

ARTICLE

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

Yitan Zhu*, Yanxun Xu*, Donald L. Helseth Jr*, Kamalakar Gulukota, Shengjie Yang, Lorenzo L. Pesce, Riten Mitra, Peter Müller, Subhajt Sengupta, Wentian Guo, Jonathan C. Silverstein, Ian Foster, Nigel Parsad, Kevin P. White, Yuan Ji

Affiliations of authors: Program of Computational Genomics & Medicine (YZ, SY, SS, YJ), Center for Molecular Medicine (DLH Jr, KG), and Center for Biomedical Research Informatics (JCS, NP), NorthShore University HealthSystem, Evanston, IL; Department of Mathematics, The University of Texas at Austin, Austin, TX (YX, PM); Computation Institute (LLP, IF) and Institute for Genomics and Systems Biology (KPW), The University of Chicago and Argonne National Laboratory, Chicago IL; Department of Bioinformatics & Biostatistics, University of Louisville, Louisville, KY (RM); School of Public Health, Fudan University, Shanghai, P. R. China (WG); Department of Human Genetics and Department of Ecology & Evolution (KPW) and Department of Public Health Sciences (YJ), The University of Chicago, Chicago, IL.

*Authors contributed equally to this work.

Correspondence to: Yuan Ji, PhD, Program of Computational Genomics & Medicine, NorthShore University HealthSystem, 1001 University Pl, Evanston, IL 60201 (e-mail: koeraser@gmail.com).

Abstract

Background: Genetic interactions play a critical role in cancer development. Existing knowledge about cancer genetic interactions is incomplete, especially lacking evidences derived from large-scale cancer genomics data. The Cancer Genome Atlas (TCGA) produces multimodal measurements across genomics and features of thousands of tumors, which provide an unprecedented opportunity to investigate the interplays of genes in cancer.

Methods: We introduce Zodiac, a computational tool and resource to integrate existing knowledge about cancer genetic interactions with new information contained in TCGA data. It is an evolution of existing knowledge by treating it as a prior graph, integrating it with a likelihood model derived by Bayesian graphical model based on TCGA data, and producing a posterior graph as updated and data-enhanced knowledge. In short, Zodiac realizes “Prior interaction map + TCGA data → Posterior interaction map.”

Results: Zodiac provides molecular interactions for about 200 million pairs of genes. All the results are generated from a big-data analysis and organized into a comprehensive database allowing customized search. In addition, Zodiac provides data processing and analysis tools that allow users to customize the prior networks and update the genetic pathways of their interest. Zodiac is publicly available at www.compgenome.org/ZODIAC.

Conclusions: Zodiac recapitulates and extends existing knowledge of molecular interactions in cancer. It can be used to explore novel gene-gene interactions, transcriptional regulation, and other types of molecular interplays in cancer.

Complex diseases such as cancer are rarely caused by an abnormality in a single gene, but rather reflect multigene abnormalities that perturb molecular interaction networks and attract cells to new malignant and carcinogenic states (1,2). For example, dysregulation of molecular signaling pathways, such as the PI3K/AKT/mTOR signaling pathway and the MAPK signaling

pathway, leads to excessive cell proliferation and evasion of apoptosis, forming the basis of carcinogenesis. Therefore, understanding genetic interactions inside cells is critical to elucidating the molecular mechanisms of cancer (3,4).

The Cancer Genome Atlas (TCGA) (5–8) has generated multimodal genomic measurements, including DNA copy number

Received: August 4, 2014; Revised: January 13, 2015; Accepted: April 10, 2015

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com.

(CN), DNA methylation (ME), gene expression (GE), and protein expression (PE), on a large collection of human tumors. Functional interactions between these features of the same gene (intragenic) or of different genes (intergenic) regulate important cellular and molecular events like replication, transcription, and cellular signaling. Therefore, TCGA data provide a unique opportunity for understanding cancer at multiple regulatory and molecular layers. Statistical models that integrate information from different modalities are powerful tools for studying a biological system. Currently there are few resources based on integrative analysis of multiple genomics features using TCGA data. Most available tools focus on reporting existing data and knowledge (9,10) instead of analysis results about cancer genetic interactions. Integrating multilayer and pan-cancer TCGA data using Bayesian graphical models, we aim to generate a comprehensive database and provide powerful computational tools for the investigation of molecular interactions in cancer.

Methods

Zodiac is a result of a big-data integrative analysis on multimodal TCGA data with a goal to generate new knowledge of molecular interactions in cancer. Unlike correlation-based methods such as weighted correlation network analysis (WGCNA) (11), Zodiac is powered by Bayesian graphical models (12,13) that simultaneously account for conditional independence among multiple genomic features (Supplementary Materials, section “Bayesian Graphical Models,” available online). This gives Zodiac some advantage in capturing the network-type relationship in biological systems (Supplementary Figures 1–3, Supplementary Table 1, Supplementary Materials, section “Bayesian Graphical Models,” available online). Zodiac provides a whole-genome and pair-wise interaction map, which contains intragenic and intergenic interactions of all pairs of genes in cancer. Innovatively, Zodiac integrates multiple layers of molecular features such as CN, ME, and GE, which is markedly different from other systems such as STRING (14) that focuses on a single feature at a time for multiple genes. All the results in Zodiac are publicly accessible at <http://www.compgenome.org/ZODIAC>.

The construction of Zodiac is divided into four sequential steps (Figure 1), including data retrieval, computation, results assembly, and results dissemination. Step 1 is data retrieval (Figure 1A) that collected and assembled TCGA data into a well-formatted data matrix via our own open-source software pipeline, called TCGA-Assembler (15). This step allows for automatic data retrieval directly from the TCGA data portal, ensuring the reproducibility of subsequent statistical analysis. Only samples with four types of molecular features (CN, GE, ME, PE) were used, and data from different cancer types were combined together as a pan-cancer analysis (Supplementary Materials, section “Analysis Details,” available online). In the end, we obtained a data set of 1448 samples from 11 cancers (Table 1), each sample characterized by up to four types of features on each of 19 304 genes (Supplementary Table 2, Supplementary Materials, section “Analysis Details,” available online). In the second step (Figure 1B) we launched massively parallel computations and applied the Bayesian graphical models (BGMs) (12,13) to the multimodal data for each gene and each gene pair using all the 1448 samples. The computation consisted of about 200 parallel jobs and was executed on the super computer Beagle (16). The entire computation took roughly 2.5 million CPU hours. Results from the analyses were stored and organized into a database (Figure 1C) in the third step. In the final step, a web interface was constructed at <http://www.compgenome.org/ZODIAC> to

enable fast queries and visualizations of the analysis results (Figure 1D).

Statistical Analysis

Treating a known genetic map as a prior graph, Zodiac combines the prior with a likelihood function that describes TCGA data and produces posterior and data-enhanced interaction maps. The key innovation is the use of random graphs that allows for changes of the connectivities in graph. Some edges present in the prior graph can be eliminated in the posterior graph, and new edges absent in the prior may be added, all depending on the information contained in the data. In summary, Zodiac realizes the evolution of:

Prior Interaction Map + TCGA data → Posterior Interaction Map in Cancer

Statistical inference of genetic interactions is based on a class of Bayesian graphical models, treating the interaction networks of genomics features as random graphs. In this case, the network topology of features is considered random, and a rigorous Bayesian posterior inference (Supplementary Materials, section “Bayesian Graphical Models,” available online) is performed to estimate the network topology as data-enhanced and new knowledge supported by TCGA data. This is different from tools that assume known and fixed network topology (17,18). More importantly, graphical models are known to be powerful at describing interactions of multiple variables (19), or genomics features in our case, which is otherwise challenging for traditional statistical models based on correlations or regressions. In addition, Bayesian statistical inference enforces stringent false discovery rate (FDR) thresholds in reported results based on posterior probabilities, which is different from P value-based inference (20).

For validating the genetic interactions inferred by Zodiac, one-sided Fisher’s exact test has been used to evaluate the enrichment of known genetic interactions in Zodiac. The one-sided test is used because scientifically we are only interested in the one-sided alternative. That is, interactions in Zodiac are more enriched with known genetic interactions than by random chance. Such one-sided enrichment tests are frequently used in the literature (21). A cutoff of .01 on P value is used to call statistically significant enrichment.

Results

Overview

The main function of Zodiac is to generate new hypotheses and validate existing discoveries in cancer based on posterior and data-enhanced inference on molecular interactions. For example, depending on user query, Zodiac returns: 1) an intragenic interaction network (Figure 2A-i) for each single gene, 2) an intra- and intergenic interaction network (Figure 2A-ii) for each pair of genes, and 3) a list of genes that possess intergenic interactions with any single gene of interest (Figure 2A-iii). Each of the three types of queries provided by Zodiac returns important information. For example, Figure 2A-iii shows the list of top genes that could be potentially targeted by the regulator gene EZH2. Similar lists can be returned for other genes of interests, such as transcription factors, tumor suppressors, or oncogenes through the single gene query. Alternatively, a search of a gene pair in Zodiac returns a graph containing statistically significant interactions between genomic features of the two genes

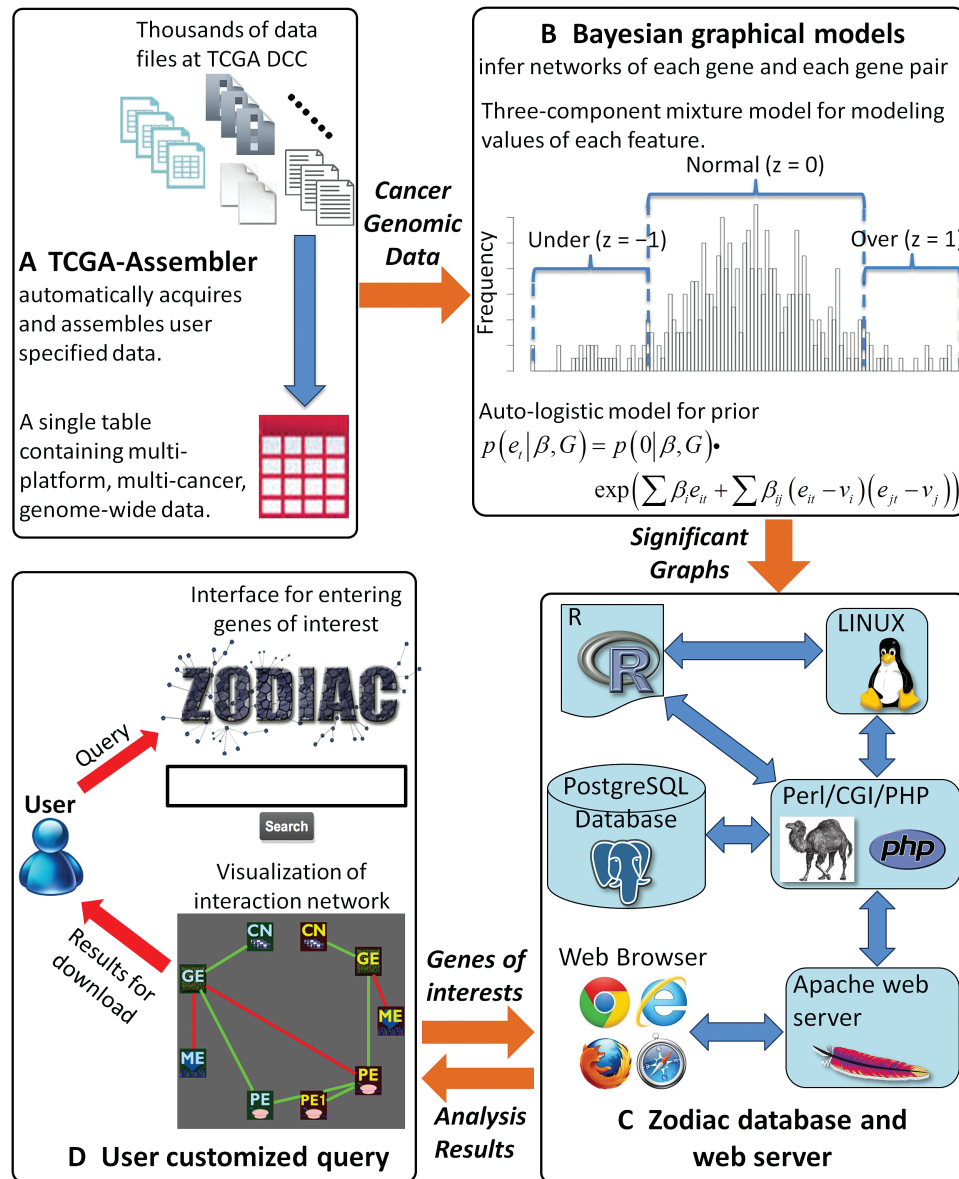


Figure 1. Four steps to build Zodiac. **A)** Multimodal, multicancer The Cancer Genome Atlas (TCGA) data are prepared by TCGA-Assembler (15), open-source software designed to retrieve and process TCGA public data from TCGA Data Coordinating Centers (DCC). Each patient sample has multiple data files generated by different array or sequencing platforms. TCGA-Assembler automatically retrieves thousands of data files from TCGA DCC and assembles them into a single mega data table for subsequent analysis. **B)** Massively parallel computational jobs, one for each gene or gene pair, are conducted using Bayesian graphical models (12–13). Statistical inferences for a total of 19,304 genes and all possible gene pairs are performed based on posterior probabilities. A functional network involving multiple modalities of a gene or gene pair is treated as a Markov random field and Markov chain Monte Carlo simulations are used for statistical estimation (Supplementary Materials, section “Bayesian Graphical Models,” available online). **C)** All statistically significant intragenic and intergenic networks from the Bayesian inference are stored in database and accessible through a web server. **D)** A user-friendly web interface is built to allow fast queries and visualization of analysis results. CN = copy number; GE = gene expression; Me = methylation; PE = protein expression; TCGA = The Cancer Genome Atlas.

(Figure 2A-ii). In all cases, statistical significance is assessed by posterior probabilities for any edges and any graphs, and stringent false discovery rate (FDR) thresholds are applied to select a list of reliable edges or graphs. The prior network used when inferring the intra- and intergenic networks applies the common biological consensus, assuming that within each gene copy number and methylation are connected with gene expression and gene expression is connected with protein expression. These edges in the prior network can be reinforced, weakened, or removed in the posterior network, depending on the information contained in the observed data. This is an essential feature of the proposed Bayesian statistical inference.

The Bayesian graphical inference is distinctly different from correlation-based inference and provides insight on concurrent interactions between multiple genetic features. For example, the two intragenic graphs in Figure 2A-i represent the conditional independence between CN and GE (left) or between ME and GE (right). The two different graphical topologies provide two different transcription regulatory patterns. The left pattern implies that change in gene expression is independent of copy number variations given methylation events, and the right pattern implies the reverse relationship that change in gene expression is independent of methylation given copy number variations. These patterns may be associated with two hypothetical and

different coregulation mechanisms of ME and CN on GE. For example, copy number amplification for a gene coupled with copy-ubiquitous methylation (Figure 2B-i) could explain the ME-dependent and CN-independent regulation on GE, while

Table 1. Sample counts from 11 cancer types used in Zodiac analysis

| Cancer types | Number of samples |
|--|-------------------|
| Bladder urothelial carcinoma [BLCA] | 50 |
| Breast invasive carcinoma [BRCA] | 151 |
| Colon adenocarcinoma [COAD] | 20 |
| Glioblastoma multiforme [GBM] | 26 |
| Head and neck squamous cell carcinoma [HNSC] | 202 |
| Kidney renal clear cell carcinoma [KIRC] | 252 |
| Lung adenocarcinoma [LUAD] | 184 |
| Lung squamous cell carcinoma [LUSC] | 103 |
| Skin cutaneous melanoma [SKCM] | 162 |
| Thyroid carcinoma [THCA] | 204 |
| Uterine corpus endometrioid carcinoma [UCEC] | 94 |
| Total | 1448 |

a copy number amplification with copy-specific methylation (Figure 2B-ii) may result in CN-dependent and ME-independent regulation on GE. We discovered many genes possessing these two patterns in Zodiac results (Supplementary Tables 3-4, Supplementary Materials, section “Additional Results,” available online). Hypotheses of this nature can only be generated under graphical models and the corresponding conditional dependence/independence between all three genomics features. In contrast, a correlation-based inference of any two features (ME/GE, or CN/GE) might yield misleading conclusions. For example, when the true coregulation pattern agrees with Figure 2B-i, measurements between copy numbers and gene expression could be negatively correlated because copy number gains actually do not lead to increased gene expression as the transcription of each gene copy is silenced by DNA methylation.

Overlap With Existing Databases of Genomic Regulations

The multimodal integration of genomics features in Zodiac can be used to discover and validate regulatory relationships

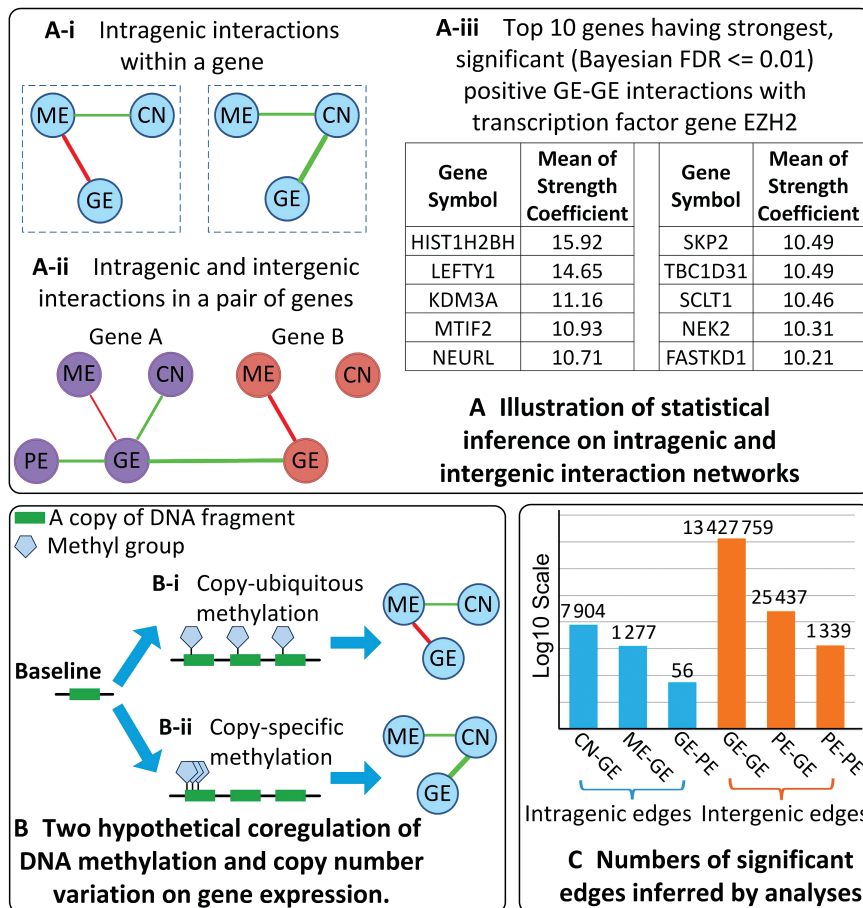


Figure 2. Illustrations and summary of inference results on interaction networks. **A)** Illustrations of potential interactions. **A-i)** Illustrations of intragenic interactions between genomic features of a single gene. Two potential intragenic networks are presented, implying different coregulation of copy number (CN) and methylation (ME) on gene expression (GE). A green or red edge, ie, line connecting two genomics features, indicates a positive or negative interaction between them, respectively. The thickness of an edge indicates the interaction strength (Supplementary Materials, section “Bayesian Graphical Models,” available online). **A-ii)** illustrates a hypothetical intergenic network for two genes. One gene does not have any PE measurement, because in The Cancer Genome Atlas not all genes have protein expression (PE) data. Table (A-iii) gives an illustration of the gene list returned by searching for genes interacting with an input gene, EZH2 in this case (more details can be found in the Results section). **B)** Two hypothetical models that might explain the two different intragenic networks in panel (A-i), which are copy number amplifications coupled with (B-i) copy-ubiquitous methylation or (B-ii) copy-specific methylation. The two models require experimental validation. **C)** Numbers of statistically significant interactions (Bayesian false discovery rate ≤ 0.01) inferred by the analyses and grouped by interaction types. CN = copy number; GE = gene expression; Me = methylation; PE = protein expression.

between genes. To examine the biological relevance of the computational results in Zodiac, we chose 16 cancer related pathways from KEGG (22), categorized as Cancer Overview, Signal Transduction, and Cell Growth and Death (Supplementary Table 5, Supplementary Materials, section “Additional Results,” available online) and assessed the overlap between KEGG and Zodiac on these pathways (Supplementary Tables 6–7, Supplementary Materials, section “Additional Results,” available online). Since Zodiac reports results based on the integration of TCGA data across multiple cancer types, we expect an abundant overlap between Zodiac results and the type of findings in KEGG that are conserved and common across cancers. Therefore, cancer-specific KEGG pathways were not included for this assessment. KEGG records various types of molecular interactions, such as protein-protein interaction and transcriptional regulation. Similarly, Zodiac investigates intergenic GE-GE, GE-PE, and PE-PE interactions. Table 2 shows that the genomic regulations indicated by KEGG are statistically significantly enriched ($P \leq 1 \times 10^{-8}$) in the corresponding interactions inferred by Zodiac.

We also assessed genomic interactions in Zodiac using transcriptional regulations provided by the Transcriptional Regulatory Element Database (TRED) (23), which includes target genes of cancer-related transcription factor families. Statistically significant intergenic PE-GE or GE-GE edges in Zodiac are considered as evidence supporting transcription factor regulation on its target genes. A total of 11 (out of 45) transcription factors (with their target genes) were found statistically significantly enriched ($P \leq .01$) in the PE-GE and GE-GE edges inferred by Zodiac (Supplementary Table 8, Supplementary Materials, section “Additional Results,” available online).

Data-Enhanced Network Inference Using Zodiac Tools

Zodiac provides a unique functionality that allows researchers to perform data-enhanced Bayesian network inference. Existing results in Zodiac, although useful, might not be suitable for all users. For example, a user might be interested in a few genes and their interaction network in prostate cancer. The reported network in Zodiac does not match such an interest. Instead, using the TCGA-Assembler (15) and BGM tools provided at the Zodiac website, the user can obtain posterior network for those genes enhanced by the prostate cancer data in TCGA. TCGA-Assembler allows for automatic and user-specified TCGA data retrieval and processing, while BGM performs subsequent network inference based on the retrieved data. The knowledge enhancement is in the fact that the posterior estimated network combines the prior network provided by the user (if available) and the information contained in TCGA data. Such an enhancement is

realized by rigorous statistical inference through BGM analysis. Figure 3 provides an example based on our own investigation. A signaling cascade in the prostate cancer pathway of the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Figure 3A) was used as the prior information in a BGM analysis, through which a posterior network was obtained (Figure 3B; Supplementary Tables 9–10, Supplementary Materials, section “Additional Results,” available online). Therefore, using the provided tools, Zodiac allows for highly customizable computation aiming for data enhanced network inference. The comprehensive TCGA data and flexibility of the BGM tools will facilitate the community to generate posterior networks as data-enhanced and renewed knowledge about cancer. More broadly, the BGM tools can be applied to general data, such as users’ in-house data, in which any genes or genomic features are measured across matched biological samples.

Some Use Cases

EZH2 is frequently found overexpressed in a variety of cancers (24,25). It is a Polycomb protein that adds methyl groups to Lysine 27 of histone protein H3, leading to chromatin condensation and thus transcriptional repression of genes. Histone modification induced by EZH2 promotes cancer development by silencing the expression of tumor suppressor genes (26–28). A previous study shows that EZH2 is a downstream target of E2F1, a transcription factor in control of cell cycle and also interacting with tumor suppressor proteins (29). This indicates a possible connection between EZH2-mediated histone modification and E2F1-mediated cellular functions in cancer. Zodiac shows a statistically significant (Bayesian FDR ≤ 0.01) GE-GE edge connecting these two genes, confirming these previous findings (29,30).

Responding to a single-gene search, Zodiac reports all the other genes with statistically significant intergenic interactions with the given gene. Using this function for EZH2 (Figure 2A-iii; Supplementary Table 11, Supplementary Materials, section “Additional Results,” available online), we notice that the top gene on the list is HIST1H2BH, which encodes a member of the histone H2B family. Interestingly, investigation on the link between EZH2 and histone genes has been centered on H3K27 with numerous publications, but a recent paper (31) for the first time demonstrates the methylation by EZH2 on Lysine 120 of H2B—ie, EZH2 is related to H2BK120 methylation.

As a follow-up investigation, we examined the E2F family. Bracken et al. (2003) reported several potential downstream targets of E2F1 (29), including genes essential for the entry into S phase of cell cycle, such as CCNE1 and CCNA2, and genes that are involved in the regulation of DNA replication, such as CDC6,

Table 2. Overlaps between KEGG and Zodiac indicated by enrichments of two kinds of interactions recorded by KEGG pathways amongst the corresponding statistically significant (Bayesian false discovery rate ≤ 0.01) intergenic interactions inferred by Zodiac*

| KEGG relationship (Corresponding zodiac relationship) | Enrichment fold | Enrichment P |
|--|-----------------|--------------|
| Transcriptional Regulation † (PE-GE or GE-GE) | 2.31 | 3.24E-17 |
| Protein Phosphoregulation ‡ (PE-PE(phos) or GE-PE(phos)) | 6.95 | 2.29E-09 |

* One-sided Fisher’s exact test and P value are used to measure the statistical confidence of enrichment. Enrichment fold evaluates the improvement on the abundance of Kyoto Encyclopedia of Genes and Genomes–indicated interactions by selecting corresponding statistically significant interactions in Zodiac. See the “Additional Results” in the Supplementary Materials (available online) for details of enrichment analysis. GE = gene expression; KEGG = Kyoto Encyclopedia of Genes and Genomes; PE = protein expression; PE(phos) = expression of phosphorylated protein.

† Transcriptional regulation in KEGG (involving a transcription factor and a target gene) can be characterized by a statistically significant intergenic PE-GE or GE-GE edge in Zodiac between the transcription factor and its target gene.

‡ Protein phosphoregulation between two genes in KEGG can be indicated in Zodiac by a statistically significant PE-PE(phos) or GE-PE(phos) edge between the kinase gene and its substrate.

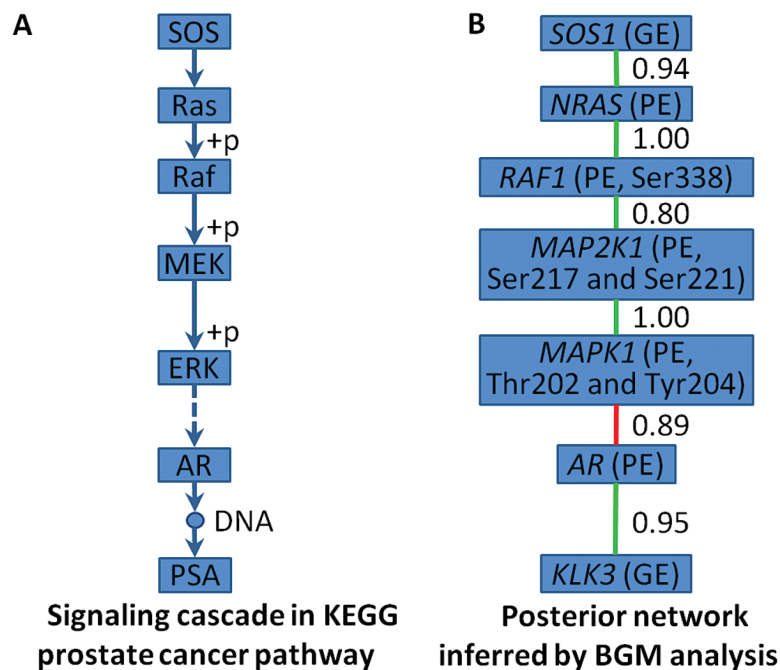


Figure 3. Illustration of data-enhanced Bayesian network inference. **A)** A signaling cascade in the Kyoto Encyclopedia of Genes and Genomes (KEGG) prostate cancer pathway. Nodes in the signaling cascade indicate proteins or protein families. **Arrow with a solid line** indicates activation. **+p** indicates phosphorylation. **Arrow with a dashed line** indicates indirect effect. The androgen receptor (AR) regulation of prostate specific antigen (PSA) is through binding to DNA and transcriptional activation of *KLK3*, which encodes PSA. **B)** Data-enhanced Posterior network inferred by Bayesian graphical model (BGM) analysis. One measurement feature is selected to represent each of the proteins or protein families in the signaling cascade (Supplementary Materials, section “Additional Results,” available online). The network includes only edges with the highest posterior probabilities. The numbers besides the edges are the posterior probabilities of including the edges inferred by BGM analysis (Supplementary Materials, section “Bayesian Graphical Models,” available online). All the other potential interaction edges have very low posterior probabilities (≤ 0.1) and not presented. Gene expression (GE) or protein expression (PE) in the parenthesis of each feature indicates gene expression or protein expression of the genes, respectively, which are used for analysis. Some of the protein expressions measure the abundances of proteins phosphorylated at specific positions. Positive interactions are indicated by **green color** and negative interactions are indicated by **red color**. Ser338, Ser217, and Ser221 indicate phosphorylations on serine-338, -217, and -221, respectively. GE = gene expression; PE = protein expression; Thr202 = phosphorylation on threonine-202; Tyr204 = phosphorylation on tyrosine-204.

DHFR, and TK1 (32,33). All of these genes have statistically significant intergenic GE-GE edges with *E2F1* in Zodiac (Supplementary Figure 4, Supplementary Materials, section “Additional Results,” available online).

Zodiac contains nearly two hundred million interaction networks, and the above examples give promising indications of the potential utility and impact of Zodiac to the community.

Web Interface

The web interface (Figure 4) of Zodiac is simple with only one input box supporting multiple query modes (Figure 4A). Users type gene symbols separated by spaces or line breaks. If users enter one gene symbol, Zodiac returns a summary of the list of genes that possess statistically significant intergenic interactions with the input gene (Figure 4B). If users input the same gene symbol twice, Zodiac returns a graph of intragenic interactions for the gene (Figure 4C). Providing two different gene symbols in the input box returns a graph with both intragenic and intergenic interactions of the two genes (Figure 4D). If multiple gene symbols are entered, Zodiac displays all statistically significant intergenic interactions between any pair of the input genes in a Circos plot (34) (Figure 4E). Details are provided in the Supplementary Materials, section “Web Server and Interface” (available online).

Discussion

The cancer biology community does not lack large databases and tools that provide interaction networks across genes (9,22,35,36).

However, few perform integration of multiple genomic features or collect computational results from a coherent probability model with FDR control. To our knowledge, Zodiac is the largest online resource reporting computational results based on the analyses of multimodal TCGA data. As a big-data integration, Zodiac is expected to generate new discoveries that penetrate deeper into the molecular mechanisms of cancer. Biologically, Zodiac yields a comprehensive map of interplays between multiple genetic factors of different genes. Such a map can provide critical information to disentangle the competing and confounding regulatory relationships. For example, the coregulation of CN and ME on GE within a gene can be revealed by examining the intragenic graph provided in Zodiac. Furthermore, investigators can perform customized network inference using provided software of TCGA-Assembler (15) and BGM on the Zodiac website. These tools enable the community to utilize TCGA data and update existing knowledge about cancer genomic interactions.

Because of its integration over multiple genomics features, Zodiac can be used to facilitate research in a variety of areas, such as transcription regulation, gene co-expression, long-range interaction between two genes, protein interaction, and DNA methylation (37). Zodiac can be incorporated and expanded to facilitate many other research efforts. Recent studies try to identify driver mutations (17,38) and stratify tumors (39) by imposing mutation profiles on existing functional interaction networks obtained from databases, like KEGG (22) and Pathway Commons (40). Highly mutated subnetworks can also be identified from the global network using this scheme (18). Zodiac can facilitate the identification of highly mutated subnetworks or driver mutations by providing high confidence interaction networks.

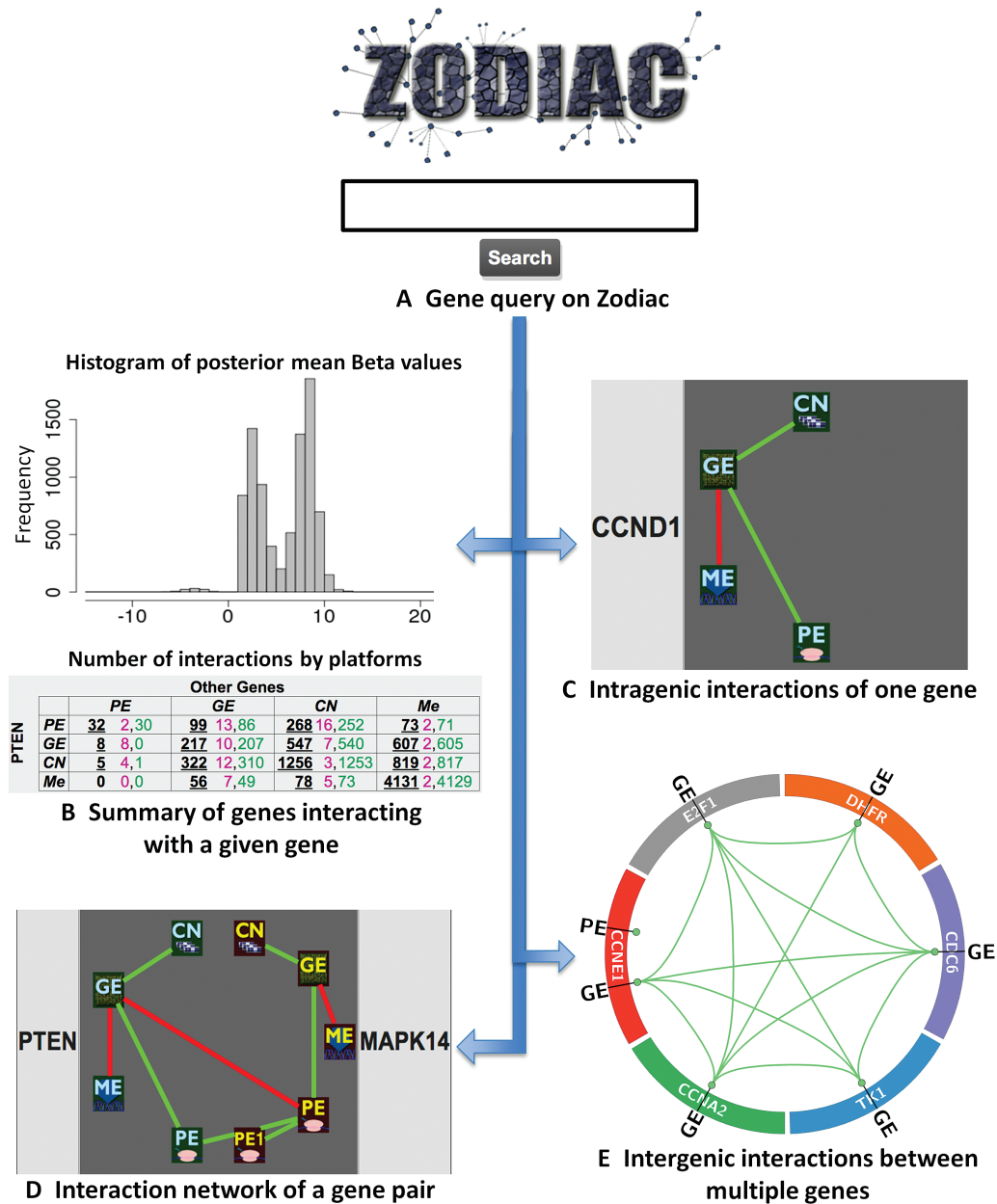


Figure 4. Web interface for querying and viewing results in Zodiac. A) The main query page is at <http://www.compgenome.org/ZODIAC>. B) Summary of genes that have statistically significant intergenic interactions with a single input gene. C) Intragenic interactions of a single input gene. D) Interaction network of two input genes. E) Statistically significant intergenic interactions of all gene pairs among multiple input genes are displayed as a Circos plot (34). The following features are provided in different query modes: First, a table of brief annotations of genes is provided with clickable links on gene symbols that connect to the genes' descriptions at the National Center for Biotechnology Information. Second, by default the graphs only report edges that have the highest posterior inclusion probabilities so that the estimated false discovery rate is no larger than a statistical significance threshold. Third, all the results will be returned on screen and a link will be provided for downloading the results. CN = copy number; GE = gene expression; Me = methylation; PE = protein expression.

When combined with mutation data, which are also available through TCGA, subnetworks specific to mutations of genes can be elucidated.

Discoveries based on BGM analyses (12,13) produce formal probability inference of the interaction networks. The benefit of such inference is that it naturally produces FDR-type of uncertainty measures on the network itself using posterior probabilities. There is a fundamental difference in the interpretation of posterior probabilities and *P* values, as the latter cannot be directly interpreted as the probability of an event (41). Posterior inference based on FDR control is applied within the analysis

of each gene or each gene pair over all potential interactions in the analysis. The use of FDR ensures the reliability of the statistical inference for each gene or gene pair over all the possible networks for the gene or gene pair. It does not, however, control the false discoveries across the approximately 200 million analyses, each subject to its own multiplicity. Controlling the overall false discoveries over all the analyses is not computationally feasible based on the graphical models because it would require evaluation of the uncertainty of roughly 1×10^{15} networks, the computation cost for each of which is substantially more expensive than many standard test-based analysis

(such as differential expression). Therefore, we choose to control the FDR for the interactions within individual analysis of a gene and a gene pair. The results presented in the manuscript are based on thresholding Bayesian FDR at 0.01. The graphical model devises conditional independence of multiple genomics features, providing a systemic and accurate depiction of co-existing interactions of multiple factors, such as CN, ME, and GE. In conclusion, the use of BGM and conditional independence inference highlights the analytical differences between Zodiac and many existing quantitative models for integration of multi-modal cancer genomics data.

The inferred edges in Zodiac are subject to many potential noises that are intrinsic to TCGA data, such as batch/lab bias. We have performed careful data preprocessing (Supplementary Materials, section “Analysis Details,” available online) to minimize these biases, but they might still have effects on certain gene pairs. Nevertheless, such biases might not severely affect the statistical inference on interactions of the genomics features. For example, we have performed additional investigation and found that the inferred network is robust to random shifts or scaling changes to the simulated data. It is important to note that results in Zodiac are based on computational analyses rather than experimental validations. A big-data computation like this will surely encompass unexpected or unconventional findings. Also, many relationships (such as intergenic CN-ME edges) are results of indirect interactions or confounders. Nonetheless, we believe that Zodiac, as an expanding resource, will benefit researchers in different areas because of its rigor in statistical modeling and fidelity to TCGA data. More importantly, the proposed integration of TCGA data will motivate more efforts and help improve understanding of the cancer biological systems using computational means.

There are some limitations of our current work. Zodiac currently is a pan-cancer project that investigates molecular interactions shared across multiple tissues. It lacks cancer-specific analysis of genetic interactions. In addition, not all molecular measurements are included in analysis, such as miRNA expression. Despite the large computational effort, Zodiac is merely an initial step towards more ambitious and exciting future investigations.

Our future plan includes using BGM to identify genetic interactions of individual cancer types and differential interactions between cancer types. Such analyses will require substantially more computation depending on the number of cancer types selected. In addition, we are looking into expanding the number of data modalities (such as including miRNA expressions and gene mutations) and the number of genes in each graph (to be greater than two), although such expansions result in an exponential increase in the computational burden, as the size of graphical space increases exponentially over the numbers of modalities and genes. We are attempting various optimization approaches to speed up the computation aiming for a future launch of genome-wide gene triplet analysis with more data modalities. Nevertheless, for important cancer-related pathways that involve far fewer genes, analysis of three or more genes is feasible based on our current BGM software, available for download on the Zodiac website. Zodiac currently is based on TCGA data retrieved and assembled in April, 2013, which are also available for download at the Zodiac website. Since then, TCGA has been adding new data with more patient samples. We plan to carry out analyses incorporating the new data in future Zodiac expansion and make results available. Especially, The Clinical Proteomic Tumor Analysis Consortium (CPTAC) (42) has been generating whole-genome proteomics data of TCGA

patient samples, which can be used for studying interactions involving protein expressions for future Zodiac analysis of specific cancer types.

Funding

YJ's and PM's research was partially supported by the US National Institutes of Health through grant R01 CA132897. This work was also supported in part by the US National Institutes of Health through resources provided by the Computation Institute and the Biological Sciences Division of the University of Chicago and Argonne National Laboratory, under grant S10 RR029030-01.

Notes

The study sponsors had no role in design of the study, the collection, analysis, or interpretation of the data, the writing of the manuscript, nor the decision to submit the manuscript for publication.

Author contributions: YJ and YZ conceived the main idea and participated in all aspects of the project. YX, RM, and PM worked on methodology development. DLH, KG, SY, SS, and WG worked on database and web portal construction. LLP, IF, NP, and JCS helped computation on Beagle. All participated in manuscript writing.

References

- Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011;12(1):56–68.
- Creixell P, Schoof E, Erler J, et al. Navigating cancer network attractors for tumor-specific therapy. *Nat Biotechnol.* 2012;30(9):842–848.
- Cancer Target Discovery Development Network. Towards patient-based cancer therapeutics. *Nat Biotechnol.* 2010;28(9):904–906.
- Chiang GG, Abraham RT. Targeting the mTOR signaling network in cancer. *Trends Mol Med.* 2007;13(10):433–442.
- Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008;455(7216):1061–1068.
- Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490(7418):61–70.
- Omberg L, Ellrott K, Yuan Y, et al. Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas. *Nat Genet.* 2013;45(10):1121–1126.
- Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013;45(10):1113–1120.
- Gao J, Aksoy B, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* 2013;6(269):p11.
- Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature.* 2014;505:495–501.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics.* 2008;9(559).
- Mitra R, Müller P, Liang S, et al. A Bayesian graphical model for ChIP-Seq data on histone modifications. *J Am Stat Assoc.* 2013;108(501):69–80.
- Xu Y, Zhang J, Yuan Y, et al. A Bayesian graphical model for integrative analysis of TCGA data. In. *IEEE Int Workshop Genomic Signal Process Stat.* Washington, DC; 2012:135–138.
- Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* 2013;41(Database issue):D808–D815.
- Zhu Y, Qiu P, Ji Y. TCGA-Assembler: open-source software for retrieving and processing TCGA data. *Nat Methods.* 2014;11(6):599–600.
- Supercomputer Beagle. <http://beagle.ci.uchicago.edu>. Accessed April 24, 2015.
- Vandin F, Upfal E, Raphael B. De novo discovery of mutated driver pathways in cancer. *Genome Res.* 2012;22(2):375–385.
- Vandin F, Upfal E, Raphael B. Algorithms for detecting significantly mutated pathways in cancer. *J Comput Biol.* 2011;18(3):507–522.
- Cowell RG, Dawid AP, Lauritzen SL, et al. *Probabilistic Networks and Expert Systems.* In: Springer; 1999.
- Nuzzo R. Scientific method: statistical errors. *Nature* 2014;506:150–152.
- Wallace TA, Prueitt RL, Yi M, et al. Tumor immunobiological differences in prostate cancer between African-American and European-American men. *Cancer Res.* 2008;68(3):927–936.
- Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 2012;40(Database issue):D109–D114.

23. Jiang C, Xuan Z, Zhao F, et al. TRED: a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Res.* 2007;35(Database issue):40.
24. Cao R, Zhang Y. The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr Opin Genet Dev.* 2004;14(2):155–164.
25. Varambally S, Dhanasekaran S, Zhou M, et al. The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature.* 2002;419(6907):624–629.
26. Yang X, Karuturi R, Sun F, et al. CDKN1C (p57) is a direct target of EZH2 and suppressed by multiple epigenetic mechanisms in breast cancer cells. *PLoS One.* 2009;4(4):e5011.
27. Fujii S, Ito K, Ito Y, et al. Enhancer of zeste homologue 2 (EZH2) downregulates RUNX3 by increasing histone H3 methylation. *J Biol Chem.* 2008;283:17324–17332.
28. Jiang X, Tan J, Li J, et al. DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell.* 2008;13(6):529–541.
29. Bracken A, Pasini D, Capra M, et al. EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer. *EMBO J.* 2003;22(20):5323–5335.
30. Wu Z, Zheng S, Li Z, et al. Polycomb protein EZH2 regulates E2F1-dependent apoptosis through epigenetically modulating Bim expression. *Cell Death Differ.* 2010;17(5):801–810.
31. Kogure M, Takawa M, Saloura V, et al. The oncogenic polycomb histone methyltransferase EZH2 methylates lysine 120 on histone H2B and competes ubiquitination. *Neoplasia.* 2013;15(11):1251–1261.
32. Helin K. Regulation of cell proliferation by the E2F transcription factors. *Curr Opin Genet Dev.* 1998;8(1):28–35.
33. Trimarchi J, Lees J. Sibling rivalry in the E2F family. *Nat Rev Mol Cell Biol.* 2002;3(1):11–20.
34. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. *Genome Res.* 2009;19(9):1639–1645.
35. Cline MS, Craft B, Swatloski T, et al. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Sci Rep.* 2013;3:2652.
36. Cline MS, Smoot M, Cerami E, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc.* 2007;2(10):2366–2382.
37. Akulenko R, Helms V. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Hum Mol Genet.* 2013;22(15):3016–3022.
38. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep.* 2013;3:2650.
39. Hofree M, Shen J, Carter H, et al. Network-based stratification of tumor mutations. *Nat Methods.* 2013;10(11):1108–1115.
40. Cerami E, Gross B, Demir E, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39(Database issue):D685–f.
41. Carlin BP, Louis TA. *Bayesian methods for data analysis.* 3rd ed: Chapman and Hall/CRC; 2008.
42. Zhang B, Wang J, Wang X, et al. Proteogenomic characterization of human colon and rectal cancer. *Nature.* 2014;513:382–387.