**Systems biology**

**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 mechanisms even across the two sub-types and revealed novel genes involved therein (e.g. TRIM5 and NCOA6).
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 5,824 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 ± (0,9,1) to ±(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 $p_{E_H}^{(N)}$ 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 $p_{E_H}^{(N)}$ if $x \leq |\text{PCC}(p, q), N| < x + 0.10$, where $|\text{PCC}(p, q), N|$ 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 $p_{E_H}^{(T)}$ for the non-interacting gene pairs, $E_{N\setminus H}$. Now, given these two distributions, the probability that a pair $(p, q)$ with correlation in $[x, x + 0.1)$ coinciding with an interaction is,

$$P_N(p, q) = P[(p, q) \in E_H| \text{PCC}(p, q), N] = \frac{|p_{E_H}^{(N)}|}{|p_{E_H}^{(N)}| + |p_{E_H}^{(T)}|},$$

where $|p_{E_H}^{(N)}|$ and $|p_{E_H}^{(T)}|$ 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 $h(p)$ (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 $MCC(p, q) = \text{max}(h(p), h(q))$. As aforementioned, we computed the mutation coefficientwise frequency distributions $p_{E_H}$ and $p_{E_{N\setminus H}}$ for the gene pairs in $E_H$ and $E_{N\setminus H}$, respectively. We infer the probability that any pair $(p, q)$ with mutation coefficient in $[x, x + 0.1)$ coinciding with an interaction as,

$$P_M(p, q) = P[(p, q) \in E_H| h(p), h(q)],$$

$$P_M(p, q) = \frac{|p_{E_H}^{(N)}|}{|p_{E_H}^{(N)}| + |p_{E_H}^{(T)}|},$$

where $|p_{E_H}^{(N)}|$ and $|p_{E_H}^{(T)}|$ are the frequencies of the bins containing $(p, q)$ in the two distributions, respectively.

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

$$r_N(p, q) = (p(p, q)\cdot P_N(p, q) + (1 - P_N(p, q))\cdot (1 - P_M(p, q)))$$

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 Mascia 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 $C$ 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 $C \in C$, we calculate its weighted interaction density $d_w(C)$,

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

We rank these cliques in non-increasing order of their weighted densities, $C_1, C_2, ..., C_m$, 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 $C_j (> i)$ such that the overlap $|C_i \cap C_j|/|C_i| \geq t_m$, a predefined overlap-threshold. If such a $C_j$ exists, we calculate the weighted inter-connectivity $I_w$ between the non-overlapping proteins of $C_i$ and $C_j$ as follows,

$$I_w(C_i, C_j) = \frac{\sum_{(p, q) \in C_i \cap C_j} P_N(p, q)}{|C_i| \cdot |C_j|}$$

If $I_w(C_i, C_j) \geq t_m$, 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 $S = \{S_1, S_2, ..., S_n\}$ and $T = \{T_1, T_2, ..., T_m\}$ be the sets of modules identified from the networks $H_N$ and $H_T$, respectively. For each $S_i \in S$, we calculate its module correlation density as follows,
follows. We first build a model and setting a high amplified genes to construct mutation profiles for 1169 genes in pancre-

We collated these scores and assigned a score of 1 to all deleted and that reflected the likelihood of these mutations in driving tumorigenesis. and driver mutations were assigned mutation scores (between 0 and 1) authors identified four kinds of mutations, among 5824 proteins (average node degree 10.165). We gathered 39 these module pairs and rank them likewise. Further, we expect cancer genes to be harbored within these module pairs and rank them likewise.

3.4 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 from 36 PDAC patients, from studies by Badia 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 pancre-atic 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 $k = 4$, $t_0 = 0.50$ 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 $t_0 = 1/3$ 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: $D_{KS} = 0.1556 < K_{0.05} = 1.36$). However, examining these interactions more carefully, we found 3746 interactions that showed changes in scores. Of these, we extracted those with score changes $\geq 0.10$, 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 $\sim 8700$ interactions showing lower correlations and $\sim 2100$ interactions showing higher correlations in tumor than normal (Kolmogorov-Smirnov test: $D_{KS} = 23.11 > K_{0.05} = 1.36$).

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, chro-matin 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 TGFB1-TGFB2, showing an increase from 0.34 to 0.56, is involved in transforming growth factor (TGF)-$\beta$ 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 $S$ and tumor $T$ 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 $\geq 0.4$ from normal to tumor. DAVID-based analysis of these 45 modules (Fig. 5) showed significant enrichment ($P < 10^{-10}$) for core signaling pathways, including KRAS signaling, TGF-$\beta$-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 $\Gamma(S, T)$, giving $|\Gamma(S, T)| = 452$, for $t_1 = 0.67$ and $\delta = 0.10$. We further
divided $\Gamma(S, T)$ into $\Gamma'(S, T) \subseteq \Gamma(S, T)$ of module pairs showing higher correlation in normal than tumor, and $\Gamma''(S, T) \subseteq \Gamma(S, T)$ of module pairs showing lower correlation in normal than tumor, giving $|\Gamma'(S, T)| = 240$ and $|\Gamma''(S, T)| = 212$. We computed the absolute differential correlation $\Delta_{cc}$ 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 $\Gamma'(S, T)$ and $\Gamma''(S, T)$, 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 $\Gamma(S, T)$, 431 had the same gene compositions, and of these, 226 showed decrease and 205 showed increase in their correlations in 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 (RUVEBL1, RUVEBL2, 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...
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 rescored (adjusted $P < 10^{-10}$) 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-$\beta$ 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, CDK2NA 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 $\geq 0.10$, 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 ($P < 10^{-08}$): 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 failed to be drawn into
Dysregulated modules and novel 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 amplification/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. Lower 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 RNAi-based depletion of ZWINT impaired homologous recombination (HR).

We found literature evidence for five of the genes, PSMA4, SF3A2, PCNA, PSM2 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 ‘spliceosome’ 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 prostate cell transformation and a possible hallmark of cancer progression (Scaggsante 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.
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 complex 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.
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 Pyush 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


