Characterizing Cancer-Specific Networks by Integrating TCGA Data

Libertas Academica
FREEDOM TO RESEARCH

Yanxun Xu[1], Yitan Zhu[2], Peter Müller[3], Riten Mitra[4] and Yuan Ji[2,5]

[1]Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX, USA. [2]Northshore University HealthSystem, Evanston, IL, USA. [3]Department of Mathematics, The University of Texas at Austin, Austin, TX, USA. [4]School of Public Health and Information Sciences, The University of Louisville, Louisville, KY, USA. [5]Department of Public Health Sciences, The University of Chicago, Chicago, IL, USA.

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

ABSTRACT: The Cancer Genome Atlas (TCGA) generates comprehensive genomic data for thousands of patients over more than 20 cancer types. TCGA data are typically whole-genome measurements of multiple genomic features, such as DNA copy numbers, DNA methylation, and gene expression, providing unique opportunities for investigating cancer mechanism from multiple molecular and regulatory layers. We propose a Bayesian graphical model to systemically integrate multi-platform TCGA data for inference of the interactions between different genomic features either within a gene or between multiple genes. The presence or absence of edges in the graph indicates the presence or absence of conditional dependence between genomic features. The inference is restricted to genes within a known biological network, but can be extended to any sets of genes. Applying the model to the same genes using patient samples in two different cancer types, we identify network components that are common as well as different between cancer types. The examples and codes are available at https://www.ma.utexas.edu/users/yxu/software.html.

KEYWORDS: Bayesian graphical model, differential networks, genomics, integration, network, TCGA

SUPPLEMENT: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

CITATION: Xu et al. Characterizing Cancer-Specific Networks by Integrating TCGA Data. *Cancer Informatics* 2014:13(S2) 125–131 doi: 10.4137/CIN.S13776.

TYPE: Methodology

RECEIVED: December 03, 2014. **RESUBMITTED:** May 11, 2015. **ACCEPTED FOR PUBLICATION:** May 12, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,383 words, excluding any confidential comments to the academic editor.

FUNDING: Funding was provided by the National Institutes of Health under NIH 2R01 CA132897. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: koaeraser@gmail.com

Introduction

The Cancer Genome Atlas (TCGA)[1,2] records multilayer genomic measurements on thousands of patient samples from more than 20 types of cancer. The genomic features recorded by TCGA include gene expression (GE), protein expression (PE), DNA methylation (ME), gene copy number (CN), and more. Such a wealth of information enables the study of interaction networks of these features for a systematic investigation of molecular compositions and dynamics of cancer. Integration of cross-platform genomics data provides opportunities on learning the complex relationships in the biological system among different features.

Identifying the genomic interaction networks and understanding their behaviors in cancer conditions are critical to the elucidation of the molecular mechanisms of cancer, identification of cancer genes and pathways,[3–5] and discovery of network-based biomarkers that improve cancer diagnosis and prognosis.[6,7] TCGA provides unique opportunities for cancer genomics research. Specifically, it allows scientists to access measurements of different genomic and epigenetic features for matched patient samples. Consequently, a core challenge is to develop effective tools that integrate the multi-platform data and systematically screen the analysis results in hopes of novel discoveries that would not have been available by only looking at data from a single feature.

Some efforts have been directed to the integration of multi-platform genomic data. Many analyses mainly focus on pairwise integration of data generated by different platforms, for example, between DNA CN and GE or between DNA ME and GE. Waaijenborg et al.[8] proposed a penalized canonical correlation analysis to study genome-wide association between DNA CN and GE. Menezes et al.[9] modeled the relationship of DNA CN and GE by a linear model based on a modified correlation coefficient and an explorative Wilcoxon test. Choi et al.[10] described a Bayesian double-layered mixture model that directly modeled the stochastic nature of CN variation and identified abnormally expressed genes because of aberrant CN change. Etcheverry et al.[11] investigated the effect of ME on GE in glioblastoma and identified 13 genes that display an inverse correlation between ME and mRNA expression using Pearson's correlation coefficient. More sophisticated methods have been designed to integrate more than two types of genomic data. Shen et al.[12] developed a clustering method to integrate heterogeneous genomic data for identifying tumor subtypes. Li et al.[13] introduced a sparse multiblock partial least squares regression method to analyze multi-platform genomic data for identifying regulatory modules containing regulatory factors from different regulation layers that are likely to jointly regulate GE. They applied the method on TCGA ovarian cancer datasets and showed that DNA CN, DNA ME, and miRNA can have a coupled impact on the expression of important oncogenes and tumor suppressor genes.[14] Another attempt by Setty et al.[15] used regularized regression via a lasso constraint[16] to study GE regulation by miRNA, DNA ME, and DNA CN based on TCGA glioblastoma data.

We propose a Bayesian graphical model that imposes a probability distribution on the unknown networks and applies an autologistic prior to learning the dependence structure between genomic features. In the graph, nodes represent features (different assay platforms) of genes, and edges indicate the conditional dependence between the features. Through probabilistic and model-based inference, we formalize a framework that integrates multi-platform TCGA data and investigate relationship of different features within and between genes. The Bayesian graphical model provides full probabilistic inferences allowing for automatic adjustment of multiplicity and false discovery rate (FDR) assessment on the network, its subnetworks, and any edges in the network. In the next section, we give a brief overview of the breast invasive carcinoma (BRCA) data and head and neck squamous cell carcinoma (HNSC) data from TCGA, to which we apply our integration analysis. In Methods section, we introduce the proposed Bayesian graphical model along with Markov Chain Monte Carlo (MCMC) simulation details. Simulation study section presents several simulation studies to evaluate the performance of the proposed model. In Results section, we report results based on the analyses of TCGA BRCA data and HNSC data. We conclude with a discussion in Discussion and Conclusion section.

BRCA and HNSC Data

TCGA has collected over 1,000 cases of BRCA and 500 cases of HNSC. We downloaded level 3 TCGA data for analysis. These samples were measured for GE, DNA CN, DNA ME, or PE, although only a portion of them possess all four measurements. GE data were generated using Illumina HiSeq 2000 RNA sequencing platform. DNA CNs were produced using Affymetrix Genome-Wide Human SNP Array 6.0 platform and SNP array and Illumina HiSeq 2000 DNA sequencing platform. DNA ME data were generated using Illumina Infinium Human DNA ME 27 and 450 beadchips. PE was produced using reverse-phase protein arrays. GEs, CNs, and MEs are whole-genome measurements, including data of more than 20,000 genes. For DNA ME data, we computed the average ME value of the CpG sites that were within 1,500 nucleotide base pairs of transcription start site and were in DNase hypersensitive region. PE data generated by reverse-phase protein array measured protein abundances of about 150 important and cancer-related genes.

Methods

Sampling model. For a single gene, data are arranged in an $N \times P$ matrix $\boldsymbol{X} = [x_{ij}]$. Rows represent different samples, columns represent features such as GE or PE, and each element x_{ij} represents the measurement of each feature per sample, $i = 1,\ldots, N$ and $j = 1, 2,\ldots, P$. We introduce a latent trinary variable $c_{ij} \in \{-1, 0,1\}$ denoting under-, regular, and overexpression of the corresponding measurement, respectively. Given different values of c_{ij}, we assume the distribution of x_{ij} as follows

$$x_{ij} \mid c_{ij}, \theta_j = \begin{cases} \text{Uniform}(\mu_j - k_{j-}, \mu_j) & \text{when } c_{ij} = -1 \\ \text{Normal}(\mu_j, \sigma_j^2) & \text{when } c_{ij} = 0 \; , \\ \text{Uniform}(\mu_j, \mu_i + k_{j+}) & \text{when } c_{ij} = 1 \end{cases}$$

where Uniform(A) denotes a uniform distribution over the set A, Normal(·,·) denotes normal distribution, and the vector $\theta_j = (\mu_j, \sigma_j^2, k_{j-}, k_{j+})$ collects all the other parameters.

Parmigiani et al.[17] proposed the idea to describe the probability of GE data. This model is called the probability of expression model. In other words, we assume a mixture model with uniform, normal, and uniform components corresponding to under-, regular, and overexpression respectively, given by

$$\begin{aligned} x_{ij} \mid c_{ij}, \theta_j \sim{} & I[c_{ij} = -1] U(x_{ij} \mid \mu_j - k_{j-}, \mu_j) \\ & + I[c_{ij} = 0] N(x_{ij} \mid \mu_j, \sigma_j^2) + I[c_{ij} = 1] U(x_{ij} \mid \mu_j, \mu_j + k_{j+}), \end{aligned} \tag{1}$$

where $I[\cdot]$ is the indicator function.

To set up upcoming graphical models, we subsequently convert the trinary variable c_{ij} to a binary variable z_{ij} with $p(c_{ij}|z_{ij} = 0) = \delta_{-1}(c_{ij})$, and

$$p(c_{ij} = 0|\pi_j, z_{ij} = 1) = \pi_j, \; p(c_{ij} = 1|\pi_j, z_{ij} = 1) = 1 - \pi_j.$$

Graphical model. We use a graph to characterize the dependence structure across different features. Denote the graph by $G = \{V, E\}$ where $V = \{1,\ldots, P\}$ represents the set of P nodes standing for P features, E is a set of undirected edges $\{u,w\}$, $u,w \in V$. A graph G can be used to describe the conditional independence structure of a set of variables indexed by V, for example, the binary indicators $\{z_{ij}, j \in V\}$ in the case of our application. The absence of an edge $\{u,w\}$ indicates conditional independence of z_{iu} and z_{iw} given the remaining variables z_{ik}, $k \neq u$, $k \neq w$. Any joint probability model $p(z_{i1},\ldots, z_{iP})$ that respects the dependence structure G with up to second-order interactions can be written as[18]:

$$p(z_i \mid \beta, G) = p(0 \mid \beta, G) \times \exp\left\{ \sum_{j=1}^{P} \beta_j z_{ij} + \sum_{\{u,w\} \in V; u<w} \beta_{uw} z_{iu} z_{iw} \right\} \quad (2)$$

where $Z_i = (z_{i1},\ldots, z_{iP})$ and $\beta = (\beta_1,\ldots, \beta_P, \beta_{12},\ldots, \beta_{(P-1),P})$. Coefficients β_{uw} are nonzero only when the corresponding edge is included in the graph. Model (2) is known as the autologistic model.

For better mixing of the posterior simulation and simplicity of prior setup, Caragea and Kaiser,[19] Hughes et al.[20], and Mitra et al.[21] proposed an alternative MCMC scheme called centered parameterization. The centered version is used in the form of

$$\begin{aligned} p(z_i \mid \beta, G) &= p(0 \mid \beta, G) \\ &\times \exp\left\{ \sum_{j=1}^{P} \beta_j z_{ij} + \sum_{\{u,w\} \in V; u<w} \beta_{uw} (z_{iu} - h_u)(z_{iw} - h_w) \right\}, \end{aligned} \quad (3)$$

where $h_u = \exp(\beta_u)/\{1 + \exp(\beta_u)\}$. The sign of β_{uw} has an intuitively appealing interpretation. We can show that β_{uw} is the log odds ratio of z_u and z_w through simple algebra, where $\beta_{uw} > 0$ implies that $p(z_u = 1 \mid z_w = 1, z_{-uw}) > p(z_u = 1 \mid z_w = 0, z_{-uw})$. That is, a positive β value implies a positive interaction between the two nodes.

$$\begin{aligned} &p(X, c, z, \pi, \theta, \beta, G) \\ &\quad = p(X \mid c, \theta) p(c \mid z, \pi) p(z \mid \beta, G) p(\theta) p(\beta \mid G) p(\pi) p(G) \end{aligned} \quad (4)$$

We introduce the priors $p(\theta)$, $p(\beta \mid G)$, and $p(G)$ next. Let Ga(a,b) denote a gamma distribution with mean a/b. We assume conditionally conjugate priors

$$\mu_j \sim N(0, \tau_\mu), \quad \frac{1}{\sigma_j^2} \sim \mathrm{Ga}(\gamma_\sigma, \lambda_\sigma),$$
$$\frac{1}{k_{j-}} \sim \mathrm{Ga}(\gamma_{k_{j-}}, \lambda_{k_{j-}}), \quad \frac{1}{k_{j+}} \sim \mathrm{Ga}(\gamma_{k_{j+}}, \lambda_{k_{j+}}),$$
$$\beta_* \sim N(0, \sigma_\beta^2), \quad \pi_j \sim U(0,1),$$

where β_* stands for the coefficients β_j, β_{uw} in (3).

Finally, we define a model $p(G)$ by constructing an informative prior. Let $G_0 = (V, E_0)$ be a prior guess of the dependence structure. For example, if there exist mRNA expression and CNV in a single network, we could connect edges between E and C, since CNV is biologically known to be positively related to mRNA expression. Then, the prior of G is based on the number of changes to G_0 by assuming

$$p(G) \propto \rho^{d(G, G_0)}, \quad (5)$$

where $d(G,G_0) = |E \cap E_0^c| + |E^c \cap E_0|$. This prior setting imposes less weight on graphs that are more distant from G_0, and the weight decreases exponentially when d increases. Mitra et al.[21] discussed the convergence property of the proposed network model. We set $\rho = 0.8$ after careful calibration. Figure 1 shows the graphical model representation of the proposed network model.

MCMC simulations. We carry out posterior inference for model (4) using MCMC simulations. Each iteration of the MCMC scheme includes the following transition probabilities,

$$p(z \mid X, \pi, \theta, \beta, G), p(c \mid \pi, z), p(\pi \mid c),$$
$$p(\theta \mid X, c), p(\beta \mid z, G), p(G \mid z, \beta).$$

We start by generating Z_{ij} from its complete conditional posterior. Following the update of z, we generate values for c from complete conditional posterior $p(c \mid \pi, z)$. If $z_{ij} = 0$, the update is deterministic, $c_{ij} = -1$. If $z_{ij} = 1$, the update requires a Bernoulli draw for $c_{ij} = 0$ versus $c_{ij} = 1$. The update of parameter θ is straightforward. Resampling G and the regression coefficients β could be challenging in large graphs, essentially because of the difficult evaluation of the normalization constant $p(0 \mid \beta, G)$ in (3).[21] For small graphs, which means that the number P of features in one single network is smaller than 12, we calculate the normalization constant $p(0 \mid \beta, G)$ directly since $p(G)$ is only supported over 2^P possible graphs, making

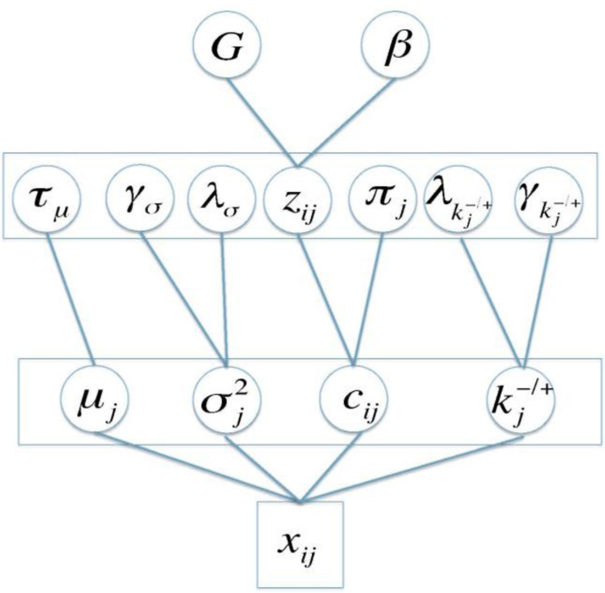

Figure 1. The graphical model representation of the proposed network model.

the evaluation of the normalization constant straightforward. Thus, resampling G and β reduces to straightforward transdimensional MCMC as in Ref. 22. When $P > 12$, we implement an importance sampling estimate[21] to approximate the ratio of the normalizing constants required for the evaluation of acceptance probabilities in the Metropolis–Hastings steps to update β and G.

FDR for edge inclusion. Statistically, owing to our fully model-based inference using posterior probabilities, we can easily assess the noise associated with the genomic data. This is a major advantage of our proposed Bayesian modeling approach over other algorithm-driven methods. Adopting the methods introduced by Newton et al.[23] and Müller et al.[14], we compute the FDR for a given cutoff β_0, given by

$$\mathrm{FDR}(\beta_0)=\frac{\sum_{u=1}^{P}\sum_{w>u}(1-\hat{\beta}_{uw})I(\hat{\beta}_{uw}\le\beta_0)}{\sum_{u=1}^{P}\sum_{w>u}I(\hat{\beta}_{uw}\le\beta_0)},$$

where $I(\cdot)$ is the indicator function, $\hat{\beta}_{uw}$ is the posterior mean of β_{uw} in (3). We use 0.01 for the cutoff on FDR to call significant edges in the analyses.

Simulation Study

To evaluate the proposed model, we examined the performance of our model with three simulated data sets, each with $N = 300$ samples and $P = 4,5,6$ features, respectively. For each of the three simulations, a simulation truth G was first generated. For each pair of nodes $\{u,w\}$, we included the edge with a probability 0.5. For each imputed edge $\{u,w\}$, we generated values of β_{uw} from $N(\mu_1, 0.5^2)$, $\mu_1 \sim U(-4, 4)$. For the autologistic intercepts j, we first generated β_j from $N(\mu_2,0.5^2)$, $\mu_2 \sim U(-0.4, 0.4)$, and subsequently generated z for $N = 300$ samples. Since $p(c_{ij} \mid z_{ij} = 0) = \delta_{-1}(c_{ij}), p(c_{ij} = 0|\pi_j, z_{ij} = 1) = \pi_j$ and $p(c_{ij} = 1|\pi_j,z_{ij} = 1) = 1 - \pi_j$, we first generated $\pi_j \sim U(0.2,0.8)$, and then generated c. Furthermore, we let $\mu_j = 0$, $\sigma_j = 0.548$, $k_{j-} = 4$, and $k_{j+} = 4$ for each node. The hyper-parameters were fixed and set at $\tau_\mu = 1, \gamma_\sigma = 2, \lambda_\sigma = 0.3, \gamma_{k+} = 11, \lambda_{k+} = 40, \gamma_{k-} = 11, \lambda_{k-} = 40$, and $\sigma_\beta = 3$.

We implemented our model to compute the posterior summaries for each simulated data set. The posterior estimates were obtained by MCMC posterior simulation with 20,000 iterations, of which 6,000 are burn-in thinned by 10. This left us 1,400 posterior samplers. The computations were completed in less than 30 seconds on a MacBook Pro with 3 GHz Intel Core i7. We calculated the posterior inclusion probability q_{uw} for each possible edge $\{u,w\}$ that is connected in the graph, which is defined as

$$q_{uw}=\frac{1}{1400}\sum I(\{u,w\}\in E)$$

substituting the edge set E of the imputed graph for each iteration of the MCMC. We then estimated the posterior graph $\hat{G}$ using the FDR control based on the criterion $\{q_{uw} > q_0\}$. The threshold q_0 was chosen such that the posterior expected FDR was smaller than 0.1. We also reported parameter estimates $\bar{\beta} = E(\beta| X)$ denoting the posterior mean for the autologistic coefficients.

Figure 2 plots the simulated true graphs and posterior estimated graphs for three simulated data sets. Edge colors black and red represent positive and negative relationships. The numbers next to edges stand for either the true β_{uw} values or the posterior mean. The estimated graphs based on posterior inference are identical to the simulation truth, and the posterior estimated $\hat{\beta}_{uw}$ are close to the true values.

Results

We implemented the proposed Bayesian graphical model and performed integrative inference based on BRCA and HNSC data described in BRCA and HNSC data section. We aim to discover the unknown dependence structure among different genomic features. The inference was obtained using the MCMC posterior simulation. The Markov chain included 20,000 iterations with 6,000 burn-in thinned by 10. The chain converged and mixed well. Three examples are showed in this section.

Multiple-genes network analysis. Transcription is an important genetic process in which DNA is transcribed to RNA. Perturbation of transcription may affect GE, and hence the subsequent protein production, leading to pathological states. Using the measurements of GE, CN and ME, one can learn the potential regulations within and between genes.

We chose two famous pathways including the MTOR signaling pathway and the p53 signaling pathway to show the posterior inference.

We selected measurements of GE, CN, and ME of three genes *PTEN*, *INS*, and *MAPK* in MTOR signaling pathway in BRCA data. So, the graph has nine nodes. Figure 3 shows the posterior estimated network. Edge colors black and red represent positive and negative relationships, respectively. The numbers

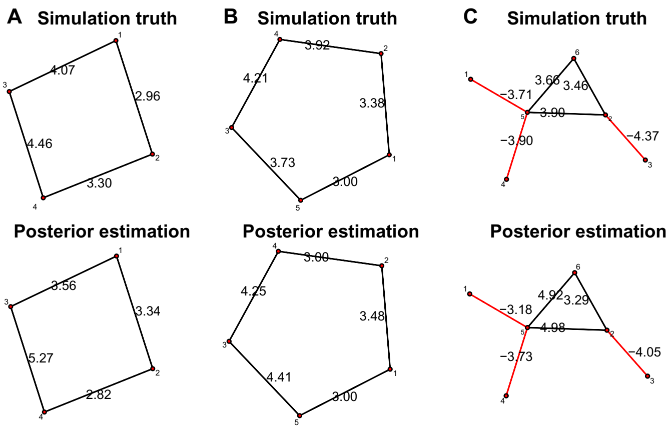

Figure 2. The simulation truth versus the estimated graph for three simulated data sets. Black edges represent positive relationship and red edges represent negative relationships. The number next to each edge represents either the true value or the posterior mean of the autologistic coefficients β_{uw}s. (**A**) Data set 1; (**B**) data set 2; (**C**) data set 3.

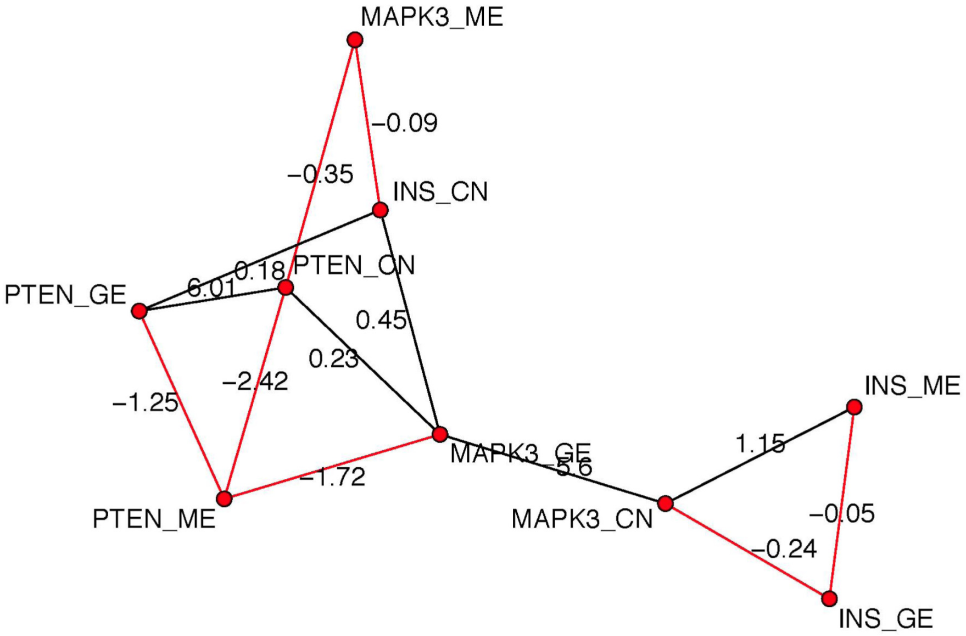

Figure 3. The posterior estimated graph for GE, CN, and ME of three genes *PTEN*, *INS*, and *MAPK* in MTOR signaling pathway in BRCA data. Black edges represent positive relationship and red edges represent negative relationships. The number next to each edge represents the posterior mean of the autologistic coefficients β_{uw}s.

next to edges stand for the posterior mean of autologistic coefficient β_{uw}. We find for gene *PTEN*, GE is positively related to CN and negatively related to ME, which is consistent with the biology finding that usually CN is positively related to GE while DNA ME is negatively related to GE. Similarly, GE and CN have positive relationship within gene *MAPK3*. In addition, given CN and GE of gene *MAPK3*, ME and GE of gene *INS* are conditionally independent with all other features.

For p53 signaling pathway, we chose features GE, CN, and ME of genes *ATM*, *TP53*, and features GE, CN of

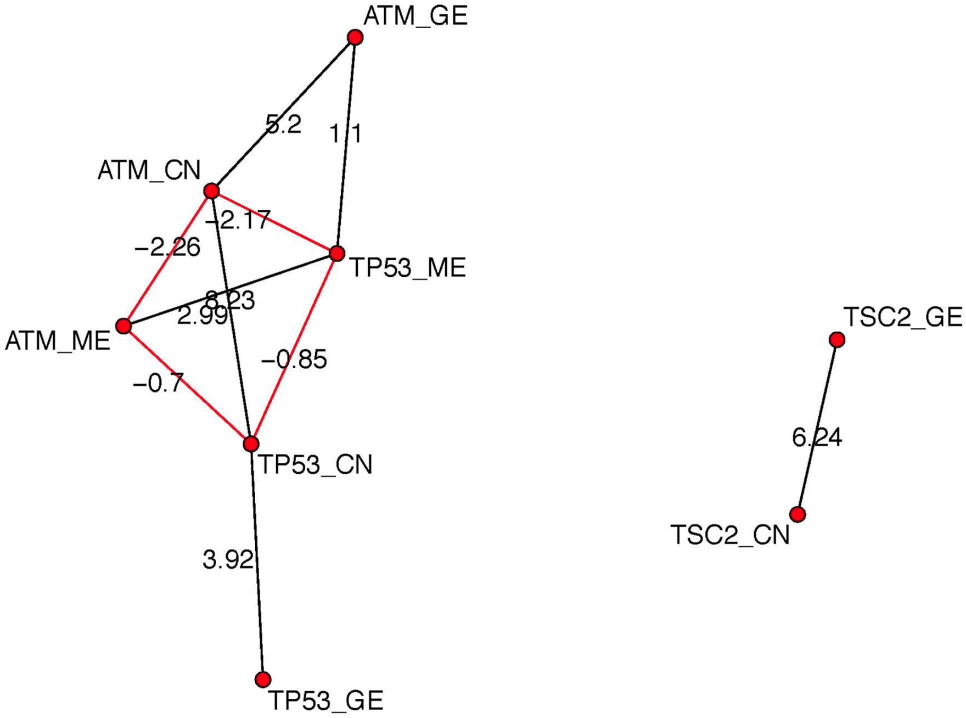

Figure 4. The posterior estimated graph for GE, CN, and ME of three genes *ATM*, *TP53*, and *TSC2* in p53 signaling pathway in BRCA data. Black edges represent positive relationship and red edges represent negative relationships. The number next to each edge represents the posterior mean of the autologistic coefficients β_{uw}s.

TSC2. Thus, our network has eight nodes. Figure 4 shows the posterior estimated network. Within one single gene, we can observe that CN is positively related to GE, while ME is negatively related to GE, which agrees with known biology knowledge. There is no edge between features of gene *TSC2* and features of other two genes *TP53* and *ATM*, indicating that *TSC2* functions independently with *TP53* and *ATM*.

Cancer-specific network analysis. A main goal of the proposed network model is to learn the cancer-specific networks so that the constructed networks can be used for the identification of different cancer types.

We selected two features GE and PE of four genes *MTOR*, *IRS1*, *PTEN*, and *AKT1* in two different cancer types – BRCA and HNSC. All four genes function in MTOR signaling pathway. Figure 5 plots the posterior estimated graph. Comparing Figure 5A and B, we observe similarities and differences. In both cancer types, PE of gene *MTOR* is actively connected to other features. However, in BRCA, GE of *MTOR* does not have an edge with any other features, while in HNSC, GE of *MTOR* has an edge with four other features. Three of them are negative relationships. We observed edges consistent with existing understanding about the pathway. For example, PE of *MTOR* has a positive edge with PE of *AKT1* in BRCA, which is consistent with the fact that *MTOR* is an activator of *AKT1*. Also, in HNSC, PE of *PTEN* has a negative edge with PE of *AKT1* showing the effect of *PTEN* as an indirect repressor of *AKT1* through *PIP3* and *PDK1*. However, this edge becomes positive in BRCA condition indicating perturbations to MTOR signaling pathway in breast cancer. Comparing the topologies between two networks in different cancer types will be an interesting future direction.

Discussion and Conclusion

To study dependence structure of different genomic features within genes and between genes, we propose a Bayesian graphical model. The inferred graph gives a clear representation of the regulatory relationships. For example, GE of *MTOR* is positively related to PE of *MTOR*, but negatively related to PEs of *AKT1*, *PTEN*, and *IRS1* in MTOR signaling pathway in HNSC. Our model can be easily applied to more features to learn the regulation relationship between features of different genes. For example, gene A could be a transcription factor for gene B if we detect a strong relationship between PE of gene A and mRNA expression of gene B. The proposed graphical model can well integrate multimodal genomic data from TCGA, including GE, PE, DNA ME, and gene CN, as shown in the analyses. It can also work on miRNA expression data, and we will extend the model to integrate DNA mutation data. In addition, we plan to implement the module linking the network topology with patient survival as a future extension. We are making a comprehensive list of these relationships using the entire TCGA data, expanding the effort to include more cancer types and more features such as microRNA and transcription factor. It is important to note that some interactions (such as CN–ME interactions) computationally derived by our proposed model might not be physical interactions, but could be indirect interactions through potentially multiple signaling steps that are not included in the current analysis. Confounding factors not included in analysis model can also contribute to the identified interactions. Nonetheless, we believe that the Bayesian graphical model will benefit researchers in different areas because of its rigor in statistical modeling for inferring interaction networks based on TCGA data.

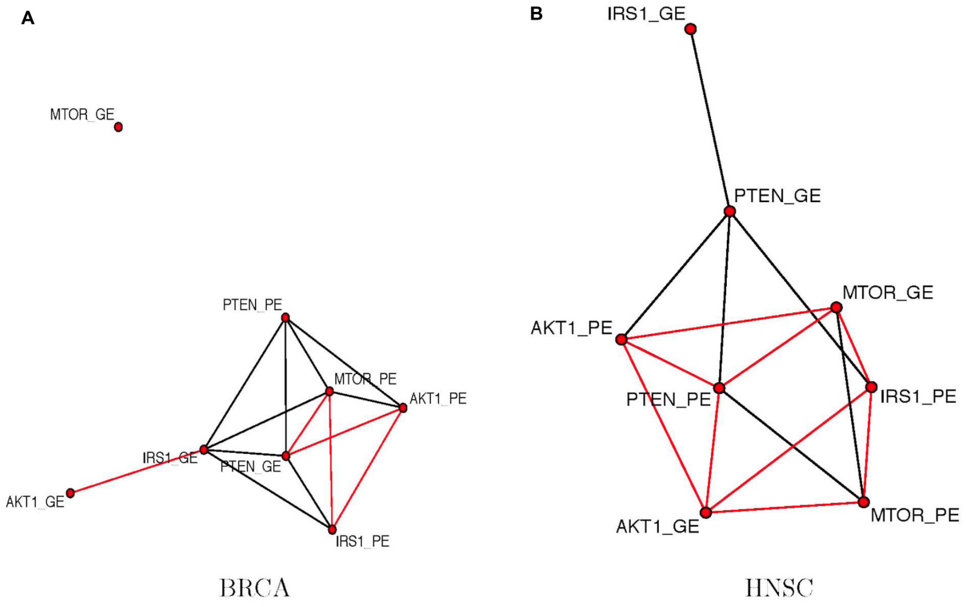

Figure 5. The posterior estimated graphs of features (GE and PE) among genes *MTOR*, *IRS1*, *PTEN*, and *AKT1* in two different cancer types. Black edges represent positive relationship and red edges represent negative relationships. (**A**) BRCA and (**B**) HNSC.

Author Contributions

Conceived and designed the experiments: YJ, PM. Analyzed the data: YX, YZ. Wrote the first draft of the manuscript: YX, YZ. Contributed to the writing of the manuscript: YX, YZ, YJ. Agree with manuscript results and conclusions: YX, YZ, PM, RM, YJ. All authors reviewed and approved of the final manuscript.

REFERENCES

1. The Cancer Genome Atlas Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061–8.
2. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70.
3. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. *Bioinformatics*. 2006;22(18):2291–7.
4. Xu J, Li Y. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*. 2006;22(22):2800–5.
5. Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. *PLoS Comput Biol*. 2014;10(10):e1003908.
6. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7(3–4):601–20.
7. Development Network; Schreiber SL, Shamji AF, et al. Towards patient-based cancer therapeutics. *Nat Biotechnol*. 2010;28(9):904–6.
8. Waaijenborg S, de Witt Hamer V, Philip C, Zwinderman A. Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Stat Appl Genet Mol Biol*. 2008;7(3):Article3.
9. Menezes RX, Boetzer M, Sieswerda M, Van Ommen GJB, Boer JM. Integrated analysis of DNA copy number and gene expression microarray data using gene sets. *BMC Bioinformatics*. 2009;10(1):203.
10. Choi H, Qin ZS, Ghosh D. A double-layered mixture model for the joint analysis of DNA copy number and gene expression data. *J Comput Biol*. 2010;17(2):121–37.
11. Etcheverry A, Aubry M, De Tayrac M, et al. DNA methylation in glioblastoma: impact on gene expression and clinical outcome. *BMC Genomics*. 2010;11(1):701.
12. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*. 2009;25(22):2906–12.
13. Li W, Zhang S, Liu CC, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. *Bioinformatics*. 2012;28(19):2458–66.
14. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing. *J Am Stat Assoc*. 2004;99(468):990–1001.
15. Setty M, Helmy K, Khan AA, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. *Mol Syst Biol*. 2012;8(1):605.
16. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B*. 1996;58:267–88.
17. Parmigiani G, Garrett ES, Anbazhagan R, Gabrielson E. A statistical framework for expression-based molecular classification in cancer. *J R Stat Soc Series B Stat Methodol*. 2002;64(4):717–36.
18. Besag J. Spatial interaction and the statistical analysis of lattice systems. *J R Stat Soc Series B*. 1974;36:192–236.
19. Caragea PC, Kaiser MS. Autologistic models with interpretable parameters. *JABES*. 2009;14(3):281–300.
20. Hughes J, Haran M, Caragea PC. Autologistic models for binary data on a lattice. *Environmetrics*. 2011;22:374–8.
21. Mitra R, Müller P, Liang S, Yue L, Ji Y. A Bayesian graphical model for chip-seq data on histone modifications. *J Am Stat Assoc*. 2013;108(501):69–80.
22. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*. 1995;82(4):711–32.
23. Newton MA, Noueriry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics*. 2004;5:155–76.