OXFORD

*Briefings in Bioinformatics*, 17(6), 2016, 1044–1059

**doi: 10.1093/bib/bbw076**
Advance Access Publication Date: 24 August 2016
Paper

# A network medicine approach to build a comprehensive atlas for the prognosis of human cancer

Fan Zhang, Chunyan Ren, Kwun Kit Lau, Zihan Zheng, Geming Lu, Zhengzi Yi, Yongzhong Zhao, Fei Su, Shaojun Zhang, Bin Zhang, Eric A. Sobie, Weijia Zhang and Martin J. Walsh

Corresponding authors: Martin J. Walsh, Department of Pharmacological Sciences, Icahn School of Medicine at Mount Sinai, 1468 Madison Ave., New York, NY 10029, Tel.: +1 212 2419711; Fax: +1 212 4261972; E-mail: martin.walsh@mssm.edu; Weijia Zhang, Department of Medicine, Division of Nephrology, Laboratory of Bioinformatics, Icahn School of Medicine at Mount Sinai, 1468 Madison Ave., New York, NY 10029, Tel.: +1 212 2412883; Fax: +1 212 8492643; E-mail: weijia.zhang@mssm.edu; Fan Zhang, Department of Medicine, Division of Nephrology, Laboratory of Bioinformatics, Icahn School of Medicine at Mount Sinai, 1468 Madison Ave., New York, NY 10029, Tel.: +86 15045863491; Fax: +86 451 86615922; E-mail:fan.zhang@mssm.edu

## Abstract

The Cancer Genome Atlas project has generated multi-dimensional and highly integrated genomic data from a large number of patient samples with detailed clinical records across many cancer types, but it remains unclear how to best integrate the massive amount of genomic data into clinical practice. We report here our methodology to build a multi-dimensional subnetwork atlas for cancer prognosis to better investigate the potential impact of multiple genetic and epigenetic (gene expression, copy number variation, microRNA expression and DNA methylation) changes on the molecular states of networks that in turn affects complex cancer survivorship. We uncover an average of 38 novel subnetworks in the protein–protein interaction network that correlate with prognosis across four prominent cancer types. The clinical utility of these subnetwork biomarkers was further evaluated by prognostic impact evaluation, functional enrichment analysis, drug target annotation, tumor stratification and independent validation. Some pathways including the dynactin, cohesion and pyruvate dehydrogenase-related subnetworks are identified as promising new targets for therapy in specific cancer types. In conclusion, this integrative analysis of existing protein interactome and cancer genomics data allows us to systematically dissect the molecular mechanisms that underlie unexpected outcomes for cancer, which could be used to better understand and predict clinical outcomes, optimize treatment and to provide new opportunities for developing therapeutics related to the subnetworks identified.

**Fan Zhang** is a postdoctoral research fellow in the Department of Medicine, Division of Nephrology, Laboratory of Bioinformatics and Department of Genetics and Genomic Sciences at Icahn School of Medicine at Mount Sinai.
**Chunyan Ren** is a postdoctoral research fellow in the Department of Pharmacological Sciences at Icahn School of Medicine at Mount Sinai.
**Kwun Kit Lau** is a PhD candidate in the Department of Developmental and Regenerative Biology and The Black Family Stem Cell Institute at Icahn School of Medicine at Mount Sinai.
**Zihan Zheng** is a student in the College of Arts and Sciences at University of North Carolina at Chapel Hill.
**Geming Lu** is a postdoctoral research fellow in the Department of Medicine, Immunology Institute at Icahn School of Medicine at Mount Sinai.
**Zhengzi Yi** is an analyst in the Department of Medicine, Division of Nephrology, Laboratory of Bioinformatics at Icahn School of Medicine at Mount Sinai.
**Yongzhong Zhao** is a postdoctoral research fellow in the Department of Genetics and Genomic Sciences at Icahn School of Medicine at Mount Sinai.
**Fei Su** is a lecturer in the College of Bioinformatics Science and Technology at Harbin Medical University, China.
**Shaojun Zhang** is an associate professor in the College of Bioinformatics Science and Technology at Harbin Medical University, China.
**Bin Zhang** is an associate professor in the Department of Genetics and Genomic Sciences at Icahn School of Medicine at Mount Sinai.
**Eric A. Sobie** is an associate professor in the Department of Pharmacological Sciences at Icahn School of Medicine at Mount Sinai.
**Weijia Zhang** is an associate professor in the Department of Medicine, Division of Nephrology, Laboratory of Bioinformatics at Icahn School of Medicine at Mount Sinai.
**Martin J. Walsh** is an associate professor in the Department of Pediatrics, Pharmacological Sciences and Genetics and Genomic Sciences at Icahn School of Medicine at Mount Sinai.

**Submitted:** 27 February 2016; **Received (in revised form):** 26 April 2016

**Key words:** landmark for cancer prognosis; multi-dimensional cancer genomic data; network medicine; system medicine

## Introduction

Cancer is a complex disease arising from the combined effects of multiple genetic and epigenetic changes, leading to the dysregulation of critical signaling pathways [1–5]. Owing to recent advances in next-generation sequencing technology and its emerging application in various clinical settings, it is now possible to evaluate cancers beyond the traditional clinical variables (i.e. age and tumor stage) by incorporating data profiled on multiple layers of genomic activities, such as gene expression profiles, genetic aberrations [somatic mutations and copy number variants (CNVs)], microRNA (miRNA) expression and methylation signatures. These additions in tumor profiling and stratification further improve the predictions in prognosis and outcomes [6].

Highly integrated analyses using multi-layered molecular information to help understand cancer outcomes have also been demonstrated [7–13]. Study by Xu *et al*. [7] showed that some associations between DNA copy number and gene expression have clinical or pathogenic relevance. Masica and Karchin [8] identified genes required for tumor's survival by examining the correlations among somatic mutation and gene expression. Kim *et al*. [9] integrated information from miRNA and mRNA expression profiles to improve the prediction of cancer survival time. Kim *et al*. [10] experimentally revealed an oncomir/oncogene cluster through integrative genome analysis, which could regulate glioblastoma survivorship by targeting RB1, PI3K/AKT and JNK pathways. Meanwhile, more studies started to include both genetic and epigenetic alterations in tumors in decision-making processes in clinical practice. For example, Zhang *et al*. [13] uncovered seven previously uncategorized subtypes of ovarian cancer that differentiate significantly in median survival time by integrating four types of molecular data related to gene expression. In light of these pioneering computational and experimental works, we seek to explore the cooperative effect of multi-layered genetic and epigenetic regulatory mechanisms.

Moreover, recent studies have focused on how multiple genes interact in a particular pathway or network to explain a complex clinical outcome [14–17]. For example, human protein–protein interaction (PPI) networks have been used to identify subnetwork signatures or functional modules that contribute to the positive or negative prognosis of glioblastoma multiforme (GBM), breast, colon, rectal, as well as ovarian cancers [14, 18], and the regulatory relationships of miRNAs and their target genes have been used in survival analysis of GBM and ovarian cancer [16]. A priori defined gene sets from MSigDB or KEGG pathway have also been associated with patient survival in breast cancer [15] and serious ovarian cancer [17].

The Cancer Genome Atlas (TCGA) project has provided resources for multi-platform genomic profiling from a large number of patient samples across many cancer types [1–4], resulting in multi-dimensional and highly integrated genomic data. Combined with improvements in the quality of interactome data, network analysis has made significant advancements in cancer biology. However, how to translate such multi-omics data into clinical application is still challenging. In this work, we propose a systematic approach to (i) evaluate the contribution of genes to patient survival taking into account multi-layered regulatory mechanisms including CNV, DNA methylation, mRNA and miRNA expression; (ii) identify subnetworks of the survival-related genes in PPI network; (iii) and generate multi-dimensional subnetwork-derived prognostic models. Finally, we uncover an average of 38 new featured subnetworks linked with prognosis across four cancer types. Further functional enrichment analysis, drug target annotation, tumor stratification and independent validation were used to evaluate the clinical utility of these subnetwork-derived models in cancer prognosis. Our study demonstrates a novel method for integrating human genomics and interactome data that proves useful for refining our biological understanding of cancer prognosis and potentially improving outcomes.

## Material and methods

### Study design

The aim of our study was to detect the potential impact of multiple genetic and epigenetic changes on the molecular states of networks that in turn affects complex cancer outcome. We reported the methodology to build a multi-dimensional subnetwork atlas for cancer prognosis through integrating the multi-type cancer genomics data from 1027 samples of four cancer types from TCGA project and the interactome data including PPI and miRNA–gene interaction. We further assessed the clinical utility of these multi-dimensional subnetwork biomarkers through prognostic impact evaluation, functional enrichment analysis, drug target annotation, tumor stratification and independent validation.

### Multi-dimensional genomic data

The multi-dimensional cancer-associated data sets containing clinical information, copy-number variation (CNV), promoter DNA methylation, mRNA-gene and miRNA expression data were collected from TCGA Cancer Browser (https://genome-cancer.ucsc.edu/proj/site/hgHeatmap/). A brief summary of the data information is provided in Table 1. Overall survival data of patients in four TCGA cancer types were considered in our article: lung squamous cell carcinoma (LUSC), GBM, kidney renal clear cell carcinoma (KIRC) and ovarian serous cystadenocarcinoma (OV).

### Protein–protein interaction data

The PPIs data from Human Protein Reference Database (HPRD) [19] were used in this study. HPRD contains over 36 700 manually curated interactions between 9205 human proteins.

### Identification of the miRNA-regulators for target genes

Two miRNA target databases (miRTarBase (Release 4.5) [20] and TarBase v6 [21]), which provide experimentally validated miRNA–target interactions, were used. Because the biologically relevant targets of each miRNA may vary from one tissue to the next, depending on the expression of the target mRNAs and the cellular context, we selected those miRNAs whose expression was inversely correlated ($r < -0.15$, $P < 0.01$), with mRNA expression in each cancer type as the regulators for the target genes [10].

**Table 1.** Summary of specimens derived from TCGA by high-throughput analysis of the four primary molecular features for each cancer type

| Cancer | CNV | Methylation | mRNA | miRNA | Core set |
|---|---|---|---|---|---|
| LUSC | GISTIC2 | 450k | HiseqV2 | HiSeq | |
| | 468 × 11 878 genes | 341 × 197 569 probes | 467 × 11 442 genes | 316 × 794 miRNAs | 313 |
| GBM | GISTIC2 | 27k | AgilentG4502A | Not used | |
| | 420 × 11 878 genes | 214 × 14 445 probes | 348 × 10 097 genes | | 156 |
| KIRC | GISTIC2 | 450k | HiseqV2 | HiSeq | |
| | 517 × 11 878 genes | 309 × 197 940 probes | 522 × 11 442 genes | 244 × 716 miRNAs | 169 |
| OV | GISTIC2 | 27k | HiseqV2 | HiSeq | |
| | 559 × 11 878 genes | 579 × 13 628 probes | 414 × 11 479 genes | 487 × 673 miRNAs | 398 |

## Data processing

We downloaded a list of expressed genes from syn1734155, including 12 081 genes with at least 3 RNA-Seq reads per sample in at least 70% of samples. We restricted the downstream analysis to these shared genes plus 18 well-known cancer genes (*AR*, *CDH4*, *EGFR*, *EPHA3*, *ERBB4*, *FGFR2*, *FLT3*, *FOXA1*, *FOXA2*, *MECOM*, *MIR142*, *MSH4*, *PDGFRA*, *SOX1*, *SOX9*, *SOX17*, *TBX3*, *WT1*) that have low-transcript detection levels, as used in [22]. As such, 12 099 genes were considered in total.

CNV profiling was estimated using the GISTIC2 method, annotated to genes using UCSC cgData HUGO probeMap, and further filtered with 12 099 expressed genes. Finally, CNV of 11 878 genes was consolidated in 468 LUSC, 420 GBM, 517 KIRC and 559 OV samples.

The DNA methylation profile was measured experimentally either using the Illumina Infinium Human DNA Methylation 450K (for LUSC and KIRC) or 27K (for GBM and OV) platform. After filtering out all probes with missing values in the DNA methylation profile, the probes were mapped onto the human genome coordinates using cgData probeMap derived from GEO GPL13534 record and further filtered with the selected 12 099 expressed genes. In all, we mapped 197 569 probes to 11 350 genes in 341 LUSC samples, 14 445 probes to 8966 genes in 214 GBM samples, 197 940 probes to 11 355 genes in 309 KIRC samples and 13 628 probes to 8790 genes in 579 OV samples.

The mRNA expression profile was measured using either Agilent 244K Custom Gene Expression G4502A (for GBM) or Illumina HiSeq 2000 RNA Sequencing V2 (for LUSC, KIRC and OV). Finally, there were 11 442 genes used both in 467 LUSC and 522 KIRC samples, 10 097 genes in 348 GBM samples and 11 479 genes in 414 OV samples.

The miRNA expression profile was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform. miRNAs that were expressed in less than five patients were removed. In the end, there were 794, 716 and 673 miRNAs used in LUSC, KIRC and OV, respectively.

For each cancer type, we defined the sample intersection across all platforms as the core sample set.

## Identification of novel subnetwork signatures of survival-related genes

For each gene, we evaluated its effect on the survival time of patients by taking into account all four types of molecular features (mRNA expression, CNV, promoter DNA methylation status and the expression of its regulatory miRNA). The R program 'coxph' was used to fit a univariate Cox proportional hazards model between each molecular feature and patient survival time, with the likelihood ratio test being used to estimate the significance. Only the features that passed the cutoff of $P < 0.05$ were considered to be related to survival time. From this analysis, we derived a score (heat) for each gene calculated through the Equation (1), which was summarized as the sum negative natural logarithm of single molecular feature $P$-values (Figure 1A). This sum corresponded to the statistic of Fisher's Method for combining $P$-values for (independent) statistical tests [23] and got comparable gene scoring results with Fisher's Method (Supplementary Figure S1).

$$score = -\sum_m \log_e(p_m), m = mRNA, CNV, methy, miRNA \qquad (1)$$

For mRNA-gene expression or CNV, $P$ value was defined as:

$$p_m = \begin{cases} p, & p < 0.05 \\ 1, & p \geq 0.05 \end{cases}, m = mRNA, CNV \qquad (2)$$

For miRNA expression or DNA methylation, considering the fact that one gene may have multiple methylation loci or several miRNA regulators, we only retained one CpG methylation probe or one of its miRNA regulators that was most correlated with survival time, and the $P$ value was defined as:

$$p_m = \begin{cases} \min(p), & \min(p) < 0.05 \\ 1, & \min(p) \geq 0.05 \end{cases}, m = methy, miRNA \qquad (3)$$

The genes with a score$>$0 were identified as survival-related genes. Then, the heat score was used as the input into HotNet2 [22, 23], which uses a heat diffusion process and a statistical test-based algorithm to discover subnetwork signatures in PPI network (Figure 1B). Thus, subnetwork signatures of survival-related genes were determined both by the scores of their genes and the interactions between the genes.

## Training and evaluation of multi-dimensional subnetwork-derived prognostic models

For each subnetwork, we first assembled a multi-dimensional molecular profile by extracting all four types of molecular features of its gene members from the core sample set of a particular cancer type. We then explored the predictive power of the subnetwork on patient overall survival using a Monte Carlo cross-validation and permutation testing procedure. Briefly, for the core sample set, we randomly split the samples into two groups: 80% as the training set and 20% as the test set. To fairly and accurately evaluate the prognostic power of each subnetwork (with different number of genes), on the training set, we used the Cox proportional hazards model with L1 penalized log partial likelihood (LASSO) [24] for feature selection to train the models based on the molecular profile of individual

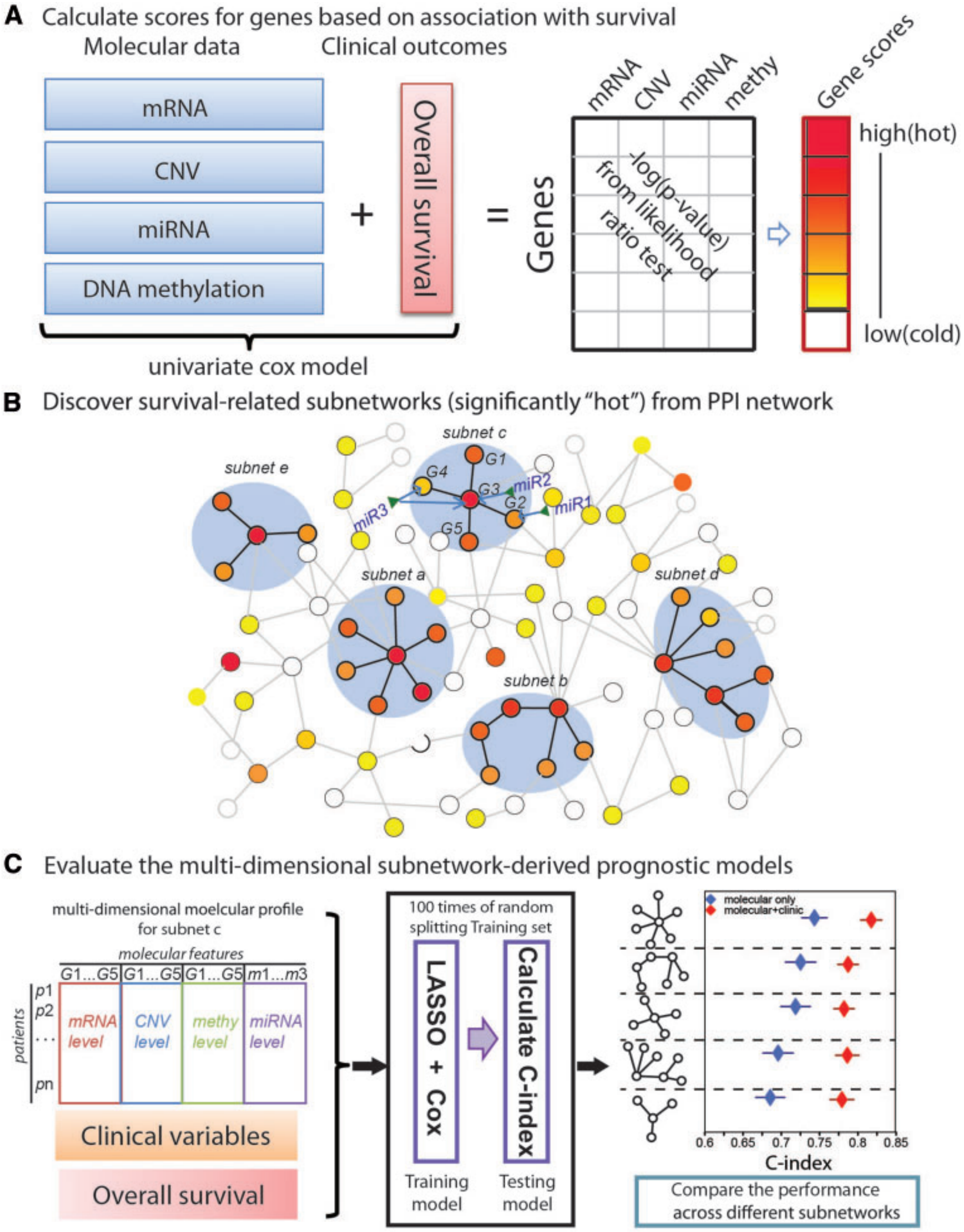

**Figure 1.** An overview of the computational approach. (**A**) A score (heat) was derived for each gene to evaluate the collaborative effect of different molecular features (gene expression, CNV, miRNA expression and DNA methylation) on patient survival time. First, *P*-values representing the significance of each molecular feature correlated with patient overall survival time were calculated using the likelihood ratio test of univariate cox model. The score was then calculated as the negative sum of the natural logarithm of the single molecular feature *P*-values (Red: high score; Yellow: low score; White: score = 0). (**B**) Subnetworks were identified using HotNet2 algorithm in a PPI network. HotNet2 used a heat diffusion process and a statistical test to derive significant subnetworks based on both the score of the genes and the local topology of the subnetwork. (**C**) Monte Carlo cross-validation and the concordance index (C-index) were applied to assess the predictive power of each subnetwork signature, based on either the multi-dimensional genomics data alone (blue) or in combination with clinical variables (red).

subnetwork. The prognostic outcomes for the training set were used to determine the regression coefficients. These coefficients were then used to predict outcomes for patients in the test set and calculate the concordance index (C-index). The above procedure was repeated 100 times to generate 100 C-indexes, and the median C-index was used as the predictive value for each subnetwork (Figure 1C). Furthermore, to test if the models built from each subnetwork showed statistically significant predictive power, 100 survival-permuted data were used to calculate *P*-values based on the comparison of the median C-index values of the original survival data with the distributions of the median C-indexes of the 100 survival-permuted data. Our survival predictive models were evaluated based on a research framework, which could be accessed in Synapse (doi:10.7303/syn1710282).

Accordingly, to assess the predictive power of integrating molecular data with clinical variables, we combined the molecular features with clinical variables to build a new multivariate Cox model. To compare the performance across different prognostic models, the one-tailed Wilcoxon signed rank test was used to calculate the *P*-value ($P < 0.05$ as the significance cutoff).

### Selection and characterization of important molecular features in the prognostic models

When building the predictive model using the molecular features of each subnetwork, LASSO was used to select a small number of 'important' features. Basically, 100 samplings of the training set could extract 100 important feature sets and the occurrence of each molecular feature was counted (Figure 3A). Because the possibility of random selection bias for any given feature could be ruled out if the feature was consistently selected for, we only kept features occurring more than five times to construct our final predictive model. The selected features were fitted in a multivariable Cox regression model using all the samples. A risk score formula was then established by weighting each of these selected features by their estimated regression coefficients in the multivariable Cox regression analysis. With this risk score formula, patients in each set were classified into high-risk or low-risk groups using the median risk score as the cutoff. Survival differences between the low-risk and high-risk groups identified in each set were assessed by the Kaplan–Meier estimate and compared using the log-rank test. Z-score transformation was used to adjust the data scale among different molecular data sets when generating heatmaps.

## Results

### Collaborative effect of genetic or epigenetic molecular features on cancer patient survival

We first investigated how multiple layers of cellular activities (either genetic or epigenetic) may contribute to clinical outcomes (i.e. patient survival time) in different cancer types. From the TCGA, we collected and pre-processed the data sets for four cancer types (LUSC, GBM, KIRC and OV), including the clinical records for each patient, and four types of high-throughput molecular data related to gene expression ((i) CNV; (ii) DNA methylation; (iii) mRNA expression; (iv) miRNA expression, hereafter denoted as diverse molecular features of a gene). An important step in the process, described in Materials and Methods, was to map each molecular feature to one or more genes, thereby allowing us to examine subnetworks of interacting genes. For each cancer type, we removed the samples with incomplete information for overall survival time or clinical variables (e.g. gender, age, tumor stage and grade). We also eliminated genes with low expression in all tumor types [22]. The information on the final number of molecular features and samples used in downstream analysis is listed in Table 1.

The significance of each molecular feature correlated with patient overall survival time was measured based on a univariate Cox proportional hazards model (likelihood ratio test $P < 0.05$ as cutoff, Figure 1A). In particular, we confirmed the association of diverse molecular features with the survival time of patients using 121 clinically relevant genes [6] (13 genes were excluded by low expression filtering), and showed that the contribution of these genes to survival involved multi-layered regulatory mechanisms that may vary in different types of cancers (Supplementary Figure S2). To further evaluate the contribution of each gene, we derived a score that indicated the collaborative effect of all its molecular features on patient survival (Figure 1A, and Materials and Methods section). Genes with a score >0 were identified as survival-related genes (genes with at least one of the four molecular features associated with patient survival time). Generally, mRNA expression and DNA methylation features gave the most contribution to patient survival variation, followed by CNV and miRNA features. In KIRC, 73% of genes were associated with patient survival on multiple molecular layers followed by LUSC (∼30%), GBM (∼21%) and OV (∼12%), and well-studied genes (genes with more molecular features tested) tended to have higher scores (Supplementary Table S1 and Supplementary Figure S3).

### Generating multi-dimensional subnetwork atlas for the prognosis of human cancer

The initial analysis examined association with patient survival on a gene-by-gene basis. Although this approach can correctly identify critical genes, it is also likely to produce false positives. We hypothesized that more robust and predictive results could be obtained by examining subnetworks of interacting genes. Therefore, after each gene was assigned a score (heat), HotNet2 [22, 23] was used to discover the survival-related subnetworks or network modules from a large PPI network obtained from HPRD [19] (Figure 1B, Materials and Methods). As a result, 30 subnetworks with at least four connected survival-related genes were identified for OV, 87 for LUSC, 134 for KIRC and 52 subnetworks for GBM, respectively.

To assess the predictive power of these candidate multi-dimensional subnetwork signatures, we performed Monte Carlo cross-validations with 100 randomizations of training and testing sample groups, and the median C-index across 100 randomizations was calculated for each subnetwork (Figure 1C, see also Materials and Methods). The nonparametric C-index is scaled such that a C-index of 1 indicates perfect prediction accuracy, whereas a C-index of 0.5 is equal to random guess. We observed that >97% of the subnetworks of OV, KIRC and LUSC had a median C-index >0.5, but only 65% from GBM had C-indexes >0.5. Furthermore, 100 survival-permuted data were used to test if the subnetwork-derived models showed statistically significant predictive power. At a $P\text{-value} < 0.05$ level, we finally determined 20 subnetworks in OV (O1–O20), 30 in LUSC (L1–L30), 98 in KIRC (K1–K98) and 7 subnetworks in GBM (G1–G7) as prognostic biomarkers. These subnetworks were numbered from 1 to *N* according to the predictive power in descending order (Supplementary Figure S4).

To understand and characterize the biological roles of each subnetwork biomarker underlying the complex clinical phenotype, we further performed functional enrichment analysis based on the known pathways or functional categories using Enrichr [25]. This analysis identified several pathways known to be involved in cancer, such as DNA repair pathway (K71, O17) [26, 27], the mTOR pathway (K84, L8) [28], Vesicle (Lysosome, Golgi, ER) and cytoskeleton regulation (K23, K67, L21, L22, G6, O15) [29–31], Notch signaling (K93, L15, O7), the VEGF signaling pathway (L20, O18) [32] and DNA damage response (O8) [33]. In addition, we identified less characterized but interesting subnetworks, such as the dynactin-related subnetwork (O1) in OV, the miRNA-regulated kinetocore subnetwork in KIRC (K1) and the PDKs-regulated metabolism switch subnetwork in LUSC (L2) (Figure 2, Supplementary Figure S5 and Supplementary Table S2). The roles of those subnetworks in cancer are worth further investigation.

### Molecular insights from the top prognostic model

To further understand why the multi-dimensional subnetworks identified above may be related to patient survival, we used the top-ranked subnetworks in OV, LUSC and KIRC as examples (Figure 2 highlighted in red), while more examples can be found

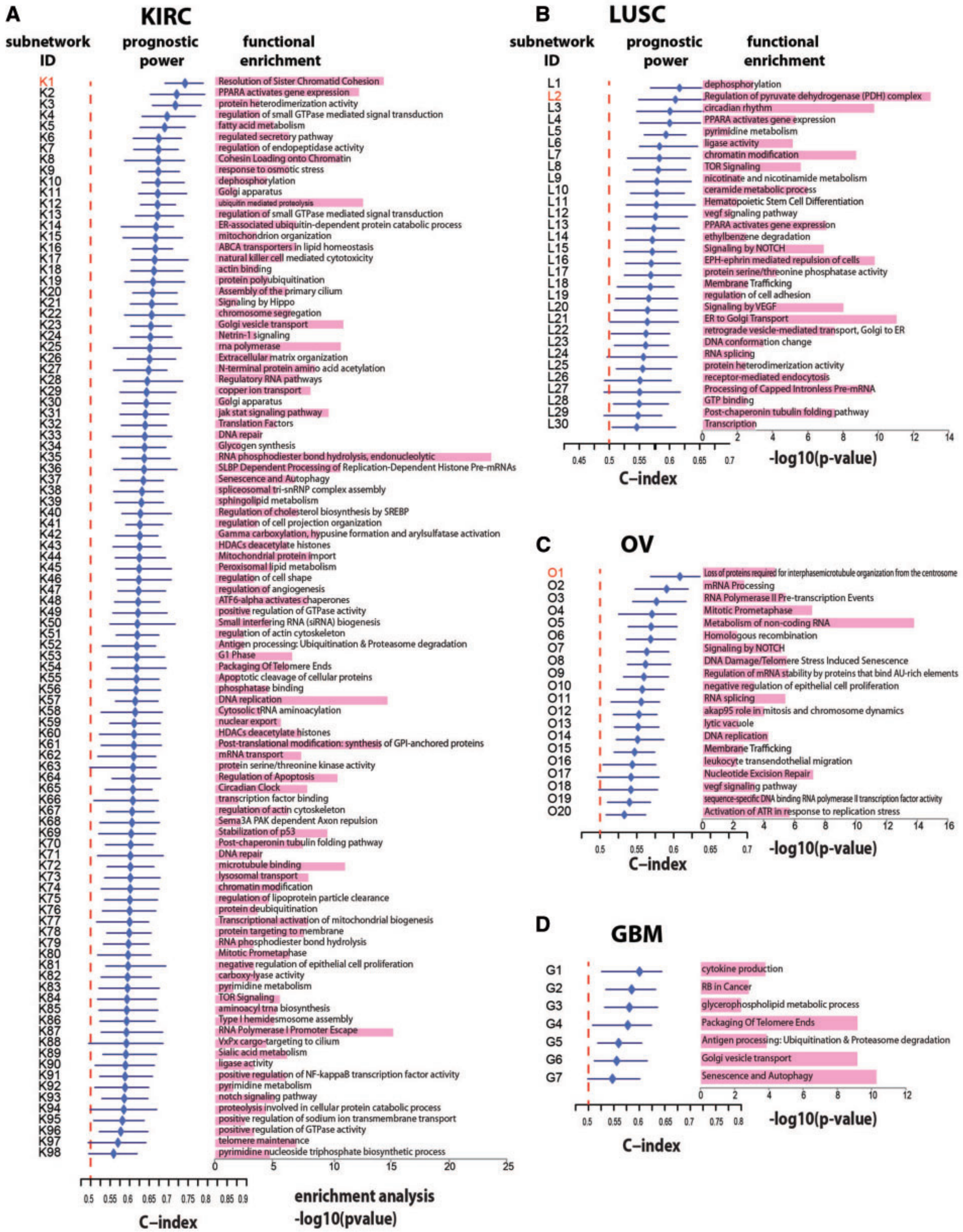

**Figure 2.** Evaluation and characterization of the multi-dimensional subnetwork-derived prognostic models. (**A–D**) C-indexes of models trained from the multi-dimensional molecular profiles of the subnetworks in KIRC (Ntotal = 169) (**A**), LUSC (Ntotal = 313) (**B**), OV (Ntotal = 398) (**C**), GBM (Ntotal = 156) (**D**). For each cancer type, during each of the 100 times of random splitting, 80% of the total samples were used to train the model and the remaining of 20% were used as the test set for C-index calculations. (The whiskers mark the 25th and 75th quartiles, with the median in the center). The red dashed lines marked the C-index equivalent to a random guess (C-index = 0.5). Functional enrichment analysis was performed using Enrichr to characterize the biological role of each subnetwork biomarker based on the known pathways or functional categories. See also Supplementary Figure S5.

in Supplementary Figure S6. Our subnetwork signatures contain important genes, interaction partners and regulation patterns that offer potential insight into mechanisms associated with tumor behavior. For each subnetwork, a consensus prognostic model was built using features that were selected in the LASSO Cox regression model at least five times across the 100 Monte Carlo cross-validations (Figure 3A, B(a) and C (a)) (information for all the prognostic models in Supplementary Table S3). To demonstrate the usefulness of these models, we evenly divided patients into high-risk and low-risk groups according to each patient's predicted risk score, and observed significant survival differences between the two groups (Figure 3B(b) and C(b), Materials and Methods). We also tested the performance of the models against an independent sample set derived from another TCGA-based study [6] (Figure 3B(c) and C(c)). The difference of the molecular profile between the two risk groups was shown in the heatmap (Figure 3B(d) and C(d)).

### *Dynactin-related subnework in OV—O1*

Overall, O1 is dominated by genes related to dynactin and its interacting partners, which play key roles in cytoskeleton reorganization and spindle assembly. The hazard ratio (HR) was used to estimate the association of individual molecular features with survival (better or worse), where an HR >1 represented a worse prognosis. For example, mRNA_*APC* had an HR of 1.543 and mRNA_*DCTN1* had an HR of 1.35, indicating that a higher level of expression of these genes was associated with shorter survival, while mRNA_*CASP2* (HR: 0.626) indicated that the higher level of expression was associated with longer survival. Likewise, higher level copy number of *DCTN1* (CNV_*DCTN1*), *APC* (CNV_*APC*) and *MAPRE1* (CNV_*MAPRE1*) was associated with a worse prognosis and higher level of DNA methylation at *PGAM1* promoter (methy_*PGAM1*) was associated with longer survival.

Prior knowledge helps us to understand how these molecular features influence cell function and may affect survival in cancer. *DCTN1* is the largest subunit of the dynactin complex that binds to microtubules and cytoplasmic dynein, which is required for cellular structures and motor functions [34]. *DCTN1* can be cleaved by caspases during apoptosis, possibly explaining why *CASP2* upregulation is linked to better survival. *MAPRE1* is a binding partner of both *DCTN1* and *APC*, which combine together to regulate microtubule polymerization, spindle dynamics and chromosome alignment [35–38]. Overexpression of *MAPRE1* has been found to occur in tumors and its oncogenic role has been shown to promote the $\beta$-catenin/T-cell factor pathway [39, 40]. Though the link between *PGAM1* and *DCTN1* has not yet been conclusively shown, *PGAM1* provide a metabolic advantage to promote tumor growth by coordinating glycolysis and biosynthesis [41]. Therefore, our prediction showed that low presence of dynactin protein *DCTN1* and its binding partners, such as *MAPRE1* (with *APC*) by low mRNA, low copy number or high DNA methylation, and high *CASP2* (cleave *DCTN1*) were associated with low risk of relapse, which matched well with previously reported roles in cancer progression (Figure 3B and Table 2).

### *Pyruvate dehydrogenase-related subnetwork in LUSC—L2*

Aerobic glycolysis over oxidative phosphorylation (the Warburg effect) is a hallmark of metabolic reprogramming in cancer cells [42]. We predict a group of genes that are critical to regulate this metabolic switch: Pyruvate dehydrogenase (*PDHs*; oxidative phosphorylation related genes to convert pyruvate to acetyl-CoA for TCA cycle) and *PDKs* (aerobic glycolysis related genes to inhibit PDHs by phosphorylation) [42]. For example, mRNA_*PDHA1* by lower *miR-326* is predicted to link to better survival. *MiR-326* has been shown to be a tumor suppressor in colorectal cancer [43] and gastric cancer [44], and elevated *miR-326* could down-regulate *MRP-1* (multidrug resistance-associated protein) and sensitize drug resistant cells to VP-16 and doxorubicin treatment [45]. Moreover, we predicted that lower mRNA_*PDK4* targeted by higher *miR-103-1* and *miR-16-2* was also linked to better survival, by which cancer-favored aerobic glycolysis would be inhibited and the tumorigenic role of *PDK4* has been demonstrated at least through activating *CREB-mTORC1* signaling cascade [46]. As to why elevated higher *PDK2* was predicted with better survival, it is probably because cancer cells are responsive to *PDK2* inhibitors such as dichloroacetate [47]. In contrast, *PDPs* promote PDH activity by dephosphorylation and inhibition of Warburg effect [48], which agrees well with our prediction that lower methy_*PDP1* is linked to better prognosis (Figure 3C and Table 2). Thus, our prediction agrees well with the metabolic reprogramming theory

### *Kinetocore-related subnework in KIRC—K1*

Another top ranking yet less characterized subnetwork (K1) in KIRC was related to 'sister chromatid Cohesion', which is critical to accurately segregate chromosomes throughout cell cycle [49]. Low levels of *miR-149* and *miR-16a*, which target *KNTC1* and *ZW10*, respectively, were associated with better prognosis, implying that these genes improve prognosis. Both *KNTC1* and *ZW10* are mitotic checkpoint proteins binding to kinetocores, and cells lacking these proteins fail to arrest in mitosis when exposed to microtubule inhibitors [50]. Meanwhile, a clear pattern of lower mRNA of *DSN1*, *MIS12*, *NSL1*, *ZWILCH* and higher methy_*ZWINT* is associated with better prognosis. *DSN1* (*MIS13*) and *NSL1* (*MIS14*) are components of *MIS12* complex, an unstable complex that may restrict kinetochore assembly to specific chromosomal regions [49]. *ZWINT* is a *ZW10* and *MIS12* interacting protein, and *ZWILCH* is a component of *KNTC1*/*ZW10* complex, both of which are able to bridge kinetochore proteins [51, 52]. The recruitment of these proteins (i.e. *KNTC1*, *ZW10*, *ZWILCH*, *ZWINT1*) to kinetochores can be affected by Aurora B kinase activity, and *CASC5* can promote Aurora B activity to phosphorylate the outer kinetochore, serve as a scaffold for kinetocore protein assembly and increase kinetochore–microtubule dynamics [53, 54]. Although the roles of each protein in this subnetwork have not been well-characterized in cancer, the importance of the kinetocore in cancer has been attributed to genomic instability and aneuploidy formation, which are common features of tumors [55] (Supplementary Figure S6 and Table 2).

## Understanding and evaluating the clinical utility of the subnetwork atlas

Above all, we identified the subnetworks within the comprehensive PPI network perturbed (or affected) by multiple genetic and epigenetic events associated with survival and further generated prognostic models from these multi-dimensional subnetworks. These subnetworks as a whole function as an atlas or landmark for cancer prognosis and reflect the dys-regulation of diverse cellular events underlying cancer outcome, including cell cycle, cellular response to stress, metabolism, signal transduction, gene expression, developmental biology, metabolism of protein, DNA repair and replication and others, thus providing clues about which cellular functions and biochemical pathways contribute to cancer outcome (Figure 4A).

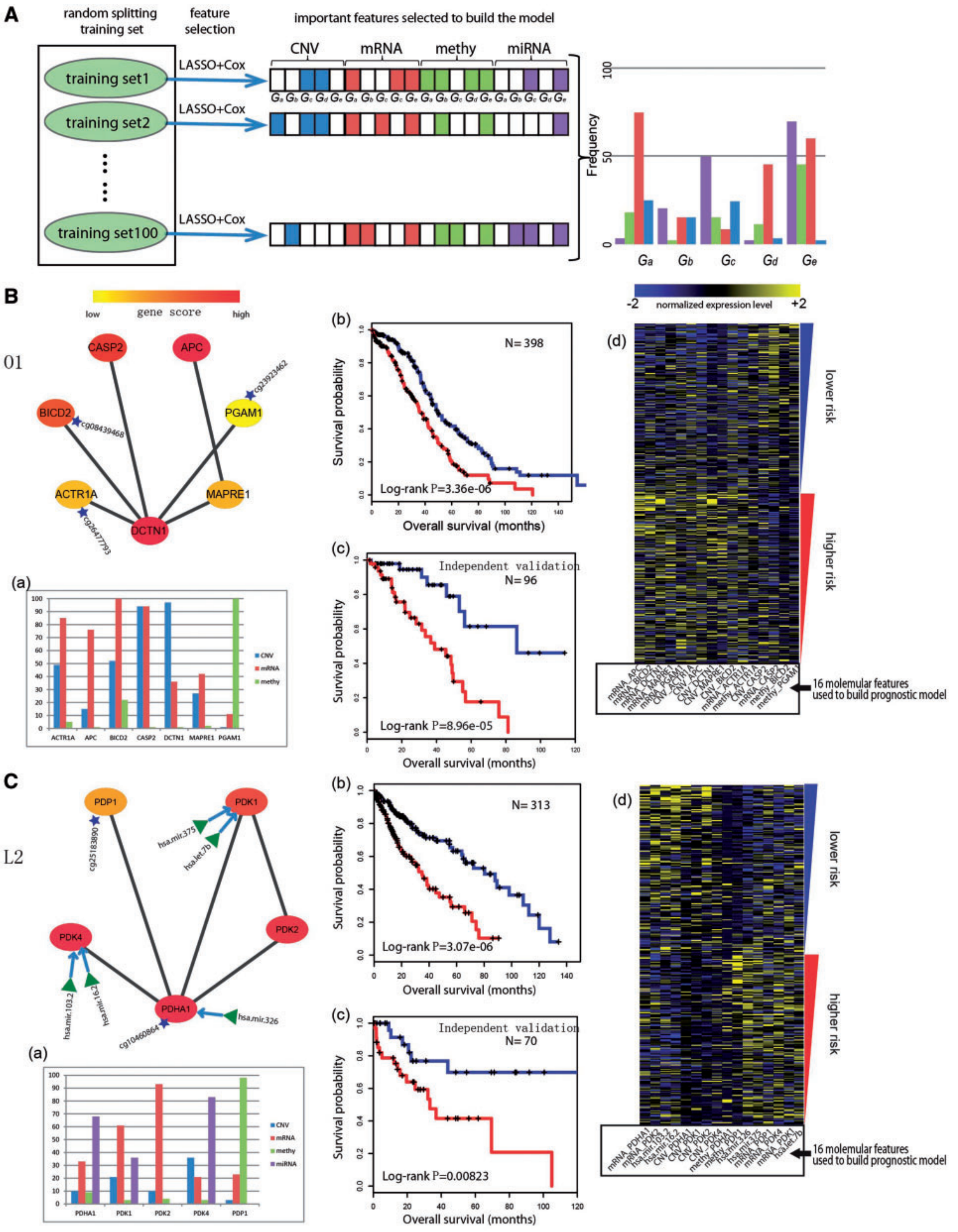

**Figure 3.** Top-ranked subnetwork biomarkers in OV and LUSC as examples to illustrate the molecular insights from the prognostic models. (**A**) Selection procedure for the important molecular features used to build the prognostic model for each subnetwork. (**B and C**) (a) Frequency of the molecular features selected by LASSO during the 100 samplings of training data; (b) Kaplan–Meier analysis according to the subnetwork-derived prognostic model, comparing overall survival time for predicted higher-risk patients versus lower-risk patients. The differences between the two curves were determined by the two-sided log-rank test; (c) Kaplan–Meier estimates of overall survival in independent test data sets. The differences between the two curves were determined by the two-sided log-rank test; (d) The heatmap of the molecular profile for the subnetwork biomarker: Rows represent patients (grouped as higher-risk or lower-risk), and columns represent selected molecular features used for making predictions. Z-score transformation was used to adjust the data scale among different molecular data sets when generating heatmaps. See also Supplementary Figure S6.

**Table 2.** Important molecular features identified from the top-performing and less characterized multi-dimensional subnetwork-derived prognostic models

| Subnet | Molecular features | HR | 95% CI of HR | P value |
|---|---|---|---|---|
| O1 | mRNA_CASP2 | 0.626 | 0.478–0.82 | 0.00068 |
| Dynactin-microtubule-related | CNV_DCTN1 | 1.738 | 1.226–2.463 | 0.00189 |
| sub-network | CNV_APC | 1.4495 | 1.0555–1.99 | 0.0218 |
| | Methy_ PGAM1 | 4.35E-05 | 6.14E-09–0.31 | 0.0264 |
| | mRNA_APC | 1.543 | 1.051–2.264 | 0.0268 |
| | CNV_MAPRE1 | 1.2386 | 1.022–1.5 | 0.0292 |
| | mRNA_DCTN1 | 1.3482 | 1.0116–1.7967 | 0.04145 |
| L2 | hsa.mir.326 | 1.29 | 1.11–1.5 | 0.00092 |
| Regulation of pyruvate dehydrogenase | methy_PDP1 | 1.79E+10 | 10395–3.08E+16 | 0.00127 |
| (PDH) complex related sub-network | hsa.mir.16.2 | 0.7156 | 0.5648–0.9067 | 0.00558 |
| | CNV_PDK2 | 0.3744 | 0.185–0.7573 | 0.0063 |
| | mRNA_PDK4 | 1.1225 | 1.0228–1.232 | 0.015 |
| | hsa.mir.103.1 | 0.6958 | 0.5145–0.94 | 0.0185 |
| K1 | hsa.mir.149 | 1.6438 | 1.3039–2.0723 | 2.6E-05 |
| Resolution of sister chromatid | methy_ZWINT | 6.29E-101 | 1.95E-154–2.03E-47 | 2.42E-04 |
| cohesion-related sub-network | CNV_ZWILCH | 0.031 | 0.0037–0.2534 | 0.00122 |
| | hsa.mir.18a | 1.9256 | 1.2915–2.871 | 0.0013 |
| | hsa.mir.192 | 0.778 | 0.66–0.9124 | 0.002 |
| | CNV_CASC5 | 0.109 | 0.0266–0.4467 | 0.002 |
| | mRNA_ZWILCH | 2.335 | 1.2446–4.382 | 0.0083 |
| | mRNA_ZWINT | 1.782 | 1.1267–2.8179 | 0.0135 |
| | mRNA_KNTC1 | 1.6255 | 1.0743–2.4596 | 0.0215 |
| | mRNA_DSN1 | 2.11 | 1.0668–4.169 | 0.03188 |

HR = hazard ratio, CI = confidence interval, two-sided *P* values were derived from the univariate cox proportional hazards model.

We further annotated the subnetworks using Drugbank (http://www.drugbank.ca) and the cancer gene index (CGI, https://wiki.nci.nih.gov/display/cageneindex) gene-compound database to identify genes that have Food and Drug Administration (FDA)-approved drugs or experimental compounds available (Figure 4A genes labeled as red and yellow, respectively). We captured 35 (22.6%) subnetworks (KIRC:26, LUSC:3, OV:5, GBM:1) with FDA-approved drugs available, such as VEGF-related subnetwork (L20), proteasome-related subnetwork (K64) and cholesterol synthesis/sterol response-related subnetwork (K40) that are well-known to be related to cancer. Moreover, we also identified 123 (79.3%) subnetworks (KIRC:79, LUSC:22, OV:18, GBM:4) with available experimental compounds, such as those related to kinetochore (K72), Notch signaling (L15) and dynactin (O1), that are currently being scrutinized as targets in cancer studies. There are also a number of subnetworks (KIRC:17, LUSC:7, OV:2, GBM:2) without any compound available, whose roles in cancer have been either intensively (i.e. G7) or barely characterized (i.e. O5, K87) (Figure 4A, highlighted in box, and Supplementary Figure S5).

We also compared the predictive power between using standard clinical variables alone and in combination with our multi-dimensional subnetwork biomarkers. Improved predictive power was shown in 22 subnetwork models in KIRC (one-sided Wilcoxon signed rank test, $P < 0.03$, $fdr < 0.05$). However, the quantitative gains (in terms of the median value of Somers' D rank correlation coefficient across the 100 splits, Somers' D equals to 2*C-1 where C denotes C-index) were limited (2.1–14.4% for 22 subnetwork-based models). Improved predictive power was also observed for all 30 subnetwork models in LUSC (one-sided Wilcoxon signed rank test, $P < 0.007$, $fdr < 0.001$) with apparent gains (Somers' D, 36.4–132%) (Figure 4B and Supplementary Table S2). As shown in Figure 4B, if the 313 patients with LUSC were stratified based on clinical variables (age and tumor stage), there was no clearly difference in median survival time (high-risk group: 55 months versus low-risk group: 64 months, log-rank $P = 0.3020$). In contrast, combing the subnetwork biomarker L4, as an example, we observed that the low-risk group had a median survival of 108 months, whereas the high-risk group had a median survival of 33 months (log-rank $P < 0.0001$). Additionally, the top two in GBM (G1, related to cytokine production; G2, containing RB-related genes) and one in OV (O3) showed improved prediction. Thus, the addition of our subnetwork models could facilitate traditional cancer management merely based on clinical variables.

To further facilitate clinical application, we narrowed down the enormous genome-wide molecular features into a smaller set of key subnetworks associated with survival and provided the ranking of the most important molecular features, which won out from our stepwise analysis in Supplementary Table S3. In addition, we developed a freely accessible web-based resource of our results to allow researchers in basic science and translational medicine to use the prognostic models directly to uncover specific genes or markers of interest, and avoid the time-consuming genome-wide screening. The homepage of the resources can be accessed via http://fanlabresources.org/.

## Identifying tumor subtypes associated with patient survival

We further assessed if our subnetwork atlas could help to stratify patients into distinct clusters or subtypes that were associated with survival. In all, 169 KIRC patients were divided into three clusters via non-negative matrix factorization (NMF) based on the similarity of their molecular profiles, which included 897 molecular features (CNV: 186, methylation: 163, mRNA: 425, miRNA: 123) derived from the 98 subnetwork models (Figure 5A). The three subtypes of patients were predicted consistently as low-,

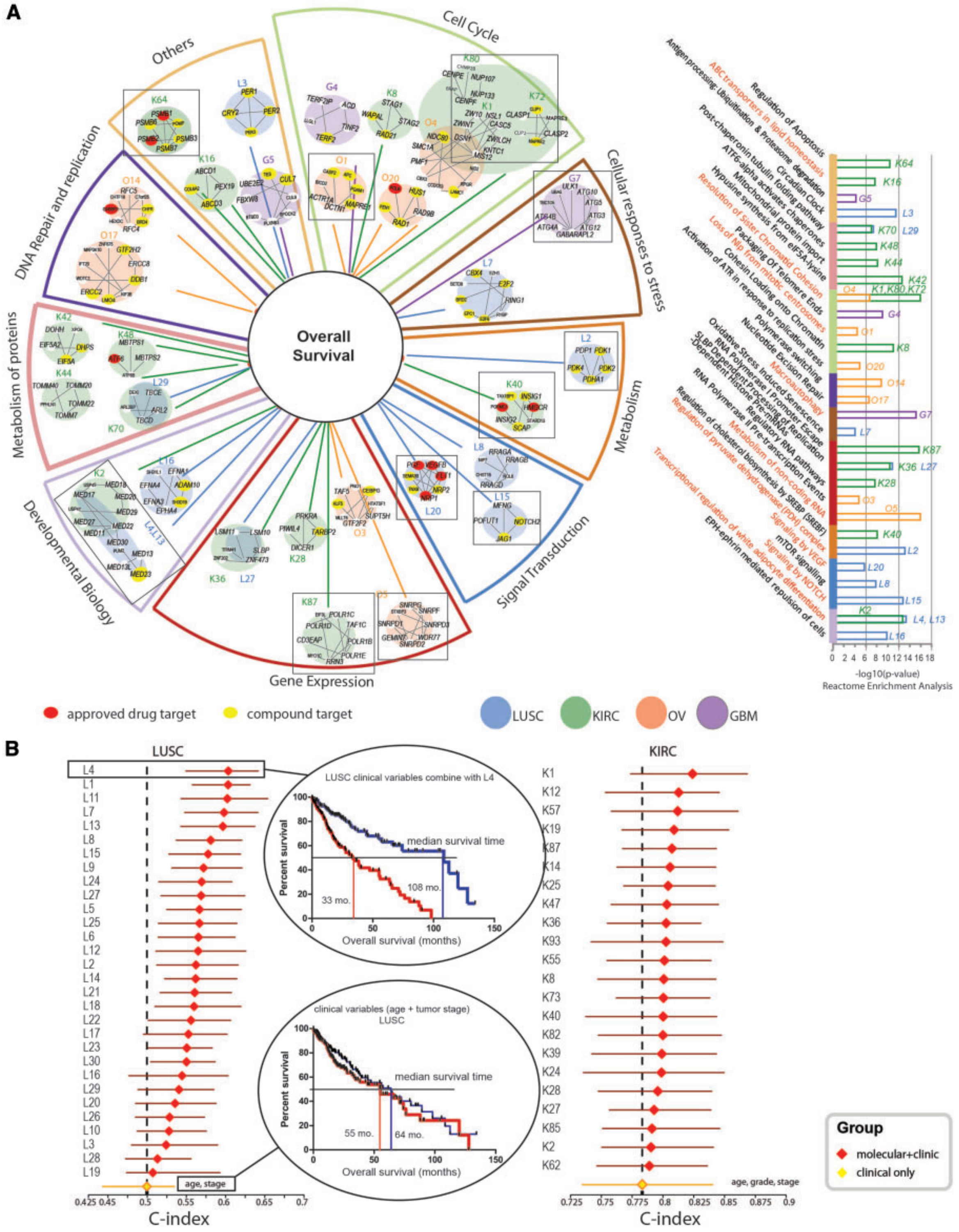

**Figure 4.** Clinical utility of the subnetwork-based prognostic biomarkers. (**A**) The subnetwork biomarkers from four cancer types were grouped into multiple cellular events based on Reactome database annotation, including cell cycle (green), cellular response to stress (brown), metabolism (orange), signal transduction (blue), gene expression (cognac), developmental biology (light purple), metabolism of protein (rouge), DNA repair and replication (purple) and others (yellow), represented by nine pie slices. Inside each pie slice, subnetworks from each cancer type were plotted (Blue: LUSC; Green: KIRC; Orange: OV; Purple: GBM), and the length of branch is reversely correlated to enrichment analysis *P*-value (longer distance, more significant). Genes annotated to each functional category were shown with a larger font size. Genes were also annotated using Drugbank (http://www.drugbank.ca) and the CGI (https://wiki.nci.nih.gov/display/cageneindex) gene-compound database. The genes targeted by FDA-approved drugs or with experimental compounds available in cancer studies were labeled red and yellow, respectively. Unlabeled genes have not yet been clearly targeted. The enriched Reactome pathway and significant *P*-value for each subnetwork were shown in the bar chart on the right panel. The bar height corresponds to the enrichment *P*-value. All subnetworks were highly enriched in Reactome pathways with *P*-values ranging from 10E-4 to 10E-18. (**B**) C-indexes by models trained from clinical variables alone or in combination with each subnetwork biomarker in KIRC (Ntotal = 169) and LUSC (Ntotal = 313). The black dotted line highlights the integrated models of subnetwork molecular data and clinical variables (red) that show better performance than that based on clinical variables alone (yellow). (The whiskers mark the 25th and 75th quartiles, with the median in the center). See also Supplementary Figure S5.

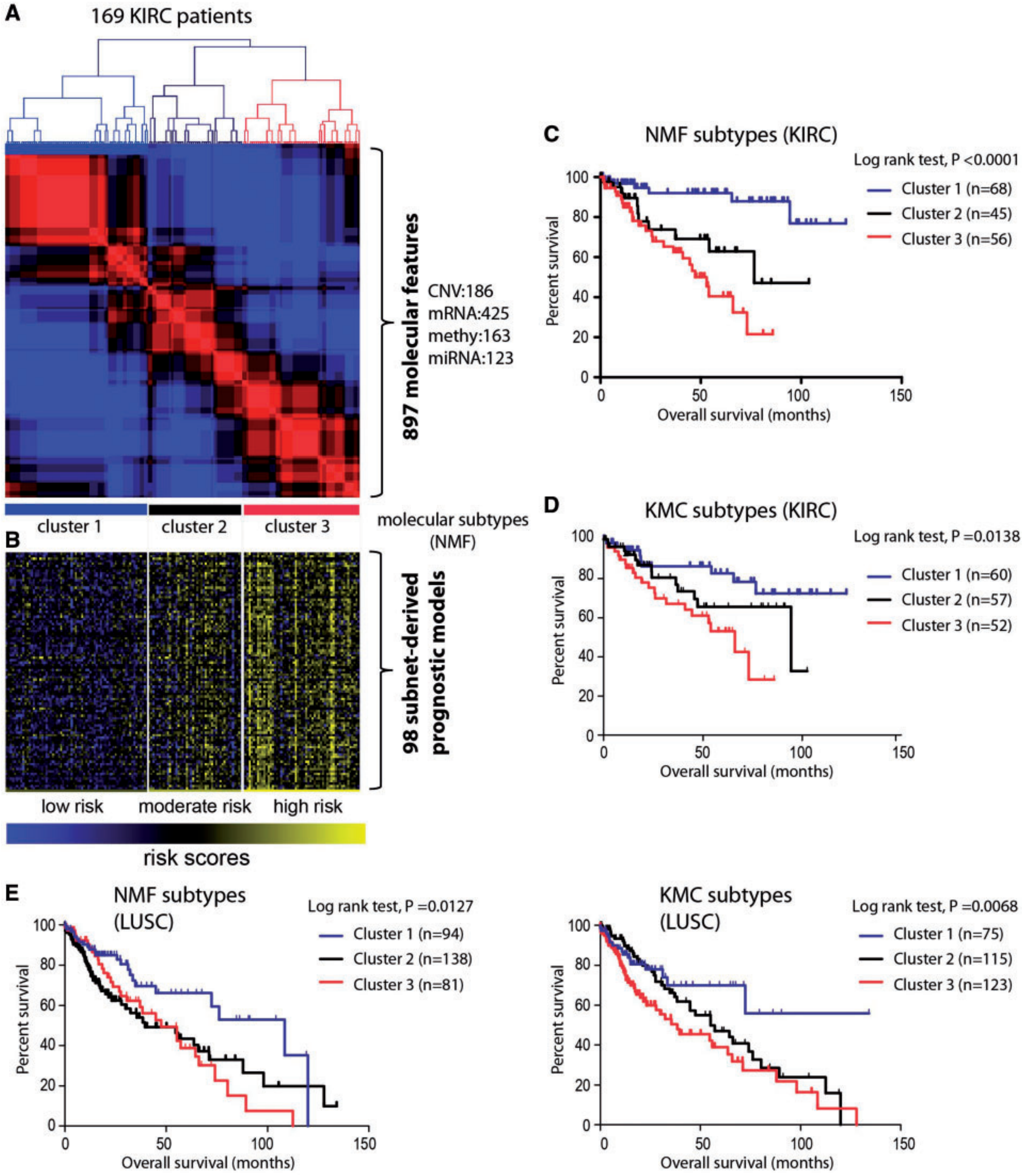

**Figure 5.** Survival-related tumor stratification in KIRC and LUSC. (**A**) Three molecular subtypes (clusters) were revealed by consensus non- NMF clustering of 169 KIRC patients based on 897 molecular features (CNV: 186, methylation: 163, mRNA: 425, miRNA: 123) derived from the 98 subnetwork-based prognostic models reveals three molecular subtypes (clusters). (**B**) The three subtypes of patients were predicted consistently as low-, moderate- and high-risk by most of the subnetwork-derived prognostic models. (**C** and **D**) Kaplan–Meier curves of overall survival for the three clusters of KIRC patients, identified by NMF or KMC. (**E**) Kaplan–Meier curves of overall survival for three clustered LUSC patients identified by NMF or KMC based on 351 molecular features (CNV: 95, methylation: 88, mRNA: 122, miRNA: 46) derived from 30 subnetwork models.

moderate- and high-risk by most of the subnetwork-derived prognostic models and showed distinct survival patterns (Figure 5B and C). Meanwhile, k-means clustering (KMC) algorithm was used to find the molecular subtypes of the KIRC patients, which yielded similar stratification of patients (Figure 5D). Likewise, 351 molecular features (CNV: 95, methylation: 88, mRNA: 122, miRNA: 46) derived from the 30 subnetworks were used to discover molecular subtypes of 313 LUSC patients that showed different survival patterns (Figure 5E).

## Independent validation of the subnetwork-derived prognostic models

According to the work of Tibishirani and Efron [56], our models might fit more to the data set that we used to train and generate the models. We think one way to solve this problem is to validate these models using completely independent data sets. Given the limited availability of suitable independent data providing the highly integrated multi-dimensional genomic data, we evaluated the performance of the prognostic models obtained from LUSC or KIRC using two independent data sets—Lung Adenocarcinoma (LUAD) or Kidney renal papillary cell carcinoma (KIRP) from TCGA, which are two histological distinct lung or kidney cancer, respectively. We got 414 samples from LUAD and 233 samples from KIRP with complete patient survival information and all four types of molecular data that were generated by the same platform as in LUSC or KIRC. We found that 27 of 30 (90%) prognostic models in LUSC were confirmed in LUAD (log-rank *P*-values ranging from 0.0495 to 3.15E-08) and 83 of 98 (85%) prognostic models in KIRC were confirmed in KIRP (log-rank *P*-values ranging from 0.0463 to 1.31E-06) (Figure 6A and B). Furthermore, we did another independent validation using the data set from a recently published International Cancer Genome Consortium (ICGC) ovarian cancer project [57], where we got 107 samples from the Australia OV cancer study with complete patient survival information and all four types of molecular data. We found that 17 of the top 20 models trained from TCGA OV were validated in the ICGC Australia OV data set (log-rank *P*-values ranging from 0.0434 to 1.38E-06) (Figure 6C). In addition, considering the fact that in regular clinical practice physicians may not be able to make all of these measurements for every patient as the TCGA has done, we further analyzed which type of molecular feature was the most informative. We counted the frequency of each type of the four molecular features, which were selected by LASSO + Cox when building the prognostic models, and observed that the mRNA feature got the highest selection probability in all four cancer types (Supplementary Figure S9A). Therefore, we further tested if using mRNA features alone in our subnetwork models could also predict patient survival. We found that 66 models (67%) in KIRC and 20 models (67%) in LUSC could also effectively predict patient survival in KIRP and LUAD, respectively (log-rank *P*-value $< 0.05$, Supplementary Figure S9B), though the combined feature set gave better performance.

## Discussion

In contrast to previous studies driven by a single molecular data type or assumptions that genes act independently, in this study, we focused on the potential impact of multiple genetic and epigenetic (gene expression, CNV, miRNA expression and DNA methylation) changes on the molecular states of networks that in turn affects complex cancer outcome. Here we report our methodology to generate a multi-dimensional subnetwork atlas for cancer prognosis through integrating cancer genomic and interactome data. Through this approach, we uncovered an average of 38 subnetwork-derived prognosis biomarkers in four cancer types. The subnetworks identified are involved in many pathways associated with cancer prognosis and include several promising targets for precision cancer therapy. Interestingly, a number of subnetworks with less characterized roles in cancer stood out, thus providing extra clues to the biological pathways that may contribute to cancer outcome. The integrative analysis not only explores the gene–gene relations but also helps to better understand how multiple regulatory mechanisms are orchestrated together to affect cancer survivorship and narrow down the enormous molecular features into a smaller set of key subnetworks (modules) associated with survival.

Subnetwork signatures provide potential strategies for clinical cancer treatment. By annotating genes with FDA-approved drugs or experimental compounds (Figure 4A and Supplementary Figure S5), we not only demonstrate that our subnetworks contain drug targets but also provides new insights into previously underestimated factors that may be crucial in cancer prognosis, such as *VEGF*-related subnetwork (L20), a well-studied cancer-related signaling pathway that promotes angiogenesis [58], However, anti-*VEGF* monotherapy (mainly anti-*VEGFA*) is not as efficient as conventional chemotherapy, and the survival beneficial effect comes when combining anti-VEGF with cytotoxic agents for patients with particular cancer subtypes [59]. Our prediction reinforces the previously underestimated signaling mediated by *VEFGB* and *PGF* (placental growth factor), whose receptors are *FLT1* (*VEGFR*-1) and neuropilin NRPs (not *VEGFR-2*), with no obvious role in angiogenesis but important in altering cytokine release and immune cell chemotaxis [58]. We also captured a number of subnetworks without any compound available, whose roles in cancer are either intensively (i.e. G7) or barely characterized (i.e. O5, K87). G7 is an autophagy-related subnetwork with paradoxical effects: either as a tumor suppressor or inducer of apoptosis (i.e. *ATG6*/*Beclin1*) or a pro-survival signal to protect tumor from metabolic stress particularly induced by chemo/radio therapy [60]. Here, we predict that high CNV and mRNA of *GABARAPL2* (*ATG8*) and high CNV_*ATG5* are associated with better survival, while high CNV_*ULK1* (*ATG1*) and high methylation at *ATG3*/10/12 are associated with worse prognosis, providing a potential guideline to selectively activate or inhibit individual autophagy-related proteins using small molecules or antibodies. In contrast, O5 and K87 are barely studied. O5 mainly contained small nuclear ribonucleo proteins (SNRPs), components of the spliceosome complex to process pre-mRNA to its mature and functional form [61]. Interestingly, *WDR77* in O5 may function as a final target of SNRPs, which has been shown to stimulate ovarian cancer cell proliferation [62]. Similarly, K87 contained various RNA polymerase I (Pol I) and their associated factors, such as *RRN3* and *CD3EAP*, which can interact with TAF and play important roles in Pol I recruitment [63, 64]. CD3EAP's effect in prognosis has also been linked to NF-$\kappa$B activity in myeloma patients [65]. Moreover, several subnetworks with experimental compounds available are targets currently being investigated in cancer studies, such as K1/80/72 and L15. K1/80/72 are three kinetochore-relate subworks: K80 contains *CENPE*/*CENPF*, which are centromere proteins; K1 contains kinetochore-associated proteins including RZZ (*ROD*, *ZW10*, *ZWILCH*) complex and MIS12 (*MIS12*, *DSN1*, *NSL1*) complex, which are crucial for kinetochore assembly and microtubule interactions; and K72 contains conserved microtubule binding protein *CLIP* and associating protein *CLASP*, which can promote the growth of kinetochore-bound microtubules [49]. Thus, our prediction agrees well with the understanding that kinetochore regulation and higher order chromatin structure play important roles in cancer [22]. L15 is a Notch signaling-related network. Elevated *NOTCH1* and its ligand *JAG1* have been detected and linked to poor prognosis in breast cancer [66] and using $\gamma$-secretase inhibitors to block Notch1 signaling may sensitize colon cancer to chemotherapy [67]. In conclusion, our subnetwork signatures provide possibilities to identify drug targets and guidelines of how to modulate functions of the identified subnetworks.

The multi-dimensional subnetwork biomarkers could advance the predictive power of cancer prognosis. Among the four cancer types, using the clinical-variables alone showed substantial predictive power for three cancers (OV, KIRC, GBM), with

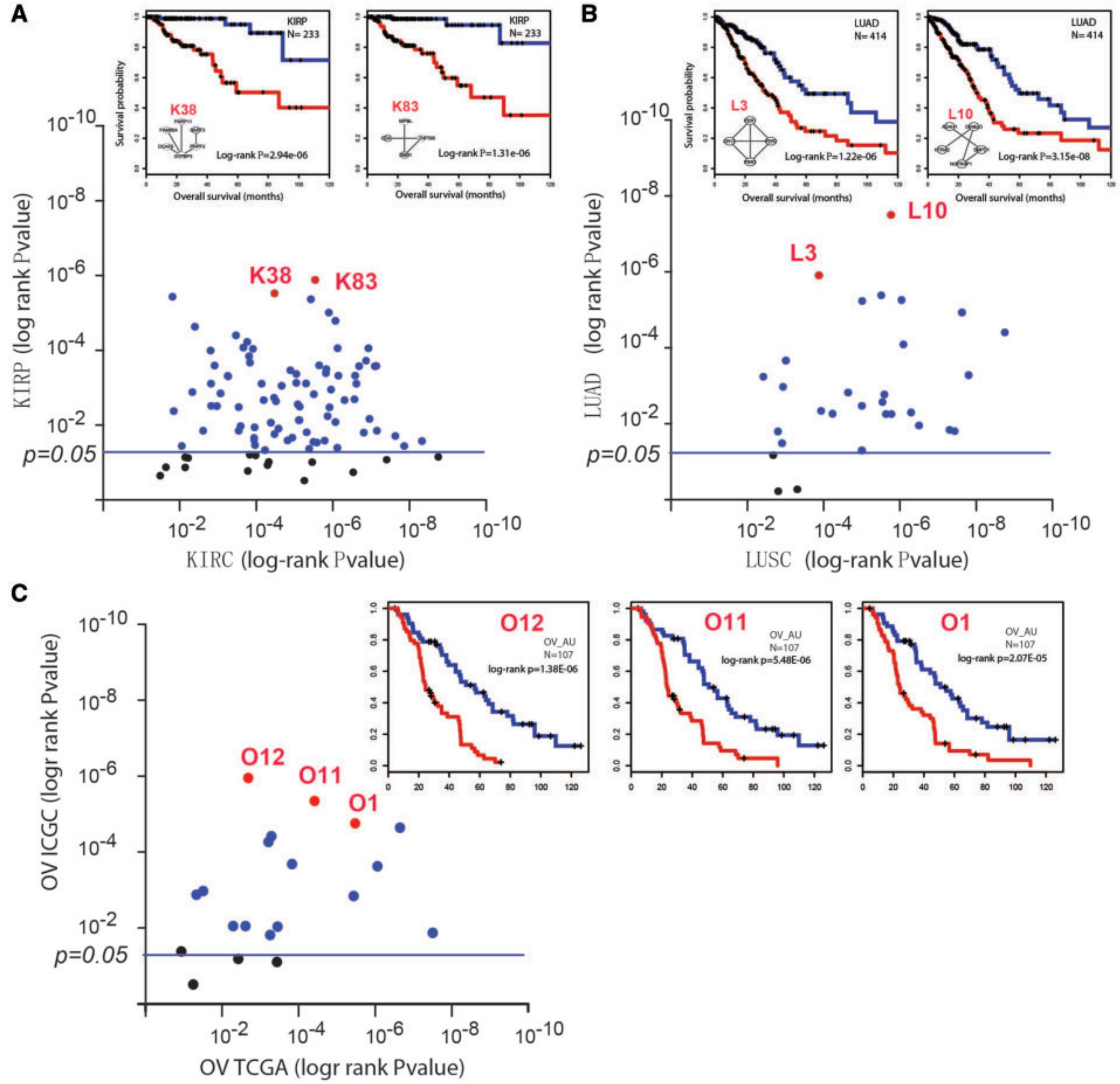

**Figure 6.** Independent validation of the prognostic models obtained from LUSC, KIRC or OV. (**A**) 83 of 98 (85%) subnetwork-derived prognostic models in KIRC were also significant predictors of patient survival in KIRP based on log-rank test (blue points). Kaplan–Meier survival plots for the top two prognostic models for 233 KIRP patients were shown. (**B**) 27 of 30 (90%) subnetwork-derived prognostic models in LUSC were also significant predictors of patient survival in LUAD (blue points). Kaplan–Meier survival plots for the top two prognostic models for 414 LUAD patients were shown. (**C**) 17 of the top 20 (85%) prognostic models trained from TCGA OV were also validated using ICGC Australia OV cancer data set (blue points). Kaplan–Meier survival plots for the top three prognostic models for 107 ICGC Australia OV patients were shown.

median C-indexes significantly >0.5 (0.629–0.783), comparable with a previously published result [6], despite that the samples we used are slightly different (<30% different samples were used). However, in combination with our subnetwork biomarkers, the quantitative gains were limited in KIRC (Somers' D, 2.1–14.4% for 22 subnetwork-based models in KIRC; a 2.1% gain in Somers' D corresponds to a 2.1% increase of rank correlation coefficient between predicted risk score and observed survival time), suggesting largely redundant information between clinical variables and molecular data in terms of patient survival stratification as discussed in previous study [6]. In contrast, though using clinical variables alone yielded poor prediction with a C-index of only 0.4917 for our 313 core LUSC samples, addition of molecular data provided crucial complementary information and significant gains in predictive power (Somers' D, 36.4–132%). This indicates that the predictive power of the clinical variables may partly depend on the data quality, such as patient population and tumor type. Therefore, clinical variables and multiple types of molecular data may provide crucial complementary information to achieve more robust predictive power when building prognostic models (Figure 4B).

Meanwhile, we would like to point out the importance of the data quality and the choice of parameter cutoff. First, the incompleteness of interactome data and some other protein interaction types (i.e. genetic) or the predicted miRNA–mRNA interactions not used in this study might limit our findings of subnetwork signals. Second, in the procedure of genome-wide screening of all potential molecular features that may correlate with patient survival, we used a less stringent cutoff $P < 0.05$ (likelihood ratio test) to keep more information. When testing a more stringent cutoff $P < 0.01$, a multiple comparison adjustment $P$-value cutoff (fdr $<$0.05) or an additional combined $P$-value based on Fisher's method (combined $P < 0.05$) in KIRC and LUSC, we found that most robust networks continued to be identified (Supplementary Figure S7). On the other hand, when using a more stringent cutoff, some survival-related subnetworks were lost, such as K17 (CST, KLRK1, MICA, MICB) related to natural killer cell-mediated cytotoxicity pathway (using $P < 0.01$) and L15 (JAG1, MFNG, NOTCH2, POFUT1) related to NOTCH signaling pathway (using fdr $< 0.05$).

Additionally, it should be noted that though our study provides important insights of translating molecular data into clinical utility, it still has some limitations that could provide guidance for future work. First, current gene scoring method treated the molecular features of individual gene as independent and had a bias toward well-studied genes (genes with more molecular features tested) (Supplementary Figure S3). Second, though LASSO was good for selecting most important features to overcome the over-fitting problem and make a fair evaluation to different subnetworks (with different number of genes), it would lose some equally important features when high pairwise correlations occurred. Though our subnetworks were exacted from physical PPI network and overall weak correlations were observed among the majority of interacting genes (Supplementary Figure S8), the potential intrinsic relations among biologically relevant genes will lead to some level of correlations. In future study, more effective gene scoring method and feature selection strategies should be applied, such as elastic net [68], which combines penalty terms of LASSO and Ridge, to compromise between variable selection and group effect.

In conclusion, here we reported our three-step approach to build a multi-dimensional subnetwork atlas or landmark for cancer prognosis by integrating cancer genomics and interactome data, represented as PPI modules perturbed by multiple genetic and epigenetic events that correspond to patient survival. Besides narrowing down the enormous genome-wide molecular features into a smaller set of key subnetworks associated with survival, we also provided the ranking of the most important molecular features and developed a freely accessible web-based resource of our results to allow researchers in basic science and translational medicine to use the prognostic models directly to uncover specific genes or markers of interest, and avoid the time-consuming genome-wide screening. Therefore, our study provides a new analytical tool to systematically dissect the comprehensive infrastructures that guide patient outcomes and a new paradigm of how genomic data can be used to better inform clinicians for advancing cancer care management.

### Key Points

- We proposed an integrative systems biology approach to build an atlas or landmark for cancer prognosis.
- We identified the subnetworks within the comprehensive PPI network perturbed (or affected) by multiple genetic and epigenetic events associated with survival and further generated prognostic models from these multi-dimensional subnetworks.
- The clinical utility of the multi-dimensional subnetwork atlas was evaluated by prognostic impact evaluation, functional enrichment analysis, drug target annotation, tumor stratification and independent validation.

## Supplementary Data

**Supplementary data are available online at** https://academic.oup.com/bib.

## Acknowledgement

We thank Anne Schaefer, Patrizia Casaccia, Shuang Wang and Tong Shen for their review of the manuscript.

## Funding

This research was supported by Senior Scholar Award in Aging from the Ellison Medical Foundation [AG-SS-2482-10 to M.J.W.]; Awards from the National Institutes of Health [HL103967 and CA154809 to M.J.W.]; and National Natural Science Foundation of China [31200998 to F.Z.].

## References

1. Cancer Genome Atlas Research N. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013;**499**:43–9.
2. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**:1061–8.
3. Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012;**489**:519–25.
4. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. *Nature* 2011;**474**:609–15.
5. Liu Z, Zhang S. Toward a systematic understanding of cancers: a survey of the pan-cancer study. *Front Genet* 2014;**5**:194.
6. Yuan Y, Van Allen EM, Omberg L, *et al*. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 2014;**32**:644–52.
7. Xu C, Liu Y, Wang P, *et al*. Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Mol Cancer* 2010;**9**:143.
8. Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res* 2011;**71**:4550–61.
9. Kim S, Park T, Kon M. Cancer survival classification using integrated data sets and intermediate information. *Artif Intell Med* 2014;**62**:23–31.
10. Kim H, Huang W, Jiang X, *et al*. Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. *Proc Natl Acad Sci USA* 2010;**107**:2183–8.
11. Zhao Q, Shi X, Xie Y, *et al*. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief Bioinform* 2015;**16**:291–303.
12. Juergens RA, Wrangle J, Vendetti FP, *et al*. Combination epigenetic therapy has efficacy in patients with refractory advanced non-small cell lung cancer. *Cancer Discov* 2011;**1**:598–607.
13. Zhang W, Liu Y, Sun N, *et al*. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. *Cell Rep* 2013;**4**:542–53.

14. Das J, Gayvert KM, Bunea F, *et al*. ENCAPP: elastic-net-based prognosis prediction and biomarker discovery for human cancers. *BMC Genomics* 2015;**16**:263.
15. Varn FS, Ung MH, Lou SK, *et al*. Integrative analysis of survival-associated gene sets in breast cancer. *BMC Med Genomics* 2015;**8**:11.
16. Joung JG, Kim D, Lee SY, *et al*. Integrated analysis of microRNA-target interactions with clinical outcomes for cancers. *BMC Med Genomics* 2014;**7(Suppl 1)**:S10.
17. Crijns AP, Fehrmann RS, de Jong S, *et al*. Survival-related profile, pathways, and transcription factors in ovarian cancer. *PLoS Med* 2009;**6**:e24.
18. Patel VN, Gokulrangan G, Chowdhury SA, *et al*. Network signatures of survival in glioblastoma multiforme. *PLoS Comput Biol* 2013;**9**:e1003237.
19. Peri S, Navarro JD, Amanchy R, *et al*. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* 2003;**13**:2363–71.
20. Betel D, Wilson M, Gabow A, *et al*. The microRNA.org resource: targets and expression. *Nucleic Acids Res* 2008;**36**:D149–53.
21. Vergoulis T, Vlachos IS, Alexiou P, *et al*. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res* 2012;**40**:D222–9.
22. Leiserson MD, Vandin F, Wu HT, *et al*. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics* 2015;**47**:106–14.
23. Vandin F, Clay P, Upfal E, *et al*. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac Symp Biocomput* 2012;55–66.
24. Tibshirani RJ. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B* 1996;**58**:267–88.
25. Chen EY, Tan CM, Kou Y, *et al*. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 2013;**14**:128.
26. Dienstmann R, Jang IS, Bot B, *et al*. Database of genomic biomarkers for cancer drugs and clinical targetability in solid tumors. *Cancer Discov* 2015;**5**:118–23.
27. Dietlein F, Reinhardt HC. Molecular pathways: exploiting tumor-specific molecular defects in DNA repair pathways for precision cancer therapy. *Clin Cancer Res* 2014;**20**:5882–7.
28. Rodon J, Dienstmann R, Serra V, *et al*. Development of PI3K inhibitors: lessons learned from early clinical trials. *Nat Rev Clin Oncol* 2013;**10**:143–53.
29. Appelqvist H, Waster P, Kagedal K, *et al*. The lysosome: from waste bag to potential therapeutic target. *J Mol Cell Biol* 2013;**5**:214–26.
30. Stehn JR, Haass NK, Bonello T, *et al*. A novel class of anticancer compounds targets the actin cytoskeleton in tumor cells. *Cancer Res* 2013;**73**:5169–82.
31. Groth-Pedersen L, Aits S, Corcelle-Termeau E, *et al*. Identification of cytoskeleton-associated proteins essential for lysosomal stability and survival of human cancer cells. *PLoS One* 2012;**7**:e45381.
32. Ellis LM, Hicklin DJ. VEGF-targeted therapy: mechanisms of anti-tumour activity. *Nat Rev Cancer* 2008;**8**:579–91.
33. Karanika S, Karantanos T, Li L, *et al*. DNA damage response and prostate cancer: defects, regulation and therapeutic implications. *Oncogene* 2015;**34**:2815–22.
34. Schroer TA. Dynactin. *Annu Rev Cell Dev Biol* 2004;**20**:759–79.
35. Green RA, Wollman R, Kaplan KB. APC and EB1 function together in mitosis to regulate spindle dynamics and chromosome alignment. *Mol Biol Cell* 2005;**16**:4609–22.
36. Berrueta L, Tirnauer JS, Schuyler SC, *et al*. The APC-associated protein EB1 associates with components of the dynactin complex and cytoplasmic dynein intermediate chain. *Curr Biol* 1999;**9**:425–8.
37. Askham JM, Vaughan KT, Goodson HV, *et al*. Evidence that an interaction between EB1 and p150(Glued) is required for the formation and maintenance of a radial microtubule array anchored at the centrosome. *Mol Biol Cell* 2002;**13**:3627–45.
38. Nakamura M, Zhou XZ, Lu KP. Critical role for the EB1 and APC interaction in the regulation of microtubule polymerization. *Curr Biol* 2001;**11**:1062–7.
39. Wang YH, Zhou XB, Zhu HX, *et al*. Overexpression of EB1 in human esophageal squamous cell carcinoma (ESCC) may promote cellular growth by activating beta-catenin/TCF pathway. *Oncogene* 2005;**24**:6637–45.
40. Liu M, Yang SB, Wang YH, *et al*. EB1 acts as an oncogene via activating beta-Catenin/TCF pathway to promote cellular growth and inhibit apoptosis. *Mol Carcinog* 2009;**48**:212–19.
41. Hitosugi T, Zhou L, Elf S, *et al*. Phosphoglycerate mutase 1 coordinates glycolysis and biosynthesis to promote tumor growth. *Cancer Cell* 2012;**22**:585–600.
42. Kroemer G, Pouyssegur J. Tumor cell metabolism: cancer's Achilles' heel. *Cancer Cell* 2008;**13**:472–82.
43. Wu L, Hui H, Wang LJ, *et al*. MicroRNA-326 functions as a tumor suppressor in colorectal cancer by targeting the nin one binding protein. *Oncol Rep* 2015;**33**:2309–18.
44. Li Y, Gao Y, Xu Y, *et al*. Down-regulation of miR-326 is associated with poor prognosis and promotes growth and metastasis by targeting FSCN1 in gastric cancer. *Growth Factors* 2015;**33**:267–74.
45. Liang ZX, Wu H, Xia J, *et al*. Involvement of miR-326 in chemotherapy resistance of breast cancer through modulating expression of multidrug resistance-associated protein 1. *Biochem Pharmacol* 2010;**79**:817–24.
46. Liu Z, Chen X, Wang Y, *et al*. PDK4 protein promotes tumorigenesis through activation of cAMP-response element-binding protein (CREB)-Ras homolog enriched in brain (RHEB)-mTORC1 signaling cascade. *J Biol Chem* 2014;**289**:29739–49.
47. Bonnet S, Archer SL, Allalunis-Turner J, *et al*. A mitochondria-K+ channel axis is suppressed in cancer and its normalization promotes apoptosis and inhibits cancer growth. *Cancer Cell* 2007;**11**:37–51.
48. Sugden MC, Holness MJ. Recent advances in mechanisms regulating glucose oxidation at the level of the pyruvate dehydrogenase complex by PDKs. *Am J Physiol Endocrinol Metab* 2003;**284**:E855–62.
49. Cheeseman IM, Desai A. Molecular architecture of the kinetochore-microtubule interface. *Nat Rev Mol Cell Biol* 2008;**9**:33–46.
50. Chan GKT, Jablonski SA, Starr DA, *et al*. Human Zw10 and ROD ave mitotic checkpoint proteins that bind to kinetochores. *Nat Cell Biol* 2000;**2**:944–7.
51. Obuse C, Iwasaki O, Kiyomitsu T, *et al*. A conserved Mis12 centromere complex is linked to heterochromatic HP1 and outer kinetochore protein Zwint-1. *Nat Cell Biol* 2004;**6**:1135. U1137.
52. Williams BC, Li ZX, Liu ST, *et al*. Zwilch, a new component of the ZW10/ROD complex required for kinetochore functions. *Mol Biol Cell* 2003;**14**:1379–91.
53. Kasuboski JM, Bader JR, Vaughan PS, *et al*. Zwint-1 is a novel Aurora B substrate required for the assembly of a dynein-binding platform on kinetochores. *Mol Biol Cell* 2011;**22**:3318–30.

54. Caldas GV, DeLuca KF, DeLuca JG. KNL1 facilitates phosphorylation of outer kinetochore proteins by promoting Aurora B kinase activity. *J Cell Biol* 2013;**203**:957–69.
55. Kops GJ, Weaver BA, Cleveland DW. On the road to cancer: aneuploidy and the mitotic checkpoint. *Nat Rev Cancer* 2005;**5**:773–85.
56. Tibshirani RJ, Efron B. Pre-validation and inference in microarrays. *Stat Appl Genet Mol Biol* 2002;**1**: Article1.
57. Patch AM, Christie EL, Etemadmoghadam D, *et al*. Corrigendum: whole-genome characterization of chemoresistant ovarian cancer. *Nature* 2015;**527**:398.
58. Ferrara N, Gerber HP, LeCouter J. The biology of VEGF and its receptors. *Nat Med* 2003;**9**:669–76.
59. Jain RK, Duda DG, Clark JW, *et al*. Lessons from phase III clinical trials on anti-VEGF therapy for cancer. *Nat Clin Pract Oncol* 2006;**3**:24–40.
60. Kondo Y, Kanzawa T, Sawaya R, *et al*. The role of autophagy in cancer development and response to therapy. *Nat Rev Cancer* 2005;**5**:726–34.
61. Maniatis T, Reed R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* 1987;**325**:673–8.
62. Ligr M, Patwa RR, Daniels G, *et al*. Expression and function of androgen receptor coactivator p44/Mep50/WDR77 in ovarian cancer. *Plos One* 2011;**6**:e26250.
63. Miller G, Panov KI, Friedrich JK, *et al*. hRRN3 is essential in the SL1-mediated recruitment of RNA polymerase I to rRNA gene promoters. *Embo J* 2001;**20**:1373–82.
64. Yamamoto K, Yamamoto M, Hanada K, *et al*. Multiple protein-protein interactions by RNA polymerase I-associated factor PAF49 and role of PAF49 in rRNA transcription. *Mol Cell Biol* 2004;**24**:6338–49.
65. Vangsted AJ, Klausen TW, Gimsing P, *et al*. The importance of a sub-region on chromosome 19q13.3 for prognosis of multiple myeloma patients after high-dose treatment and stem cell support: a linkage disequilibrium mapping in RAI and CD3EAP. *Ann Hematol* 2011;**90**:675–84.
66. Reedijk M, Odorcic S, Chang L, *et al*. High-level coexpression of JAG1 and NOTCH1 is observed in human breast cancer and is associated with poor overall survival. *Cancer Res* 2005;**65**:8530–7.
67. Meng RD, Shelton CC, Li YM, *et al*. gamma-secretase inhibitors abrogate oxaliplatin-induced activation of the Notch-1 signaling pathway in colon cancer cells resulting in enhanced chemosensitivity. *Cancer Res* 2009;**69**:573–82.
68. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 2005;**67**:301–20.