Zodiac: A Comprehensive Depiction of Genetic Interactions in Cancer by Integrating TCGA Data

(Supplementary Information)

Yitan Zhu[§1], Yanxun Xu[§2], Donald L. Helseth Jr[§3], Kamalakar Gulukota[3], Shengjie Yang[1], Lorenzo L. Pesce[4], Riten Mitra[5], Peter Müller[2], Subhajit Sengupta[1], Wentian Guo[6], Jonathan C. Silverstein[1], Ian Foster[4], Nigel Parsad[1], Kevin P. White[7,8], Yuan Ji[*1,9]

1. Center for Biomedical Research Informatics, NorthShore University HealthSystem, Evanston, Illinois, USA
2. Department of Mathematics, The University of Texas at Austin, Austin, Texas, USA
3. Center for Molecular Medicine, NorthShore University HealthSystem, Evanston, Illinois, USA
4. Computation Institute, The University of Chicago and Argonne National Laboratory, Chicago, Illinois, USA
5. Department of Bioinformatics & Biostatistics, University of Louisville, Louisville, Kentucky, USA
6. School of Public Health, Fudan University, Shanghai, P.R. China
7. Institute for Genomics and Systems Biology, The University of Chicago and Argonne National Laboratory, Chicago, Illinois, USA
8. Department of Human Genetics and Department of Ecology & Evolution, The University of Chicago, Chicago, Illinois, USA
9. Department of Health Studies, The University of Chicago, Chicago, Illinois, USA

[§] Equal contribution authors
[*] Correspondence author

1 Bayesian Graphical Models (BGM)

1.1 Introduction of Graphical Models

Denoting with C, M, E, and P the four genomics features of copy number, DNA methylation, mRNA expression, and protein expression, we apply a Bayesian graphical model[1] to learn the dependence structure of these features through a graph. The vertices of the graph represent the features, and the presence or absence of edges indicates the conditional dependence or independence between the features, respectively. For example, an edge between M and E and a lack of edge between C and E implies methylation-controlled transcription, which is robust to copy number changes. In other words, the mRNA expression is sensitive to methylational variation but not copy number variation.

Exploring conditional independence among a set of random variables is a classical statistical inference problem. Bayesian inference enables a stochastic exploration of the graphical space by introducing priors on the graph itself to regularize the otherwise unstable estimation. We consider Markov random fields (MRF) models[2] , and introduce binary latent indicators of the presence of genomics variations. We define an MRF as a pair $G = (V, \mathcal{E})$, where V is a set of vertices and $\mathcal{E}$ is a set of undirected edges. The vertices correspond to the variables, in our case genomic features, C, M, E, or P for a single gene. The edges in $\mathcal{E}$ are a subset of $\{\{i, j\},\ i \neq j \in V\}$. A path is defined as an ordered set of vertices $(i_0, i_1, \ldots i_n)$ such that $\{i_{k-1}, i_k\} \in \mathcal{E}$ for $k = 1, \ldots, n$.

For TCGA applications, instead of using directed graphs, MRFs that do not consider directionality are suitable for two reasons. First, there are not time-course data in TCGA thus virtually eliminating the possibility of performing formal statistical inference based on directed graphs. Second, if needed, most edge directionalities can be easily deduced from biological knowledge. For example, an edge between C and E of the same gene implies that the copy number variations (CNVs) of that gene affect the mRNA expression of the gene, i.e., $C \rightarrow E$.

1.2 Methods

1.2.1 Probability Model

For each gene, data are arranged in an $S \times T$ matrix $\boldsymbol{Y} = [y_{it}]$ where rows i represent the genomic features of the gene, columns t represent different biological samples, and each element y_{it} represents the measurement of each feature for each sample, $i = 1, \ldots, S$ and $t = 1, 2, \ldots, T$. The proposed model introduces latent trinary indicators $z_{it} \in \{-1, 0, 1\}$ with interpretation as under-, regular and over-expression of the corresponding measurement as follows:

$$z_{it} = \begin{cases} -1 & \text{abnormally low measurement,} \\ 0 & \text{normal measurement ,} \\ 1 & \text{abnormally high measurement.} \end{cases}$$

Using z_{it} we apply the mixture model[3] for y_{it}, given by

$$\begin{aligned} (y_{it} - \mu_i) \mid z_{it}, \boldsymbol{\theta}_i \;\; \sim \;\; & I[z_{it} = -1] U(y_{it} \mid -k_{i-}, 0) + I[z_{it} = 0] N(y_{it} \mid 0, \sigma_i^2) \\ & + I[z_{it} = 1] U(y_{it} \mid 0, k_{i+}), \end{aligned} \tag{1}$$

where $I[\cdot]$ is the indicator function, μ_i is the random effect of feature i, $U(A)$ denotes a uniform distribution over the set A, and $N(\cdot \mid \mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. In words, we assume a mixture model with uniform, normal and uniform components corresponding to under-, regular and over-expression. The vector $\boldsymbol{\theta}_i = (\mu_i, \sigma_i^2, k_{i-}, k_{i+})$ collects all the other parameters.

We subsequently convert the trinary variable z_{it} to a binary variable e_{it} with $p(z_{it}|e_{it} = 0) = \delta_{-1}(z_{it})$, and

$$p(z_{it} = 0|\pi_i, e_{it} = 1) = \pi_i, \quad p(z_{it} = 1|\pi_i, e_{it} = 1) = 1 - \pi_i.$$

This conversion is needed to set up the following graphical model.

Denote $V = \{1, \ldots, S\}$ the set of S vertices representing S genomic features. Recall that a graph is a pair $G = \{V, \mathcal{E}\}$ where $\mathcal{E}$ is a set of undirected edges $\{i, j\}$, $i, j \in V$. A graph G

is used to describe the conditional independence structure of a set of variables indexed by V, for example the binary indicators $\{e_{it},\ i \in V\}$. The absence of an edge $\{i,j\}$ indicates conditional independence of e_{it} and e_{jt} given the remaining variables e_{kt}, $k \neq i, k \neq j$. Any joint probability model $p(e_{1t}, \ldots, e_{St})$ that respects the dependence structure G can be written as[4] :

$$p(\boldsymbol{e}_t \mid \boldsymbol{\beta}, G) = p(0 \mid \boldsymbol{\beta}, G) \times \exp\left\{ \sum_{i=1}^{S} \beta_i e_{it} + \sum_{\{i,j\} \in V; i<j} \beta_{ij} e_{it} e_{jt} \right\}, \tag{2}$$

where $\boldsymbol{e}_t = (e_{1t}, \ldots, e_{St})$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_S, \beta_{12}, \ldots, \beta_{S-1,S})$. For instance, we have $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_{12}, \beta_{23}, \beta_{13})$ when $S = 3$. Coefficients β_{ij} are non-zero only when the corresponding edge is included in the graph. Model (2) is known as the autologistic model.

An alternative scheme called centered parametrization[5−6] improves mixing of posterior simulation and simplifies prior specification. The centered version is used in the form of

$$p(\boldsymbol{e}_t \mid \boldsymbol{\beta}, G) \;\; = \;\; p(0 \mid \boldsymbol{\beta}, G) \times \exp\left\{ \sum_{i=1}^{S} \beta_i e_{it} + \sum_{\{i,j\} \in V; i<j} \beta_{ij} (e_{it} - \nu_i)(e_{jt} - \nu_j) \right\}, \tag{3}$$

where $\nu_i = \exp(\beta_i)/\{1 + \exp(\beta_i)\}$.

The joint model factors as

$$p(\boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{e}, \boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\beta}, G) = p(\boldsymbol{Y} \mid \boldsymbol{z}, \boldsymbol{\theta}) p(\boldsymbol{z} \mid \boldsymbol{e}, \boldsymbol{\pi}) p(\boldsymbol{e} \mid \boldsymbol{\beta}, G) p(\boldsymbol{\theta}) p(\boldsymbol{\beta} \mid G) p(G). \tag{4}$$

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

$$\mu_i \sim N(0, \tau_\mu), \;\; \frac{1}{\sigma_i^2} \sim \mathrm{Ga}(\gamma_\sigma, \lambda_\sigma),$$

$$\frac{1}{k_{i-}} \sim \mathrm{Ga}(\gamma_{k_{i-}}, \lambda_{k_{i-}}), \quad \frac{1}{k_{i+}} \sim \mathrm{Ga}(\gamma_{k_{i+}}, \lambda_{k_{i+}}),$$

$$\beta_i, \beta_{ij} \overset{indep.}{\sim} N(0, \sigma_\beta^2), \quad \pi_i \sim U(0,1).$$

Lastly, we define a model $p(G)$. Let $G_0 = (V, \mathcal{E}_0)$ be a prior guess of the dependence structure. For genomic inference, G_0 can be often easily elicited. For example, one could connect the edge between C and E, since CNV is biologically known to be positively related to gene expression. Therefore, G_0 woud be a graph with three vertices C, M, and E and an edge set $\mathcal{E}_0$. Knowing G_0, the first option of the prior of G is based on the number of changes to G_0 by assuming a geometric kernel

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

where $d(G, G_0) = |\mathcal{E} \bigcap \mathcal{E}_0^c| + |\mathcal{E}^c \bigcap \mathcal{E}_0|$ and $\rho \in (0, 1)$. This prior setting imposes less weight on graphs that are more distant from G_0 and the weights decreases exponentially when the distance d increases. The prior (5) works well for large graphs with say, > 10 vertices. In real data applications for TCGA, we specify G_0 according to biological knowledge: we assume an edge between C and E, M and E, and E and P for the same gene. We did not assume any other edges in the prior graph G_0. Therefore, G_0 reflects the biological belief that copy number and DNA methylation affect gene expression, and gene expression and protein expression are dependent. We assign parameter ρ different values according to the size of the graph. In particular, $\rho = 0.9$ for graphs with 3 nodes, 0.8 for graphs with 4-5 nodes, 0.3 for graphs with 6 nodes, and 0.1 for graphs with 7+ nodes. We perform extensive simulations to evaluate the model when ρ take these values; results of the simulation are presented in Section1.3.

1.2.2 Markov Chain Monte Carlo Simulations

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

$$p(\boldsymbol{e} \mid \boldsymbol{Y}, \boldsymbol{\pi}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \boldsymbol{\beta}, G), p(\boldsymbol{z} \mid \boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{e}), p(\boldsymbol{\pi} \mid \boldsymbol{z}),$$

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}, \boldsymbol{z}, \boldsymbol{\alpha}), p(\boldsymbol{\beta} \mid \boldsymbol{e}, G), p(G \mid \boldsymbol{e}, \boldsymbol{\beta}).$$

We start by generating $\boldsymbol{e}$ from its complete conditional posterior. Following the update of $\boldsymbol{e}$, we generate values for $\boldsymbol{z}$ from complete conditional posterior $p(\boldsymbol{z} \mid \boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{e})$. If $e_{it} = 0$, the update is deterministic, $z_{it} = -1$. If $e_{it} = 1$, the update requires a Bernoulli draw for $z_{it} = 0$ versus $z_{it} = 1$. The update of parameters $\boldsymbol{\theta}$ is straightforward. Resampling G and the regression coefficients $\boldsymbol{\beta}$ could be challenging in large graphs, essentially because of the difficult evaluation of the normalization constant $p(0 \mid \boldsymbol{\beta}, G)$ in (3)[1].

1.2.3 Posterior Inference Using False Discovery Rates

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. Denoting λ a generic symbol for a probability of interest, and adopting the methods introduced by Newton et al.[7] and Müller et al.[8], we compute the posterior expected false discovery rate ($p\overline{\text{FDR}}$) for a given cutoff λ_0, given by

$$p\overline{\text{FDR}}(\lambda_0) = \frac{\sum_{i=1}^{S}\sum_{j>i}(1-\hat{\lambda}_{ij})I(\hat{\lambda}_{ij} \leq \lambda_0)}{\sum_{i=1}^{S}\sum_{j>i}(\hat{\lambda}_{ij} \leq \lambda_0)},$$

where $I(\cdot)$ is the indicator function, $\hat{\lambda}_{ij}$ is a posterior estimate of λ_{ij}. Different cutoff values λ_0 can be used for FDR control such that $p\overline{\text{FDR}} < p_0$ for a desirable rate p_0.

1.3 Simulation Study

1.3.1 A Small Study

Here we examined the performance of the graphical models with three simulated data sets, each with $T = 350$ samples and $S = 3, 4, 5$ features, respectively. Hence, the number of vertices in a graph was between 3 and 5. For each simulation, a true graph G was first generated. For each pair of vertices $\{i, j\}$, we generated the edge with probability 0.5. For each imputed edge $\{i, j\}$, we generated values of β_{ij} from $N(\mu_1, 0.5^2)$, with $\mu_1 \sim U(-4, 4)$. We generated β_i, the autologistic intercept in (3) from $N(\mu_2, 0.5^2)$, and $\mu_2 \sim U(-0.4, 0.4)$. Then, we generated $\boldsymbol{e}$ for $T = 350$ samples. Since $p(z_{it} \mid e_{it} = 0) = \delta_{-1}(z_{it})$, $p(z_{it} = 0 | \pi_i, e_{it} = 1) = \pi_i$, and

$p(z_{it} = 1|\pi_i, e_{it} = 1) = 1 - \pi_i$, we first generated $\pi_i \sim U(0.2, 0.8)$ and then generated $\boldsymbol{z}$. Furthermore, we set $\mu_i = 0, \sigma_i = 0.316, k_{i-} = 4$, and $k_{i+} = 4$ for each feature i. The hyper-parameters were $\tau_\mu = 1$, $\gamma_\sigma = 2, \lambda_\sigma = 0.1$, $\gamma_{k_+} = 11, \lambda_{k_+} = 40$, $\gamma_{k_-} = 11, \lambda_{k_-} = 40$, and $\sigma_\beta = \sqrt{10}$.

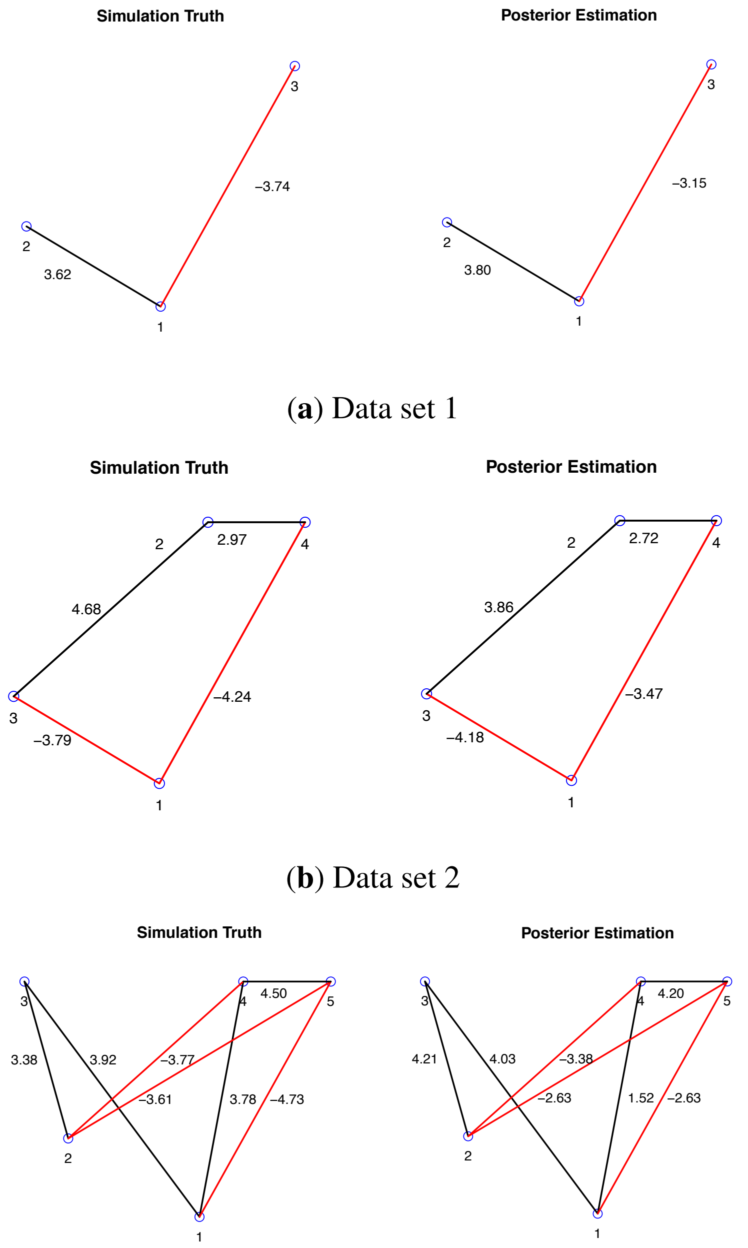
In Fig. S1, the plots in the left panel present the three simulated true graphs. Black edges and red edges represent positive and negative interactions, respectively. For instance, in data set 1, features 1 and 2 are positively related, and features 1 and 3 are negatively related. Features 2 and 3 are conditionally independent given feature 1, i.e., $\beta_{23} = 0$ in the autologistic model (3).

We implemented the proposed graphical model to compute the posterior estimates of parameters for each simulated data set. The posterior estimates were obtained through MCMC sampling with 10,000 iterations, of which the first 6,000 were discarded as burn-in (thinning every 10 iterations). We calculated the posterior inclusion probability q_{ij} for each possible edge $\{i, j\}$, defined as

$$q_{ij} = \frac{1}{B} \sum I(\{i, j\} \in \mathcal{E})$$

substituting the edge set $\mathcal{E}$ of the imputed graph for each iteration of the MCMC. Here B is the number of MCMC samples kept for analysis. We obtained the posterior estimated graph $\hat{G}$ by thresholding, based on a criterion $\{q_{ij} > q_0\}$, using the posterior inclusion probability q_{ij} for each edge. The threshold q_0 was chosen so that the posterior expected false discovery rate $p\overline{\text{FDR}}(q_0) \leq 0.01$. We also reported parameter estimates of regression coefficients $\bar{\boldsymbol{\beta}} = E(\boldsymbol{\beta} \mid \boldsymbol{Y})$, the posterior mean for the autologistic coefficients. Fig. S1 plots the posterior estimated graph for the simulated data. The number next to each edge represents either the true value (left panel) or the posterior mean β_{ij}'s (right panel). We can see that the estimated graph match the simulation truth for all three datasets, with similar estimated values of the β's.

Since graph G is modeled as a random variable, we also reported the inference $r = P(G = G_0|data)$, where G_0 is the simulation truth. For the three data sets $r = 0.534, 0.52$, and 0.236, respectively. The last r value is smaller since the true graph in that last simulation had more edges, thus increasing the complexity in the estimation. In other words, the less sparse the true graph is, the less powerful the inference.



(**a**) Data set 1

(**b**) Data set 2

(**c**) Data set 3

Figure S1 The simulation truth versus the estimated graph for three simulated data sets. Edge colors black and red represent positive and negative relationships, respectively. The number next to each edge represents either the true value (left graph) or the posterior mean (right graph) of the autologistic coefficients β_{ij}'s. The estimated graph based on posterior inference is identical to the simulation truth.

The sign of β_{ij} has an intuitively appealing interpretation related to the effect of the j-th feature on the probability of presence of i-th feature, keeping the other feature fixed. Let $e_{-ij} = \boldsymbol{e} \backslash \{e_{it}, e_{jt}\}$. It can be easily shown that β_{ij} is the log odds ratio of e_{it} and e_{jt} through simple algebra, where $\beta_{ij} > 0$ implies that $p(e_{it} = 1 \mid e_{jt} = 1, e_{-ij}) > p(e_{it} = 1 \mid e_{jt} = 0, e_{-ij})$. Due to this nice interpretation, we use the magnitude of the β values as a measure for the line thickness when reporting the posterior networks in Zodiac.

1.3.2 A Large Study

Building on the encouraging results of the small simulation study, we conducted a large study involving many data sets and configurations of ρ. For a fixed number of vertices ranging from 3 to 8 and a value of $\rho \in \{0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ we generated 20 simulation data sets based on the scheme in the previous subsection. Fig. S2 summarizes the false nondiscovery rates (FNR) and false discovery rates (FDR) across the 20 data sets for each graph size and ρ value. Also, the number of samples was 500, 650, 800, 950, 1,100, and 1,250 for graphs with 3 – 8 vertices, respectively. Examining the results, we chose selected ρ values reported in Subsection 1.2.1 for Zodiac runs.

1.3.3 Comparison to Partial Correlation

We have explained the differences and advantages of the Bayesian graphical models versus standard correlation-based inference. Here, we further show that the proposed Bayesian graphical models are different and arguably more powerful than the partial correlation, which also computes the association of two random variables in the presence of other random variables. We considered a special case in which we had four features and the true graph was a rhombus, as in data set 2 in the simulation: feature 1 connected with 2, 2 connected with 3, 3 connected with 4, 4 connected with 1. There were no edges between 1 and 3, or between 2 and 4 as shown in Fig. S1 (data set 2). After implementing the proposed graphical model, the posterior estimated graph was the same as the true graph, and the posterior estimates of $\boldsymbol{\beta}$ were close to the truth.

To compare with inference based on partial correlation, we used R function PCOR.TEST to infer conditional independence between two variables. According to the simulation truth, features 1 and 3 were conditionally independent given both 2 and 4, but dependent conditional on either 2 or 4, but not both. Our graphical model obtained the right inference, providing the identical graphical presentation (Fig. S1, data set 2). However, using partial correlation we would conclude that features 1 and 3 were conditionally independent given feature 2 (P value < 0.01), and also conditionally independent given feature 4 (P value < 0.01). Therefore, partial correlation analysis failed to capture the true conditional independence structure of the graph, and provided wrong estimates of the conditional dependence relationship involving the four vertices.

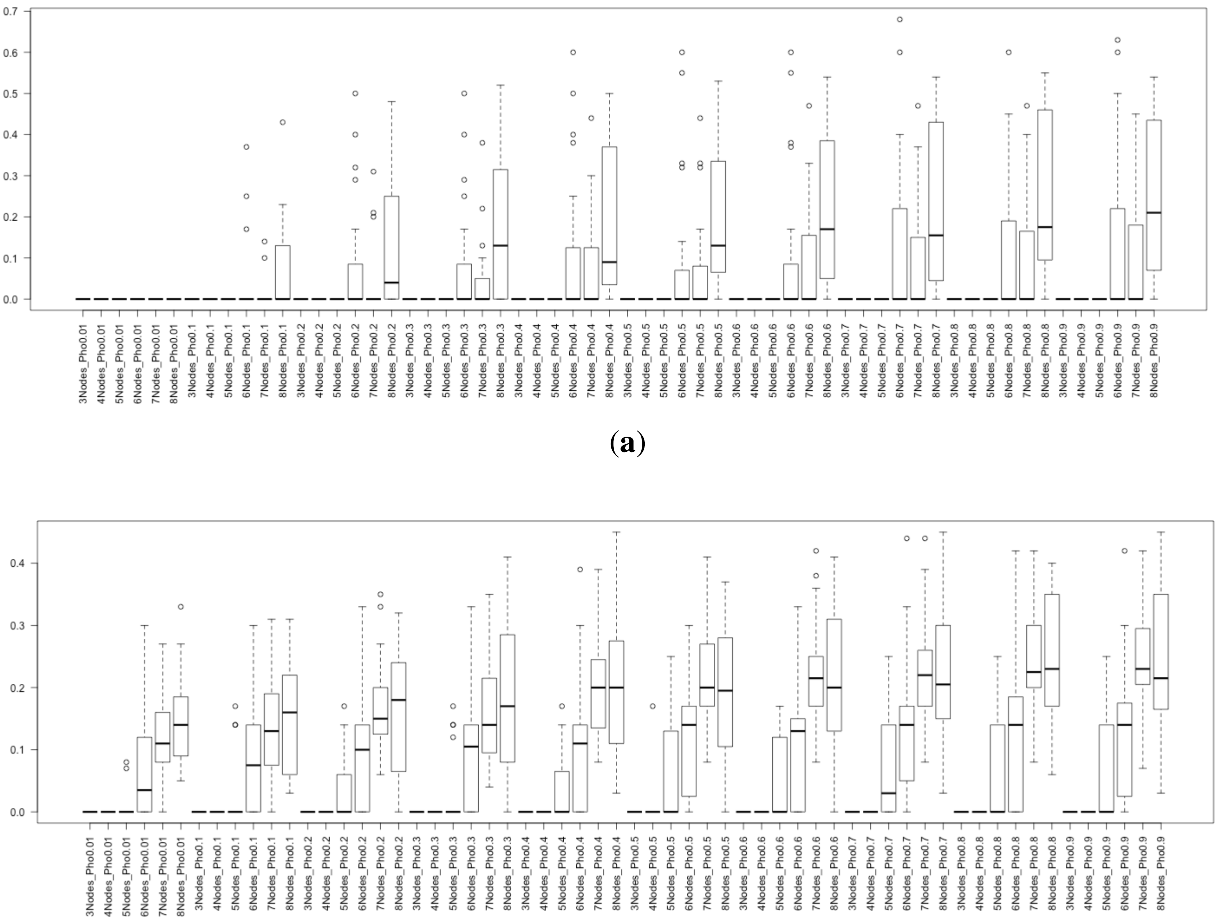


Figure S2 Summary of simulation results. Each boxplot summarizes FNRs (**a**) and FDRs (**b**) across 20 simulation data sets for a fixed ρ and graphs with a fixed number of vertices.

2. Analysis Details

2.1 Data Preparation

We analyzed multimodal and pan-cancer TCGA data that are publicly accessible. We utilized TCGA-Assembler[9] to retrieve and process data from TCGA Data Coordinating Center (DCC, http://cancergenome.nih.gov/abouttcga/overview/howitworks/datasharingmanagement). TCGA-Assembler is a software package to automatically download, assemble, and process public TCGA data. Using TCGA-Assembler, data retrieval was fully automatic and reproducible.

Multimodal TCGA data were retrieved and processed using TCGA-Assembler from TCGA DCC in April 2013, as described below. First, for each cancer type, we downloaded gene expression (GE) data generated by RNA-Sequencing, protein expression (PE) data measured by Reverse Phase Protein Array (RPPA), copy number (CN) data produced by Affymetrix SNP array 6.0, and DNA methylation data measured by Infinium HumanMethylation 450 BeadChip. Only tumor samples measured by all four assay platforms were kept for analysis. Second, a mean DNA copy number value and a mean DNA methylation value were calculated for each gene in each sample using TCGA-Assembler, resulting in gene-level summaries for CN and ME. TCGA GE and PE data were already organized by genes. Third, for RNA-Seq data, zero values were first replaced with the smallest positive value in the data, and then $\log_2$ transformation was taken for all RNA-Seq data; for methylation data, we took a transformation of $\log_2(x/(1-x))$, where x represented an input methylation value. Fourth, each genomic feature was standardized within each cancer type, so that it had a zero mean and a unit standard deviation. Fifth, data of different cancer types were combined together into a mega table, and any genomic feature with missing values in more than 25% of samples over all cancer types was removed. Lastly, we required that a gene must have measurements for at least CN, ME, and GE to be included in our analyses; otherwise the gene was not included.

In TCGA data, while CN, ME, and GE measurements were available genome-wide, measurements of PE by RPPA were available for less than 200 genes, which correspond to important cancer-related proteins. As part of quality control, we compared TCGA gene symbols with the official NCBI gene symbols using the R package *HGNChelper*[10] and corrected obsolete and ambiguous gene symbols. Table 1 in the main text and Table S1 here summarize the 1,448 samples (across 11 cancer types) and the 19,304 genes that were used in the analysis. The number of samples for each cancer type varies depending on how many samples from each cancer type were profiled by TCGA for all of the four genomics platforms. Kidney renal clear cell carcinoma has the largest number of samples, followed by head and neck squamous cell carcinoma and lung adenocarcinoma. Table S1 shows the breakdowns of gene numbers by available features. Out of the 19,304 genes, 19,172 genes have one measurement for each of CN, ME, and GE for the gene, but no PE measurement; 129 genes have PE data, and 28 of them have more than one protein expression values; 3 genes have more than one CN measurements due to their alternative locations in the genome.

Table S1. Number of genes with different combination of available measurements.

Number of CN Readouts	Number of GE Readouts	Number of ME Readouts	Number of PE Readouts	Number of Genes
1	1	1	0	19,172
1	1	1	1	101
1	1	1	>1	28
>1	1	1	0	3
Total				**19,304**

2.2 Computation

We applied the Bayesian graphical models (BGMs)[11-12] to each of the 19,304 genes and each of the 19,304×19,303/2=186,312,556 possible gene pairs to infer intragenic interactions and intergenic interactions. Each gene or gene pair was analyzed as an independent and separate computational job using the BGM, resulting in a total of 19,304+186,312,556=186,331,860 individual computational jobs. On average, each job took about 47 CPU seconds to compute, consisting of 10,000 Markov chain Monte Carlo (MCMC) iterations. The whole study including all analyses took about 2,432,666 CPU hours. We carried out these computational jobs on Beagle, a supercomputer at the Computation Institute of The University of Chicago and the Argonne

National Laboratory[13]. Beagle is a Cray XE6 supercomputer system with 17,424 CPUs, 23TB of memory, and 600TB of hard-disk space. The total computation power of Beagle can achieve 150 teraflops.

3. Additional Results

3.1 Summary of Intragenic Interactions and Graphs

We summarized the intragenic interactions of individual genes. Only genes with non-zero mRNA-seq read counts in more than 50% of the samples were included in summary. A total of 17,157 genes were kept for result summary after filtering. Table S2 summarizes the numbers of genes with different types of significant (FDR ≤ 0.01) intragenic interactions inferred by Zodiac.

Table S2 Numbers of genes with significant intragenic interactions.

Interaction Type	Number of Genes with Such Significant Intragenic Interactions (FDR ≤ 0.01)
CN-GE	7,904
ME-GE	1,277
GE-PE	56
CN-ME	5,019
CN-PE	4
ME-PE	2

Next we considered different types of graphs with only three genomics features, CN, GE, and ME, as an investigation of transcription co-regulation by copy number variation and DNA methylation. There were eight distinct intragenic interaction graphs formed by CN, GE, and ME for a single gene, as shown in the first column of Table S3. We calculated the posterior probability of each of the eight graphs in the MCMC simulation for each gene, and report the mean of these posterior probabilities across all genes (Column 2 of Table S3). The most frequent graph is co-existence of CN-GE and ME-GE interactions with a mean posterior probability of 22.24%, indicating joint regulation of copy number variation and DNA methylation on gene expression. Another type of graph, ME-GE and CN-ME, corresponds to ME-dependent regulation on GE (Fig. 2a-i left and Fig. 2b-i in the main text), which may be caused by a copy-ubiquitous methylation mechanism. This type of graphs consists of 5.85% in Table S3. In contrast, CN-dependent regulation indicated by a graph including CN-GE and CN-ME edges (Fig. 2a-i right and Fig. 2b-ii in the main text) is more prevalent (16.76% in Table S3).

Table S3 Frequencies of different intragenic interaction graphs

Graph Type	Frequency of Graph Over All Genes
No Interaction	5.03%
CN-GE edge only	19.09%
ME-GE edge only	6.81%
CN-ME edge only	4.31%
CN-GE and ME-GE edges	22.24%
ME-GE and CN-ME edges	5.85%
CN-GE and CN-ME edges	16.76%
CN-GE, ME-GE, and CN-ME edges	19.93%

3.2 Enrichment of Intergenic Interactions in KEGG Pathways

Sixteen cancer-related pathways from the KEGG Pathway database[14] were selected for the validation of interactions inferred by Zodiac. These pathways belong to three different categories including *Cancer Overview*, *Signal Transduction*, and *Cell Growth and Death* (Table S4). Genomic interactions in Zodiac were inferred based on integrating data of multiple cancer types. Thus they are expected to characterize conserved, common molecular mechanism between cancer types. KEGG pathways related to specific cancer types were not included for validation. Only genes with non-zero mRNA-seq read counts in more than 50% of the samples were included in validation to ensure a high quality validation. A total of 17,157 genes were kept for result validation after filtering. Enrichment analyses were conducted for two kinds of genomics functions, including transcriptional regulation and protein phosphoregulation.

Table S4. KEGG pathways used for validation of inferred interactions

Cancer Overview	Signal Transduction	Cell Growth and Death
Pathways in cancer	MAPK signaling pathway	Cell cycle
Transcriptional misregulation in cancer	PI3K-Akt signaling pathway	Apoptosis
Proteoglycans in cancer	Notch signaling pathway	p53 signaling pathway
	mTOR signaling pathway	
	Wnt signaling pathway	
	TGF-beta signaling pathway	
	ErbB signaling pathway	
	VEGF signaling pathway	
	Jak-STAT signaling pathway	
	NF-kappa B signaling pathway	

(1) **Evidence of transcriptional regulation**. In Zodiac, we considered significant (FDR ≤ 0.01) intergenic PE-GE and GE-GE interactions as evidence of potential transcriptional regulation. The first gene involved in the PE-GE or GE-GE interaction could be transcriptional factor and the second gene could be its target gene. There were 17,157×17,156/2 = 147,172,746 gene pairs with potential PE-GE or GE-GE interactions. Among them, 540 gene pairs had transcriptional regulation relations recorded in the selected KEGG pathways. Zodiac inferred significant PE-GE or GE-GE interactions between 13,449,210 gene pairs, 114 of which were among the 540 gene pairs and have inferred interactions consistent with the transcriptional activations or repressions indicated by KEGG. Calculated using the Fisher's exact test based on hypergeometric distribution, the enrichment is statistically significant (*P*-value 3.24e-17) with an enrichment fold of (114/13,449,210)/(540/147,172,746) = 2.31.

(2) **Evidence of protein phosphoregulation**. In Zodiac, we considered significant (FDR ≤ 0.01) intergenic PE-PE(phos) and GE-PE(phos) interactions as evidence indicating protein phosphoregulation, in which the first gene promotes or reduces the phosphorylation of the protein encoded by the second gene. Only 37 genes in Zodiac had expression values of phosphorylated proteins and totally there were 17,156×37 = 634,772 possible gene pairs that could possess PE-PE(phos) or GE-PE(phos) interactions. Among these, 234 gene pairs were

indicated by the KEGG pathways to have protein phosphoregulation relations. Zodiac inferred 6,242 gene pairs with statistically significant PE-PE(phos) or GE-PE(phos) interactions, among which 16 gene pairs had interactions consistent with the protein phosphorylation or de-phosphorylation relations indicated by KEGG pathways. The enrichment is statistically significant (*P*-value 2.29e-9) with an enrichment fold of (16/6,242)/(234/634,772) = 6.95.

3.3 Enrichment of Intergenic Interactions in TRED

We also assessed intergenic interactions in Zodiac using transcriptional regulations provided by the Transcriptional Regulatory Element Database (TRED)[15]. Significant intergenic PE-GE and GE-GE interactions (FDR ≤ 0.01) were considered evidence supporting potential transcription factor regulations on genes, where the first gene with PE or GE readout can be a transcription factor and the second gene with GE readout is its target gene. A total of 45 cancer-related transcription factors and their target genes indicated by TRED were involved in the enrichment analysis. Again, we only included 17,157 genes whose mRNA-seq read counts were non-zero in more than 50% of the samples. We found significant enrichment between Zodiac and TRED on 11 transcription factors (TFs) and their targeted genes, with the Fisher's exact test evaluating enrichment significance and a *P*-value cutoff of 0.01. See Table S5.

Table S5 Transcription factors with transcriptional regulations (recorded by TRED) significantly enriched in interactions inferred by Zodiac

Transcription Factor (TF)	Number of genes that are TF targets (by TRED) and have significant interactions with TF (by Zodiac)	Number of genes that have significant interactions with TF (by Zodiac)	Number of target genes of TF (by TRED)	Enrichment p-Value	Enrichment Fold
SPI1[a]	37	3693	47	9.13E-17	3.66
ETS1	36	3355	75	2.54E-08	2.45
JUN	13	699	109	4.87E-04	2.93
ETS2	7	865	30	5.93E-04	4.63

CEBPB	6	497	44	1.59E-03	4.71
EGR1	10	1135	58	4.39E-03	2.61
POU2F2	7	2470	16	4.40E-03	3.04
CEBPD	4	821	15	4.65E-03	5.57
MYC	69	1303	665	4.91E-03	1.37
ERG	7	2811	15	5.97E-03	2.85
STAT1	5	1057	22	9.61E-03	3.69

[a] We take *SPI1* as an example to show how the enrichment significance and fold are calculated. There are totally 17,156 genes with potential significant PE-GE or GE-GE interactions inferred by Zodiac between *SPI1* and the gene. 47 of these genes are transcriptionally regulated by *SPI1* as indicated by TRED. Zodiac inferred significant PE-GE or GE-GE interaction between *SPI1* and 3693 genes, which includes 37 out of the 47 genes with TRED recorded transcriptional regulations by *SPI1*. Using the Fisher's exact test based on hypergeometric distribution, the p-value of the enrichment is 9.13E-17. And the enrichment fold is (37/3693)/(47/17,156) = 3.66.

3.4 Genes Interacting With EZH2 and E2F1

We used Zodiac to identify the genes that had significant GE-GE interactions with *EZH2*. There were 644 genes (FDR ≤ 0.01) and Table S6 summarizes the top 40 genes positively interacting with *EZH2* sorted by the posterior mean of strength coefficient β (see SI Bayesian Graphical Models).

Fig. S3 uses a Circos plot[16] to show the significant (FDR ≤ 0.01) intergenic interactions between *E2F1*, *CCNE1*, *CCNA2*, *CDC6*, *DHFR*, and *TK1*. It can be seen that *E2F1* is connected to all other genes through significant GE-GE edges, which supports the notion that the other genes are potential downstream targets of *E2F1*[17].

Table S6 Top 40 genes with significant strong positive GE-GE interactions with *EZH2*

Gene Symbol	Posterior Mean of β	Gene Symbol	Posterior Mean of β
HIST1H2BH	15.92175	*CDC6*	9.62499
LEFTY1	14.64688	*KIF14*	9.60699
KDM3A	11.16383	*MCM10*	9.56511
MTIF2	10.92703	*DLGAP5*	9.50962
NEURL	10.71264	*NCAPG2*	9.48744
SKP2	10.49488	*KIF4A*	9.46784
TBC1D31	10.49417	*NUSAP1*	9.46527
SCLT1	10.45962	*RAD51*	9.45719
NEK2	10.31364	*POLQ*	9.444
FASTKD1	10.20907	*WDR62*	9.39296
CDT1	10.17701	*CCNA2*	9.39158
HELLS	10.15806	*SPC25*	9.36061
KIF15	10.14204	*FOXM1*	9.3148
XRCC2	9.93363	*SASS6*	9.29762
CENPE	9.78229	*VASH2*	9.29377
MCM8	9.76895	*KIFC1*	9.2771
CENPA	9.72743	*CCNF*	9.2743
HMMR	9.71635	*CDCA5*	9.26555
BUB1B	9.6758	*BLM*	9.26354
SKA3	9.64015	*KIF2C*	9.24806

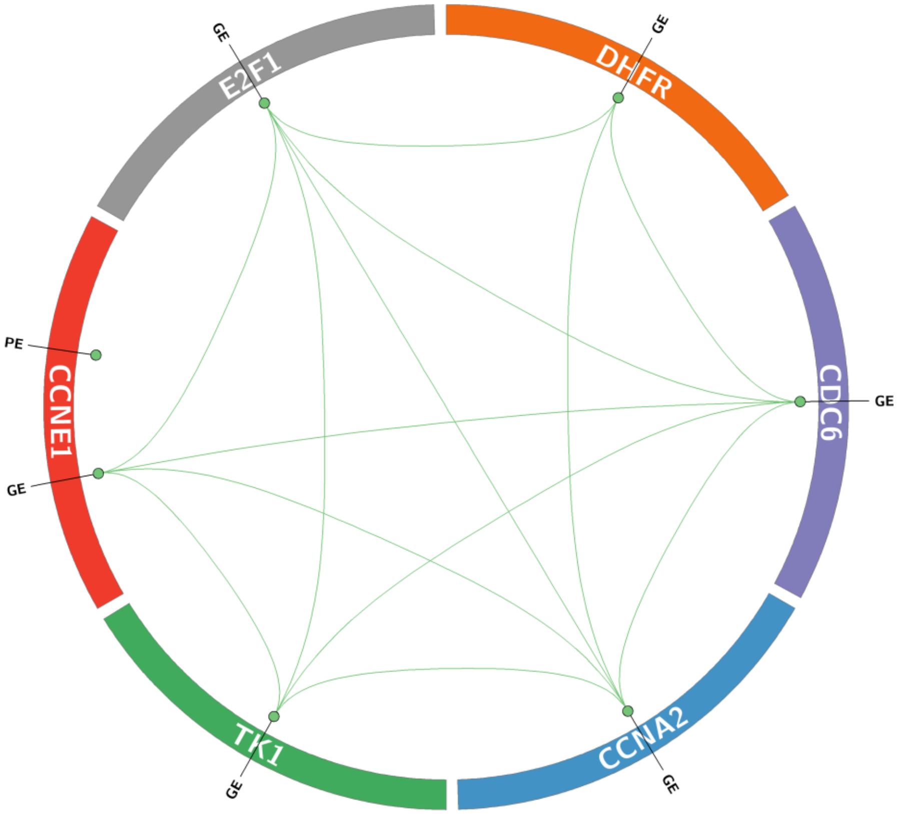


Figure S3 Significant (FDR ≤ 0.01) intergenic interactions between *E2F1*, *CCNE1*, *CCNA2*, *CDC6*, *DHFR*, and *TK1*. Green lines indicate positive interactions and red lines indicate negative interactions.

3.5 Example of Data-enhanced Network Inference

We selected a signaling cascade from the KEGG prostate cancer pathway as an example to demonstrate using BGM software and TCGA data for examining existing knowledge about genomic interactions and producing data-enhanced network inference. The selected signaling cascade includes 6 steps as indicated by Fig. 3a in the main text.

Step 1: SOS converts Ras into its active conformation.
Step 2: Ras activates Raf.
Step 3: Phosphorylated Raf activates MEK.

Step 4: MEK phosphorylates and activates ERK.

Step 5: ERK indirectly interacts with Androgen Receptor (AR).

Step 6: AR acts as a transcription factor and activates the transcription of *KLK3*, whose protein product is PSA.

We used TCGA-Assembler[9] to retrieve and preprocess TCGA prostate cancer data and obtained 162 samples measured for both gene expressions and protein expressions. For the simplicity of analysis, one measurement feature was selected to represent each of the nodes involved in the signaling cascade (see Fig. 3a in the main text). For SOS, we used the GE of *SOS1*, as TCGA does not provide PE of *SOS1*. For Ras, we used PE of *NRAS*. For Raf, we used PE of *RAF1* with Ser338 phosphorylation. For MEK, we used PE of *MAP2K1* with Ser217 and Ser221 phosphorylations. For ERK, PE of *MAPK1* with Thr202 and Tyr204 phosphorylations was used. For androgen receptor and PSA, we used PE of *AR* and GE of *KLK3*, respectively. Thus, the data involved in analysis include 162 samples and 7 features. The prior network used in analysis was a single thread cascade of *SOS1-NRAS-RAF1-MAP2K1-MAPK1-AR-KLK3*. And the parameter ρ controlling the strength of prior network was set at 0.1.

Table S7 Posterior probabilities of all potential interactions between the selected features

Measurement Feature	***NRAS* (PE)**	***RAF1* (PE, Ser338)**	***MAP2K1* (PE, Ser217 and Ser221)**	***MAPK1* (PE, Thr202 and Tyr204)**	***AR* (PE)**	***KLK3* (GE)**
***SOS1* (GE)**	**0.94**	0.09	0.05	0.05	0.07	0.06
***NRAS* (PE)**		**1.00**	0.04	0.04	0.07	0.05
***RAF1* (PE, Ser338)**			**0.80**	0.05	0.07	0.06
***MAP2K1* (PE, Ser217 and Ser221)**				**1.00**	0.08	0.05
***MAPK1* (PE, Thr202 and Tyr204)**					**0.89**	0.04
***AR* (PE)**						**0.95**

Bold font indicates relatively high posterior probabilities with posterior FDR ≤ 0.1 if all the edges are selected.

Table S7 gives the posterior probabilities inferred by BGM for all potential edges between the selected measurement features. Clearly, only edges representing interactions in the prostate cancer pathway signaling cascade (Fig 3.a in the main text) have high posterior probabilities (≥ 0.75) compared to all other potential edges, which are indicated by bold font in Table S7. This means the posterior inference network is consistent with existing knowledge about this signaling transduction mechanism. Also, all interactions in this posterior inference network are positive, which is consistent with existing understanding about the pathway, except the interaction between *MAPK1* (PE, Thr202 and Tyr204) and *AR* (PE), which according to KEGG is an indirect effect.

4 Web Server and Interface

Zodiac allows investigators to query and view the evidence for inferred interactions through web browser. Currently, four query procedures are available.

(1) The first procedure focuses on interactions of other genes to a single gene of interest, a one-versus-rest query. In the search box of Zodiac users input one gene symbol and Zodiac returns all significant intergenic interactions with this gene (Fig. 4b in the main text). Such a query procedure is particularly useful for identifying important genes that interact with the input gene. Details about specific interaction types and lists of the strongest interactions can be viewed by clicking on hyperlinks within the result table.

(2) The second query procedure is a query of significant intragenic interactions for a single gene. This procedure is initiated by entering the same gene symbol twice and returns the intragenic interactions of that gene (Fig. 4c in the main text).

(3) The third query procedure deals a pair of genes. Users enter two different gene symbols and Zodiac displays the interaction network between all features of the two genes (Fig. 4d in the main text).

(4) The fourth procedure allows users to enter multiple gene symbols and returns a Circos plot[16] showing all significant intergenic interactions between input genes, which are a collection of all significant intergenic interactions obtained by pair-wise analyses.

In the first three query procedures, the input gene symbols are validated using a gene list including gene IDs, symbols, and aliases obtained from NCBI[18]. If there is an error or ambiguity in the input gene symbol(s), the user is prompted to choose their desired gene from a list of relevant genes. Alternative gene symbols and descriptions of genes are provided to help in the selection. Gene symbol is shown with a hyperlink to its NCBI webpage that provides a detailed description of the gene. The result webpage also provides a link to download the graphs and inference statistics from the current query for additional analysis off-line. In the multi-gene query, unrecognized gene symbols that are not uniquely identifiable are ignored for drawing the graph and are listed below it. See details in the online Zodiac Tutorial (http://www.compgenome.org/TCGA/tutorial.html).

We also provide Application Programming Interfaces (APIs) that allow users to send a URL request to the Zodiac server for direct visualization of network inference results without performing the step of entering gene symbols or IDs. Using the APIs, developers can hyperlink from outside programs to directly access Zodiac interaction networks. For the first query procedure (one gene versus the rest), the URL request reads as

http://www.compgenome.org/ZODIAC?Gene_List=*GeneID*

where *GeneID* in the URL should be replaced by the NCBI gene symbol or ID of the gene in query. For all other query procedures (e.g. gene pair and multi-gene), the URL request is

http://www.compgenome.org/ZODIAC?Gene_List=*GeneID+GeneID…*

where … represents that more gene symbols or IDs with '+' for delimitation can be added to the end of the URL.

References

1. Mitra, R., Müller, P., Liang, S., Yue, L. & Ji, Y. A bayesian graphical model for chip-seq data on histone modifications. *Journal of the American Statistical Association* **108**, 69–80 (2013).

2. Havard, R. & Leonard, R. Gaussian Markov random fields: theory and applications (CRC Press, Boca Raton, FL USA, 2005).

3. Parmigiani, G., Garrett, E., Anbazhagan, R. & Gabrielson, E. A statistical framework for expression-based molecular classification in cancer. *Journal of the Royal Statistical Society: Series B* **64**, 717–736 (2002).

4. Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B* 192–236 (1974).

5. Caragea, P. & Kaiser, M. Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics* **14**, 281–300 (2009).

6. Hughes, J., Haran, M. & Caragea, P.C. Autologistic models for binary data on a lattice. *Environmetrics* **22**, 857–871 (2011).

7. Newton, M.A., Noueiry, A., Sarkar, D. & Ahlquist, P. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176 (2004).

8. Müller, P., Parmigiani, G., Robert, C. & Rousseau, J. Optimal sample size for multiple testing. *Journal of the American Statistical Association* **99**, 990–1001 (2004).

9. Zhu, Y., Qiu, P. & Ji, Y. TCGA-Assembler: open-source software for retrieving and processing TCGA data. *Nature Methods* **11**(6), 599-600 (2014).

10. R package HGNChelper, http://cran.r-project.org/package=HGNChelper.

11. Mitra, R., Müller, P., Liang, S., Yue, L. & Ji, Y. A Bayesian graphical model for ChIP-Seq data on histone modifications. *Journal of the American Statistical Association* **108**, 69–80 (2013).

12. Xu, Y. et al. in IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS) 135–138 (Washington, DC, USA; 2012).

13. Supercomputer Beagle, http://beagle.ci.uchicago.edu.

14. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40**, D109-114 (2012).

15. Jiang, C., Xuan, Z., Zhao, F. & Zhang, M. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Research* **35**, 40 (2007).

16. Krzywinski, M. et al. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**, 1639-1645 (2009).

17. Bracken, A. et al. EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *The EMBO Journal* **22**, 5323-5335 (2003).

18. ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz, accessed 8/13/2013.