BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 12 2013, pages 1553–1561
doi:10.1093/bioinformatics/btt191

Systems biology

Advance Access publication April 23, 2013

Systematic tracking of dysregulated modules identifies novel genes in cancer

Sriganesh Srihari and Mark A. Ragan*

Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072 Australia

Associate Editor: Gunnar Ratsch

ABSTRACT

Motivation: Deciphering the modus operandi of dysregulated cellular mechanisms in cancer is critical to implicate novel cancer genes and develop effective anti-cancer therapies. Fundamental to this is meticulous tracking of the behavior of core modules, including complexes and pathways across specific conditions in cancer.

Results: Here, we performed a straightforward yet systematic identification and comparison of modules across pancreatic normal and cancer tissue conditions by integrating PPI, gene-expression and mutation data. Our analysis revealed interesting *change-patterns* in gene composition and expression correlation particularly affecting modules responsible for genome stability. Although in most cases these changes indicated impairment of essential functions (e.g. of DNA damage repair), in several other cases we noticed strengthening of modules possibly abetting cancer. Some of these *compensatory* modules showed *switches* in transcription regulation and recruitment of tumor inducers (e.g. SOX2 through overexpression). In-depth analysis revealed novel genes in pancreatic cancer, which showed susceptibility to copy-number alterations (e.g. for USP15 in 17 of 67 cases), supported by literature evidence for their involvement in other tumors (e.g. USP15 in glioblastoma). Two of the identified genes, YWHAE and DISC1, further supported the nexus between neural genes and pancreatic carcinogenesis. Extension of this assessment to BRCA1 and BRCA2 breast tumors showed specific differences even across the two sub-types and revealed novel genes involved therein (e.g. TRIM5 and NCOA6).

Availability: Our software CONTOURv1 is available at: http://bioinformatics.org.au/tools-data/.

Contact: m.ragan@uq.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on November 2, 2012; revised on April 17, 2013; accepted on April 18, 2013

1 INTRODUCTION

Cancer is the outcome of an intricate interplay of dysregulated mechanisms that are otherwise responsible for maintaining the genomic integrity of the cell. Although our current knowledge of these mechanisms is inadequate to fully understand cancer, immense efforts are underway to identify novel 'cancer genes' (oncogenes) and implicate known ones for novel roles in cancer—for example, the recent implication of SOX2 as a frequently amplified gene resulting in fusion and amplification of transcription factor MYCL1 involved in small-cell lung cancer (Rudin *et al*., 2012).

Methods for computational identification of disease genes look mainly for genes that are differentially expressed, have similar expression profiles with known disease genes, are 'central' or 'reachable' in disease molecular networks or have disease associations in literature (reviewed in Doncheva *et al*., 2012). Some examples include ENDEAVOUR (Tranchevent *et al*., 2008) and GeneRanker (Gonzalez *et al*., 2008).

A crucial distinguishing factor of cancer genes is that they belong to core mechanisms responsible for genome stability and cell proliferation (e.g. DNA damage repair and cell cycle) and function as highly synergetic or coordinated groups. Therefore, critical to implicating novel genes is the identification of core *modules* including pathways and complexes dysregulated in cancer. For example, Lage *et al*. (2007) identified disease complexes and used them in a Bayesian predictor to rank genes involved in epithelial ovarian cancer. Kim *et al*. (2009) traced back paths through the human protein interaction (PPI) network from differentially expressed genes (target) to genes harboring mutations (causal) and successfully applied this approach to identify disrupted pathways in glioblastoma multiforma. On the other hand, Liu *et al*. (2012a) used an interaction enrichment analysis to identify pairs of genes whose relationships differed between normal and cancer. Through their Gene Interaction Enrichment and Network Analysis (GENIA), the authors categorized interaction profiles as cooperative (expressions correlated), competitive (expressions anti-correlated), redundant (suppression of both causes dysfunction) and dependent (expression of one is dependent on the other), and then mined these profiles in breast and pancreatic cancers to identify dysregulated pathways. Liu *et al*. (2012b) incorporated gene expression into annotated pathways to quantify 'pathway activity' patterns in cancer. Subsequently, the authors extracted these patterns using a feature-selection model on a PPI network to construct a 'pathway interaction network'. Sub-networks in this pathway network represented distinct pathways and their cross-talk in cancer.

Apart from identifying modules, it is necessary to understand the *modus operandi* or the underlying 'program' driving these disruptions, and this requires systematic *tracking* of the behavior of these modules under cancer conditions. For example, Zhang *et al*. (2012) mined tightly connected gene co-expression sub-networks across >30 cancer networks (cell lines) and tracked aberrant modules as frequent sub-networks appearing across these cancers. However, studying multiple cancers simultaneously makes it challenging to discern clearly the intricate underlying

*To whom correspondence should be addressed.

mechanisms; different genes are involved in different cancers and even different cancer sub-types, and their roles across these cancers are also different. What is required, therefore, is a systematic method to track gene and module behavior across specific conditions in a controlled manner (e.g. between normal and a cancer type or between specific cancer sub-types).

In addition, it is important to effectively integrate 'multi-omics' data into such an analysis—for example, from The Cancer Genome Atlas (TCGA): http://tcga-data.nci.nih.gov/. Multi-omics data integration as such has attracted attention during the past few years—for example, Chu and Chen (2008) combined PPI and gene expression data to construct a cancer-perturbed PPI network in cervical carcinoma to study gain- and loss-of-function genes as potential drug targets. Masica *et al.* (2011) correlated somatic mutations and gene expression to identify novel genes in glioblastoma multiforma (e.g. SYNE1, KLF6, FGFR4 and EPHB4). Zhang *et al.* (2012) integrated DNA methylation, gene expression and microRNA expression data in 385 ovarian cancer samples from TCGA and performed 'multi-dimensional' analysis to identify disrupted pathways. Magger *et al.* (2012) combined PPI and gene expression data to construct tissue-specific PPI networks for 60 tissues and used them to prioritize disease genes. Finally, Zhao *et al.* (2012) proposed an iterative model to combine mutation and expression data and used it to identify mutated driver pathways in multiple cancer types.

Putting all these findings together, we note (i) it is crucial to study the behavior of modules across specific conditions in a controlled manner to understand the *modus operandi* of cancer mechanisms and to implicate novel genes; and (ii) although most existing methods concentrate on 'mountain' genes that show distinct aberrant behavior in cancers, there are many more 'hills' that often do not display such drastic changes (Wood *et al.*, 2007). These 'hills' are *contours*, and they may not be identifiable through their own behavior, but their changes are quantifiable when considered in conjunction with other genes (e.g. as modules); these genes may not be differentially expressed, but they are differentially *co-expressed* with other genes. These points are substantiated further in the following analyses.

1.1 An initial analysis

We performed an integrated analysis combining 29 600 interactions among 5824 proteins from Biogrid (Stark *et al.*, 2011) and 39 paired normal and tumor gene expression samples from pancreatic ductal adenocarcinoma (PDAC) patients (Badea *et al.*, 2008) to study the behavior of genes and their modules in the tumor *vis-a-vis* normal (more details later).

We computed the gene expression correlationwise distribution of physically interacting gene (protein) pairs for normal and tumor conditions (co-expression is measured as Pearson correlation; we use the terms *genes* and *proteins* interchangeably) (Fig. 1a). We noticed a significant reduction in correlation of gene pairs in tumor *vis-a-vis* normal—a reduction in 8701 highly correlated interactions (of absolute correlation ≥ 0.50). This comprised a possible loss of positively correlated 'accelerators' (interactions driving normal cellular processes) and negatively correlated 'brakes' (interactions suppressing tumor inducers and genome instability). Interestingly, the analysis of 'jumps' (increase or decrease) in correlation revealed two interactions, RBPMS–RHOXF2 and SMN1–TMSB4X, displaying extreme jumps [from $\pm(0.9,1)$ to $\pm(0.9,1)$] (Supplementary Files). Among these, RHOXF2, with no noticeable change in expression level (mean of 4.67 and 4.34, respectively), has been recently implicated as a cancer promoter (Shibata-Minoshima *et al.*, 2012).

We next extracted 529 modules of interacting genes from the PPI network, and we computed a correlationwise distribution of these modules for normal and tumor (Fig. 1b). We noticed a reduction in correlation for 79 modules (of correlation ≥ 0.4), two of which included the cancer-promoter SOX2 contributing to the reduction.

Taking these findings into account, we hypothesize that modules displaying differential behavior constitute cancer modules, and they harbor cancer genes. We devise a systematic method to identify these modules and genes by tracking their behavior across specific conditions. We apply our method to two case studies—normal versus PDAC and BRCA1 versus BRCA2 breast tumors. We call our method CONTOUR [Cancer (Onco) geNes from disrupTed mOdUles and their Relationships].

2 METHODS

Using the human PPI network as a backbone, we infer two *tissue condition-specific* PPI networks, one for *normal* and one for *tumor*, by incorporating expression and mutation profiles of genes in the two conditions

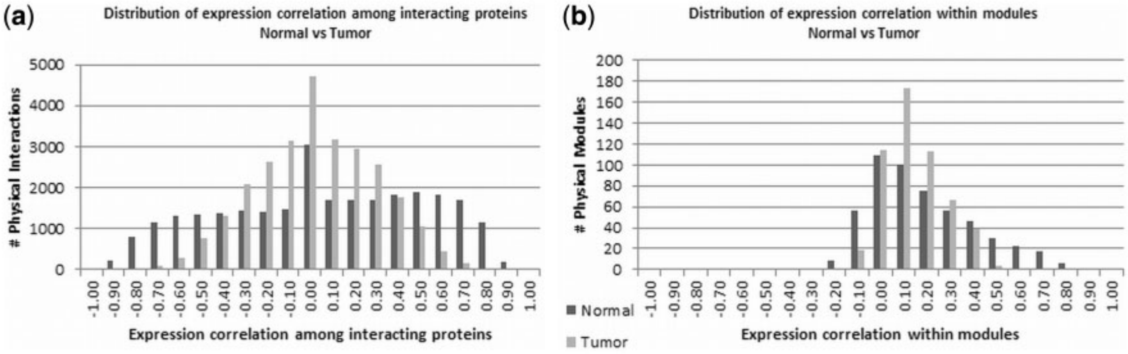

Fig. 1. An initial analysis: expression correlationwise distribution of (**a**) physical interactions and (**b**) modules in normal versus PDAC tumor

(Fig. 2). Next, adopting a maximal clique-merging approach, we extract distinct *modules* from the two conditional networks. We then match normal and tumor modules to identify key ones displaying changes in gene composition or co-expression, and we isolate genes involved therein. We evaluate these genes for potential roles in cancer.

2.1 Inferring normal and tumor PPI networks

Using human protein interactions, we first construct the *generic* PPI network $H = (V_H, E_H)$, where V_H is the set of proteins and E_H is the set of interactions. Every interaction $e = (p, q) \in E_H$ has a weight $r(p, q)$ (between 0 and 1) reflecting the reliability of the interaction.

2.1.1 Expression profiles By integrating gene expression and mutation profiles, we construct two *tissue condition-specific* PPI networks from H—the normal H_N and tumor H_T networks as follows. Let E be the set of all possible gene pairs. First, consider the normal (N) condition. We compute the *correlationwise frequency distribution* ρ_{E_H} for the interacting gene pairs $E_H \subseteq E$: we construct bins of size 0.10 in the range [0,1] by including all gene pairs $(p, q) \in E_H$ into bin $\rho^{(x)}_{E_H}$ if $x \le |PCC((p, q), N)| < x + 0.10$, where $|PCC((p, q), N)|$ is the absolute Pearson correlation of (p,q) under the normal condition (the last bin also includes pairs with correlation 1). Similarly, we calculate a second distribution $\rho_{E\setminus E_H}$, for the non-interacting gene pairs, $E\setminus E_H$. Now, given these two distributions, the probability that a pair (p,q) with correlation in $[x, x + 0.1)$ coinciding with an interaction is,

$$P_\rho(p, q) = Pr[(p, q) \in E_H|\rho] = \frac{|\rho^{(x)}_{E_H}|}{|\rho^{(x)}_{E_H}| + |\rho^{(x)}_{E\setminus E_H}|}, \tag{1}$$

where $|\rho^{(x)}_{E_H}|$ and $|\rho^{(x)}_{E\setminus E_H}|$ are the frequencies of the bins containing (p,q) in the two distributions, respectively (for derivation, see Supplementary Files).

2.1.2 Mutation profiles For every gene, $p \in V_H$, let $l(p)$ (between 0 and 1) represent the likelihood of p to be involved in tumorigenesis by undergoing mutations. We expect this likelihood to be related to the interactivity of p. So, we infer the *mutation coefficient* for every pair $(p, q) \in E$ as $MCC(p, q) = \max\{l(p), l(q)\}$. As aforementioned, we compute the *mutation coefficientwise frequency distributions* ψ_{E_H} and $\psi_{E\setminus E_H}$ for the gene pairs in E_H and $E\setminus E_H$, respectively. We infer the probability that any pair (p,q) with mutation coefficient in $[x, x + 0.1)$ coinciding with an interaction as,

$$P_\psi(p, q) = Pr[(p, q) \in E_H|\psi] = \frac{|\psi^{(x)}_{E_H}|}{|\psi^{(x)}_{E_H}| + |\psi^{(x)}_{E\setminus E_H}|}, \tag{2}$$

where $|\psi^{(x)}_{E_H}|$ and $|\psi^{(x)}_{E\setminus E_H}|$ are the frequencies of the bins containing (p,q) in the two distributions, respectively.

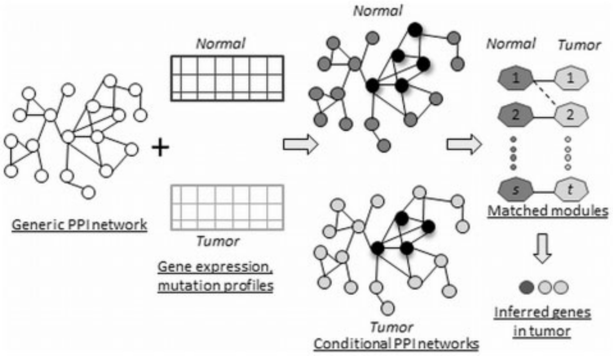

Fig. 2. The CONTOUR workflow

We then infer the weight $r_N(p, q)$ in H_N by scaling $r(p, q)$ with a combined naïve Bayesian probability (see Supplementary Files for derivation) as,

$$r_N(p, q) = r(p, q)* \frac{P_\rho(p, q) * P_\psi(p, q)}{P_\rho(p, q) * P_\psi(p, q) + (1 - P_\rho(p, q)) * (1 - P_\psi(p, q))}. \tag{3}$$

The weights in the tumor PPI network H_T are inferred similarly.

We reflect the expression and mutation 'landscapes' of genes onto their interactions in the two conditions. We capture these landscapes through the expressionwise and mutationwise frequency distributions of gene pairs. So, by re-weighting the interactions in the generic PPI network based on these distributions, we infer the two conditional PPI networks. We do not change the topology of the network, only re-weight the interactions. Note here that (i) we assume independence between expression and mutation profiles of genes, although this is not necessarily the case (see Masica *et al.*, 2011); and (ii) in the normal condition, mutation profiles may not be available; therefore, setting $MCC(p, q) = 0$, Equation (2) evaluates to a constant $|E_H|/|E|$ for every $(p, q) \in E_H$. But, in general, this formulation covers conditions where mutation profiles are available.

2.2 Identifying modules from the PPI networks

Our module-identification algorithm is based on clique-merging, similar to the ones previously proposed for identifying complexes from PPI networks (Liu *et al.*, 2009; Srihari and Leong, 2013). Our algorithm works in two steps: in the first step, it finds all maximal cliques from the PPI network and ranks them in non-increasing order of their weighted interaction densities, and in the second step, it merges highly overlapping cliques to build modules.

We identify the set $\mathcal{C}$ of all *maximal cliques* of size at least k in the PPI network using a fast depth-first search with pruning-based algorithm (CLIQUES) by Tomita *et al.* (2006). Next, for every clique $C \in \mathcal{C}$, we calculate its *weighted interaction density* $d_w(C)$ as,

$$d_w(C) = \frac{\sum_{(p,q)\in C} r(p, q)}{\binom{|C|}{2}} \tag{4}$$

We rank these cliques in non-increasing order of their weighted densities, $\{C_1, C_2, \ldots, C_m\}$, and go through this ordered list repeatedly merging highly overlapping cliques to build *modules*. Specifically, for every clique C_i in the list, we repeatedly look for a clique C_j $(j > i)$ such that the overlap $|C_i \cap C_j|/|C_j| \ge t_o$, a predefined overlap-threshold. If such a C_j is found, we calculate the *weighted inter-connectivity* I_w between the non-overlapping proteins of C_i and C_j as follows,

$$I_w(C_i, C_j) = \frac{\sum_{p\in\{C_i\setminus C_j\},\, q\in\{C_j\setminus C_i\}} r(p, q)}{|C_i\setminus C_j|.|C_j\setminus C_i|} \tag{5}$$

If $I_w(C_i, C_j) \ge t_m$, a predefined merge-threshold, then C_j is merged into C_i forming a module, else C_j is discarded.

We capture the effect of differences in interaction weights between normal and tumor through the weighted density-based ranking of cliques. Weighted density assigns higher rank to larger and stronger cliques. Therefore, we expect cliques with lost proteins or weakened interactions (expected in tumor) to go down the rankings resulting in *altered* module generation, thereby capturing changes in modules between normal and tumor.

2.3 Comparing modules across conditions

Let $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ and $\mathcal{T} = \{T_1, T_2, \ldots, T_m\}$ be the sets of modules identified from the networks H_N and H_T, respectively. For each $S_i \in \mathcal{S}$, we calculate its *module correlation density* as follows,

$$d_{cc}(S_i) = \frac{\sum_{p,q \in S_i} PCC((p,q),N)}{\binom{|S_i|}{2}}. \qquad (6)$$

The correlation densities for tumor modules $\mathcal{T}$ are calculated similarly.

We identify the set $\Gamma(\mathcal{S},\mathcal{T})$ of *disrupted or altered module pairs* by modeling it as a *maximum weight bipartite matching* (Gabow, 1976) as follows. We first build a *similarity graph* $M = (V_M, E_M)$, where $V_M = \{\mathcal{S} \cup \mathcal{T}\}$, and $E_M = \bigcup \{(S_i, T_j) : J(S_i, T_j) \geq t_J, \Delta_{cc}(S_i, T_j) \geq \delta\}$, where $J(S_i, T_j) = |S_i \cap T_j|/|S_i \cup T_j|$ is the Jaccard similarity and $\Delta_{cc}(S_i, T_j) = |d_{cc}(S_i) - d_{cc}(T_j)|$ is the *differential correlation density* between S_i and T_j, and t_J and δ are thresholds. Every edge (S_i, T_j) is weighted by $J(S_i, T_j)$. We next identify the disrupted module pairs $\Gamma(\mathcal{S},\mathcal{T})$ by finding the maximum weight matching in M, and we rank them in non-increasing order of their differential density Δ_{cc}.

Finally, we infer genes involved in cancer as $\mathcal{O} = \{g : g \in S_i \cup T_j, (S_i, T_j) \in \Gamma(\mathcal{S},\mathcal{T})\}$ ranked in non-increasing order of $\Delta_{cc}(S_i, T_j)$.

To identify altered modules, we match normal and tumor modules by setting a high t_J, which ensures that the module pairs either have the same gene composition or have lost or gained only a few genes (e.g. if $|S_i| = 8$, $|T_j| = 9$, then $t_J = 2/3$ requires at least an overlap of 7). Among these, the module pairs showing higher differential correlation are ranked higher. Further, we expect cancer genes to be harbored within these module pairs and rank them likewise.

2.4 Preparation of experimental data

We gathered *Homo sapiens* PPI data inferred from multiple low- and high-throughput experiments deposited in Biogrid v3.1.93 (Stark *et al.*, 2011). To minimize false-positives, we used a scoring scheme, Iterative-CD (Liu *et al.*, 2009) with 40 iterations, to assign a *reliability score* to every interaction in the PPI network. The score (between 0 and 1) reflects the reliability of interactions by accounting for the number of common neighbors shared among the proteins in each pair. Discarding low-scoring interactions (<0.20) resulted in a PPI network of 29 600 interactions among 5824 proteins (average node degree 10.165). We gathered 39 matched pairs (78 total) of normal and tumor gene expression samples resected from 36 PDAC patients, from studies by Badea *et al.* (2008) (NCBI GEO accession GSE15471). We gathered pancreatic mutation profiles from the Supplementary Materials of Jones *et al.* (2008). These authors identified four kinds of mutations, *viz.* somatic, homozygous deletion, amplification and driver mutations, of which genes with somatic and driver mutations were assigned mutation scores (between 0 and 1) that reflected the likelihood of these mutations in driving tumorigenesis. We collated these scores and assigned a score of 1 to all deleted and amplified genes to construct mutation profiles for 1169 genes in pancreatic cancer.

3 RESULTS

3.1 Experimental settings and initial validations

We first tested our module-extraction procedure on the yeast *Saccharomyces cerevisiae* PPI network, as the available 'gold standard' set of complexes in yeast is reasonably complete and well-defined. After testing a range of parameters, setting $k = 4$, $t_o = 0.50$ and $t_m = 0.25$ resulted in the best recall (sensitivity) of 72% and precision (specificity) of 83% and a Gene Ontology-based functional coherence of 76% for the modules. Therefore, we maintained these settings for our experiments. We also note that by relaxing the parameters to $t_o = 1/3$ and $t_m = 1/10$, the functional coherence was maintained at 61%, indicating the modules corresponded to larger complex-groups and pathways (details in Supplementary Files).

3.2 Analyzing disruptions in tumor PPI network

The normal H_N and tumor H_T networks displayed roughly equal numbers of interactions, as well as average scores (weights)—27 277 and 27 266 interactions with average scores of 0.208 and 0.210, respectively. Figure 3a shows no significant differences in the score distributions of the two networks (Kolmogorov–Smirnov test: $D_{NT} = 0.1556 < K_{\alpha=0.05} = 1.36$). However, examining these interactions more carefully, we found 3746 interactions that showed changes in scores. Of these, we extracted those with score changes ≥ 0.10, which included 176 interactions showing decrease (weakening) and 135 interactions showing increase (strengthening) of scores from normal to tumor. Similar analysis using expression correlations (Fig. 3b) identified ~8700 interactions showing lower correlations and ~2100 interactions showing higher correlations in tumor than normal (Kolmogorov–Smirnov test: $D_{NT} = 23.11 > K_{\alpha=0.05} = 1.36$).

DAVID-based (Dennis *et al.*, 2003) functional analysis of genes involved in these 311 ($= 176 + 135$) interactions showed significant enrichment ($P < 10^{-20}$) for the following biological process terms: mitotic cell cycle, cell division, DNA repair, chromatin modification, anaphase-promoting complex-dependent mechanisms and ubiquitin-protein ligase activity during cell cycle. We found key DNA damage repair (DDR) and cell cycle players, including BRCA1 (homologous recombination during DDR), RAD21 (DDR and cell cycle), FANCA (part of Fanconi anemia group, closely related to BRCA1-pathways), INO80 (chromatin remodeling) and MCM2 (DNA replication) involved in these interactions. The pair TGFBR1-TGFBR2, showing an increase from 0.34 to 0.56, is involved in transforming growth factor (TGF)-β signaling and is implicated in pancreatic cancer (Jones *et al.*, 2008). RAD21, involved in at least four of the rescored interactions, is a key player in DDR and has been implicated in breast cancer (Yan *et al.*, 2012).

3.3 Analyzing disruptions in tumor modules

We next performed a comparative analysis of normal $\mathcal{S}$ and tumor $\mathcal{T}$ modules to understand disruptions at the module level. Overall we noticed that the total number of modules, as well as average modules sizes, were almost the same across the two conditions (Table 1). The reason is that the interaction scores were roughly the same (Fig. 3a), resulting in similar module generation in both conditions.

But Table 1 also shows an overall decrease in correlation in tumor modules, which was not entirely unexpected given our analysis in Section 1.1. Further, this decrease had affected mainly the highly correlated modules (Fig. 4). In particular, there was a reduction in 45 (of 66) modules with correlation ≥ 0.4 from normal to tumor. DAVID-based analysis of these 45 modules (Fig. 5) showed significant enrichment ($P < 10^{-03}$) for core signaling pathways, including KRAS signaling, TGF-β-signaling, Wnt-signaling, P53-apoptosis, cell cycle and DNA repair; these pathways are genetically altered in 80% of pancreatic tumors (Jones *et al.*, 2008).

Next, we computed the set of matching modules $\Gamma(\mathcal{S},\mathcal{T})$, giving $|\Gamma(\mathcal{S},\mathcal{T})| = 452$, for $t_J = 0.67$ and $\delta = 0.10$. We further

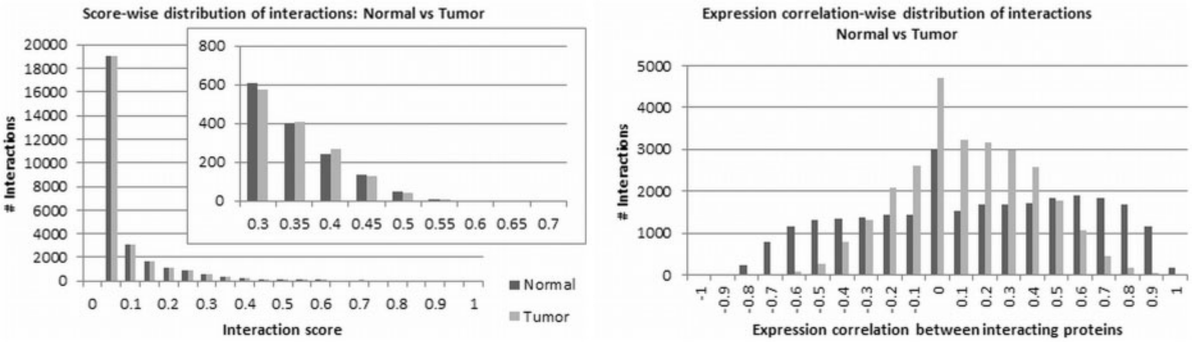

Fig. 3. (**a**) Scorewise distribution of interactions [inset: zoom into (0.3, 0.7)]; (**b**) expression correlationwise distribution of interactions in normal and tumor

Table 1. Properties of normal and tumor modules

Module set	No. of modules	Average module size	Correlation		
			Max	Avg	Min
Normal $\mathcal{S}$	574	7.285	0.898	0.147	−0.288
Tumor $\mathcal{T}$	576	7.175	0.632	0.107	−0.205

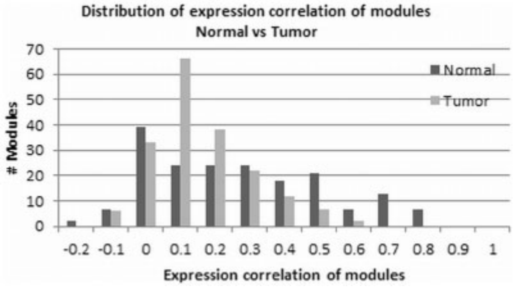

Fig. 4. Correlationwise distribution of modules in normal and tumor

divided $\Gamma(\mathcal{S},\mathcal{T})$ into $\Gamma'(\mathcal{S},\mathcal{T}) \subseteq \Gamma(\mathcal{S},\mathcal{T})$ of module pairs showing *higher* correlation in normal than tumor, and $\Gamma''(\mathcal{S},\mathcal{T}) \subseteq \Gamma(\mathcal{S},\mathcal{T})$ of module pairs showing *lower* correlation in normal than tumor, giving $|\Gamma'(\mathcal{S},\mathcal{T})| = 240$ and $|\Gamma''(\mathcal{S},\mathcal{T})| = 212$. We computed the absolute differential correlation Δ_{cc} of these subsets, as shown in Table 2. Interestingly, this demonstrated a marginal increase in correlation for 212 modules in tumor *vis-a-vis* normal, with a maximum increase of 0.353. However, DAVID-based analysis showed enrichment for similar terms in both $\Gamma'(\mathcal{S},\mathcal{T})$ and $\Gamma''(\mathcal{S},\mathcal{T})$, which was not specific enough to differentiate the roles of the two subsets and, therefore, whether *compensatory* or *tumor-driving* mechanisms coming into play. This prompted further in-depth analysis of the modules.

3.4 In-depth analysis of disrupted modules

Of the 452 module pairs in $\Gamma(\mathcal{S},\mathcal{T})$, 431 had the *same* gene compositions, and of these, 226 showed decrease and 205 showed increase in their correlations in tumor.

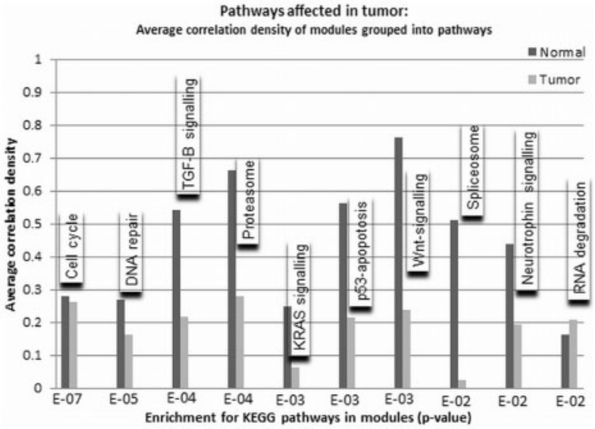

Fig. 5. Enrichment and correlations of pathways affected in PDAC tumor

3.4.1 Modules with the same gene compositions Among the 205 strengthened modules, we noticed a few interesting cases of modules in tumor being strongly (positive and negative) regulated by different transcription factors from the ones in normal. For example, the module (RUVBL1, RUVBL2, HJURP and CENPA) showed a weak correlation density of −0.086 in normal, indicating inactivity of the module. We found this module was (weakly) regulated by the transcription factor-complex TAL1–TCF4 with the gene correlations with TAL1 being (0.34, −0.21, 0.31 and 0.64) and with TCF4 being (0.48, −0.41, 0.34 and 0.47). However, in tumor, the correlation of the module had strengthened to 0.758, a steep increase of 0.844. Further, the gene correlations with TAL1 had decreased to (−0.35, 0.06, −0.64 and −0.50), whereas with TCF4 had increased steeply to (0.89, 0.69, 0.87 and 0.91), and TAL1 and TCF4 had themselves become anti-correlated (−0.25). This indicated an increase in activity of the module with strong positive regulation by TCF4. Interestingly, Jones *et al.* (2008) identified TCF4, a component of the Wnt/Notch signaling pathway, as an altered gene in 100% of pancreatic tumors.

3.4.2 Modules with changes in gene compositions Analysis of the remaining 21 module pairs that changed gene compositions revealed an interesting swapping phenomenon—new genes had

Table 2. Correlations of matched normal and tumor modules pairs

Module pairsubset	$t_J = 0.67$			$t_J = 0.50$		
	$\lvert\Gamma(\mathcal{S},\mathcal{T})\rvert = 452$			$\lvert\Gamma(\mathcal{S},\mathcal{T})\rvert = 617$		
		Δ_{cc}			Δ_{cc}	
	No. of Pairs	Max	Avg	No. of Pairs	Max	Avg
$\Gamma'(\mathcal{S},\mathcal{T})$	240	0.691	0.180	346	0.804	0.198
$\Gamma''(\mathcal{S},\mathcal{T})$	212	0.353	0.118	271	0.534	0.121

Note: $\Gamma'(\mathcal{S},\mathcal{T}) \subseteq \Gamma(\mathcal{S},\mathcal{T}), \Gamma''(\mathcal{S},\mathcal{T}) \subseteq \Gamma(\mathcal{S},\mathcal{T})$.

replaced existing genes, forming physical interactions with the remaining ones in these modules in tumor. For example, the normal module (HDAC2, HDAC1, MTA2, MTA1, CHD4 and BCL11B) had changed to (HDAC2, HDAC1, MTA2, MTA1, CHD4, SIN3A and SOX2) in tumor with BCL11B replaced by SIN3A and SOX2, and the module correlation had increased by 0.137. SOX2 regulates transcriptional network of oncogenes (Chen *et al.*, 2012) and is implicated in small-cell lung cancer (Chen *et al.*, 2012; Rudin *et al.*, 2012); SIN3A is a transcriptional repressor implicated in breast cancer (Ellison-Zelski and Alarid, 2010), whereas BCL11B is involved in lymphoid malignancies (Satterwhite *et al.*, 2001). We hypothesize that these gene-swapping events in modules were indicative of tumor disruptions; therefore, they are relevant to understand cancer.

In our computational model, modules change gene compositions only when the ranking of constituting cliques changes, resulting in altered clique-merging between conditions (Section 2.2). The cliques are re-ranked only when their constituting interactions are rescored. A total of 311 interactions were re-scored between normal and tumor (Section 3.2), and these constituted 53 cliques (of sizes 5–7) of the total ~5470 cliques present in each network. The jumps in ranks (average 4) noticed for these 53 cliques were higher than the jumps (average 1) noticed for the remaining cliques. These 53 cliques contributed to modules that showed changes in gene compositions.

DAVID-based functional analysis of the 311 rescored interactions showed significant ($P < 10^{-20}$) enrichment for key genome-stability mechanisms, including DDR and cell cycle (revisit Section 3.2). The 21 altered modules showed significant ($P < 10^{-66}$) enrichment for these mechanisms, and also for the following Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways ($P < 10^{-03}$): cell cycle, DNA replication, Wnt-signaling, TGF-β signaling and P53-based apoptosis, all of which have been implicated in the Jones *et al.* (2008) study on pancreatic cancer. In addition to this, of the 31 genes that had *swapped*, six (TP53, KRAS, SFN, SMAD4, CDKN2A and ARID1A) are implicated in pancreatic cancer, and 19 are included in KEGG pathways in cancer (Supplementary Files). These observations provide evidence that the gene-swapping events in our model are strongly indicative of genes and modules relevant to cancer.

We next describe an example of a gene-swapping event to aid interpret its relevance to cancer mechanisms. Figure 6 shows a normal-tumor module pair in which FAM175B was replaced by TP53. This was the consequence of a weakened FAM175B-clique displaced in ranking by a strengthened TP53 clique. It is known that the expression level of TP53 is low in normal cells, whereas DNA damage triggers increase in TP53 expression, which is responsible for activating transcription of DNA repair proteins. FAM175B is a member of the BRISC complex, responsible for recruitment of BRCA1–BARD1 heterodimer to sites of DNA damage at double-strand breaks. We noticed the TP53-expression level had increased from 5.86 (normal) to 6.56 (tumor) (Table 3), and the correlation of the module (RAD51, BRCA1, TP53, SUMO1 and UBC) had increased from 0.123 (normal) to 0.352 (tumor). This indicated activation of DNA repair proteins, RAD51 and BRCA1 in the presence of TP53, which are components of the homologous recombination double-strand break repair pathway. The correlation of the module (FAM175B, BRCC3, BRE and BARD1), which represented the BRISC complex, had reduced from 0.381 (normal) to 0.144 (tumor). This indicated possible impairment of DDR mechanisms observed in cancer cells.

3.5 Evaluating predicted genes in PDAC

3.5.1 Evaluation against 'gold standard' We evaluated our predicted genes against three gold standard (benchmark) lists for pancreatic cancer, namely, OMIM (pancreatic #260350) (Hamosh *et al.*, 2004), COSMIC Classic (Bamford *et al.*, 2004) and the Jones *et al.* (2008) study. We gathered our predicted genes from module pairs showing differential correlation ≥ 0.10, giving 143 genes.

We calculated recall (sensitivity) values as the fraction of benchmark genes covered by the top R-predicted genes. When $R = 143$ (taking all predicted genes), the recall values were OMIM: 78% (7/9), COSMIC: 67% (12/18) and Jones: 43% (37/86) (Fig. 7a). When $R = 25$, the recall values were OMIM: 78% (7/9), COSMIC: 55% (10/18) and Jones: 19% (16/86). The top 7 genes ($R = 7$) belonged to all the three benchmarks. Finally, five of our predicted genes, coding for transcription factors JUN, SMAD3, CEBPA, FOS and STAT3, were confirmed recently as biomarkers in PDAC (Winter *et al.*, 2012).

3.5.2 Assessment using differential expression of genes There were 2559 differentially expressed genes in the PPI network (adjusted $P < 0.001$) of which 1362 (or 53.23%) belonged to at least one module. The top three-enriched terms for these 1362 genes were the biological processes ($P < 10^{-20}$): cell cycle, DNA repair and chromatin modification, and the KEGG pathways ($P < 10^{-03}$): KRAS signaling, Wnt-signaling and P53-based apoptosis; these processes and pathways are genetically altered in at least 80% of pancreatic tumors (Jones *et al.*, 2008). The top three enriched terms for the remaining 1197 differentially expressed genes were the biological processes ($P < 10^{-10}$): cytoskeleton organization, actin filament-based process and intracellular transport, and the KEGG pathways ($P < 10^{-03}$): regulation of actin cytoskeleton, focal adhesion and endocytosis; these are relevant to cancer cell migration and metastasis (Yamaguchi and Condeelis, 2007). These genes were failed to be drawn into

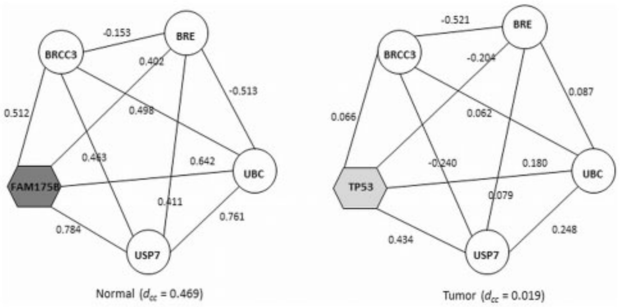

Fig. 6. Swapping behavior in the matched modules N203 and T152. Correlations are indicated on the edges and expression levels are in Table 3

Table 3. Average expression levels (39 sample pairs) and differential expression (adjusted *P*-values) of genes in normal N_{203}- and tumor T_{152}-matched modules (Fig. 6)

Genes/ condition	Common genes				Swapped genes	
	BRCC3	BRE	UBC	USP7	FAM17B	TP53
Normal	5.76	5.25	12.66	8.60	6.23	5.86
Tumor	5.93	5.00	13.07	8.80	6.68	6.56
P-value	NS	NS	4.21E-05	6E-03	1.06E-4	6.34E-06

NS, not significant at 0.001 level.

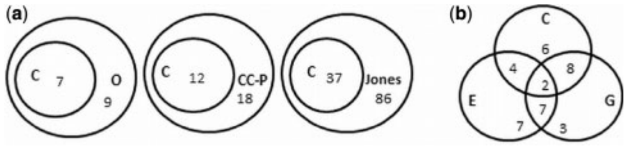

Fig. 7. (**a**) Overlap of CONTOUR (C) genes with OMIM (O), COSMIC Classic Pancreatic (CC-P) and Jones *et al.* (Jones). (**b**) Overlaps of top 20 novel genes with predictions from ENDEAVOUR (E) and GeneRanker (G)

modules because of lack of sufficient interactions, in particular the membrane interactions; these interactions are involved in intracellular transport, response to cell migratory stimuli and metastasis. On the other hand, there were 39 genes that did not show significant differential expression, yet belonged to modules that showed differential correlation of at least 0.10, and participated in cell cycle and DNA repair which are relevant to cancer—for example, BRE and BRCC3 (Table 3). This reiterates our motivation to track differentially co-expressed modules to identify genes in cancer.

3.5.3 Evaluation of novel genes We next evaluated our top-ranking novel genes for potential involvement in PDAC. Table 4 (lower portion) lists 25 genes for which we found three independent evidence for potential roles in pancreatic cancer: (i) frequency of copy number alterations (CNA) summing-up copy number gain (↑), loss (↓) and neutral loss of heterozygosity (LOH) (–) from the International Cancer Genome Consortium (ICGC) portal (http://dcc.icgc.org/web/); (ii) *Q*-value significance for amplication/deletion events at the chromosomal loci of these genes across >3000 tumors from Tumorscape-GISTIC (http://www.broadinstitute.org/tumorscape/); and (iii) preliminary literature pointing toward their roles in other cancers. Low *Q*-values (<0.25) suggest amplification/deletion events at the loci are enriched by selective pressures, and focal events affect relatively small regions of genomic DNA, spanning a few hundred kilo base pairs to a couple of mega base pairs. These two factors together are strongly indicative of presence of cancer genes (Beroukhim *et al.*, 2010). For example, USP15 showed CNA in 17/67 (25.37%) donors with a significant *Q*-value of 0.0255. FAM175B, encountered earlier in Figure 6, showed CNA in 7/64 (10.94%) donors with a significant *Q*-value of 0.00318. ZWINT showed 9/27 (33.34%) CNA, and recently Zhang *et al.* (2012) noted that RNA*i*-based depletion of ZWINT impaired homologous recombination (HR).

We found literature evidence for five of the genes, PSMA4, SF3A2, PCNA, PSMC2 and EEF1A1. These were among the top 25 identified copy-number alterations yielding to cancer liabilities owing to partial loss (CYCLOPS) genes studied by Nijhawan *et al.* (2012). These authors noted that copy-number losses that target tumor suppressor genes often involve multiple neighboring genes that may not contribute directly to cancer development, but their loss renders cancer cells highly vulnerable to further suppression of those genes. These genes form the weak links supporting cancer cells, and their targeted inhibition can be an effective anti-cancer therapy. Interestingly, Nijhawan *et al.* (2012) also found significant enrichment for 'proteasome' and 'splicesome' among CYCLOPS genes, which supported our enrichment analysis (Supplementary Files).

Further, the gene YWHAE was also found by two popular gene-prioritization methods, ENDEAVOUR (Tranchevent *et al.*, 2008) and GeneRanker (Gonzalez *et al.*, 2008) (Fig. 7b). YWHAE is involved in many vital cellular processes, such as metabolism, protein trafficking, signal transduction, apoptosis and cell cycle regulation. It acts as a tumor suppressor and its expression is upregulated coordinately with p53 and BRCA1, and it belongs to neurotrophin signaling pathway required for neurogenesis (GeneCards: http://www.genecards.org/) (Safran *et al.*, 2002). One of our predicted genes DISC1 is a regulator of multiple aspects of neurogenesis, and it participates in Wnt-mediated neural progenitor proliferation. The identification of DISC1 and YWHAE further supports the nexus between neural genes and pancreatic carcinogenesis (Biankin *et al.*, 2012).

Finally, PCNA is a known biomarker for cell proliferation in several cancers, and EEF1A1 has been recently implicated as a biomarker in prostrate cell transformation and a possible hallmark of cancer progression (Scaggiante *et al.*, 2012).

3.6 Differential genes in BRCA1 and BRCA2 tumors

We extended our investigation to one more case study involving breast cancer conditions between BRCA1 and BRCA2 tumors. Although at the onset one might not expect to see drastic differences between the two sub-types like in the case of normal versus PDAC, surprisingly our analysis revealed some interesting quantifiable differences between modules in the two tumors.

Table 4. Evaluation of novel genes in PDAC found by our method

Genes	Evidence				Properties of host modules (Normal $\mathcal{S}$, Tumor $\mathcal{T}$)				
	CNA (%)[a]	Q-value[b]	Focal	PubMed Id (year)	$\Delta_{cc}(S_i, T_j)$	$\|S_i\|$	$\|T_j\|$	$d_{cc}(S_i)$	$d_{cc}(T_j)$
SFN	9/67 (13.34) ↓	1.23E-26	Y	15073049 (2004)	0.534	4	4	−0.087	0.445
TP53	14/67 (20.90) ↓	1.00E-06	Y	11097231 (2000)	0.450	5	5	0.469	0.019
EP300	5/67 (7.46) ↓	2.43E-01	N	19569050 (2010)	0.297	4	4	0.642	0.327
KRAS	8/67 (11.94) ↓↑	1.00E-35	Y	11097231 (2000)	0.286	5	4	0.255	−0.030
SF3B1	8/67 (11.94) ↓	NS	N	21995386 (2011)	0.246	5	5	0.337	0.091
CDKN2A	19/67 (28.36) ↓	NS	N	21150883 (2011)	0.181	4	4	−0.070	0.111
SMAD4	13/67 (19.40) ↓	NS	N	15146471 (2004)	0.016	5	5	0.168	0.185
ARID1A	9/67 (13.43) ↓	6.36E-27	Y	22009941 (2012)	0.002	12	13	−0.035	−0.032
HNRNPA1	11/67 (16.42) –	2.25E-02	Y	18776302 (2008)	0.804	5	5	0.768	−0.036
TERF1	10/67 (14.93) ↑	4.68E-03	N	17430594 (2007)	0.654	5	4	0.647	−0.007
SUMO1	4/67 (5.97) ↓	NS	N	15881673 (2005)	0.627	4	5	0.732	0.105
ANXA7	6/67 (8.96) ↓ –	NS	N	17018618 (2006)	0.617	5	5	0.723	0.105
UBC	14/67 (20.90) –	NS	N	16633365 (2006)	0.450	5	5	0.469	0.019
FAM175B	7/64 (10.94) ↓↑	3.18E-03	Y	NA	0.450	5	5	0.469	0.019
YWHAE	16/67 (23.88) ↓ –	NS	N	18561318 (2008)	0.357	6	6	0.546	0.189
SUMO4	11/67 (16.42) ↓ –	1.25E-01	Y	NA	0.286	4	5	0.255	−0.030
SLC25A6	9/27 (33.34) ↓ –	4.28E-07	Y	NA	0.286	4	5	0.255	−0.030
NSL1	4/67 (5.97) ↑	NS	N	16585270 (2006)	0.258	8	9	0.652	0.394
ZWINT	9/27 (33.34) ↓	NS	N	22724020 (2012)	0.258	8	9	0.652	0.394
BHLHE40	7/67 (10.45) ↓	NS	N	22825629 (2012)	0.230	4	4	0.431	0.201
POLB	5/67 (7.46) ↓↑	1.42E-24	Y	19147782 (2009)	0.216	4	4	0.021	0.238
PSMA4	6/67 (8.96) ↓↑ –	NS	N	20587604 (2010)	0.210	6	6	0.742	0.532
DST	7/67 (10.45) ↓	4.50E-02	N	NA	0.210	4	4	−0.003	0.206
SF3A2	9/67 (13.43) ↓↑	1.41E-15	Y	NA	0.174	6	6	0.107	0.281
PCNA	3/67 (4.48) ↓	NS	N	23019409 (2012)	0.167	4	4	−0.031	0.136
USP15	17/67 (25.37) ↓	2.55E-02	Y	22344298 (2012)	0.161	8	6	0.226	0.064
DISC1	9/67 (13.43) ↓↑	2.08E-02	Y	NA	0.151	5	5	−0.174	−0.023
PSMC2	3/27 (11.11) ↓↑	1.01E-01	Y	16317774 (2006)	0.146	10	10	0.020	0.166
CHD3	17/67 (25.37) ↓↑ –	2.02E-01	N	21472101 (2011)	0.081	6	6	−0.119	−0.037
ANAPC7	10/67 (14.93) ↓↑ –	NS	N	NA	0.149	8	8	0.165	0.016
EEF1A1	13/67 (19.40) ↓	2.24E-01	Y	22355332 (2012)	0.146	4	5	−0.075	−0.221
BCL11B	8/67 (11.94) ↓ –	NS	N	11719382 (2001)	0.137	6	7	0.250	0.387
PARD6G	11/67 (16.42) ↓ –	5.63E-22	N	22576693 (2012)	0.027	7	7	0.172	0.144
WAS	43/67 (64.18) ↓ –	2.20E-23	N	18505064 (2008)	0.015	5	5	−0.036	−0.020

NA, not available in PubMed abstract search; NS, not significant at 0.25.

[a]Donors affected/donors analyzed from ICGC 'Pancreatic cancer' datasets (http://dcc.icgc.org/web/) as of October 2012. Hundred randomly chosen genes showed 1.14/67 (1.08%) CNA on average from the same cohort. ↑ gain, ↓ loss, – neutral loss of heterozygosity.

[b]Q-value for amplication/deletion across 'all cancers' (Tumorscape).

We obtained gene expression data from 19 BRCA1 and 30 BRCA2 familial breast tumor samples from the study by Waddell *et al.* (2010). The application of our method generated 541 and 537 modules in the two tumors, respectively, with 22 module pairs showing changes both in gene composition and expression correlation. We list here top five novel candidate genes from these modules (Table 5) along with their proportion of somatic mutations from ICGC 'breast cancer' cases (note that ICGC is not restricted to only these two tumors; therefore, the actual proportion might be higher) and also the Q-value significance from Tumorscape.

Hatakeyama (2011) discusses the roles of tripartite motif (TRIM) proteins in regulating carcinogenesis. Several TRIM members have been implicated in different cancer types, including TRIM68, co-located with TRIM5 on chromosome 11, is noted for overexpression in pancreatic cancer. Mahajan *et al.* (2008) note possible involvement of NCOA6 in breast, colon and lung cancers.

4 CONCLUSION

Modules including complexes and pathways work in additive, compensatory and alternative ways to counter genome destabilizing agents. Cancer is an outcome of coordinated dysfunctioning of these *very* modules; therefore, countering it necessitates an even more coordinated and systematic approach. The considerable differences in module behavior between normal and cancer and even between two sub-types of the same cancer depict the complexity and specificity of roles that genes undertake in these conditions. In this context, the point highlighted in Section 1, *viz.*

Table 5. Genes from differential modules in BRCA1 and BRCA2 tumors

Genes	Host modules			Somatic mutations[a]	Q-value
	$\lvert S_i \rvert$	$\lvert T_j \rvert$	$\Delta_{cc}(S_i, T_j)$		
TRIM5	4	4	0.411	15/507	2.04E-02 ($\mathcal{F}$)
MED12	5	6	0.374	11/507	NS
NCOR2	5	4	0.351	5/507	NS
NCOA6	5	4	0.351	2/42	3.26E-02 ($\mathcal{F}$)
ITSN2	6	5	0.211	6/507	NS

$\mathcal{F}$, focal event; NS, not significant.

[a] From ICGC breast cancer data as of October 2012.

it is critically important to study cancer in a systematic and controlled manner so as to precisely measure and characterize the roles of genes in cancer, makes even more sense.

ACKNOWLEDGEMENTS

The authors thank Dr Sarah Song and Piyush B. Madhamshettiwar for the datasets, Dr Guimei Liu for the Iterative-CD software, Professors Sean Grimmond and Kum Kum Khanna for valuable discussions and the anonymous reviewers for their valuable feedback to improve this work.

Funding: National Health and Medical Research Council project (1028742 to Dr Peter Simpson and M.A.R.).

Conflict of Interest: none declared.

REFERENCES

Badea,L. *et al.* (2008) Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies gene specifically overexpressed in tumor epithelia. *Hepatogastroenterology*, **55**, 2015–2026.

Bamford,S. *et al.* (2004) The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer*, **91**, 355–358.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Biankin,A.V. *et al.* (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature*, **491**, 399–405.

Chen,S. *et al.* (2012) SOX2 gene regulates the transcriptional network of oncogenes and affects tumorigenesis of human lung cancer cells. *PLoS One*, **7**, e36326.

Chu,L.H. and Chen,B.S. (2008) Construction of a cancer-perturbed protein-protein interaction network for discovery of apoptosis drug targets. *BMC Syst. Biol.*, **2**, 56.

Dennis,G. *et al.* (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, R60.

Doncheva,N.T. *et al.* (2012) Recent approaches to the prioritization of candidate disease genes. *Wiley Interdiscip Rev. Syst. Biol. Med.*, **4**, 429–442.

Ellison-Zeski,S. and Alarid,E.T. (2010) Maximum growth and survival of estrogen receptor-alpha positive breast cancer cells requires the Sin3A transcriptional repressor. *Mol. Cancer*, **9**, 263.

Gabow,H.N. (1976) An efficient implementation of Edmonds' algorithm for maximum matching on graphs. *J. ACM*, **23**, 221–234.

Gonzalez,G. *et al.* (2008) GeneRanker: an online system for predicting gene-disease associations for translational research. *Summit on Translat. Bioinfoma.*, **1**, 26–30.

Hamosh,A. *et al.* (2004) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Hatakeyama,S. (2011) TRIM proteins and cancer. *Nat. Rev.*, **11**, 792–804.

Jones,S. *et al.* (2008) Core signalling pathways in human pancreatic cancers revealed by global genomic analysis. *Science*, **321**, 1801–1806.

Kim,Y.A. *et al.* (2009) Identifying causal genes and dyregulated pathways in complex diseases. *PLoS Comput. Biol.*, **7**, e1001095.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotech.*, **25**, 309–316.

Liu,Y. *et al.* (2012a) Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. *BMC Syst. Biol.*, **6**, 65.

Liu,K.Q. *et al.* (2012b) Identifying dysregulated pathways in cancers from pathway interaction networks. *BMC Bioinformatics*, **13**, 126.

Liu,G. *et al.* (2009) Complex discovery from weighted PPI networks. *Bioinformatics*, **25**, 1891–1897.

Magger,O. *et al.* (2012) Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS Comput. Biol.*, **8**, e1002690.

Mahajan,M. and Samuels,H.H. (2008) Nuclear receptor coactivator/coregulator NCoA6 (NRC) is a pleiotropic coregulator involved in transcription, cell survival, growth and development. *Nucl. Recept. Signal.*, **6**, e002.

Masica,D.L. and Karchin,R. (2011) Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res.*, **71**, 4550.

Nijhawan,D. *et al.* (2012) Cancer vulnerabilities unveiled by genomic loss. *Cell*, **150**, 842–854.

Rudin,C.M. *et al.* (2012) Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.*, **44**, 1111–1116.

Safran,M. *et al.* (2002) GeneCards 2002: towards a complete, object-oriented, human gene compendium. *Bioinformatics*, **18**, 1542–1543.

Satterwhite,E. *et al.* (2001) The BCL11 gene family: involvement of BCL11A in lymphoid malignancies. *Blood*, **98**, 3413–3420.

Scaggiante,B. *et al.* (2012) Dissecting the expression of EEF1A1/2 genes in human prostate cancer cells: the potential of EEF1A2 as a hallmark for prostate transformation and progression. *Br. J. Cancer*, **106**, 166–173.

Shibata-Minoshima,F. *et al.* (2012) RHOXF2 (PEPP2) as a cancer-promoting gene by expression cloning. *Int. J. Oncol.*, **40**, 93–98.

Srihari,S. and Leong,H.W. (2013) A survey of computational methods for protein complex prediction from protein interaction networks. *J. Bioinform. Comput. Biol.*, **11**, 1230002.

Stark,C. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.

Tomita,E. *et al.* (2006) The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor. Comput. Sci.*, **363**, 28–42.

Tranchevent,L.C. *et al.* (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res.*, **36**, W377–W384.

Waddell,N. *et al.* (2010) Subtypes of familial breast tumours revealed by expression and copy number profiling. *Breast Cancer Res. Treat.*, **123**, 661–677.

Winter,C. *et al.* (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput. Biol.*, **8**, e1002511.

Wood,L.D. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.

Yamaguchi,H. and Condeelis,J. (2007) Regulation of the actin cytoskeleton in cancer cell migration and invasion. *Biochim. Biophys. Acta.*, **1773**, 642–652.

Yan,M. *et al.* (2012) Enhanced RAD21 cohesin expression confers poor prognosis in BRCA2 and BRCAX, but not BRCA1 familial breast cancers. *Breast Cancer Res.*, **14**, R69.

Zhang,J. *et al.* (2012) Weighted frequent gene co-expression network mining to identify genes involved in genome stability. *PLoS Comput. Biol.*, **8**, e1002656.

Zhang,S. *et al.* (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.

Zhao,J. *et al.* (2012) Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*, **28**, 2940–2947.