP53, a well-known cancer gene, was selected as a seed gene to extract 2,240 proteins (within two-jump neighbors) and 11,798 PPIs from the HPRD database. Using MSIGNET to integrate TCGA gene expression data and TP53 centered PPIs, we identified a network including the top ranked 250 genes (top 10%) and top 25% edges. To improve the accuracy of network prediction, we also applied MSIGNET to the GSE3149 dataset including 33 'early' samples and 56 'late' samples (Figure 4B) and identified a gene network with 83 genes and 201 edges in common in both datasets (Figure 4C; p-value 4.6e-24).

There were several known ovarian cancer genes interacting with each other in MISGNET identified network (Figure 4C). NCOR2, whose expression is associated with several cancer types including ovarian cancer, serves as a hub gene in the network. Its neighbor gene, NFKB1, is also a known gene associated with ovarian cancer, frequently overexpressed in 'early' recurrent patients. Both genes interact with RXRA, a highly active gene in ovarian cancer. SPTBN1 is significantly overexpressed among patient samples in the 'early' group and interacts with CSNK2A1; the latter has an increased protein activity and enzyme activity in a majority of cancers. In the MSIGNET-identified network, CSNK2A1 directly interacts with BRCA1, a marker gene of ovarian cancer.

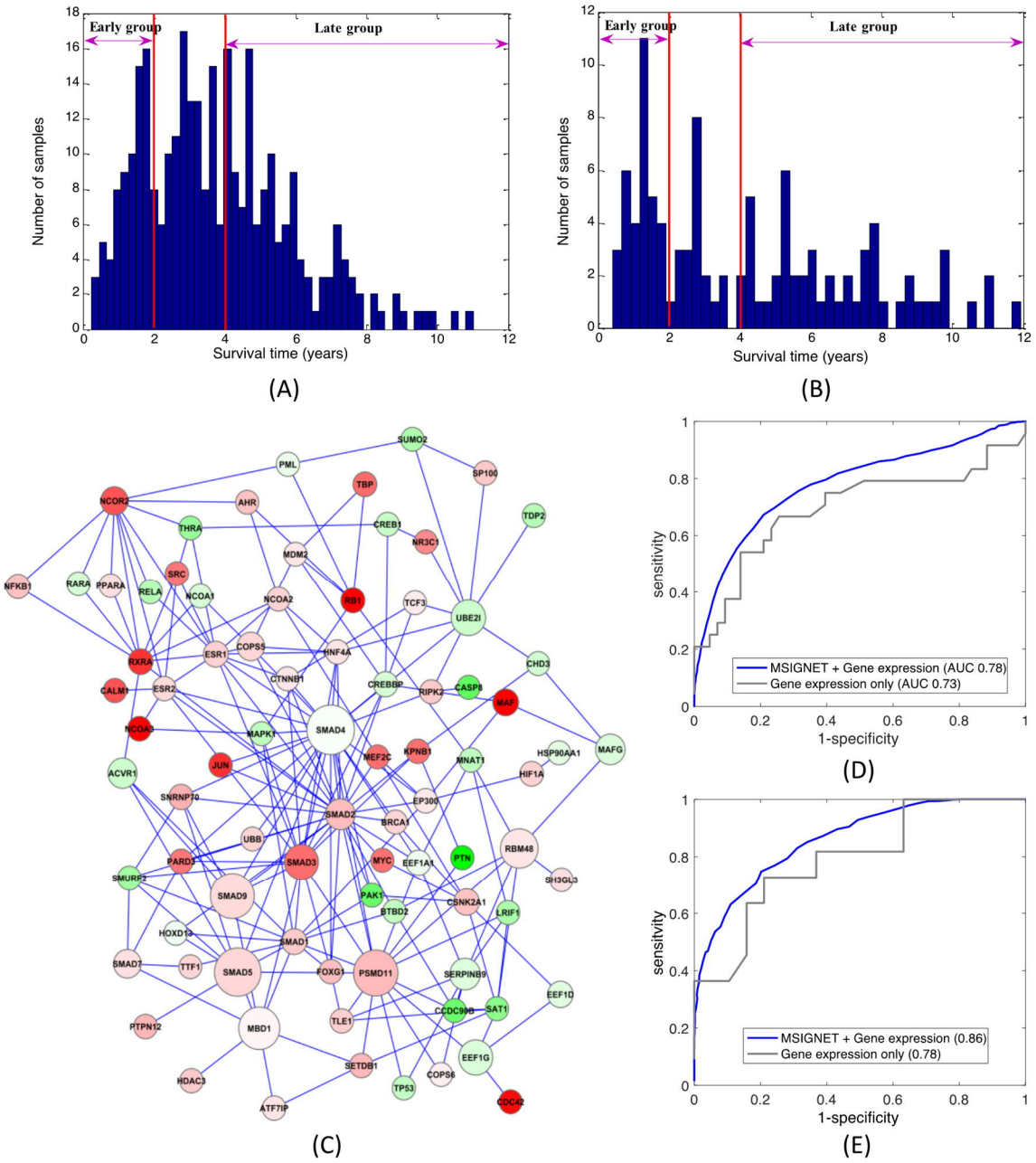

Figure 4 Ovarian cancer-recurrence associated gene network analysis. (A) The survival time distribution of TCGA ovarian cancer patients with chemotherapy. (B) The survival time distribution of GSE3149 ovarian cancer patients. (C) A common gene network with 83 genes and 201 edges. 'Red' genes are upregulated in the 'early' group while 'green' genes are downregulated. The size of each node represents the sampling frequency of MSIGNET. (D) Sample recurrence ROC curve (5-fold cross-validation) for the classifier between 'early' and 'late' samples, built upon genes in (C) using MSIGNET network and TCGA gene expression data as input. (E) Sample recurrence ROC curve (5-fold cross-validation) for the classifier built upon genes in (C) using MSIGNET learned network and GSE3149 gene expression data as input.

Gene functional enrichment analysis using DAVID reported the enrichment of several key pathways functioning in ovarian cancer (Supplementary Material, S4; Supplementary Table S4). For example, TGF-beta signaling pathway (p-value = 1.6e-11) plays a key role in ovarian cancer cell proliferation [27]. MAPK signaling pathway, a known pathway for cancer cell survival and usually resistant to drug therapy [28], is also enriched (p-value 3.6e-3). Moreover, several hallmark pathways associated with ovarian cancer are enriched, like DNA damage response and DNA repair (p-values = 6.3e-3 and 1.5e-2, respectively). Many genes in these processes are also involved in the regulation of cell proliferation and cell growth. DNA damage response plays a key role in maintaining genome stability, directly associated with cell survival and also coordinates a series of cellular processes including DNA replication, DNA repair and cell-cycle progression. Drugs like olaparib that targets DNA damage response taking advantage of clinical synthetic lethality have already shown therapeutic benefit to ovarian cancer patients [5].

The MSIGNET-identified network reflected gene differential expression between ovarian cancer 'early' and 'late' recurrent patients from the network aspect. We further checked the network genes performance in predicting patient survival time. Using the network in Figure 4C together with the gene expression data as features, for each patient we predict its sample label ('early' or 'late') using NetSVM [29], respectively for samples in TCGA and GSE3149 data sets (Figure 4D and 4E). We performed three-fold cross validation for each data set and finally obtained AUC (area under ROC curve) = 0.78 for TCGA samples and 0.86 for GSE3149 samples. While using gene expression data only, the AUC was 0.73 for TCGA and 0.78 for GSE3149 samples. Studying protein interactions of genes together with their mRNA expression contributes to the discovery of key molecules that have prediction power on ovarian cancer recurrence.

4. Discussion

We developed a Metropolis sampling-based approach for identifying a disease-associated gene network with input of disease-specific gene expression and human general PPIs. MSIGNET was specifically designed to efficiently identify a disease-associated network out of thousands of genes and their interactions. MSIGNET had a superior performance when the underlying network contained multiple local modules. Using two studies of Parkinson's disease and ovarian cancer, we demonstrated that MSIGNET can capture known disease-associated genes and additional relevant genes functioning in the disease. We were focused on identifying predictive and interpretable biomarkers associated with ovarian cancer survival. We used TCGA samples on purpose because a vast majority of them are primary tumors, with early or late local recurrence. The good performance of MSIGNET-identified features on predicting ovarian cancer recurrence revealed that a network-centric view of gene differential expression can contribute to the interpretation of short survival of ovarian cancer due to early recurrence (local recurrence).

For ovarian cancer recurrence analysis, overall survival (OS) was used in this paper for sample selection, however progression-free survival (PFS) can be another option. As most patients received multiple post-progression treatments, which can significantly confound the effects of the investigational therapy on the overall survival time endpoint. PFS is unaffected by post-progression therapies and may provide earlier evidence of efficacy of new treatments, which can expedite regulatory approval. However, correlations in the relative treatment effect between PFS and OS in first-, second- and third-line platinum-based chemotherapy trials [30, 31] for epithelial

ovarian cancer are all moderate. It is challenging yet interesting to investigate the difference of identified genomic markers when a different endpoint is used.

For cancer studies, not limited to ovarian cancer but other aggressive cancer types like breast and prostate cancer, another potential future application of MSIGNET could be studying networks of genes with differential expression between metastatic samples and local recurrent samples. Less significance of differential expression can be anticipated due to the diversity in metastatic samples. MSIGNET provides a new means of integrating those weakly differentially expressed genes with their PPIs and identifying gene modules or signalling pathways which can be potentially associated with distant metastases.

Code Availability

MSIGNET was implemented in MATLAB (2016a). Demo data and code were made publicly accessible at https://sourceforge.net/projects/msignet/.

Additional Materials

The following additional materials are uploaded at the page of this paper.

1. Supplementary Material, S1: Network Simulation using Human PPIs from the HPRD Database.
2. Supplementary Material, S2: Network-based Gene Expression Data Simulating using an MRF-GG Model.
3. Supplementary Material, S3: Functional Enrichment Analysis of Parkinson's Disease Network Genes.
4. Supplementary Material, S4: Functional Enrichment Analysis of Ovarian Cancer Network Genes.

Author Contributions

X. C and J. X proposed conceived and designed the research. X. C developed the method, performed the simulation and disease-focused studies. X. C and J.X wrote the manuscript together.

Competing Interests

The authors have declared that no competing interests exist.

References

1. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474: 609-615.
2. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science. 2019; 365: 786-793.
3. O'Neil NJ, Bailey ML, Hieter P. Synthetic lethality and cancer. Nat Rev Genet. 2017; 18: 613-623.
4. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011; 144: 646-674.

5. Gelmon KA, Tischkowitz M, Mackay H, Swenerton K, Robidoux A, Tonkin K, et al. Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: A phase 2, multicentre, open-label, non-randomised study. Lancet Oncol. 2011; 12: 852-861.
6. Greene CS, Krishnan A, Wong AK, Ricciotti E, Zelaya RA, Himmelstein DS, et al. Understanding multicellular function and disease with human tissue-specific networks. Nat Genet. 2015; 47: 569-576.
7. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic signaling pathways in the cancer genome atlas. Cell. 2018; 173: 321-337.e310.
8. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human Protein Reference Database--2009 update. Nucleic Acids Res. 2009; 37: D767-D772.
9. Qiu YQ, Zhang S, Zhang XS, Chen L. Detecting disease associated modules and prioritizing active genes based on high throughput data. BMC Bioinformatics. 2010; 11: 26.
10. Kim Y, Kim TK, Kim Y, Yoo J, You S, Lee I, et al. Principal network analysis: Identification of subnetworks representing major dynamics using gene expression data. Bioinformatics. 2011; 27: 391-398.
11. Ma H, Schadt EE, Kaplan LM, Zhao H. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics. 2011; 27: 1290-1298.
12. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015; 43: D789-D798.
13. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002; 18 Suppl 1: S233-S240.
14. Ulitsky I, Krishnamurthy A, Karp RM, Shamir R. DEGAS: De novo discovery of dysregulated pathways in human diseases. PLoS One. 2010; 5: e13367.
15. Ushijima K. Treatment for recurrent ovarian cancer-at first relapse. J Oncol. 2010; 2010: 497429.
16. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statist Sci. 1992; 7: 457-472.
17. Wei Z, Li H. A Markov random field model for network-based analysis of genomic data. Bioinformatics. 2007; 23: 1537-1544.
18. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. J Computat Biol. 2001; 8: 37-52.
19. Ulitsky I, Shamir R. Identification of functional modules using network topology and high-throughput data. BMC Syst Biol. 2007; 1: 8.
20. Dittrich MT, Klau GW, Rosenwald A, Dandekar T, Muller T. Identifying functional modules in protein-protein interaction networks: An integrated exact approach. Bioinformatics. 2008; 24: i223-i231.
21. Breitling R, Amtmann A, Herzyk P. Graph-based iterative group analysis enhances microarray interpretation. BMC Bioinformatics. 2004; 5: 100.
22. Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RK, Graeber MB. Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics. 2006; 7: 1-11.

23. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009; 4: 44-57.
24. Szczesny W, Vistad I, Kaern J, Nakling J, Trope C, Paulsen T. Impact of hospital type and treatment on long-term survival among patients with FIGO Stage IIIC epithelial ovarian cancer: Follow-up through two recurrences and three treatment lines in search for predictors for survival. Eur J Gynaecol Oncol. 2016; 37: 305-311.
25. Stalberg K, Crona J, Razmara M, Taslica D, Skogseid B, Stalberg P. An integrative genomic analysis of formalin fixed paraffin-embedded archived serous ovarian carcinoma comparing long-term and short-term survivors. Int J Gynecol Cancer. 2016; 26: 1027-1032.
26. Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, et al. Cancer treatment and survivorship statistics, 2019. CA Cancer J Clin. 2019; 69: 363-385.
27. Alsina-Sanchis E, Figueras A, Lahiguera A, Gil-Martin M, Pardo B, Piulats JM, et al. TGFbeta controls ovarian cancer cell proliferation. Int J Mol Sci. 2017; 18.
28. De Luca A, Maiello MR, D'Alessio A, Pergameno M, Normanno N. The RAS/RAF/MEK/ERK and the PI3K/AKT signalling pathways: Role in cancer pathogenesis and implications for therapeutic approaches. Expert Opin Ther Targets. 2012; 16 Suppl 2: S17-S27.
29. Chen L, Xuan J, Riggins RB, Clarke R, Wang Y. Identifying cancer biomarkers by network-constrained support vector machines. BMC Syst Biol. 2011; 5: 161.
30. Shimokawa M, Kogawa T, Shimada T, Saito T, Kumagai H, Ohki M, et al. Overall survival and post-progression survival are potent endpoint in phase III trials of second/third-line chemotherapy for advanced or recurrent epithelial ovarian cancer. J Cancer. 2018; 9: 872-879.
31. Sjoquist KM, Lord SJ, Friedlander ML, John Simes R, Marschner IC, Lee CK. Progression-free survival as a surrogate endpoint for overall survival in modern ovarian cancer trials: A meta-analysis. Ther Adv Med Oncol. 2018; 10: 1758835918788500.

Enjoy OBM Genetics by:

1. Submitting a manuscript
2. Joining in volunteer reviewer bank
3. Joining Editorial Board
4. Guest editing a special issue

For more details, please visit:
http://www.lidsen.com/journals/genetics